IOP Conf. Series: Journal of Physics: Conf. Series **803** (2017) 012095 doi:10.1088/1742-6596/803/1/012095

# A Technique for Assessment of Parallelism of Specialized Test Variants Using Statistical Methods

**O V Maruhina[1,2], O M Zamyatina[1,2], P I Mozgaleva[1] and I V Papina[1]**

[1] Tomsk Polytechnic University, Tomsk, Russia
[2] Tomsk State University, Tomsk, Russia

**E-mail:** papinayuliya@gmail.com

**Abstract.** The paper offers a technique for evaluation of parallelism of ongoing student assessment test variants on the basis of statistical methods through the example of tests in mathematics. The authors analyze the test results of National Research Tomsk Polytechnic University students and make adequate conclusions regarding parallelism of text variants.

## 1. Introduction

Independent assessment and examination of students present an important issue. A large number of techniques for independent and objective assessment of students' knowledge has already been developed and proposed [1-3]. However, the issue of selecting invariant tasks remains highly topical. One of the possible ways of solving this problem is to develop a technique for evaluation of parallelism of test variants on the basis of mathematical and statistical methods.

Presently, the education system regards independent examination of students as the most objective assessment of their knowledge and the potential of their intellectual capabilities. In 2003, Russia became a signatory of the Bologna Declaration, which pronounced the creation of the united European Higher Education Area that relies on three general principles: three-tier higher education (Bachelor – Master – Ph.D. student), a system of credit units (credits), and the use of internationally recognized education quality assessment methods [4]. The selected testing method used for education quality assessment is widely used around the world. In the USA, testing is carried out by several companies, the largest being the Educational Testing Service (ETS), operating since 1947. By requests from educational institutions, government agencies, and individuals from 180 countries, it performs over 12 million tests every year and holds more than 60 patents for various testing devices and technologies [5].

The main drawback of knowledge monitoring and assessment by means of classical examination is its inherent bias. Adolf Melezinek, an engineering pedagogy theorist renowned in Western Europe, believes that by monitoring learning outcomes using the traditional exam, we obtain a result that is subjective at worst and relative at best, but is in no way objective and absolute [3]. An exam result is subjective when a teacher puts a mark bearing in mind the factors that are not directly relevant to a testee's knowledge and skills; for example, a teacher gives a higher grade to a diligent student with less aptitude and superficial knowledge than to a more capable, but less diligent student. An exam result is relative when a teacher is a priori convinced that there exists a "natural" allocation of results: only a small part of students can achieve very good results, while another small part will necessarily have bad results, and most students will have average results. Such approach leads to a decrease of motivation in

most students, and they stop working as well as they are able to; moreover, comparing the performance of different teachers, groups, departments, and universities becomes impossible.

Education reforms in Russia along with the current strategy of Tomsk Polytechnic University (TPU), namely the integration into the international educational environment and increasing the university's competitiveness, increase the demand for a testing system. In this context, a system for independent assessment of students' knowledge in general subjects has been developed. The materials for monitoring and control of subject knowledge are presented in several variants. The technique proposed and developed by the authors was evaluated by the example of the "Mathematics" discipline, which includes twenty-one variant of test activities. This leads to the issue of parallelism ("sameness") of these test variants and, as a consequence, that of students' knowledge assessment quality and its objectivity.

This way, the object of this research is the students of TPU that participate in the process of ongoing assessment of the mathematics knowledge quality, and the subject is the variants of test activities (the present research deals with a test in higher mathematics). The test involved a total of 1001 persons, all of them are first year students of engineering majors from TPU.

Table 1 shows the results of the initial processing of the data from the mathematics test with twenty-one variants. Assessment of each variant's difficulty $\delta_j$ was carried out according to the algorithm described in [6, 7]. The research involved the LogitModels test assessment software developed by the authors.

