# Electronic testing of knowledge and factors influencing its results

***Pavel Horovčák***[1] *and* ***Beáta Stehlíková***

***Elektronické testovanie znalostí a faktory vplyvu na výsledky***

*Didactic test in electronic version presents fast and precise, modern and effective form of feedback from students to teacher. Electronic test, as a highly formalised instrument of evaluation of the students' preparation and knowledge, has its own unique place in the whole education process. The contribution suggests on practical applications how to prepare and use the electronic test, based on experiences with exploitation and operation of electronic test during verification of student's knowledge. The main consideration is devoted to testing conditions, quantification of criteria for evaluation of results, analysis of success factors like: repeatability and reproducibility of test, duration of testing, appropriate formulation of test questions, awareness of students about the questions. The article specifies selected aspects of quantification of criteria for evaluation (quality) of results of electronic knowledge testing.*

***Key words:*** *electronic test, test preparation, internet, form, database, normal distribution, Sign Test, Mann-Whitney Test*

## Introduction

Modern cybernetics states that more periodic proof of arbitrary process causes its more effective control. Testing and evaluation of students' knowledge is an organic part of every study and education process. There is the place for proofing and correction of learning process after explanation of educational material and its prospective self-study.

The functions of diagnostic processes are following: testing, feedback, classification, verification, motivation as well as education.

The primary sphere of using the didactic test is an effective form of testing in educational process. Such tests can present some feedback from teacher point of view and testing and valuation of knowledge from the student point of view. Modern Internet technologies enable didactic test realization in electronic form, what brings new possibilities of testing exploitation. The relevance of electronic testing grows with increasing number of students, with the progressive intrusion of e-learning as well as with the progressive intrusion of distance education into the educational process.

Examination of knowledge in standard way and methods (in writing) is coupled with several disadvantages, such as e.g. a sizeable teacher's time consumption for test preparation, the same questions' set for each student, a virtually zero repeatability of test and the large time consumption for review and valuation of results.

Examination of knowledge in modern way (using didactic test in electronic form) essentially minimizes above-mentioned disadvantages apart from the first one (i.e. time consumption for test preparation). Each student obtains own test's version, test correction and valuation is executed immediately and test results are saved into database tables.

The second sphere of using the test is how to evaluate learning results objectively. There are two methods for objective evaluation of learning results. The first method examines evaluation before its actual realization (a prior procedures); the second one relies on performance of the evaluation process after it has been performed (a posterior procedures). Most of the teachers combine those methods in their evaluation process, so they can achieve as objective result of the evaluation process as possible.

During the school year 2003-2004, we introduced electronic testing of knowledge and realized some planed experiments to verify influence of some factors to tests results. The mentioned electronic test was developed at our department. This article summarised our results and brings some submissions how to prepare your potential evaluation process.

## Features of electronic test

Electronic test is implemented as a form in which questions and answers are sequential presented. Respondent can select the appropriate answers by mouse click. Electronic test implementation enables variable number of answers on individual questions. The range of answers' number is 2 – 10, the number

[1] *doc. Ing. Pavel Horovčák, CSc., Ing. Beáta Stehlíková, PhD.,* Department of Applied Informatics and Process control, Institute of Production Process Control, Technical University of Košice, Letná 9, Košice 04011, Slovak Republic, Pavel.Horovcak@tuke.sk, Beata.Stehlikova@tuke.sk

(Recenzovaná a revidovaná verzia dodaná 14. 6. 2007)

of accurate answers is one or more. To answer the question with solely one accurate answer the radio buttons are used (they indicate directly this fact), or the check boxes, which does not implicitly indicate the count of valid answers. For questions with more than one accurate answer only check boxes there are used.

Presentation of valuation (of testing results) for the examinant can be realized in a few ways, dependent on purpose of test. The first way there is valuation with indication of acquired number of points only, without displaying of individual questions and answers. If the test is intended for self-study as well as for self-assessment, there is the second way – an alternative with displaying of individual answers valuation. This displaying of valuation can be done either just by statement of answer accuracy or inaccuracy (without its labelling) or by selected answer labelling together with statement of answer's accuracy or inaccuracy. An answers valuation can be realized also in more manners. The first one ensures that student obtains plus one point per each valid answer, for incorrect answer or not answered question he obtains zero points. The second manner is stricter, but more exact – for each correct answer he obtains plus one point, for incorrect answer minus one point and for not answered question zero points.

By the total valuation of test results it can be (but must not be) considered the duration of test execution in specified time range. For test implementation the solution to add some points (bonus) in the case of premature finishing of test in defined time interval (fin_time) was selected. We have chosen the linear calculation of bonus so that by finishing of test in specified time zero points are added. The test-end in specified time is automatically guaranteed by script. The time interval and the number of bonus points for test it is possible to define at configuration of test. As standard default values there are zero values for both parameters defined, which means the dismissal of premature test-end.

