Article

Changes in the Intelligence Levels and Structure in Russia: An ANOVA Method Based on Discretization and Grouping of Factors

Tatiana Avdeenko, Anastasiia Timofeeva, Marina Murtazina* and Olga Razumnikova

Abstract: In the present paper, we investigate how the general intelligence quotient (IQ) and its subtests changed for students from Russian University from 1991 to 2013. This study of the effect of such factors as gender, department, and year on the IQ response is carried out using the ANOVA model. Given the unevenness of the initial sample by years and departments, and consequently, heterogeneity of variances when divided by the original natural categories, we decided to aggregate the values of explanatory variables to build an adequate model. The paper proposes and investigates an algorithm for joint discretization and grouping, which uses the procedure of partial screening of solutions. It is an intermediate option between the greedy algorithm and exhaustive search. As a goodness function (an optimality criterion), we investigate 26 intermediate options between the AIC and BIC criteria. The BIC turned out to be the most informative and the most acceptable criterion for interpretation, which penalizes the complexity of the model, due to some decrease in accuracy. The resulting partition of the explanatory variables values into categories is used to interpret the modeling results and to arrive at the final conclusions of the data analysis. As a result, it is revealed that the observed features of the IQ dynamics are caused by changes in the education system and the socio-economic status of the family that occurred in Russia during the period of restructuring the society and intensive development of information technologies.

Keywords: human intelligence; IQ; Flynn effect; ANOVA; discretization; grouping of factors; Bayesian information criterion

1. Introduction

In the modern world, intelligence is perceived as a strategically important resource that determines the development of science and production. At the turn of the 20th–21st centuries, the mean national intelligence quotient (IQ) began to be seen as an important factor (but not the only possible) contributing to the differences in the economic growth rates [1,2]. In this regard, the regularities of changes in the mean intellectual abilities of human [3], the specificity of the factor structure for intelligence [4], and factorial invariance of intelligence structures among the members of socio-humanitarian and technical professions [5], as well as the theoretical foundations for measuring intelligence in cohorts are being improved [6]. In addition, future cognitive trends and consequences for economic, and social development are predicted [7], and mechanisms for the development of intellectual abilities are being developed [8].

Psychometric theories of intelligence are based on objective intelligence tests that measure IQ. This quantitative assessment shows the intelligence level of an individual relative to the intelligence level of an average human in some groups. The first attempts to measure mental abilities and explain the results of measurements were made at the end of the 19th century by Francis Galton [9,10]. He laid the foundations of the mathematical-statistical apparatus used in the psychometric data analysis, which was then developed by...
Karl Pearson and Charles Spearman in the procedures of correlation and factor analysis [10]. As a result of empirical research, partially based on the early experiments of Galton, in 1904 Spearman presented a factor analysis model trying to explain the differences in the results of intelligence tests by two factors: General intelligence ("g") and specific ability ("s") [11]. Subsequently, the intelligence model proposed by Spearman was challenged by many models suggesting that intellectual behavior arises from a set of factors, for example, inductive and deductive reasoning, verbal understanding, processing speed, and spatial visualization [12]. Criticism of the two-factor model led to its improvement and the creation by Spearman of an intelligence model that is a hierarchy of factors with the so-called g factor at the top and group factors at a lower level. Among the group factors, Spearman identified the most important ones as the logical, the mechanical, the psychological, and the arithmetical abilities [13]. At the end of the 20th century, John Bissell Carroll’s three-tiered theory of cognitive ability explained many individual differences in intellectual ability without contradicting the dominance of the g factor. The g factor, also called the general factor, accounts for about 40 percent of the total variance in IQ test batteries when various cognitive tests are performed on a sample of people with a wide range of cognitive abilities [14]. The existing IQ tests differ in the number and type of intelligence measured by them. Some IQ subtests correlate more with the g factor than others. For example, the vocabulary subtest has a high correlation with the g factor [15].

In the study of intelligence according to IQ tests, the variability of the mean intellectual abilities of a person over time was revealed. The English psychologist Richard Lynn recorded a difference in the national intelligences of Japan and the United States, as well as an increase in the intelligence indicators in Japan at the turn of the 1970–1980s [16,17]. In the 1980s, James R. Flynn investigated the issues of IQ change in the USA [18], reanalyzed the research data of Lynn, and pointed out some errors in comparing the national intelligences of the Japanese and Americans [19]. He also presented the research results describing the phenomenon of IQ gains in the USA and other countries [20,21]. Later, R.J. Herrnstein and C.A. Murray gave this phenomenon the name the “Flynn effect” [22] that has become generally accepted in the scientific literature. Some researchers call this phenomenon the “Lynn–Flynn effect” [23].

The discovery of the Flynn effect has spawned many studies trying to determine the factors that influence the variability of average human intellectual abilities over time. Genetic factors, environmental factors, decreased fertility, and methodological problems that distort measurements were previously considered to explain the increase in IQ over time [24]. Investigating the Flynn effect, W. Dickens and J. Flynn proposed a formal model that explained how the environment and genes can interact and influence IQ [25]. In the course of numerous studies, it was found that environmental factors predominate in the nature of the Flynn effect [26]. This is supported by the observation that the Flynn effect is most pronounced when applied to the tasks that have the least load on the g factor [27].

The Flynn effect has been summarized in meta-analytic studies. The authors of [28] determined that the Flynn effect and group differences in IQ indices were explained by different reasons. The authors of [26] discuss the stability of the Flynn effect was confirmed in different age groups, measures, samples, and levels of performance. A meta-analysis of data collected over the 1950–2014 period in 48 countries found the Flynn effect to be present in every age group, and the Flynn effect was stronger in developing countries. The authors of [29] found a decrease in IQ growth in the last decade of the 20th century, and a positive relationship between full-scale, crystallized, and spatial IQ with changes in the gross domestic product per capita was revealed (the change in the gross domestic product showed negligible effects for fluid IQ). The meta-analysis was performed based on 219 studies in a sample over 105 years (1909–2013) in 31 countries. Russia was not included in this sample.

One of the first large-scale studies of the Flynn effect in Russia was presented in 2019 [30]. To analyze the dynamics of intellectual abilities, 238,363 protocols of voluntary online testing for 2012–2018 were used. The participants in the experiment were presumably
men born in 1974–1999. The analysis of these data showed that for respondents born until the mid-1980s, there was a decrease in IQ test results replaced by a linear increase, the rate of which was estimated at about 0.19 IQ points per year. The obtained dynamics of IQ test results could be explained by the peculiarities of the socio-economic situation in Russia in the 1980s–1990s.

At the beginning of the 21st century, scientists began to observe a slowdown or reverse the development of the Flynn effect. B. Bratsberg and O. Røgeberg demonstrated a sharp drop in average IQ values in a large sample spanning three decades of Norwegian birth cohorts (1962–1991) [31]. This phenomenon is called the “negative Flynn effect” [32,33]. In the study of the reasons for the decrease in the population IQ, such factors were revealed as immigration [31,34], dysgenics [31,35], a change in the sex ratio of the sample [31], and education standards decline [31,36].

Thus, the conclusions drawn from the above studies are very diverse. However, the general result of the research can be formulated as follows: The Flynn effect is not universal and is sensitive to environmental factors. This is confirmed by numerous studies which showed that during the twentieth century, there was a rather significant increase in IQ indices, and since the late 20th–early 21st centuries, the opposite effect of weakening the intelligence, especially its individual components, can be observed. At the same time, despite a huge amount of research, there is no consensus on the reasons for an increase or decrease in the average human intellectual abilities over time. The existing studies on the issue under consideration are very different in terms of nationality, age, studied subjects, sample sizes, composition, and completeness of the intelligence tests, as well as methods of data analysis. A set of additional factors used in mathematical models, the accounting of which could more accurately describe the structure of changes in the intelligence, is also important. To create a more complete and systematic view of current research on the Flynn effect, we analyzed a number of papers [7,21,30,31,37–52] on this topic in accordance with the above attributes and presented them in an overview in Table A1 in Appendix A.

The analysis of scientific publications allows us to say that very few long-term IQ studies have been carried out in Russia, and that they have been conducted for a much shorter time than in several other countries. Our present work fills this gap to some extent. The study is based on the data of the IQ testing among young people aged 17.7 ± 1.3 years, conducted during 23 years from 1991–2013. The sample consists of 3631 young people of both genders, which were tested according to the Amthauer method. They were students from the Novosibirsk State Technical University who chose various departments for training. This university is now included in the top 20 among technical universities in Russia. The students studied at different departments and had various scores of the entrance Unified State Exam (USE), which usually correlated with their IQ.