**Table 1.** Test data initial processing results.

| Test variant | Number of test subjects | Failed the test | Completed all test activities | Test difficulty $\delta_j$ |
|---|---|---|---|---|
| *V1* | 49 | 0 | 0 | –0.01 |
| *V2* | 50 | 0 | 0 | –0.04 |
| *V3* | 59 | 0 | 0 | 0.12 |
| *V4* | 52 | 0 | 0 | 0 |
| *V5* | 53 | 1 | 0 | 0.22 |
| *V6* | 42 | 0 | 0 | 0.15 |
| *V7* | 46 | 0 | 0 | 0.04 |
| *V8* | 41 | 2 | 0 | 0.23 |
| *V9* | 42 | 1 | 0 | 0.06 |
| *V10* | 48 | 1 | 0 | 0.10 |
| *V11* | 50 | 1 | 0 | 0.09 |
| *V12* | 38 | 0 | 0 | 0.10 |
| *V13* | 45 | 0 | 0 | 0.03 |
| *V14* | 48 | 0 | 0 | 0.39 |
| *V15* | 56 | 2 | 0 | 0.11 |
| *V16* | 43 | 0 | 0 | –0.05 |
| *V17* | 45 | 1 | 0 | 0.24 |
| *V18* | 53 | 0 | 0 | 0.02 |
| *V19* | 43 | 0 | 0 | 0.38 |
| *V20* | 50 | 0 | 1 | –0.12 |

| | | | | |
|---|---|---|---|---|
| *V21* | 48 | 0 | 0 | 0.02 |

Figure 1 shows the graph of change of mathematics test variant difficulty.

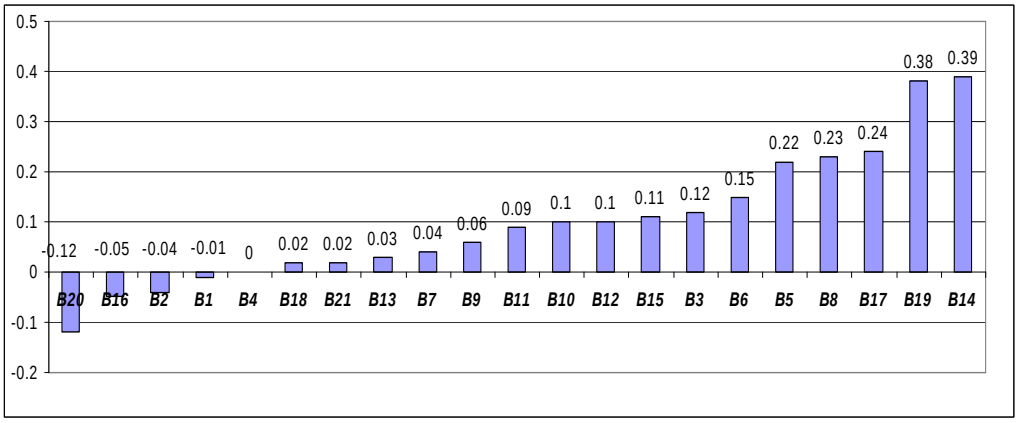


**Figure 1.** Magnetization as a function of the applied field. Note how the caption is centered in the column.

Coefficient of variation was used to assess the significance of variation of test variant difficulty values (1):

$$Cv = \frac{S_x}{\bar{x}} \cdot 100\ \%\ , \tag{1}$$

where $S_x$ is the standard deviation of test variant difficulty distribution, and $\bar{x}$ is the mean value of test variant difficulty. Different properties are characterized by different coefficients of variation. However, in relation to the same property, value *Cv* of the same indicator remains more or less stable and usually does not increase above 50 % under symmetric distributions. In case of severely asymmetrical distribution series, the coefficient of variation can reach 100 % and even above that. For the test under consideration, the coefficient will be:

$$Cv = \frac{S_x}{\bar{x}} \cdot 100\ \ \% = (0{,}099/0{,}132) \cdot 100\ \% = 134\ \%.$$

Variation is considered significant when $Cv > 25\ \%$, which means that the studied population is considered heterogeneous and, subsequently, the test variants are non-parallel in their difficulty.

