SEMANTIC NET FROM CONCEPTS AS A MODEL OF STUDENT'S KNOWLEDGE: HOW STABLE ARE THE RESULTS OF EXPERIMENTAL STUDY?

Evgeny A. Eremin

Abstract: *Original method to study students' conceptual knowledge was reported at previous ITHEA conferences. According to this method, to show their educational achievements students had to pass through some testing procedure, which requires combining of basic terms from a learnt course into interrelated pairs. Then students' answers, saved in text files, were analyzed by computer software that joined all interrelated concepts into groups, independent from each other. The larger these groups were the better students' success in digestion of learnt course was supposed.*

As human cognitive processes are very complex, it is important to make sure that experimental method is really valid and steady. The present paper examines stability of the experimental results; new implemented experiments were also aimed at estimation of method's inaccuracy. Several diversified tests were made to assure that method error do not deface observed picture: the difference between repeated tests was evaluated; the stability of student's pair selection and of knowledge fragmentation were analyzed to make sure that we can trust the results of experiments; regression towards the mean effect, which can skew the outcome of low-performed students, was estimated – the situation was found to be secure; another discipline was checked using the same method and there was no noticeable dependence of numeric scaled gauges from a learning content.

Detailed study shows that the generality of students demonstrate a growth of the proposed measure of learning success and its magnitude distinctly exceeds inaccuracy of method. The value of this inaccuracy was also estimated.

Keywords: *concept, relation, semantic network, knowledge representation, learning*

ACM Classification Keywords: *H.1.2 User/Machine Systems – Human information processing, I.2.4 Knowledge Representation Formalisms and Methods, Experimentation – Semantic networks, I.2.6 Learning, Human factors – Concept learning, Knowledge acquisition.*

Introduction

The accepted classification, described for instance in well known book [Anderson and Krathwahl, 2001], itemizes several different kinds of knowledge: factual, conceptual, procedural and meta-cognitive. It becomes more and more evident every year that remembering of numerous facts is useless without any system in acquisition. Continuous flow of new data puts forward conceptual organization of incoming information, so the level of the conceptual knowledge now comes before the level of the factual one. Systematization of learning facts is absolutely essential today because of rapidly growing amount of data about outward things. Hence problems, connected with building the structure of digesting concepts, become more and more urgent.

We can find demonstrative examples of knowledge organization role in many research works from different fields. For instance, the significance of structural knowledge in physics was reported in [Koponen et al, 2002]. The authors even established a special course, which "purpose is not to teach more physics but to organise what has already been learnt". The illustration from other knowledge domain was developed in [Loubser et al, 2001], where the conceptual base for environmental literacy was analyzed.

Experimental research also supports the discussed above theoretical thesis about the importance of knowledge systematization. Very decisive case can be found in detailed experiments [Lee et al, 2011]: this publication clearly showed that assessment based on knowledge integration can measure a wider range of knowledge levels than traditional multi-choice items.

Following ideas about importance of conceptual knowledge, in 2008 I started series of experiments to study how students organize basic concepts of a learning course in their minds and how to evaluate the level of their competency in systematization. A special testing procedure was developed to check ability to group the main terms of learnt discipline and to find some quantitative measures for this learning property. Developed experimental method was reported at the IV International Conference “Modern e-Learning” (MeL 2009) and published in its proceedings in Russian [Eremin, 2009]. Later on English version of these materials appeared in the associated international journal [Eremin, 2010]. Obtained during the study results were discussed at MeL 2011 [Eremin, 2011] and also at the “Multidisciplinary Academic Conference” (MAC 2012) [Eremin, 2012].

As human cognitive processes are very complex, it is important to make sure that experimental method is really valid and steady. So the aim of the present paper is **to consider possible method’s inaccuracy**. It describes the experimental study of a question how stable are the results we get. Different checks are made for this purpose; details of these control experiments will be discussed in the next sections.

In the issue all revisions proved that the magnitude of method errors can not deface a picture observed in experiments. The approximate value of this error was also calculated.

Basic experiment

This section briefly describes the main features of the basic experiment [Eremin, 2009-2012], which was implemented in order to study students’ skills in organization of a system from fundamental terms. Later on this experiment will be carefully checked up for stability and validity of results.

The first step to organize the original experiment was to form a full list of basic terms from the learning course, which teacher suppose good student must know. According to personal author’s preferences the educational course about computer architecture was under consideration.