The electronic test uses three groups of tables, which differs by its purpose and lifetime. The first group consists of three tables (they are always present) assigned for saving of test configuration, list of test users' (teachers) and for creation of introductory test page. Second group includes five tables per electronic test's realization. These tables consist of test questions and answers, students' results, statistics of answers' correctness on individual questions of test and test introductory page. The table of test introductory page is modified at each configuration of test. Tables of questions and answers as well as test statistics are deleted and once more created by correction, adding or deleting of test questions. It is linked to the fact that the number of columns in table of statistics is derived from number of test's questions. The third group of tables consists of two temporary tables that are created for every test's participant and after test ending and results browsing they are automatically deleted. In these tables generated test including answers marked by student and valuation of their correctness is stored.

## Practical use of electronic test

Electronic test can operate in two interlocking modes. The first mode is test administration executed by teacher; the second one is conducting of test executed by student. The both modes are executed by using web browser with several various forms. By means of this forms the needed data and information by teacher at creation, configuration, valuation or maintenance of test are entered, as well as, valid answers for electronic test's questions are selected by student. After fulfilling and submitting, the appropriate form is processed by script language PHP.

The processing consists of appropriate application function call and subsequent saving of user's data (such as test structure, questions and answers) into relating database tables by means of SQL statements. Once the electronic test configuration has finished, the introductory test page is created in the database table with the name of test.

After execution of this file, student starts to conduct the test. Student begins with his identification (surname) and application generates for him a unique electronic test. The reason of uniqueness of the test is that the questions and answers are presented in random sequence individual for each student. This test after answers fulfilling is valuated and its results are saved into tables. If the test configuration it enables, student can study a test course as well as a fruitfulness of answered questions.

## Test operation

After inputting of access parameters (login and password) test's administrator page or more simply teacher's page appears.

The administrator is able to create, remove, correct, configure, view test, view test results, execute maintenance of results, inspect contents of tables, remove and create tables, add (remove) the test's end users, change its own access password as well as to join some particular tests into one unit by using of individual function.

The teacher has fewer functions and can only configure test, review the test itself and test outline, review test results and change his access password.

The creation of test consists of preparation of questions and answers in specified form (text file), its syntactic verification, upload to web server and saving into database tables. Particular scripts support operation of verification, file upload and its saving. Creation of input file is "supported" by teacher. When the test is prepared on the server, before its operation it is necessary to realize the configuration, which enables to select the name of test, set up its parameters and save them into configuration table. Within the configuration function several parameters can be defined, e.g. time of text execution (in range 1 to 90 minutes), time interval of test validity (it is possible to execute test only within this given interval), selection of the style file for test layout from list, set up possibility of results viewing, the way of results valuation as well as total point number for test. In addition, this function enables to set the limit of test fruitfulness that is the minimal number of valid answers from which the point's assignment will start. This number is of course less then total number of test questions. Furthermore, it is possible to set time bonus in form of number of minutes (fin_time) and points (bonus) which must be less then half of test duration in minutes and less then half of test points.

After test configuration parameters selection and setting there is an inspection of their acceptability and validity executed. In the case that some parameter is not valid or not acceptable, its value is put on the screen together with acceptable range and the control (execution) returns back to the beginning of test configuration. If all test configuration parameters are correctly set up, they are saved into configuration table as configuration valid for given test.

## Information obtained in test preparation

### Technical aspect of question preparation

Creation of questions and answers in form of text file is a dominating way of preparation of test questions and answers. This preparation is easy, comfortable and quick. For file formation it is possible to use the MS Notepad, or any other simple text editor. The usage of MS Word processor is not recommended at all, (in case of necessity only with txt output). If working with diacritical characters (in the text file as well as on the web server) in the code page different from Windows-1250 it is suitable to use the freeware PSPad editor (Fiala, 2005). This editor enables to convert the text in the all standard (Central-European) coding. The question's format is given by sequential number with dot and space, text of question of maximal length of 255 characters (on one or more lines) and it is finished by three characters %qx at a new line. Character x indicates the number of valid answers as follows: x=1 one valid answer, x=2 more valid answers (however it can be also only one). The answer's format is given by y text of answer (only one line, maximum 255 characters). Character y indicates the correctness of answer as follows: y=1 denotes the valid answer, y=0 denotes incorrect answer. After the last answer's line there is on new line stated couple of characters %a, which ends the answers' set. Afterwards it follows the next question. Syntactic errors in the text file are detected by inspection script using regular expressions. The upload process to server is stopped when the error occurs and it is necessary to correct the text file. Interactive environment of test's administrator enables the individual questions and answers correction. For the entire test creation this practice is not suitable, it has only supplementary nature.

### Teacher's aspect of question preparation

The questions and answers prepared for test according to our experience should be of common nature, demanding the logical approach and thinking. They should be clear, single valued and independent.

## Test results valuation

The process of feedback between student and teacher reflect to the teacher the knowledge and accomplishments of the student as well as the effect of supplied information and the arrangements taken during educational process and during explanation of teaching subject. This process can be considered from teacher's point, as well as, from student's point of view.

### Teacher's point of view

The implementation of electronic test is dependent on selected test model. For models with one valid answer it is suitable to use radio buttons, for models with extended number of valid answers check boxes are applicable. Test is realized in the mode of electronic form, in which individual test questions together with the answers are step by step displayed. Afterwards, it is possible to mark the answer by mouse click.