The purpose of the study was to identify the influence of the explanatory variables (predictors) on the dependent variable, namely, a general IQ test. As a mathematical model for analyzing the collected data, it seems most appropriate to use the ANOVA model. The IQ value (or its subtests’ values) obtained according to the Amthauer method was a numerical output variable, with the explanatory categorical variables being the year, department, and gender. The ANOVA model not only assesses the impact of the above factors on the IQ values separately, but also their interactions.

However, there are circumstances that complicate the construction of an effective ANOVA model from the available sample data. The first difficulty is related to the inconsistency of the values of the explanatory variables with the model assumptions. Namely, the year factor is quantitative, not categorical, as the model requires, and the department factor contains too many categories. Previously, to solve this problem, explanatory variables were manually categorized in such a way that only two time periods separated by the transition from the 20th to the 21st century, and three groups of departments related, respectively, to the mathematical, engineering, or humanitarian orientation in training were distinguished [53]. However, such a rough categorization leads to constructing a
model that is not optimal from the point of view of the complete extraction of information contained in the data.

The second difficulty is the imbalance of the available data. When conducting long-term studies of intelligence, it is not always possible to develop a sample design that makes it possible to obtain optimal estimates of the effects in the ANOVA model, since it is difficult to provide the conditions under which a similar population of individuals would be tested every year. In this regard, our sample is characterized by an uneven distribution of students across the departments in each year, and there is also a large effect of factors interaction.

To solve these problems, the paper proposes and investigates a method of joint discretization and grouping of explanatory variables based on a consistent combination of the AIC and BIC optimality criteria, assessing the quality of the ANOVA model in terms of both prediction accuracy and complexity of the model that best matches the available data.

The paper has the following structure. Section 1 provides an overview of the research into changes in the structure of intelligence in the world in the 20th–early 21st centuries. In Section 2, we present a description and exploratory analysis of the collected data, formulate a mathematical model, and justify the need to categorize the values of the exploratory variables. Section 3 contains a selection of a goodness function for solving the problem of joint discretization and grouping, and the proposed algorithm for solving the problem. In Section 4, we study the algorithm, choose a compromise variant for discretization and grouping the factors, and finally, interpret the results of mathematical and statistical modeling. Section 5 provides a discussion.

2. Preliminary Research

2.1. Data Collection and Exploratory Analysis

In the period of 1991–2013, a data set of IQ indices was collected according to the Russian version of the Amthauer’s Intelligence Structure Test [54]. The testing involved 3646 students aged 17.7 ± 1.3 years, enrolled in the first year of a technical university of various specialties: From engineering and mathematical, to humanitarian and economic. The intelligence structure was represented by the performance indicators of nine subtests, including IQ1—completion of sentences; requires a sense of reality, independence of thinking; IQ2—selection of words; requires inductive verbal thinking, verbal sensitivity; IQ3—verbal analogies; requires the ability of combination, mobility in thinking; IQ4—conceptualization; requires logical thinking, ability for abstract verbal thinking; IQ5—calculations, requires practical and logical mathematical abilities; IQ6—number series completion; requires the theoretical and abstract dealing with numbers; IQ7—figure detecting; requires the ability to understand two-dimensional structures; IQ8—identification of cubes; requires technical and constructive abilities; IQ9—short-term memory; requires the ability to remember the words.

Testing was carried out in groups of 15–20 students during practical classes in the psychology course. All volunteers participated in the study for course credits. The content of the tasks and instructions for their performance was given in booklets distributed to each subject. The teacher gave the necessary explanations before testing and controlled the time for completing the subtests. In accordance with the time schedule, students completed all nine subtests during 90 min. During further processing of the results, the number of correct answers for each subtest was converted to the IQ level according to standard tables in accordance with age.

Thus, the data contains the results of testing the students for nine IQ1–IQ9 subtests and the general IQ test. There is also data on independent variables that can affect the IQ, including gender, year of testing, and department in which the student studies. The exploratory analysis allows us to understand what properties the data has, and what mathematical models can be built to study the effect of independent variables on the IQ, as well as whether the quality of the data satisfies the initial assumptions of the models.
The important requirements of supervised statistical models are the conditions of normality and homoscedasticity (homogeneity of the variances of the dependent variable). If we talk about the partitioning of exploratory variables into natural categories, namely, two levels of the gender factor, ten levels of the department factor, and twenty-three levels of the year factor, then the division into these categories of the initial data is extremely uneven. This circumstance is explained by the fact that students from different departments were not tested in every subsequent year (see Table 1).

Table 1. Natural distribution of tested students by departments and years.

<table>
<thead>
<tr>
<th>Department</th>
<th>Abbreviation</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Applied mathematics and computer science</td>
<td>AMCSF</td>
<td>1995–2004</td>
</tr>
</tbody>
</table>

The non-uniformity of observations leads to the correlation of the corresponding variables and the heterogeneity of the variances. For example, unevenness only by years of testing is well illustrated in Figure 1.

Figure 1. Violin plot of the IQ variable distributed by the years.

Hypothesis testing confirms visual assumptions about deviation from normality and homoscedasticity within natural groups. Checking the independence of the year and department factors by the contingency table for natural categories rejects the hypothesis of independence at a 0.001 significance level ($\chi^2_{198} = 6845.5$). However, the contingency table has many cells with a small number of observations, which negatively affects the correctness of the chi-square test. For confirmation, the correlation ratio $\eta^2$ was calculated, showing the influence of the department on the year factor. It equals 0.167 ($F_{9, 3621} = 80.4$), which also indicates a significant correlation at a 0.001 significance level.

2.2. ANOVA Model

As it was stated earlier, our purpose is to identify an influence of variables, such as year, department, and gender, on the response variable, i.e., the IQ test (or one of its nine subtests). Since these variables are considered categorical, it seems reasonable to use the ANOVA model to solve the problem. The exploratory analysis showed that the explanatory variables were highly correlated, especially the year and department factors. The ANOVA model allows us to estimate not only the impact of the above factors on the IQ values separately, but also their interactions.
In order to solve the above problem, the ANOVA model is formulated, as follows,

\[ y_{ktji} = \mu + \alpha_k + \beta_t + \gamma_j + (\alpha \beta)_{kt} + (\alpha \gamma)_{kj} + (\beta \gamma)_{tj} + (\alpha \beta \gamma)_{ktj} + \varepsilon_{ktji}, \]  

where \( y_{ktji} \) is the \( i \)-th observed value corresponding to the IQ level for a student of the \( k \)-th gender of the \( j \)-th department in the year \( t \), \( \alpha_k \) is the effect of the \( k \)-th gender (\( k = 1 \) for male, \( k = 2 \) for female), \( \beta_t \) is the effect of the year \( t \), \( t = 1991, \ldots, 2013 \), \( \gamma_j \) is the effect of the \( j \)-th department, \( (\alpha \beta)_{kt} \) is the interaction effect of the \( k \)-th gender and the \( t \)-th year, \( (\alpha \gamma)_{kj} \) is the effect of the interaction of the \( k \)-th gender and the \( j \)-th department, \( (\beta \gamma)_{tj} \) is the effect of the interaction of the \( j \)-th department and the \( t \)-th year, \( (\alpha \beta \gamma)_{ktj} \) is the effect of the interaction of the \( k \)-th gender, \( t \)-th year and \( j \)-th department, and \( \varepsilon_{ktji} \) is a random error.

It is impossible to estimate all the effects of the model (1). Usually, they resort to reducing the parameters estimating not the parameters themselves, but paired comparisons with some baseline levels. For example, the difference \( (\alpha_2 - \alpha_1) \) is the influence on the IQ of the female factor versus the male one. In the present research, we have taken the first levels of the factors as the baseline ones.

A critical requirement for using the ANOVA model is the condition of homoscedasticity of the dependent numerical variable for the selected levels of factors. As shown by the exploratory analysis, the prerequisites for applying the model are not achieved with natural discretization and grouping of factors. Therefore, the problem arises of aggregating the explanatory variables into larger categories. For the ordinal variable year, it is advisable to use discretization methods, and for the categorical variable department to use grouping methods. A method that implements a combination of these approaches in application to the problem being solved is presented in the next section.

3. Proposed Method
3.1. Goodness Function

When conducting long-term studies, it is not always possible to develop a sample design that makes it possible to obtain optimal estimates of the effects in the ANOVA model, since it is difficult to ensure conditions under which a similar sample of persons would be observed every year. Often, as in our case, one must be content with passive observations.