**2. Assessment of variant homogeneity**

Correlation coefficient formula was used to assess the relation between the results of performing two test tasks or test variants (2):

$$\varphi_{jl} = \frac{p_{jl} - p_j \cdot p_l}{\sqrt{p_j \cdot q_j \cdot p_l \cdot q_l}}, \tag{2}$$

where j and l are test task numbers, $p_{jl}$ is the proportion of subjects, who completed both test tasks correctly, i.e. those, who earned one score for both tasks; $p_j$ and $p_l$ are the proportions of subjects, who successfully completed tasks j and l, respectively; qj and ql are the proportions of subjects, who failed to correctly complete tasks j and l; $q_i = 1 - p_j$; $q_l = 1 - p_l$.

## 3. Estimation of parallelism on the basis of cluster analysis

In order to determine the groups of mathematics test variants that are homogeneous in difficulty, cluster analysis (Ward's method and *k*-means) was employed. Cluster analysis is intended for splitting a set of objects into a defined or unknown number of classes basing on a certain mathematical criterion of classification quality (*cluster* — bunch, beam, collection, or group of elements characterized by some common property). The criterion of the classification quality reflects to a certain extent the following informal requirements:

*A.* The objects inside the groups should be closely associated with each other;

*B.* Objects of different groups should be far from each other;

*C.* Under otherwise equal conditions, distribution of objects into groups should be uniform.

Requirements 1 and 2 express the standard concept of class partitioning compactness; requirement 3 serves to prevent the criterion from forcing the joining of separate object groups.

Many procedures in clustering are carried out in a step-by-step manner. This means that two most adjacent objects $x_i$ and $x_j$ are grouped and viewed as a single cluster. Consequently, the number of object decreases and becomes equal to n – 1, with one cluster containing two objects, and the others – only one. This process can be repeated until all of the group's objects are joined into one cluster. The most suitable partitioning is usually selected by the researcher himself, who is provided with a dendrogram reflecting the results of grouping the objects on all steps of the clusterization algorithm.

Traditionally, classification is distinguished into hierarchical and non-hierarchical (sometimes called structural). The algorithms of obtaining these classifications can be distinguished respectively.

The operating principle of hierarchical algorithms lies in successive joining into cluster of the most adjacent elements first, and then the more remote elements. Most of these algorithms rely upon the similarity (distance) matrix, and each separate element is initially viewed as a separate cluster. The general pattern of such hierarchical grouping can be represented as repetitive application of three operations to measures of distance "object (cluster) — object (cluster)":

*A.* Find minimum distance $d_{S_1,S_2}$ between object (cluster) S1 and object (cluster) S2

*B.* Join S1 and S2 into a single cluster, assigning it group index $S_1 \cup S_2$

*C.* Calculate distance $d_{S,S_1 \cup S_2}$ from cluster $S_1 \cup S_2$ to any other object (cluster) S.

Table 2 shows the results of clusterization using the k-means method (grouping of test variants by their difficulty).

The desired number of clusters remained equal to 3.

**Table 2.** Results of clusterization using the k-means method.

| **Cluster 1** | ***V2*** | ***V7*** | ***V8*** | ***V9*** | ***V11*** | ***V13*** | ***V16*** | ***V17*** | ***V18*** | ***V21*** |
|---|---|---|---|---|---|---|---|---|---|---|
| Distance to the center of the cluster | 0.70 | 0.62 | 0.46 | 0.85 | 0.74 | 0.59 | 0.75 | 0.98 | 0.62 | 0.86 |
| **Cluster 2** | ***V1*** | ***V3*** | ***V6*** | ***V10*** | ***V12*** | ***V15*** | | | | |
| Distance to the center of the cluster | 0.60 | 0.74 | 0.54 | 0.69 | 0.91 | 0.47 | | | | |
| **Cluster 3** | ***V4*** | ***V5*** | ***V14*** | ***V19*** | ***V20*** | | | | | |
| Distance to the center of the cluster | 0.69 | 0.75 | 0.82 | 0.96 | 0.84 | | | | | |

The graph below (Figure 2) shows the mean values of all variables for separate clusters.