To improve objectivity during evaluation of results of electronic testing, it is necessary to deal with success of students as well as success of individual test questions.

The **degree of student's success** evaluated through didactic test is a dynamic variable, which is affected by multiple factors. We selected some of them, and based on specific experiments, we tried to quantify them

and determine they importance (or unimportance). Repetition of the same test by the same group of student in certain time periods should demonstrate the effect of the *factor of learning*. The success of student during the test can be affected by the time period for test completion, which is the *factor of time limit*. There are more methods for motivation of students to repetition of the learning topic. One of them is ongoing testing of their preparedness for the lesson. The other, more sophisticated one is participation of students on preparation for ongoing testing, which forces students to search for the utmost meaning of the learning topic, so they "can raise questions". In this way, the teacher can measure students' feedback not only based on answers on the raised questions, but also based on the questions themselves, because he can estimate weather students got the essence of the topic. In case that students know the formulation of the questions, as well as, the "correct answers", test results can be different than in the case that students do not have those specific information, the *factor of „recognition of questions"*. Test results can be also affected by variability of possible answers. In case that there is only one correct answer among the offered answers, the choice is simpler than in case that there is one or more correct answers, while their number is not specified. The answer is correct only if all correct answers are selected from the offered alternatives. The factor was named as the *factor of answer type (number)*. It would be naïve if the teacher would assume that the student after test completion would not share his experience with his colleagues who have not taken the test, yet. They then come for testing prepared for specific questions, which have already been raised in the test. However, that approach does not guarantee mastering of the learning topic and the ability to orient in the learning material. An important factor affecting test results can be ability of students to communicate among each other. This factor was named as the *factor of communication*.

The degree of success of individual questions depends on numerous factors. Based on our experience, we included into this category the already mentioned factor of recognition of questions and the factor of answer type. The optimistic assumption of the teacher that the level of difficulty of all defined test questions is the same is not always correct. Very high degree of success of a question can mean good understanding or triviality of the question itself, which then has no added value in the test. Reasons for low degree of success of a question can be its misinterpretation, underestimation of an important part of the whole learning material, or that the question or provided answers are not clearly defined. The factor was named as the factor of quality. Analysis of the quality of questions as one of the success factors of testing would be rather difficult, but evaluation and improvement of quality of test questions contributes to improvement of the objectivity of the evaluation process.

Electronic test enables the analysis of obtained test results. One of the electronic test's functions is the presentation of the total survey (Vysledky / Results, then option Celkom / Total) about fruitfulness of answering individual questions of the whole group of tested students (see at *Tab.* 1). Based on these results it is possible to detect the incorrect formulation of both the questions and also the answers, as well as, to recognize the deficits in understanding or explanation of the teaching subject. This analysis helps and enables to improve the process by undertaking and realization of required arrangements.

*Tab. 1. Sample of table showing results of individual questions' fruitfulness.*

| Questions | | | | | |
|---|---|---|---|---|---|
| Count | | q1 | q2 | q3 | q4 |
| 79 | | 79 | 79 | 79 | 79 |
| Correct answered | 1 | 50 | 41 | 48 | 46 |
| Uncorrect answered | -1 | 20 | 27 | 22 | 20 |
| Without answer | 0 | 9 | 11 | 9 | 13 |

**Student's point of view**

By valuation of students' approach to the didactic tests execution, it is possible to assume, that:

1. Student answers the question in the case he things that he knows the correct answer.
2. Student doesn't answer the question in the case he doesn't know the correct answer.

In order to proof above-mentioned assumptions, there was questionnaire conducted on the sample of 40 students after the execution of several screening inspections in the form of electronic tests. They were asked to answer the questions about their practice during the test execution. The questions and answers were:

1. What is your sequence of answering the questions: (A) from first question to the last one, or (B) skip arbitrary from question to the question?
2. In case that you cannot choose the correct answer after the first reading of the questions: (A) are you leaving the question without answering or, (B) do you choose the random answer in any case?

3. Once you answered all the questions: (A) are you waiting till the end of time limit reviewing your answers, or (B) do you close the test and inspect its evaluation?
4. In case that you are aware, that there is only one correct answer to the particular question: (A) do you read carefully all proposed answers, (B) once you read the correct answer you do select it and do not deal with the others?
5. Results of the questionnaire are presented in the table 2.

*Tab. 2. Practice by working-out of the didactic test.*

| | Answer | | | |
|---|---|---|---|---|
| | A | | B | |
| question | number | ratio [%] | number | ratio [%] |
| 1 | 39 | 98 | 1 | 3 |
| 2 | 31 | 78 | 9 | 23 |
| 3 | 11 | 28 | 29 | 73 |
| 4 | 29 | 73 | 11 | 28 |

Inquiry presents important information:

- 23 % of the students answer the question even they do not know the correct answer,
- 28 % of the students admit that they do not sufficiently analyse all possible answers before selection of the correct one.

**Statistical evaluation methodology**

Planning of experiments (tests and their courses) was done based on anticipated effect of the mentioned factors. Individual experiments were realized during evaluation of ongoing preparation of students and their evaluation during exams. Students were not informed about the fact that "their test results" will be used for statistical purposes.