As shown by the exploratory analysis of the data, the available sample is very uneven in the distribution of observations for natural categories, namely, by 23 years of observation, 10 departments, and gender. That is, a very uneven number of observations fell into each data block. This circumstance leads to an unacceptable quality of the ANOVA model, due to the failure to fulfill the homoscedasticity condition (significant differences in the variances of individual data blocks). A natural solution to this problem can be data aggregation, that is, the transition from small categories to larger ones, in which the data will be more evenly distributed. Aggregation, on the one hand, builds an optimal statistical model, in the best possible way extracting information contained in real data. On the other hand, the unification of informationally similar groups of students from the intelligence structure point of view will allow for a better interpretation of the results and make reasoned conclusions.

In our study, we will aggregate two explanatory variables year and department. In this case, the aggregation for the ordinal variable year will be carried out by discretization methods, considering possible mergers of only adjacent intervals. For the categorical department variable, we will consider all possible mergers using the grouping method. The gender factor should not be aggregated as it includes the minimum possible number of categories.

The results of merging the levels of factors based on the application of discretization and grouping methods depend on the goodness criterion according to which the obtained
partitions are compared with each other. Most often, the quality of a regression model is judged by the coefficient of determination, calculated as

\[ R^2 = 1 - \frac{ESS}{TSS}, \]

where \( ESS \) is a residual sum of squares, \( TSS \) is a total sum of squares for the model. The choice of the determination coefficient as a goodness criterion does not give any results, since the original partition always provides a minimum residual sum of squares. Indeed, any merging of intervals leads to a decrease in \( R^2 \).

To test the significance of the model, the F-statistic is used, calculated as

\[ F = \frac{R^2 \frac{N - p}{1 - R^2}}{\frac{p}{p - 1}}, \]

where \( N \) is the number of observations, \( p \) is the number of estimated parameters of the model. It takes degrees of freedom into account, so an increase in model complexity must be offset by a sufficient decrease in the residual sum of squares. When using the F-statistic, contrary to \( R^2 \), any merging of intervals leads to an improvement of the goodness function. Thus, the work of the algorithm ended only when the intervals could no longer be merged, that is, using this criterion leaves the minimum possible number of categories. Thus, neither \( R^2 \) nor F-statistic is suitable for solving the discretization and grouping problem, since they lead to limiting degenerate partitions.

The authors of [55] studied two groups of goodness criteria, which were distinguished for solving the discretization problem, namely, information and statistical ones. Information criteria are intended to reduce the loss of information when merging separate categories into groups. They are intended to improve the quality of extracted knowledge. Statistical criteria try to solve simultaneously two opposite problems. On the one hand, these criteria pay attention to the accuracy of model prediction; on the other hand, they do not make it possible to build an overly complex model. This balances the accuracy and complexity of the model.

The authors of [55] also showed discretization methods based on theoretical information complexity, and the methods based on statistical measures of data dependency are asymptotically equivalent. That is, both those and other discretization methods achieve similar goals in different ways. Therefore, in our work, to solve the aggregation problem, we will use two statistical criteria, namely, the Akaike information criterion (\( AIC \)), and the Bayesian information criterion (\( BIC \)), as well as their combinations.

\( AIC \) is often used in feature selection tasks, for example, in a stepwise regression procedure. It provides a trade-off between the quality of fit and the complexity of the model (the number of parameters \( p \)). The indicator is calculated, as follows,

\[ AIC = 2p + N \log ESS. \]

\( BIC \) penalizes an increase in the number of parameters more severely

\[ BIC = p \log N + N \log ESS. \]

Thus, according to the Akaike criterion, a more complex model will be obtained, while the use of the Bayesian criterion will lead to a division into a smaller number of categories. We propose to consider intermediate variants of the criteria by varying the penalty term for the model complexity. Thus, we introduce the following class of evaluation functions, defined by the parameter \( 2 \leq c \leq \log N \):

\[ Q = p \cdot c + N \log ESS. \] (2)

In order to represent \( Q \) as a function of the results of the grouping and discretization of the explanatory variables, we express the \( ESS \) in terms of the intragroup sums of
squared deviations from the mean of the response. Let the model include three explanatory variables, and then the individual within-group sum of squares is calculated as

$$\text{ESS}(G_{ktj}) = \sum_i y_{ktji}^2 - \left( \frac{\sum_i y_{ktji}}{n_{ktj}} \right)^2,$$

where $G_{ktj}$ is a subset of $n_{ktj}$ students of $k$-th gender, $j$-th group of departments, $j = 1, \ldots, J$, surveyed in the period $t$, $j = 1, \ldots, T$, $y_{ktji}$ is $i$-th observed value corresponding to the IQ value of students from the subset $G_{ktj}$. Initially, each group of departments includes one department, each period includes one year. The residual sum of squares can be formulated as

$$\text{ESS}\left( \bigcup_{k,t,j} G_{ktj} \right) = \sum_{k,t,j} \text{ESS}\left( G_{ktj} \right).$$

When determining the number of parameters, it is necessary to consider that for some combinations of the factors levels there are no observations in the subset $G_{ktj}$, therefore

$$p\left( G_{ktj} \right) = I\left( G_{ktj} \neq \emptyset \right),$$

where $I(S)$ takes the value 1 if the statement $S$ is true, and takes the value 0 otherwise.

Then the total number of parameters will be

$$p\left( \bigcup_{k,t,j} G_{ktj} \right) = \sum_{k,t,j} I\left( G_{ktj} \neq \emptyset \right).$$

Thus, the goodness function is defined by

$$Q\left( \bigcup_{k,t,j} G_{ktj} \right) = p\left( \bigcup_{k,t,j} G_{ktj} \right) \cdot c + N \log \text{ESS}\left( \bigcup_{k,t,j} G_{ktj} \right). \quad (3)$$

While solving the problem of grouping and discretization, a set of subsets $G_{ktj}$ is formed that would minimize the function (3). Therefore, the optimization problem can be formulated follows

$$Q\left( \bigcup_{k,t,j} G_{ktj} \right) \rightarrow \min_{t,j,G_{t11},\ldots,G_{tTJ}} \quad (4)$$

It should be borne in mind that with a very large number of categories, building good groupings is difficult, due to the risk of overfitting the model. As a last resort, to avoid overfitting, efficient grouping algorithms can combine all values into one group, thereby excluding the variable from consideration. In order to prevent such a situation, the stopping criterion must include a condition for the minimum number of categories (for example, two).

### 3.2. Algorithm for Joint Discretization and Grouping

As follows from the previous section, to solve the problem of joint discretization (for the year variable) and grouping (for the department variable), it is necessary to develop an algorithm that solves the optimization problem (4). This algorithm refers to the so-called wrapping methods [56,57] that focus on the quality of the estimated model. The existing approaches [58,59] are designed for classification tasks; that is, they assume that the response is categorical. Therefore, their application to transform a set of exploratory variables in constructing ANOVA models requires discretizing the response, leading to the loss of significant information. The well-known algorithms based on statistical (ChiMerge [60], Chi2 [61]) and information evaluation measure (MDLP [62]) have the same disadvantage.

Compared to the problem of discretization, the grouping problem has not been studied so deeply in the literature. Most of the existing algorithms (Sequential Forward Selection
method [63], CHAID [64], MODL [65]) solve the grouping problem with a greedy heuristic based on bottom-up categorization.

Thus, most of the existing supervised discretization algorithms are designed to solve classification problems, that is, for a categorical response. They are mainly aimed at improving the quality of predicting the response (quality of classification) [66,67], while our task in the present research is not to predict IQ, but to obtain and interpret estimates of the effects of the explanatory variables.

Moreover, the existing discretization algorithms are usually univariate. In this regard, it seems relevant to develop an algorithm for the optimal categorization of several explanatory variables, considering their interrelationships, to build an ANOVA model. Here categorization includes two tasks: Discretization of quantitative variables and grouping of categorical variables.

To solve this problem, the algorithm is proposed and studied by the authors of [68], based on a standard supervised agglomerative discretization technique. It is greedy and includes the following steps:

1. **Step 1.** Set starting intervals by placing each unique value into a separate interval.
2. **Step 2.** Calculate the evaluation function $Q$ for each pair of adjacent intervals.
3. **Step 3.** Combine the pair with the best $Q$ value (if the stopping condition did not work).
4. **Step 4.** Repeat steps 2–3 until no further merging is possible.

At step 2, a goodness function is calculated, while simultaneously combining two attributes. For a categorical variable, all possible pairs of factor levels are considered, for a quantitative variable, we consider only adjacent intervals. In addition, a zero index according is assigned to the variable for which the levels are not combined. This is done in case the optimal solution is in combining the levels only for one of the variables.

At step 3, the algorithm stops if, after combining, it is not possible to achieve an improvement in the goodness function. At step 4, further merging is considered impossible if all factors have two levels.