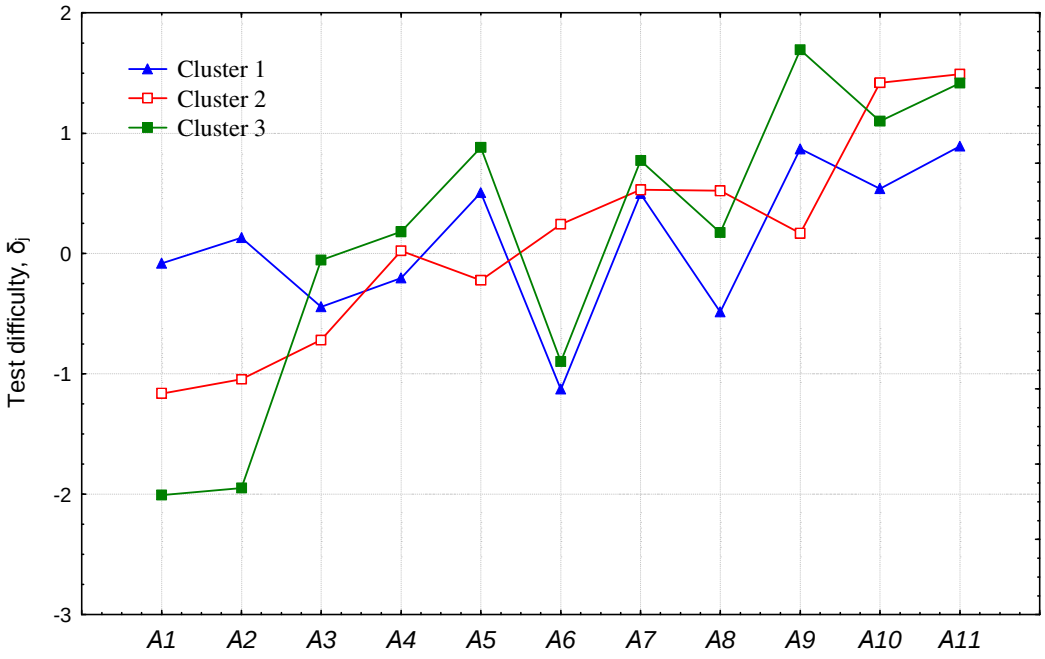


**Figure 2.** Graphs of mean values for each cluster by variables (tasks).

The result of cluster analysis indicated the difference (non-parallelism) of the variants of the mathematics test. When comparing the mean difficulty values of each cluster by task, it should be noted that tasks *A1*, *A2*, *A4*, *A5*, *A6*, and *A7* are the most heterogeneous, which is clearly seen on the graphs of each cluster's mean difficulty values.

This way, cluster analysis has defined three clusters, inside each of which the variants are parallel, but these groups are non-parallel in difficulty between each other. It is recommended to introduce two different systems for mathematics test result scaling for two clusters [10, 11].

As a result of the research, the authors of the paper developed a technique for assessment of parallelism of test variants, which includes the following stages:

*D.* Calculation of standard values of task difficulty for twenty-one variants using a specialized LogitModels software [12];

*E.* Systematization of results of standard values of difficulty levels in the form of matrices of difficulty level mean values by test tasks of test variants;

*F.* Check of test variant parallelism according to the following criteria:

a. coefficient of variation of mean standard values of variants' difficulty level, where the standard of comparison is the value of the coefficient of variation (must be $Cv < 25$ %);
b. correlation analysis of the mean standard values of tasks' difficulty level and the mean standard values of variants' difficulty level; here, the standard of comparison is the correlation coefficient (must be $r_j \rightarrow 1$), which indicates the strong positive relation of variants;
c. cluster analysis (must have one group of variants).

## 4. Conclusion of parallelism.

The flowchart of the developed technique is given in Figure 3.