For evaluation of the planned experiment, mathematical and statistical methods were used, since they allow formalisation of the experiment results and the applied procedures.

Test results, which are available to the teacher, contain information on each "tested" student and each test question. We analysed both behaviour (results) of students and behaviour (results) of different types of questions. From the available information, the following variables were created and analysed:

For analysis of „success" of students:

- number of correctly answered questions,
- number of incorrectly answered questions,
- number of not answered questions.

For analysis of "success" of questions:

- number of correct answers for the question,
- number of incorrect answers for the question,
- number of occasions when the question was not answered.

## Results of evaluation of students' success

**The base test results**

The evaluation of test results was performed on the analysis of random variables (number of correctly answered questions, number of incorrectly answered questions and number of not answered questions) was focused on type of distribution of mentioned random variables (RV) and its characteristics. The range of the variables (test results) is partitioned into $k$ intervals of equal length. Each interval we named the class which is represented by "class sign" $x_j$ and number of tests with result partitioned in the interval (class) $n_j$.

**Analyses of variable "number of correctly answered questions"**

Characteristics metrics present: (*Tab. 3*)

| characteristic | symbol | value |
|---|---|---|
| number of tests | n | 150 |
| average | $\bar{x}$ | 12.39 |
| standard deviation | s | 3.83 |
| variance ratio | Vx | 31% |
| skewness | Ax | -0.27 |
| kurtosis | Ex | 2.37 |

*Tab. 3. Statistical characteristics of variable "number of correctly answered questions".*

- Medium value (arithmetical valued average) $\bar{x} = \frac{1}{n} \cdot \sum_{j=1}^{k} xj \cdot nj \; ; \; n = \sum_{j=1}^{k} nj$ (1)
- Standard deviation $s = \sqrt{\frac{1}{n} \cdot \sum_{j=1}^{k} (xj - \bar{x})^2 \cdot nj}$ (2)
- Decrement of variability presents too characteristic variance ratio $Vx = \frac{s}{\bar{x}}$ (3)
- Data have minus sign. $Ax = \frac{1}{n} \cdot \frac{\sum_{j=1}^{k} (xj - \bar{x})^3 nj}{s^3}$ (4)
- that means, more numbers of the data are higher than average (more results of the test are better than average).
- High value of kurtosis $Ex = \frac{1}{n} \cdot \frac{\sum_{j=1}^{k} (xj - \bar{x})^4 nj}{s^4} - 3$ (5)
- reflects the intense data arrangement round average.

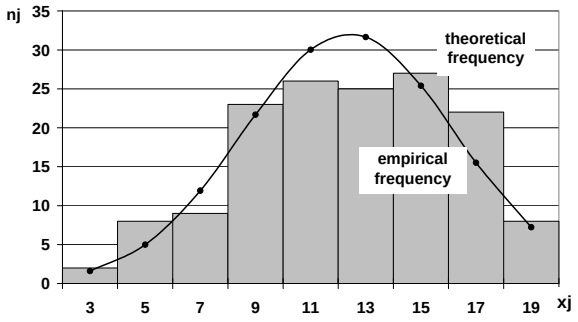


*Fig. 1. Numbers of the correctly answered questions.*

There were used the Pearson Chi-square test on significance level $\alpha=0,05$. The null and alternative hypotheses are:

$$H_0: \; o_j - e_j = 0$$
$$H_A: \; o_j - e_j \neq 0$$

Pearson's test statistic: $\chi^2 = \sum_{i=j}^{k} \frac{(o_j - e_j)^2}{e_j}$ (6)

where: $o_j$ is the observed frequency of the $j$ th class (it is nj), $e_j$ is the expected frequency of the $j$ th class, computed as $e_j = p_j n$, where $p_j$ is the probability that x lies in the $j$ th class. Rejection region: $\chi^2 > \chi^2_\alpha$, where $\chi^2_\alpha$ is Pearson's frequency function by $\alpha$ and number of freedom (Floreková and Benková, 1999).

*Conclusion*: Expected probability distribution is normal, $e_j$ is frequency for normal distribution with estimated location parameter mean xp and scale parameter standard deviation s. The test statistic $\chi^2 = 4{,}76$ doesn't fall in the rejection region $\chi^2_\alpha = 9{,}49$, we do not reject $H_0$. Therefore at $\alpha = 0{,}05$ the data do not provide sufficient evidence to indicate difference between observed frequency of the $j$ th class (empirical distribution of data) and expected frequency of the $j$ th class (normal distribution). It is presented on picture 1 theoretical frequency display line, empirical frequency display columns.

**Analyses of variable "number of incorrectly answered questions"**

Number of incorrectly answered questions, this variable demonstrates number of questions in the test, which student answered, but chooses answer was wrong.

Statistical characteristics metrics are in the table 4.