This described algorithm does not guarantee that an optimal solution will be obtained, because moving along the path that currently seems to be the best, discarding all other options, does not exclude the fact that the discarded path at the current step turns out to be optimal. Most likely, the solution obtained in this way will be locally optimal. Another extreme choice of an optimization algorithm is a complete enumeration of all possible options, which ultimately guarantees a global optimum. However, a full enumeration in the case of several explanatory variables, which involves the joint solution of discretization and more resource-intensive grouping problems, requires unacceptably large computation resources.

Based on the above reasoning, in the present work, we have chosen an intermediate version of the algorithm with a partial screening of the best solutions. To implement it, instead of one best solution, at each iteration of the algorithm, it is required to save several best solutions, and then it is for them to perform the subsequent merging of categories. To prevent the number of such solutions from growing to unacceptable values, it is proposed to limit them to a fixed value $m$, which is the same for each iteration.

In order to store information about all the best solutions at a given iteration of the relative initial (rather than current) partitions of the explanatory variables, it is proposed to use lists of clusters formed by the levels of factors as a result of their merging. For categorical factors, the initial numbering of levels is set in an arbitrary order, for quantitative variables—it is set in ascending order of their values. At the initialization stage, each level is placed in its own cluster with the corresponding number. Thus, for example, for the factor $x_1$ with $T_0$ original levels, the list of clusters $L_{11}$ at the first step contains $T_0$ elements $[1, \ldots, T_0]$. If the $k$-th best solution involves combining the levels $t_1$ and $t_2$ ($t_1 < t_2$), then in the list $L_{1k}$ the element $t_2$ is reset to zero, and the element $t_2$ is added to the element $t_1$. For a newly formed cluster consisting of levels $t_1$ and $t_2$, the goodness function is calculated according to the relation (3), since the subsets with levels $t_1$ and $t_2$ are replaced by one new subset ($t_1 \lor t_2$).
Algorithm 1 Discretization and grouping with a partial screening of the best solutions

Input: raw data includes response values, quantitative exploratory variable $x_1$ with $T_0$ levels, and qualitative exploratory variable $x_2$ with $J_0$ levels. We set a constant $c$ for the goodness function (2), the number $m$ of best solutions, fixed and stored at each iteration.

Step 1. Set $H = 1$.

Step 2. Calculate the value $Q_0$ of function (2) according to the initial data.

Step 3. For $h$ from 1 to $H$ do

Step 3.1. Form subsets $G_{k_1j_1}$ in accordance with the lists of clusters $L_{1h}$ and $L_{2h}$.

Step 3.2. If the number of non-zero clusters in the list $L_{1h}$ ($L_{2h}$) does not exceed 2, then $t_1 = t_2 = 0$ ($j_1 = j_2 = 0$).

Step 3.3. For each pair of non-zero neighboring clusters $t_1$, $t_2$ from the list $L_{1h}$ and for $t_1 = t_2 = 0$, and each pair of non-zero clusters $j_1$ and $j_2$ from the list and for $j_1 = j_2 = 0$, perform:

Step 3.3.1. If $t_1 = t_2 = 0$ ($j_1 = j_2 = 0$), then the merging for the first (second) exploratory variable is not performed.

Step 3.3.2. Calculate the values $Q_1$ of the goodness function using the Formula (3):

- if merging occurs for two variables, then subsets $G_{kt_1j_1}$, $G_{kt_2j_2}$, $t_1 \neq t_2$ are replaced by the subset $G_{k(t_1 \lor t_2)t_1j_1}$, subsets $G_{kt_1j_2}$, $G_{kt_2j_2}$, $t_1 \neq t_2$ are replaced by the subset $G_{k(t_1 \lor t_2)t_2j_2}$, subsets $G_{k(t_1 \lor t_2)j_1}$, $G_{k(t_1 \lor t_2)j_2}$, $G_{kt_1j_2}$, $G_{kt_2j_2}$ are replaced by a subset $G_{k(t_1 \lor t_2)(t_1 \lor t_2)j_1}$, $G_{k(t_1 \lor t_2)(t_1 \lor t_2)j_2}$.

- if merging occurs only on one variable $x_1$, then subsets $G_{kt_1j_1}$, $G_{kt_1j_2}$ are replaced by the subset $G_{k(t_1)j_1}$.

Step 4. Define the smallest $m$ values of $Q_1$ and the appropriate variants of merging described by the variables $t_1$, $t_2$, $j_1$, $j_2$, $h$.

Step 5. If the minimum of $m$ smallest values of $Q_1$ is less than the current value $Q_0$, go to Step 6, otherwise the stop of the algorithm occurs.

Step 6. Assign $Q_0$ minimum value found at Step 4.

Step 7. Assign $H = m$.

Step 8. For $h'$ from 1 to $H$, form new lists of clusters $L_{1h'}$, $L_{2h'}$ based on proceeding from variants of merging corresponding to the best solutions found at Step 4.

Step 9. Save lists $L_{1h'}$, $L_{2h'}$ corresponding to $Q_0$.


Output: Lists of clusters $L_{1h'}$, $L_{2h'}$, corresponding to the last achieved minimum value $Q_0$.

4. Results

4.1. Experimental Results

The performance of the Algorithm 1 largely depends on the number of best solutions $m$ left at each iteration of the algorithm. It is the value $m$ that determines the acceptable compromise solution between the greedy algorithm and exhaustive search.

In the study, we changed the value $m$ from 1 to 50. For each $m$ the values of the constant $c$ were taken with increments of 0.25 (a total of 26 values from 2 to $\log N \approx 8.2$). The behavior of the algorithm for different values of $m$ depends on what value the constant $c$ takes, that is, the effect of interaction of the two variables takes place. For example, for the AIC criterion ($c = 2$), the results are unstable. In some cases, an increase in the number of best solutions $m$ leads to a deterioration in the goodness function $Q$. For $m > 16$ better results are consistently achieved. However, in a number of cases ($2.25 \leq c \leq 3.25$ or $7.5 \leq c \leq \log N$) the best result is achieved for $m = 2$, and further, with increasing $m$, a stable deterioration in the goodness function occurs. For $6 \leq c \leq 6.75$, the minimum is reached at $m = 2$, and then the value of the goodness function did not change with an increase in the number $m$ of best solutions. In all other cases (except for $c = 3.75$ and $c = 7.5$), a stepwise decrease in the goodness function was observed.

Figure 2 shows the proportion of cases for each $m$ (among 26 values of the constant $c$), when the best result was achieved for a given $m$ (for all possible $m$ from 1 to 50).
It is clearly seen that for $m = 1$ the best solution was never achieved, that is, the original greedy version of the algorithm never gave a global optimum. Therefore, the proposed modification of the algorithm with a partial screening of solutions provides better results in comparison with a greedy version presented by the authors of [68]. However, since the proportion of cases where the best result was achieved does not grow with increasing of $m$, it is impossible to recommend a certain value $m$ at which the best results are guaranteed. Thus, with each new application of the algorithm, a preliminary adjustment is required, which consists in calculating for several consecutive values of $m$ and choosing the best solution.

Let us denote by $m^*$ minimum value of $m$ at which the best value of the goodness function $Q$ was achieved. In our study, in half of the cases corresponding to 26 values of the constant $c$, $m^*$ took the value 2, in about a third of the cases—the value 12, and for single cases—intermediate values 4, 10, and 11. Table 2 shows the values $m^*$ corresponding to the constants $c$ selected in Section 4.2.1.

Table 2. F-statistics for the ANOVA model depending on the parameter $c$.

<table>
<thead>
<tr>
<th>Factor</th>
<th>$c = 7$</th>
<th>$c = 7.5$</th>
<th>$c = 7.75; 8$</th>
<th>$c = \log N (BIC)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year</td>
<td>38.28</td>
<td>28.18</td>
<td>26.38</td>
<td>37.54</td>
</tr>
<tr>
<td>Department</td>
<td>160.27</td>
<td>202.02</td>
<td><strong>206.49</strong></td>
<td>204.51</td>
</tr>
<tr>
<td>Gender</td>
<td>23.04</td>
<td><strong>45.64</strong></td>
<td>38.53</td>
<td>39.72</td>
</tr>
<tr>
<td>Year:Department</td>
<td><strong>35.41</strong></td>
<td>22.90</td>
<td>32.64</td>
<td>31.06</td>
</tr>
<tr>
<td>Year:Gender</td>
<td>1.57</td>
<td>5.70</td>
<td>6.17</td>
<td><strong>7.58</strong></td>
</tr>
<tr>
<td>Department:Gender</td>
<td>0.20</td>
<td>3.58</td>
<td><strong>4.41</strong></td>
<td>3.24</td>
</tr>
<tr>
<td>Year:Department:Gender</td>
<td>2.61</td>
<td>2.75</td>
<td><strong>3.81</strong></td>
<td>3.38</td>
</tr>
<tr>
<td>sumF</td>
<td>261.38</td>
<td>310.77</td>
<td>318.42</td>
<td><strong>327.04</strong></td>
</tr>
<tr>
<td>Opt number of periods $T^*$</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Opt number of groups $J^*$</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>The number $m^*$ of best solutions</td>
<td>12</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

4.2. Results of Discretization and Grouping
4.2.1. Partitioning for General IQ

First, we apply the Algorithm 1 to find the optimal partitioning of the explanatory variable values for the general IQ as a response.