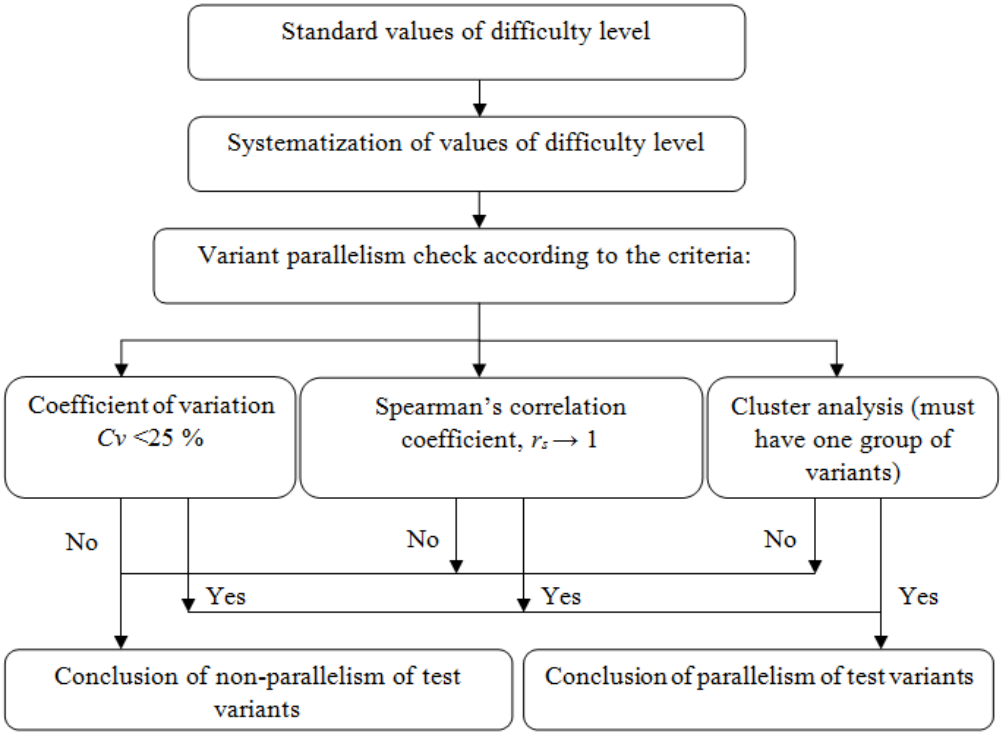


**Figure 3.** The flowchart of the developed technique for assessment of parallelism of test variants

The check of variants' parallelism according to the criteria specified in the developed technique has shown that all three criteria indicate the acceptance of the null hypothesis of non-parallelism of the test variants.

## 5. References

[1] Aderhold J, Davydov V Yu, Fedler F, Klausing H, Mistele D, Rotter T, Semchinova O, Stemmer J and Graul J 2001 *J. Cryst. Growth* **222** 701
[1] Zamyatina O M and Mozgaleva P I 2014 *Global Engineering Education Conf.: Engineering Education Towards Openness and Sustainability* (Istanbul: EDUCON) pp 114-8
[2] Fonseca P, Casanovas J, Montera J and Carmona C 2014 *4ta. Conferencia Iberoamericana en Sistemas, Cibernetica e Informatica, CISCI 2005, Memorias* **1** 158-63
[3] Melezinek A 1993 *European Journal of Engineering Education* **18** 13-6
[4] Zamyatina O M, Mozgaleva P I, Solovjev M A, Bokov L A and Pozdeeva A F 2013 *World Applied Sciences Journal* **27 (13A)** 433-8
[5] Stas N F, Mamontov V V, Knyazeva E M and Galanov A.I. 2009 *Sovremennye problemy nauki i obrazovaniya* **5** 43-8
[6] Marukhina O V 2003 *Algoritmy obrabotki informatsii v zadachakh otsenivaniya kachestva obucheniya studentov vuza na osnove ekspertno-statisticheskikh metodov: Dis. kand. tekhn. nauk: 05.13.01.* (Tomsk) p 165
[7] Neyman Yu M and Khlebnikov V A 2000 *Vvedenie v teoriyu modelirovaniya i parametrizatsii pedagogicheskikh testov* (Moscow: Prometey) p 168
[8] Pimenov DY, Guzeev V I, Koshin A A and Pashnyov V A 2015 *Russian Engineering Research* **35** 64-8
[9] Kytmanov A A 2010 *Programming and Computer Software* **36** 103-108
[10] Radkevich M M 2001 *Kuznechno-Shtampovochnoe Proizvodstvo (Obrabotka Metallov Davleniem)* pp 18-23

[11] Dinis R and Montezuma P 2016 *Electronics Letters* **52** 972-4
[12] Zamyatin A and Cabral P 2011 *DYNA (Colombia)* **78** 42-50