*Tab. 4. Characteristics of variable "number of incorrectly answered questions".*

| characteristic | symbol | value |
|---|---|---|
| number of tests | n | 150 |
| average | $\bar{x}$ | 4.92 |
| standard deviation | s | 2.80 |
| variance ratio | Vx | 57% |
| skewness | Ax | 1.19 |
| kurtosis | Ex | 4.45 |

Variables have skew plus sign (that means more data are smaller than average). Variability of the data is high. Expected probability distribution is normal, $e_j$ is frequency for normal distribution with estimated location parameter mean xp and scale parameter standard deviation s. Result of Pearson's Chi-square Test: The test statistic $\chi^2 = 24{,}85$ fall in the rejection region $\chi^2_\alpha = 9{,}49$, we reject $H_0$. Therefore at $\alpha = 0{,}05$ the data do provide sufficient evidence to indicate difference between observed frequency of the $j$ th class (empirical distribution of data) and expected frequency of the $j$ th class (normal distribution).

*Note: Results of other tests for normal distribution of data gives the same conclusion.*

**Analyses of variable "number of questions with no answer"**

These questions in the test students leave without choosing any answer. This variable is analysed to find type of distribution.

Distribution of variable number of not answered questions reminds of exponential distribution. $e_j$ is frequency for exponential distribution with probability function $f(x) = \lambda e^{-\lambda x}$ with estimated parameter $\bar{x} = \frac{1}{\lambda}$.

*Conclusion*: The test statistic $\chi^2 = 21{,}4$ doesn't fall in the rejection region $\chi^2_\alpha = 23{,}7$, we do not reject $H_0$. Therefore at $\alpha = 0{,}05$ the by Pearson's test statistic data do not provide sufficient evidence to indicate difference between observed frequency of the $j$ th class (empirical distribution of data) and expected frequency of the $j$ th class (exponential distributions). It is presented on picture 1 theoretical frequency display line, empirical frequency display columns.

It was possible to consequence this result after correction (reduction of number of tests in the class with 0 not answered question by 23 % - this amount of students' following inquiry answer all questions without knowing the correct answer), in which there were answered all questions. Picture 3 presents theoretical and empirical frequencies.

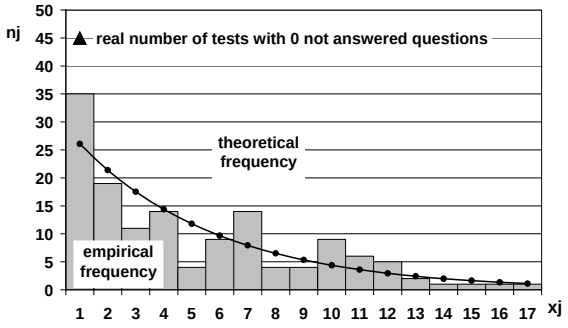


*Fig. 2. Numbers of not answered questions.*

This category of the students' probably have contributed to the specification of RV of number of incorrectly answered questions, what means, that those respondent, who answered always all the questions without knowing the correct answer, but choosing random one, can bring some uncertainty to the evaluation.

*Conclusion*: Standard statistical tests (for one or more parameter selection) require sufficient number of data and normality of data. If the selections contain sufficient number of values and the selected random variable $x =$ „multitude" does not meet the condition of normal distribution, it can be achieved through „root" transformation: $o = \sqrt{x}$ (Hebák and Hustopecký, 1987).

In opposite cases, to reject significance or verify insignificance of relevant factors, it is necessary to use non-parametrical tests. We chose the Sign Test and Two-Sample Rank test (Mann-Whitney Test).

**Factor of learning**

*Conditions*: test i_msdos had 22 questions, and focused on general information about operating system MSDOS, in the experiment participated 70 respondents, who repeatedly performed the same test in one-week interval.

*Hypothesis:* $H_0$: In repeated test the results are not statistically significant different. $H_1$: In repeated test the results are better. $H_2$: In repeated test the results are worse.

*Anticipated result*: improvement of individual results of respondents on repeated tests.

*Evaluation*: variables for investigation of „success" of students, Sign Test. On the significance level of $\alpha = 0{,}05$, the null hypothesis was verified, if repetition of test had no impact on test result of each

student. In the alternative hypothesis $H_1 \mid H_2$, we verified if through repetition of test, the results improved / deteriorated.

The Sign Test needs data in pairs, counts differences between data of each pair, discard any differences of zero; di = 0, assign + or − sign to each difference.

Test statistic: n+ is the number of times a "+" is observed and n- the number of times a "-" is observed. Ignore all *zero* or *equal* observations.

Reject H0 if either n+, n- is greater than or equal to critical value (tables for Signed Test) (Komenda and Klementa, 1981).

*Conclusion*: All values of testing criteria n+ for verify $H_1$ (testing criteria n+ for variables: number of correctly answered questions is 51; number of incorrectly answered questions is 39; number of not answered questions is 38) are greater than the critical value (25); all values of test criteria n− for verify $H_2$ (testing criteria n− for variables: number of correctly answered questions is 12; number of incorrectly answered questions is 19; number of not answered questions is 15) are smaller than the critical value. The null hypothesis $H_0$ and the alternative hypothesis $H_2$ can be rejected in favour of alternative hypothesis $H_1$. The Sign Test proved a statistically significant difference (improvement) among pairs of results of students in the first and repeated test.

**Factor of time limit**

*Conditions*: test i_spol1 had 50 questions, and focused on aggregate information from informatics, in the experiment participated 33 respondents in two groups with 12 and 21 participants respectively. The time period for the first group was 15 minutes and for the second group 20 minutes.