The result of partitioning depends a lot on a constant $c$, determining an intermediate option of the goodness function between the $AIC$, which pays more attention to the accuracy...
of the model, and \( BIC \), which penalizes more for the complexity of the model. To make a choice between the \( AIC \) and the \( BIC \) (a choice between 26 values of the constant \( c \)), it is necessary to involve some additional criteria. As such a criterion, it was decided to use F-statistic for each explanatory variable and their interactions, obtained from ANOVA tables with the analysis results. The choice in favor of a particular solution can be made based on the greater significance of the parameter of the most interest to the researcher, since the greater value of the F-statistics indicates the greater significance of the corresponding effect. Table 2 presents the results only for those \( c \) values for which the maximum F-statistics was achieved for either one variable, or interaction of variables, or for the sum of F-statistics (sumF). The resulting optimal number of periods is denoted by \( T^* \), the optimal number of department groups is denoted by \( J^* \).

For \( c = 7 \), the year variable and the effect of interaction between the year and the department are of the greatest importance in F-statistics. The influence of the gender variable is the strongest for the partition obtained at \( c = 7.5 \). The department variable, the effects of the department–gender interaction, as well as the interaction of all three factors, have the highest F-statistics at \( c = 7.75, 8 \). The maximum influence of the interaction effect between the gender and the year, as well as the total influence of all variables, is found for the model built based on the \( BIC \) criterion (\( c = \log N \)).

From Table 2, we see possible partitions by year into 3, 4, and 5 periods, depending on the value of \( c \) constant:


The solution corresponding to five periods is not typical, it occurs only once among the best partitions and corresponds to \( c = 7.5 \) with highest F-statistics for gender. Therefore, this solution with too many periods should be abandoned, unless the variable gender takes precedence over other explanatory variables. Moreover, the solutions corresponding to four and five periods contain the latest period consisting of only one 2013. However, in 2013, only 16 students were tested, so it would be considered inappropriate to separate this year into a distinct period. Therefore, it is reasonable to choose a partition into three periods by the variable year.

The decision to limit ourselves to three periods by the variable year is also confirmed by the analysis of F-statistics. Indeed, this partitioning corresponds to two values \( c = 7 \) and \( c = \log N \) (see Table 3), corresponding to the most important significant parameters of the model. In the first case, for \( c = 7 \), we obtain a significant F-statistics value for the year and department interaction, allowing to get rid of the variance inhomogeneity revealed during the exploratory analysis. In the second case, with \( c = \log N \), we get the most significant sum of F-statistics (sumF), which characterizes this model with a proportional significance for all effects and their interactions.

Table 3. Final partition by the year and the department obtained for \( BIC \) fitness function.

<table>
<thead>
<tr>
<th>Periods for Year</th>
<th>Groups for Department</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1 1991–2002</td>
<td>G1 FB, REEF, ACEF, MTF, PEF</td>
</tr>
<tr>
<td>P2 2003–2005</td>
<td>G2 AEF, HEF, PhEF, MAF</td>
</tr>
<tr>
<td>P3 2006–2013</td>
<td>G3 AMCSF</td>
</tr>
</tbody>
</table>

For the two best partitions by the year selected above, we have two corresponding partitions by the department. For \( c = 7 \), the algorithm produces four groups of departments in the first partition:

- Group I—FB, ACEF, PEF;
- Group II—AEF, HEF, MAF;
- Group III—AMCSF;
- Group IV—REEF, MTF, PhEF;
For $c = \log N$ the algorithm gives three groups of departments in the second partition:

- Group I—FB, ACEF, PEF, REEF, MTF;
- Group II—AEF, HEF, MAF, PhEF;
- Group III—AMCSF.

Comparing the resulting partitions, we see that the first three groups I, II, III of the first partition turned out to be stable. The departments included in them remain in the second partition. In contrast, the departments of the unstable IV group were distributed in the second partition into groups I and II. The most stable group turned out to be group III, consisting of one AMCSF in both partitions. This department is associated with serious requirements for mathematical training.

The partition into different groups of departments according to the level and dynamics of general IQ is associated to a greater extent with the distribution of students among the faculties of a particular university, which is not entirely interesting from the point of view of global and country changes in the structure of intelligence. Moreover, among groups I, II, and IV, we cannot single out a group of completely humanities or, say, completely engineering specialties. Rather, the group assignment is associated with a passing grade for a particular department, which determines the overall level of enrollment. Therefore, for the convenience of interpretation, we will choose a simpler second partition into three groups, especially since such a partition has a larger sum significance (according to F-statistics) for all effects and interactions of the ANOVA model.

Thus, the chosen partitioning by year and department variables is $3 \times 3$, obtained for \textit{BIC} fitness function (last column of Table 2, $c = \log N$). In Table 3, we denote three time periods for the year as P1, P2, P3, and three groups for the department as G1, G2, and G3.

### 4.2.2. Partitioning for IQ Subtests

To arrive at meaningful conclusions, the developed algorithm was applied not only for the original model with the general IQ as a response, but also for models built for all nine IQ subtests. The discretization and grouping results for the IQ subtests differ quite strongly from each other, which complicates the process of interpreting the results.

For the convenience of interpretation, it seems expedient to analyze the results for each of the nine IQ subtests, as well as for the general IQ test, with the same partition of the year and department values. As such a common partition, we have chosen the $3 \times 3$ one given in Table 4 and obtained for the general IQ with the \textit{BIC} goodness function. In order to understand how much this partition deviates from the optimal one, the values of the \textit{BIC} for each IQ component are compared for the following partitions:

- Q0: Natural partition (without discretization and grouping);
- Qopt: Optimal partition for the corresponding subtest;
- Q33: Optimal partition for general IQ ($3 \times 3$ partition in Table 3).

#### Table 4. Comparison of the \textit{BIC} values for the IQ1–Q9 subtests.

<table>
<thead>
<tr>
<th>Subtest</th>
<th>Q0</th>
<th>Qopt</th>
<th>Q33</th>
<th>$(Q33-\text{Qopt})/(Q0-\text{Qopt})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>IQ1</td>
<td>45,347.10</td>
<td>44,410.53</td>
<td>44,540.93</td>
<td>13.92%</td>
</tr>
<tr>
<td>IQ2</td>
<td>47,304.71</td>
<td>46,246.06</td>
<td>46,292.62</td>
<td>4.40%</td>
</tr>
<tr>
<td>IQ3</td>
<td>46,440.84</td>
<td>45,381.67</td>
<td>45,453.85</td>
<td>6.81%</td>
</tr>
<tr>
<td>IQ4</td>
<td>46,376.30</td>
<td>45,580.15</td>
<td>45,622.01</td>
<td>5.26%</td>
</tr>
<tr>
<td>IQ5</td>
<td>48,866.38</td>
<td>47,991.48</td>
<td>48,178.74</td>
<td>21.40%</td>
</tr>
<tr>
<td>IQ6</td>
<td>48,135.41</td>
<td>47,103.49</td>
<td>47,249.93</td>
<td>14.19%</td>
</tr>
<tr>
<td>IQ7</td>
<td>45,913.20</td>
<td>44,893.94</td>
<td>44,945.67</td>
<td>5.07%</td>
</tr>
<tr>
<td>IQ8</td>
<td>48,095.02</td>
<td>47,085.38</td>
<td>47,159.35</td>
<td>7.33%</td>
</tr>
<tr>
<td>IQ9</td>
<td>46,556.39</td>
<td>45,620.72</td>
<td>45,703.43</td>
<td>8.84%</td>
</tr>
</tbody>
</table>

As can be seen from Table 4, the \textit{BIC} value for the $3 \times 3$ partition chosen for interpretation is close enough to the \textit{BIC} value for optimal partition, built specifically for the
IQ subtests. The average deviation is about 10% of the difference between the optimal partition for a particular subtest and the $3 \times 3$ partition presented in Table 3 (the maximum deviation reaches 21.4 percent for the subtest IQ5).