In our case the list of concepts, used as experimental base, contained more than 120 terms. The most general concepts – like *computer*, *software* and *hardware*, *theoretic basis* etc., complemented by the terms that expand the previous ones – *operating system*, *processor*, *memory*, *DMA*, *principle of hierarchy*, *byte* and many others, were included into this base. Some related terms from connected disciplines such as microelectronics, logics and number notations were also added to the list. Contrary, the list did not include the names of concrete operating systems, external devices and their manufacturers, and other similar data, less essential from the position of learning the main course’s regularities. Using the standard terminology from object-oriented programming, we may say that classes of the concepts were under consideration but not their instances.

The next preparatory step was to build a semantic network from selected concepts. This is an optional operation, but it is very useful because gives us possibility to picture to ourselves what we want to get from a scrupulous student. The possible variant of such network, imaging the teacher’s view on the selected knowledge domain, is presented in fig. 1. Note that not all terms are visible in the picture: the numbers in the right top corner of rectangles with various concepts indicate how many related terms are hidden.

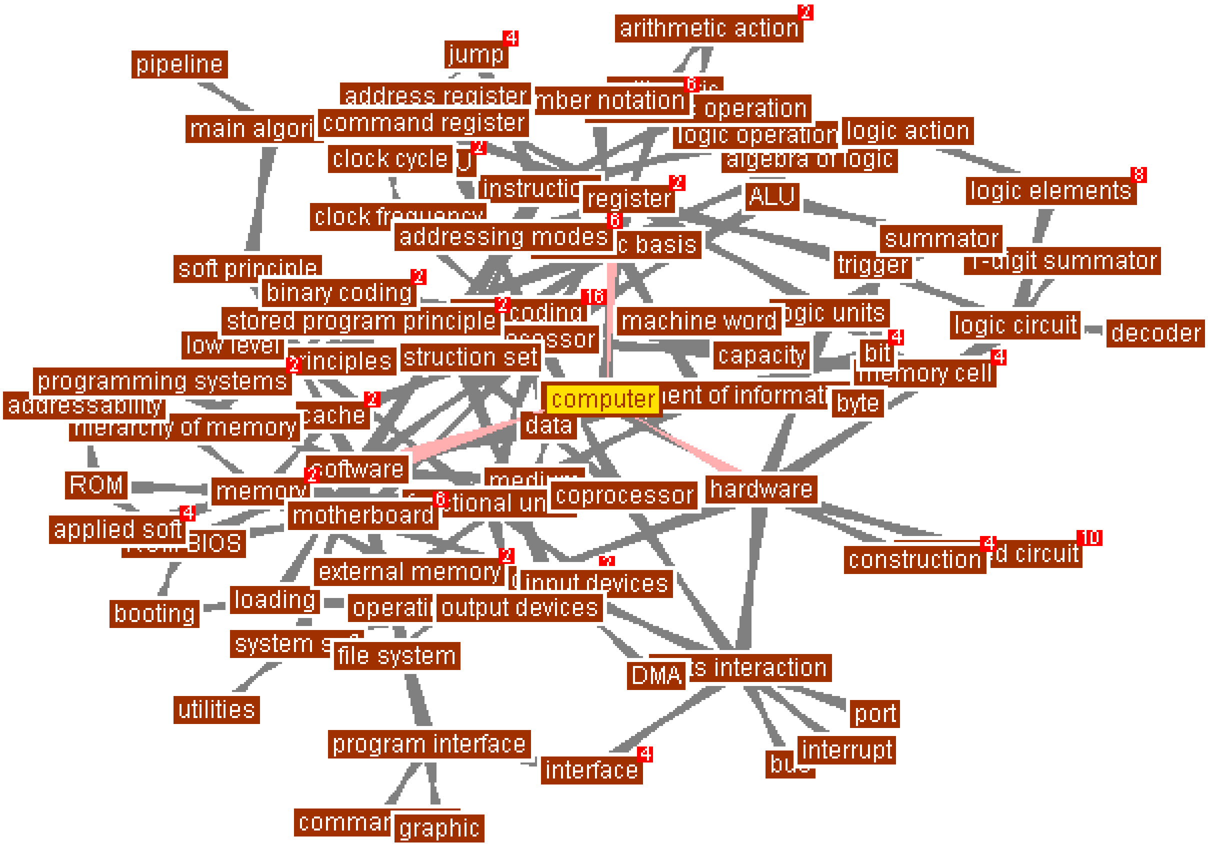

Figure 1

Semantic network in the picture has minimum two features that are important for further discussion. Firstly, this is really *network* but not a simple tree structure: many key concepts in our scheme are interlinked with large number of neighbors. Secondly, we see *connected* graph in the picture, because at least one way between two arbitrary terms always exists. From pedagogical point of view it means than all terms from the learning course are interrelated and form one single structure. As experiments showed, students could not obtain suchlike fully connected network even if they tried to do it.