*Hypothesis:* $H_0$: The results for first and second group are not different. $H_1$: The results for shorter and longer time limit are different.

*Anticipated result*: better results of respondents in case of longer time limit (the two samples come from different populations.)

*Evaluation*: variables for investigation of „success“ of students, Mann-Whitney Test. On the significance level $\alpha = 0{,}05$ the test statistic doesn't fall in the rejection region $U = 73$, we do not reject $H_0$ against the alternative hypothesis $H_1$: The data of test results with various time limits consist of two independent samples taken from the two populations.

Mann-Whitney Test (U Test).

$H_0$: The two samples have the same median. $H_1$: The two samples have not the same median.

The first step is combine two samples and rank the total set of scores. Then it is needed to count sum of scores for each samples identify this sums by R1 and R2.

Test statistic: $U = \min\{U_1, U_2\}$ where $U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$, and

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 \qquad (7)$$

Reject $H_0$ if value of test statistic either is smaller than or equal to critical value (tables for Mann-Whitney Test) (Komenda and Klementa, 1981).

*Conclusion*: $H_0$ about consistency of selections cannot be rejected (value of testing criteria for variables: number of correctly answered questions is 110,5; number of incorrectly answered questions is 92,0; number of not answered questions is 98,5). Result of test did not prove statistically significant impact of the factor of long setting time limit on success of students in tests. However, it is necessary to notice that the time limit was in both cases set as „real“, and the necessary time was not underestimated.

**Factor of recognition of questions**

*Conditions*: test oop9 had 40 questions, focused on development of a database application and basics of SQL language, in the experiment 13 respondents participated. The test consisted of 20 known questions and 20 unknown questions. Known questions were prepared by students and submitted to their teacher. Those questions as well as the correct answers were available to their authors and also their colleagues. Unknown questions covered the same topic, and prepared by the teacher. Students did know neither their wording, nor the correct answers.

*Hypothesis:* $H_0$: The results for known and unknown questions are not statistically significant different. $H_1$: The results for known questions are different.

*Anticipated result*: better individual results of respondents in case of known questions.

*Evaluation*: variables for investigation of „success“ of students, Mann-Whitney Test.

For the Mann-Whitney Test, the null hypothesis $H_0$ was verified on the significance level of $\alpha = 0{,}05$, if the students results for known and unknown question are about the same have the same median, against the alternative hypothesis $H_1$, the results for known and unknown questions are different have not the same median.

*Conclusion:* The validity of hypothesis $H_0$ against the validity of $H_1$ was verified on significance level $\alpha = 0{,}05$. For variables number of correctly answered questions and for number of not answered questions, hypothesis $H_1$ can be accepted. For variable number of incorrectly answered questions, $H_0$ cannot be rejected (the critical value is 45; testing criteria for variables: number of correctly answered questions is 1,5; number of incorrectly answered questions is 55; number of not answered questions is 0).

The variable number of incorrectly answered questions in case of known questions and not known questions statistically does not differ significantly.

Results of students in case of known questions were better in relation to the number of correctly answered questions (more) and in relation to the number of not answered questions (less).

The overall result of evaluation of impact of factor of question recognition through difference of achieved score by the respondent for known and unknown questions (the score was calculated as the difference between the correctly and incorrectly answered questions): The result for all respondents is better for known questions.

**Factor of communication**

*Conditions*: test i_spol1 had 50 questions, and focused on the areas of informatics, in the experiment participated 73 respondents. The test was performed four times by 12, 21, 21 and 19 students.

*Hypothesis:* Box's test $H_0$: Matrix of Covariance are identical, variability of selections are identical. $H_1$: Between matrix of covariance there is statistically significant difference. Analysis of variance $H_0$: Between averages vectors there is not statistically significant difference. $H_1$: Between averages vectors there is statistically significant difference.

*Anticipated result*: consistent improvement of results.

*Evaluation*: For the four three-dimensional selections, the consistency of covariance matrix was verified (Box' test). Consequently, through analysis of variance, the hypothesis on consistency of average vectors was verified.

*Conclusion*: $H_0$ on consistency of covariance matrix cannot be rejected with (testing criteria = 25,09 < critical value = 28.87). The samples, consistency of variability were confirmed.

During the analysis of variance $H_0$ on consistency of vectors of averages can be rejected (testing criteria = 2,62 > critical value = 2,17). Analysis of variance confirmed the difference between test results in mid values of variables. Improvement of results was confirmed. Rather significant consistent improvement of partial as well as overall results (even always new) of respondents indicates that it is necessary to carefully choose repeated usage of same tests. We consider it more appropriate to use variation of two or more tests.

## Results of evaluation of success of questions

Variables defined for evaluation of success of questions were transformed (root transformation). For variables transformed from variables related to questions, it was verified and fulfilled the assumption of normality for one-dimensional selection of each variable as well as for tree-dimensional selection.

**Factor of question recognition**

*Conditions*: test oop9 had 40 questions, and focused on creation of database application and basics of SQL language, in the experiment participated 13 respondents. Test consisted of 20 known and 20 unknown questions. Known questions were prepared by students and submitted to their teacher. Those questions as well as the correct answers were available to their authors and also their colleagues. Unknown questions covered the same topic, and were prepared by the teacher. Students did know neither their wording, nor the correct answers.