In addition to the larger value of the goodness function, we observe an increase in the homogeneity of variances in the groups, due to the use of the proposed method. Table 5 shows the $p$-value when testing the hypothesis of uniformity of variances across the groups using Levene’s Test. The most striking increase in uniformity is observed for the general IQ, which indicates the effectiveness of the proposed approach. A negative result of this test is observed only for IQ9, which can be explained by deviations from normality for this subtest.

Table 5. P-values for Levene’s Test for Homogeneity of Variance (center = median).

<table>
<thead>
<tr>
<th>Response</th>
<th>Initial Natural Partition</th>
<th>Optimal Partition for the Test/Subtest</th>
<th>Partition $3 \times 3$ (Optimal for General IQ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IQ</td>
<td>$2.67 \times 10^{-5}$</td>
<td>0.314689</td>
<td>0.314689</td>
</tr>
<tr>
<td>IQ1</td>
<td>$1.63 \times 10^{-6}$</td>
<td>0.050505</td>
<td>0.00808</td>
</tr>
<tr>
<td>IQ5</td>
<td>$3.01 \times 10^{-12}$</td>
<td>$1.87 \times 10^{-10}$</td>
<td>0.00367</td>
</tr>
<tr>
<td>IQ8</td>
<td>$3.51 \times 10^{-6}$</td>
<td>0.00069</td>
<td>0.056866</td>
</tr>
<tr>
<td>IQ9</td>
<td>$1.98 \times 10^{-17}$</td>
<td>$7.55 \times 10^{-23}$</td>
<td>&lt;1.0 $\times 10^{-16}$</td>
</tr>
<tr>
<td>Number of groups</td>
<td>176</td>
<td>-</td>
<td>15</td>
</tr>
</tbody>
</table>

4.3. Changes in the IQ Levels and Structure Over Time Periods

Now we apply ANOVA for general IQ and its subtests IQ1–Q9 as a response and year, department, and gender as explanatory variables whose values are categorized according to partitioning $3 \times 3 \times 2$ obtained using the proposed algorithm with BIC fitness function for year and department and natural grouping for gender. Some results of modeling are shown in Figures 3–7.

![Figure 3](image-url)  
Figure 3. Dynamics of general IQ test among male (a) and female (b).
Figure 4. Dynamics of verbal abilities (subtest IQ1) among male (a) and female (b).

Figure 5. Dynamics of mathematical abilities (subtest IQ5) among male (a) and female (b).

Figure 6. Dynamics of spatial generalization (subtest IQ8) among male (a) and female (b).
Changes in general IQ are shown in Figure 3. If we compare the first observation, period P1, with the last, period P3, then the results suggest a slight decrease in general IQ. This decline is significantly observed in the G2 group ($t = -2.06, p\text{-value} < 0.05$), including a wide range of engineering and humanitarian specialties. In the G1 group, including computer, economic, and engineering specialties, a significant change is not observed in both the male ($t = -0.65, p\text{-value} > 0.1$) and female group.

As for the P2 period, it stands out strongly in the level of intelligence in comparison with the neighboring periods P1 and P3, in all groups G1, G2, G3. Moreover, in the G1 group, there is a rather strong decrease ($t = -1.97, p\text{-value} < 0.05$), and in the G2 and G3 groups—a significant increase, which is typical for both genders, especially for the female part of group G3 ($t = 2.13, p\text{-value} < 0.05$).

A more detailed consideration of the intelligence structure is provided by ANOVA models for the subtests IQ1–IQ9. The dynamics of the subtests partly repeat the dynamics of general IQ, but there are also differences both in the groups of departments and by gender. For example, in the subtest IQ1 that reveals verbal abilities (see Figure 4), in the P2 period, there is no significant decrease for the IQ1 value as for general IQ among men in the G1 group ($t = -0.15, p\text{-value} > 0.1$), in contrast to women, where such a decrease is observed. In other groups, the dynamics of verbal abilities are similar to that of general IQ.

The dynamics of mathematical abilities in the G1 group differed both from that of the verbal abilities and general IQ in this group. Thus, in Figure 5, we see a significant increase in IQ5 from the P1 period to the P3 period among men ($t = 2.94, p\text{-value} < 0.005$), as well as a significant increase from the P2 to the P3 period among women ($t = 1.85, p\text{-value} < 0.1$). In the G2 group among women, on the contrary, a stable and rather significant decrease in mathematical abilities is observed from the P1 to the P3 period, without an increase in the P2 period. In the G3 group, there was no significant increase in mathematical ability among men, such as among women.

According to the dynamics of spatial generalization subtest presented in Figure 6, we notice a significant decrease in the P3 period for the G2 group ($t = -3.47, p\text{-value} < 0.001$), with a certain rise in the P2 period, which is typical for other results of this group. In the G1 group, there is also a decrease in IQ8 in the P3 period compared to P1, while we see a multidirectional change in this value in the P2 period for women and men. That is, the G1 group is characterized by significantly greater differences in the structure of intelligence among women and men than in other groups.

Finally, the biggest difference we see in changing short-term memory in Figure 7. Here we see a significant decrease in IQ9 from period P1 to period P3 for men in groups G1 ($t = -5.39, p\text{-value} < 0.001$) and G2 ($t = -4.57, p\text{-value} < 0.001$). In the G3 group, as for other tests, we again observe a slight increase in the P2 period compared to the period P1.
Thus, we can conclude that only for students from group G3 there is a stable increase in the intellectual ability both for general IQ and its subtests.

5. Discussion

In the present paper, we conducted a study of changes in the IQ and its subtests using the ANOVA model. The main problem in conducting long-term IQ studies is that it is not always possible to develop a sample design. As a result, the analyzed sample is characterized by an uneven distribution for a number of features that characterize the participants in the experiment. Ultimately, this does not obtain optimal estimates of the effects in the ANOVA model. To overcome this problem, the algorithm for joint discretization and grouping of explanatory variables in the ANOVA model has been proposed. It is based on generalized statistical criteria that include, as special cases, the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). These criteria, although they belong to statistical ones, are rather closely related to information-theoretic criteria. Therefore, the results of discretization and grouping based on them cannot only ensure the optimal balance in accuracy and complexity of the mathematical model, but also consider all the information contained in the collected data.

A standard supervised agglomerative discretization algorithm has been significantly improved. Instead of one best solution at each iteration of the algorithm, it is suggested to save a fixed number $m$ of best solutions and then to perform the subsequent merging of categories for them. It is revealed that the original greedy version of the algorithm ($m = 1$) never gave a global optimum. Consequently, the proposed modification of the algorithm with a partial screening of solutions provides the best result. However, with increasing $m$, a worse solution can be obtained, so it is impossible to recommend a certain value $m$ at which the best results are guaranteed. Therefore, with each new application of the algorithm, a preliminary adjustment is required, which consists of calculating for several consecutive values of $m$ and choosing the best solution $m^*$. In present study $m^*$ took the value between 2 and 12 (equals 2 in half of the experiments).

5.1. Interpretation of the Results

When interpreting the results obtained, of particular interest is separating a relatively short period of time 2003–2005 (P2) in the optimal partitioning by the year. This result suggests that the period contains significant information about the change in the intelligence of students at that time. Indeed, this relatively short period of IQ testing is represented by young people at the age of 17–20 years, whose school education began in 1991–1995, that is, in the most active period of the restructuring of Russian society.

In 1992, the market was liberalized, as a result, production fell in almost all spheres of population activity, and prices rose almost 30 times. The fall in GNI led to a significant decrease or even cessation of financing science, education, health care, and culture. The period of 1993–1994 represented an initial stage of privatization related to the formation of the private sector of the economy, which, however, did not result in an increase in the production and welfare of the population.

As a result of the crisis in industry, trade, and the economy in general, the standard of living has dropped significantly. The population in various ways adapted to these extraordinary transformations in all spheres of life: Political, economic, and social, which could not but affect the consciousness of children aged 6–10 years. Thus, the first distant reason for identifying the periods of the discovered IQ variability in the period 2003–2005 may be a complex impact of changes in the socio-cultural environment on the development of children’s intellectual abilities and their ideas about the role of higher education and lifestyle choices.

Stable (regardless of the analysis method) increase in the IQ when comparing the 2003–2005 period (P2) with the 1991–2002 period (P1) is observed for the group represented by the students majoring in mathematics (group of departments G3), and a more pronounced rise was characteristic of the female part of group G3 (see Figure 3).
A greater increase in the IQ (Flynn effect) for women than for men is explained by indicators of improvement in living conditions and the social status of the corresponding population groups [69]. However, the effect we discovered for the period of 2003–2005 (P2) can be more likely caused by changes in socio-cultural stereotypes of the behavior of women who are forced to master new professions and expand their scope of activity when adapting to new living conditions. This explanation of the results obtained is supported by a noticeable decrease in the disparity between men and women in science, engineering, and mathematics in the sample of talented students in the period of 1981–2010, which is also considered a consequence of the influence of socio-cultural factors [70].

Another group of departments with an increase in the IQ during the P2 period is made up of students of very different areas of training: Physical and engineering, aircraft, and humanities education (group G2). With a general decline in the prestige of science and engineering knowledge during the Russian transition period (usually called “perestroika”), due to the collapse of the economy, these departments could be chosen by young people with an idealistic and romantic outlook on the future, in contrast to more pragmatic majors presented in the economic, power engineering, mechanical-technological or radio engineering departments (group G1). Against the background of fluctuations in life values and ideals during the “perestroika” period, the desire to implement one’s dreams in the G2 group could support the motivation of the educational process and the formation of the knowledge system necessary for successful IQ testing. This hypothesis can be confirmed by an increase in IQ in the G2 group according to the IQ1 subtest designed to assess not only verbal ability, but also general awareness. However, the P2 period did not affect the IQ level for men from the G1 group, and there was even a decrease in the IQ1 component in the female group (see Figure 4).

Considering the dynamics of IQ in the period of 2006–2013 (P3), it should be noted that significant changes took place at that time in the system of Russian school education. The Unified State Exam (USE) was introduced, the results of which began to be used gradually from 2003 to 2007, depending on the region, for final assessment of secondary school education quality and subsequent competitive admission to higher educational institutions.

Thus, the identified time periods of different IQ dynamics based on the application of the proposed method can be related to different transformations in Russian society, including changes in the education system. The stable allocation of a specific period of 2003–2005 in the variability of IQ dynamics, corresponding to various combinations of optimality criteria, apparently reflects the effect of cardinal transformations in both the socio-economic sphere of life and the education system on the development and assessment of the intellectual abilities of young people.

In the period of 1991–2005 (P1 and P2), IQ testing was passed by students who had knowledge acquired according to a unified curriculum at school and served as the basis for entrance exams to higher educational institutions. The next period of 2006-2013 (P3) was characterized by the consequences of pedagogical diversity in the use of school curricula, however, with a single regulation for assessing knowledge using the USE. One of the psychological factors in testing with a time limit may be the development of problem-solving skills based on guessing the answers, rather than on a strategy of thinking about possible alternatives. Such a guessing strategy when testing intellectual abilities will inevitably lead to an increase in the number of incorrect answers, and accordingly, to a low IQ. As the obtained data show, only in the female sample of students who chose economic and engineering departments (G1), there is an increase in IQ in the P3 period. In other cases, there are either no temporal differences, or the IQ in P3 is lower than in P2 or P1 (see Figure 5).

It should be noted that a comparison of the features of the IQ dynamics under the influence of the year and department variables when performing various Subtests IQ1–IQ9 indicates the relative stability of the indicators of visual-spatial abilities (IQ8, see Figure 6) or short-term memory (IQ9, see Figure 7), in comparison with greater variability of the
verbal or mathematical components of intelligence (see, for example, Figures 4 and 5). The relatively large variability of these components in P3, therefore, can be caused by significant changes in school curricula (introduction of the USE) and the assessment of the knowledge acquired during this training.

An additional cause of a decrease in memory observed in the P3 period (see Figure 7) can probably be the ever-growing informatization of society and the use of personal computers and mobile devices by students as a reference book. Along with making it easier to obtain information, systematic access to the Internet as a mnemonic assistant reduces the load on memory, leading to an observed decrease in memorization efficiency.

5.2. Conclusions

The obtained results of the analysis of the intellectual abilities of young people in the period of 1991–2013 (i.e., children born in 1974–1997) allow us to conclude that: The observed features of the temporal dynamics of the IQ are caused by changes in the education system and the socio-economic status of the family, which occurred in Russia during the period of restructuring the society, and an intensive development of information technologies. The revealed periods of 2001–2002 and 2003–2005 with a pronounced variability of the IQ indicators of students who chose different areas for training, apparently reflect the transitional conditions in the awareness of the value of knowledge and the prestige of different professions, due to a significant reorganization of the socio-cultural environment of education and training of children born in 1974–1997.

The results obtained in the present study are consistent with the conclusions presented by the authors of [30] on the dynamics of the IQ for children born in 1974–1999. The respondents in the study were men aged 18–40 years with a level of education not lower than the general secondary education. The period 2001–2002 (1983–1984 year of birth) corresponds to the lowest IQ values, while for 1985–1987, a significant jump in IQ values was found. Similarly, in our study, for men, such a jump in average IQ values was revealed during the observation period of 2003–2005 for students of very different areas of training: Physical and engineering, aircraft, and humanitarian and mathematical.

Moreover, our results are generally consistent with the trends found in European countries. For example, in Norway [31], in cohorts born in 1975–1991, a continuous decrease in the IQ values was revealed. At the same time, some fluctuations in values are observed in the early 1980s of birth. The authors of [31] used data on conscription, thus, it mainly covered the male population. In our studies, if we exclude the unstable period of 2003–2005 (1985–1987 years of birth), then, in general, there is a decrease in IQ for men of different specializations (except for mathematical orientation, for which there is not enough data after 2006).


Funding: This research was funded by Ministry of Science and Higher Education of the Russian Federation (project No. FSUN-2020-0009).

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and with the principles of the Code of Ethics of the Russian Psychological Society.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data are not publicly available. The data that support the findings of this study can be obtained from the data curation author upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.
 Appendix A

<table>
<thead>
<tr>
<th>Study Group</th>
<th>Sample Size</th>
<th>Years</th>
<th>Country</th>
<th>IQ Tests</th>
<th>Factors</th>
<th>Model and/or Method</th>
<th>Effect</th>
<th>Authors (Reference)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Psychometric IQs, student assessment tests</td>
<td>Different Groups at different periods in different countries</td>
<td>262 periods between 1909 and 2013</td>
<td>13–28 nations: Argentina, Belgium, Bulgaria, China, Denmark, Finland, Ireland, Kenya, Norway, Switzerland, Sudan, and Turkey</td>
<td>Wechsler tests</td>
<td>Education vs. GDP per capita, growth</td>
<td>n-weighted Fishers-z-transformation to calculate a mean and correlations (r), and standardized regression coefficients (β).</td>
<td>In developed countries, only minor increases of cognitive ability, but the increases at low cognitive ability levels</td>
<td>[7]</td>
</tr>
<tr>
<td>Three military authorities in charge of psychological testing, 21 educational research institutes in Western Europe</td>
<td>Different groups per country</td>
<td>1952–1982</td>
<td>European country, plus Australia, Canada, Greenland, Iceland, and New Zealand</td>
<td>Ravens Progressive Matrices Test and the Wechsler Adult Intelligence Scale</td>
<td>Different age groups</td>
<td>The difference between the means, dividing by the average of the standard deviations</td>
<td>Flynn effect: massive gains on all kinds of IQ tests</td>
<td>[21]</td>
</tr>
<tr>
<td>Russia-born males, aged 18–40 years</td>
<td>238,363</td>
<td>2012–2018</td>
<td>Russia</td>
<td>Short screening test (SST-30)</td>
<td>Age, Education, Territorial affiliation</td>
<td>Descriptive statistics</td>
<td>The dynamics of IQ scores by year of birth showed a decrease in IQ scores by the mid-1980s. XX century (relative to the level of the late 1970s), replaced by growth, close to linear.</td>
<td>[30]</td>
</tr>
<tr>
<td>Norwegian-born males, aged 18–19 years</td>
<td>736,808</td>
<td>1962–1991</td>
<td>Norwegian</td>
<td>Three speeded tests of arithmetic (30 items), word similarities (54 items), and figures (36 items).</td>
<td>Birth-order effect within-family Flynn effects</td>
<td>Standard fixed-effects model Bayesian model for sibling pairs</td>
<td>For the 1962–1975 Flynn increase period. For the 1975–1991 decrease period</td>
<td>[31]</td>
</tr>
<tr>
<td>Study Group</td>
<td>Sample Size</td>
<td>Years</td>
<td>Country</td>
<td>IQ Tests</td>
<td>Factors</td>
<td>Model and/or Method</td>
<td>Effect</td>
<td>Authors (Reference)</td>
</tr>
<tr>
<td>----------------------------------------------------------------------------</td>
<td>-------------</td>
<td>------------------------</td>
<td>-------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>Individuals aged 50–84 from the Survey of Health, Aging, and Retirement</td>
<td>92,739</td>
<td>2004/2005 and 2013</td>
<td>10 European countries: Austria, Belgium, Denmark, France, Germany, the Netherlands, Sweden, and Switzerland, Italy, and Spain</td>
<td>Immediate word recall</td>
<td>Changes in socio-demographic and health conditions, including decreases in cardiovascular disease, physical activity, and educational achievement</td>
<td>Multivariate random intercept models</td>
<td>Word recall improved 4.40 to 5.08 words, p &lt; 0.05</td>
<td>[37]</td>
</tr>
<tr>
<td>Thirty-three American adolescents, aged 13–18</td>
<td>10,073</td>
<td>1989 and 2003</td>
<td>USA</td>
<td>The Kaufman Brief Intelligence Test (nonverbal sections)</td>
<td>Age, sex, ability level, parental age, and SES</td>
<td>Effect sizes were also calculated by dividing the adjusted mean difference by the overall sample standard deviation</td>
<td>Overall, the Flynn effect was not significant but the effects varied substantially by age and ability level</td>
<td>[38]</td>
</tr>
<tr>
<td>Cognitively normal participants</td>
<td>204 and 177</td>
<td>1991–1997 and 2008–2009</td>
<td>France</td>
<td>Neuropsychological test battery</td>
<td>Age, 2 samples born in 1923 on average (the younger—from the 1991 sample and the older—from the 2008 sample)</td>
<td>t-test</td>
<td>The Flynn-like effect</td>
<td>[39]</td>
</tr>
<tr>
<td>American adult household members of all ages</td>
<td>25,555</td>
<td>1972–2008</td>
<td>USA</td>
<td>Vocabulary test from the twenty-item Gallup-Thorndike Verbal Intelligence Test</td>
<td>The groups were combined by decade, giving four decade groups: 1970s, 1980s, 1990s, and 2000s</td>
<td>Exploratory factor analysis, confirmatory factor analysis</td>
<td>A decline in the 1980s, and a steady increase throughout the 1990s and 2000s</td>
<td>[40]</td>
</tr>
</tbody>
</table>
Table A1. Cont.