*Hypothesis:* Box's test $H_0$: Covariance matrix are the same, dispersion of selects are identical. $H_1$: Between matrix of covariance there is statistically significant difference. Analysis of variance $H_0$: Between average vectors there is not statistically significant difference. $H_1$: Between average vectors there is statistically significant difference.

*Anticipated result*: better results in case of known questions.

*Evaluation*: Variables defined for evaluation of success of questions were transformed (root transformation). For variables transformed from variables related to questions, it was verified and fulfilled the assumption of normality for three-dimensional selection. Consequently, selections of known and unknown questions were compared: Box's test, test of consistency of vectors of averages for heteroscedasticity.

*Conclusion*: The result of comparison of two three-dimensional selections is inconsistency of covariance matrix. Selections of known and unknown questions do not correspond in variability. $H_0$ can be rejected $S_1 \neq S_2$ (testing criteria = 25,75 > critical value = 12,59). In the test of consistency of vectors of averages, $H_0$ can be rejected as well in favour of $H_1$ (testing criteria = 70.94 > critical value = 7.81). The difference in averages is caused by variables number of correct answers to the question and number of occasions that the questions have been answered. From the overall achieved average results, it follows that the success of answering question is definitely better (multiple times) for known questions.

**Factor of answer type**

*Conditions*: test kontrol1 had 34 questions, and focuses on the area of informatics, in the experiment participated 113 respondents. The test consisted of 25 questions with the possibility of only one correct answer (radiobox) and 9 with the possibility of selection of not defined number of correct answers (checkbox).

*Hypothesis:* $H_0$: the results of questions radiobox type and checkbox type are about the same have the same median. $H_1$: the results of questions radiobox type and checkbox type are different have not the same median.

*Anticipated result*: better results in case of one answer questions.

*Evaluation*: Due to the low number of checkbox type of question (9), we used the non-parametric Mann-Whitney Test, for evaluation of impact of type of answers on success of the questions. At the significance level of $\alpha = 0{,}05$, we verified the validity of hypothesis $H_0$ against $H_1$ about consistency of selections for variables for questions of radiobox and checkbox type.

*Conclusion*: All values of testing criterion for $H_1$ are smaller than the critical value (critical value is 60; testing criteria for variables number of correct answers for the questions is 25; number of incorrect answers for the questions is 23,5; number of occasions when the question was not answered is 57,5). The null hypothesis $H_0$ can be rejected in favour of the alternative hypothesis $H_1$. For all variables, the impact of the type of question was evaluated as statistically significant. Better results in case of one-answer type questions were confirmed.

## Evaluation of quality of questions

The optimistic assumption of the teacher that the level of difficulty of all defined test questions is the same is not always correct. Evaluation of quality of questions was performed on couple of applied tests. For quality evaluation purposes, special attention should be given to questions for which the ordered n-set up {number of correct answers; number of incorrect answers; number of occasions when the question has been answered} differs. Very high degree of success of a question can mean good understanding and learn-ability or triviality of the question itself. Reasons for low degree of success of a question can be its misinterpretation, underestimation of an important part of the whole learning material, or that the question or provided answers are not clearly defined.

## Factor of quality

*Conditions*: test i_spol1 had 50 questions, and focused on the areas of informatics, in the experiment participated 50 respondents.

*Anticipated result*: Identification of questions with extremely high or extremely low success.

*Evaluation*: Variables defined for evaluation of success of questions were transformed (root transformation). For variables transformed from variables related to questions, it was verified and fulfilled the assumption of normality for three-dimensional selection. In the three-dimensional file, distance observations were searched.

*Conclusion*: Through the test of distant observations, we identified as different (in the structure of correct, incorrect and missed answers, presented at *Tab.* 5) questions 11, 12, 14 and 22. Out of them, it is possible to classify question 11 as too simple, and questions 12, 14 and 22 as too complicated. For those three questions, it is necessary to check the formulation of the question, or in the future, stress related areas during the learning process.

*Tab. 5. Structure of question's results.*

| | Question | | | | | |
|---|---|---|---|---|---|---|
| | q 11 | q 12 | q 13 | q 14 | q 21 | q 22 |
| Number of correct answers for the question | 48 | 10 | 42 | 8 | 39 | 1 |
| Number of incorrect answers for the question | 1 | 37 | 6 | 33 | 4 | 34 |
| Number of occassionc when the question was not answered | 1 | 3 | 2 | 9 | 7 | 15 |
| Sum | 50 | | | | | |
| | easyy | difficultt | normal | difficultt | normal | difficultt |

Evaluation and improvement of quality of test questions is a part of evaluation process objectivity improvement.

## Conclusion

New form of communications, which rise between student and teacher by exploitation of didactic tests in electronic form enables improvements of educational process „on both sides of teacher's desk" (Benková and Floreková, 2000). Communications quality (through test) determines accomplishment of examiner at actual subject area and his psychological and pedagogical abilities with one leg standing on the other side of knowledge examination. Exploitation of comprehensive knowledge for specific usage of didactic tests is possible to that extant, in which the proposed test model is able to justify the reality of examination. To study means to remember and respond in different way according to experiences.