<table>
<thead>
<tr>
<th>Study Group</th>
<th>Sample Size</th>
<th>Years</th>
<th>Country</th>
<th>IQ Tests</th>
<th>Factors</th>
<th>Model and/or Method</th>
<th>Effect</th>
<th>Authors (Reference)</th>
</tr>
</thead>
<tbody>
<tr>
<td>American adult, aged 40–45</td>
<td>WAIS-R ( n = 1800; ) WAIS-III ( n = 2450; ) WAIS-IV ( n = 2200 )</td>
<td>1981, 1997, 2008</td>
<td>USA</td>
<td>The Wechsler Adult Intelligence Scale</td>
<td>WAIS-R, III, IV racial/ethnic groups 13 age groups</td>
<td>The latent variable model of the subtest scores; likelihood ratio test</td>
<td>An increase in intelligence depended on instrument of testing</td>
<td>[41]</td>
</tr>
<tr>
<td>Ninety-three Danish males</td>
<td>1988 ( n = 33,833, ) 1998 ( n = 25,020 ) 2003/4 ( n = 23,398 )</td>
<td>1998, 2004</td>
<td>Denmark</td>
<td>Letter Matrices, Verbal Analogies, Number Sequences, Geometric Figures</td>
<td>Year</td>
<td>The intercorrelations (Pearson’s Product-Moment and Point-Biserial, as appropriate)</td>
<td>Scores on all four tests declined</td>
<td>[43]</td>
</tr>
<tr>
<td>Adolescents, aged 14–21</td>
<td>12,686</td>
<td>1986–2000</td>
<td>USA</td>
<td>PIAT-M</td>
<td>Year Education household income</td>
<td>Regression models</td>
<td>Children with more educated mothers and/or children born into higher income households had an accelerated Flynn effect</td>
<td>[44]</td>
</tr>
<tr>
<td>Six categories of participants, aged 8–15</td>
<td>4932</td>
<td>1977, 2010</td>
<td>Saudi Arabia</td>
<td>Raven’s Standard Progressive Matrices</td>
<td>Age</td>
<td>t-statistic</td>
<td>Scores on all four tests declined</td>
<td>[46]</td>
</tr>
</tbody>
</table>
Table A1. Cont.

<table>
<thead>
<tr>
<th>Study Group</th>
<th>Sample Size</th>
<th>Years</th>
<th>Country</th>
<th>IQ Tests</th>
<th>Factors</th>
<th>Model and/or Method</th>
<th>Effect</th>
<th>Authors (Reference)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twelve UK regions</td>
<td>not indicated</td>
<td>2011–2013</td>
<td>UK</td>
<td>Immediate word recall Serial subtraction Number series</td>
<td>Socioeconomic variables</td>
<td>Pearson correlation Principal Components Analysis</td>
<td>Factor of socioeconomic development is correlated with regional IQ</td>
<td>[47]</td>
</tr>
<tr>
<td>Subjects, aged 30–63</td>
<td>79</td>
<td>1999 2008</td>
<td>France</td>
<td>WAIS III and WAIS IV</td>
<td>The subtests in the French WAIS III and WAIS IV</td>
<td>Scaled score means and standard deviations</td>
<td>IQ declined by 3.8 points</td>
<td>[48]</td>
</tr>
<tr>
<td>Students tested for special education services from 10 school districts</td>
<td>10,800</td>
<td>1964–2000</td>
<td>USA</td>
<td>WISC-R and WISC-III</td>
<td>Age Testing</td>
<td>D-score approach Individual growth modeling</td>
<td>The relationship between IQ and age is complex</td>
<td>[49]</td>
</tr>
<tr>
<td>Nine children, aged 13–16</td>
<td>1025</td>
<td>1971 and 2015</td>
<td>Czech Republic</td>
<td>Amthauer’s IST</td>
<td>Age</td>
<td>Welch’s t-test</td>
<td>The Flynn effect</td>
<td>[50]</td>
</tr>
<tr>
<td>Nine children, aged 5–6</td>
<td>1628 and 1195</td>
<td>1984 and 2006</td>
<td>China</td>
<td>The Chinese Preschool and Primary Scale of Intelligence</td>
<td>The 1984 norm sample and 2006 sample; gender</td>
<td>t-statistic</td>
<td>A gain of 2.06 IQ points per decade</td>
<td>[51]</td>
</tr>
<tr>
<td>Meta-analysis</td>
<td>13,172</td>
<td>1977–2014</td>
<td>German-speaking countries</td>
<td>Test for spatial perception</td>
<td>population vs. mixed samples vs. university students</td>
<td>multiple weighted regression</td>
<td>IQ score decrease</td>
<td>[52]</td>
</tr>
</tbody>
</table>
References

2. Lynn, R.; Meisenberg, G. National IQs calculated and validated for 108 nations. Intelligence 2010, 38, 353–360. [CrossRef]
11. Spearman, C. ‘General intelligence’ objectively determined and measured. Am. J. Psychol. 1904, 15, 201–293. [CrossRef]
17. Lynn, R. IQ in Japan and the United States shows a growing disparity. Nature 1982, 297, 222–223. [CrossRef]
18. Flynn, J.R. Race, IQ and Jensen. Intelligence 1977, 49, 1–9. [CrossRef]
20. Williams, R.L. Overview of the Flynn effect. Intelligence 2013, 41, 753–764. [CrossRef]
27. Williams, R.L. Overview of the Flynn effect. Intelligence 2013, 41, 753–764. [CrossRef]


42. Teasdale, T.W.; Owen, D.R. Continuing secular increases in intelligence and a stable prevalence of high intelligence levels. *Intelligence* 1989, 13, 255–262. [CrossRef]

43. Teasdale, T.W.; Owen, D.R. Secular declines in cognitive test scores: A reversal of the Flynn Effect. *Intelligence* 2008, 36, 121–126. [CrossRef]

44. Ang, S.; Rodgers, J.L.; Wänström, L. The Flynn Effect within subgroups in the US: Gender, race, income, education, and urbanization differences in the NLSY-Children data. *Intelligence* 2010, 38, 367–384. [CrossRef]


50. Laciga, J.; Cigler, H. The Flynn effect in the Czech Republic. *Intelligence* 2017, 61, 7–10. [CrossRef]


70. Wai, J.; Putallaz, M. The Flynn effect puzzle: A 30-year examination from the right tail of the ability distribution provides some missing pieces. *Intelligence* 2011, 39, 443–445. [CrossRef]