From test results, it is possible to draw three output categories – evaluation of students through scoring, factors affecting success of results, and indication of different questions (factor of quality).

Simple configuration of test parameters enables modification of test set up according to specific conditions and requirements of the teacher. The key benefit of electronic test is the fact that test results stored in database table are available to the teacher immediately after test completion.

We paid attention to examination of impact of selected factors on success of students (evaluated factors included: factor of quality, factor of time, factor of question recognition, and factor of communication among students), and on success of answering individual questions (evaluated factors included: factor of question recognition and factor of answer type). Factors of repetition, question recognition and communication have statistically significant effect on success of students. Through repetition of test, the student acquires certain level of security about the learning topic during the ongoing evaluation process. On contrary, repetition of test more times in a row is not appropriate during the final examination of students. It is more suitable to have a group of tests (3 or 4), which can be cyclically repeated. Assignment of question creation to students (and consequent recognition of questions) positively motivates ongoing preparation of students and repetition of the learning material, however, it does not guarantee deep understanding of the essence of the topic, since it leads more to memorization of specific correct answers than to in-depth study of the topic. Because of that reason, it is advisable to optimise structure of questions, and carefully choose the proper percentage of such questions on the test.

Factor of time has no significant statistical impact (it depends more on formulation and length of questions and answers). Our experience proved approximately 20 minutes for 60 questions, while the minimum completion time was below 10 minutes.

Statistically significant impact on success of questions was proved for the factor of question recognition and the factor of answer type. Based on this information, the teacher can design tests with appropriate level of difficulty for evaluation of students' knowledge.

The overall evaluation of results for individual questions (factor of quality) provides feedback to the teacher about the difficulty of questions: too simple, or too difficult (or misunderstood). This feedback allows correction of test questions and answers on one side, and, amendment or extension of the learning topic on the other side. Improvement of objectivity of different questions can be achieved through test of distant observations.

Main advantage, which brings didactic test in electronic form, is rationalization of pedagogical process. Exploitation of tests provides feedback to the teacher about level of knowledge of student. Other advantage is the fact, that the student is able to use the PC to find out or to validate the correct answer during the test conduction within the time limit. The plumbless fact is that the test can be executed by student also outside of school room, as it is placed on the internet, and thus the test removes the space limits.

*Notice: The contribution was solved in the scope of projects KEGA 3/3084/05 (H), KEGA 3/3125/05 (M), KEGA 1/3126/05 (B), VEGA 1/2179 /05 (D), VEGA 1/2160 /05 (K), VEGA 1/2603/05 (S) and ABILITIES Co 027306 (6RP)*

## References

Benková, M., Floreková, Ľ.: Informatizácia vysokoškolského štúdia (in Slovak, Informatics in academic study). In: Informatika a algoritmy '2000 Vedecká konferencia s medzinárodnou účasťou Prešov 7. a 8. septembra 2000. *Prešov: TU-FVT, 2000. s. 9-11. ISBN 80-88941-13-X.*

Fiala, J.: PSPad freeware editor http://www.pspad.com, 2005

Benková, M., Floreková, Ľ.: Komunikácia človek-stroj a medziľudská komunikácia vo výučbe na technických univerzitách (in Slovak, Man-machine communication and human communication in education on technical universities). In: Zborník príspevkov z konferencie informatika a algoritmy 1998, *s.129-131. Košice: FVT TU, 1998. ISBN 80-88941-00-8.*

Benková, M., Floreková, Ľ.: Štatistické metódy (in Slovak, Statistical methods). Košice : FBERG TU, *1999. 120 s. ISBN 80-7099-411-8*.

Hebák, P. Hustopecký, J.: Vícerozměrné statistické metody s aplikacemi (in Czech, Multidimensional statistical methods with applications), *SNTL Alfa, Praha 1987*.

Komenda, S., Klemenda, J.: Analýza náhodného v pedagogickém experimentu a praxi (in Czech, Analyses of random in educational experiment and praxis). *SPN Praha, 1981*.

Lewy, A.: Postmodernism in the Field of Achievement Testing. In Studies in Educational Evaluation. *Volume 22, Number 3, 1996, pp. 223-244 (22). ISSN 0191-491X, accessed 2. Feb. 2006* <http://www.ingentaconnect.com/content/els/0191491x>.

NIST/SEMATECH e-Handbook of Statistical Methods. [online] [quoted 31.1.2007] Available from <http://www.itl.nist.gov/div898/handbook/prc/section3/prc35.htm>, <http://www.itl.nist.gov/div898/handbook/>

Turek, I.: Úvod do didaktiky vysokej školy. *Košice: KIPTU, 2005. 318 s. ISBN 80-7099-882-2*.

U.S. Department of Education, National Center for Education Statistics. The NPEC Sourcebook on Assessment, *Volume 1: Definitions and Assessment Methods for Critical Thinking, Problem Solving, and Writing, NCES 2000—172. [online] [quoted 31.1.2007] Available from* <http://nces.ed.gov/pubs2000/2000195.pdf>