When this complete list of terms had being formed, it became a base for experimental testing. The procedure of testing was the following. A student had to point out two terms and fix a link of definite kind between them. The testing software saved all constructed pairs into a text file that was analyzed after accomplishment of experiments by means of another computer program. Details of testing procedure can be found in [Eremin, 2010].

At the last step all resulting files fell under final computer processing. Its main aim was to educe linked groups of concepts for every student. For example, fig. 2 shows how computer combines concepts *functional units*, *processor*, *memory*, *input devices* and *output devices* into one group, using four associations *class/subclass* for central category *functional units*. Then *ALU* and *CU* (arithmetic and logic unit and control unit) will be joined to this group through the relations *whole/part* with term *processor*. Thus we get a set of seven interrelated concepts as a result.

As we already mentioned, in ideal case all concepts of a course must be interrelated. But experiments showed, that real students' files represented more scattered picture, which consisted of several isolated groups from concepts, and some groups were very small (only 2-3 terms). Every such small group can be interpreted as a separate fact that student did not associate with other facts from the course. You must also note that an increased rate of fragmentation indicates that student's knowledge is sparser.

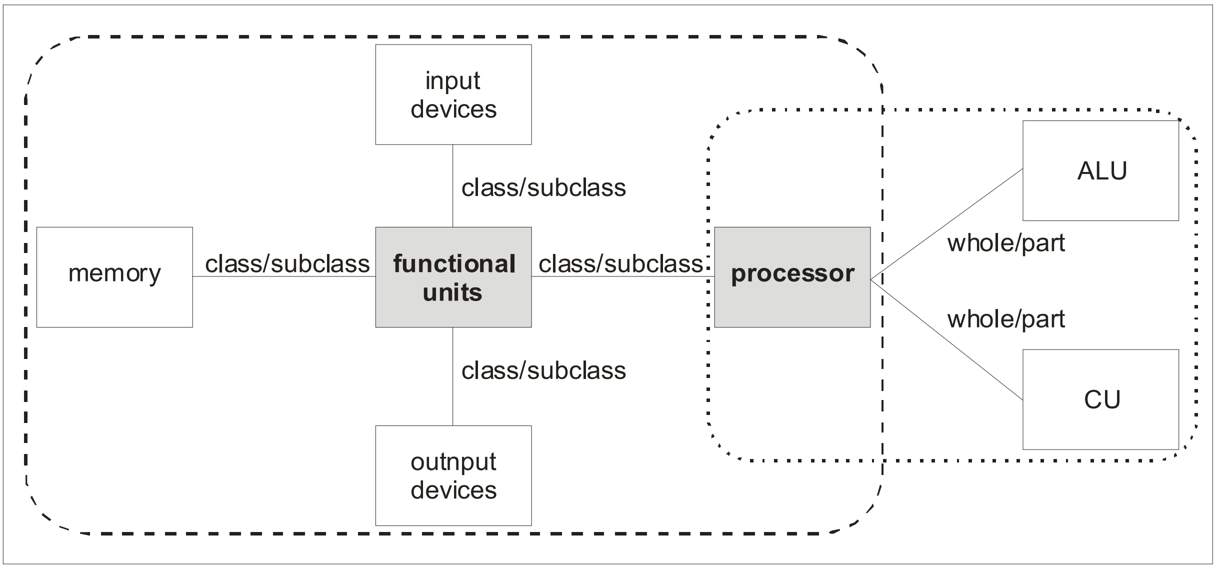

Figure 2

Typical result for ordinary student is shown in fig. 3, where 3 largest groups of concepts are depicted (there are also two groups from 3 terms and two from 2 ones, but they are not included into fig.3 as non-essential). Experiment shows that student's knowledge is fragmental and not clearly recognized as common picture (compare with the network in fig. 1). The most essential limitation is that general concepts like *hardware* and *software* seem to exist isolated from most of others and in similar way theoretical subjects like *scale of notation* are absolutely separated.

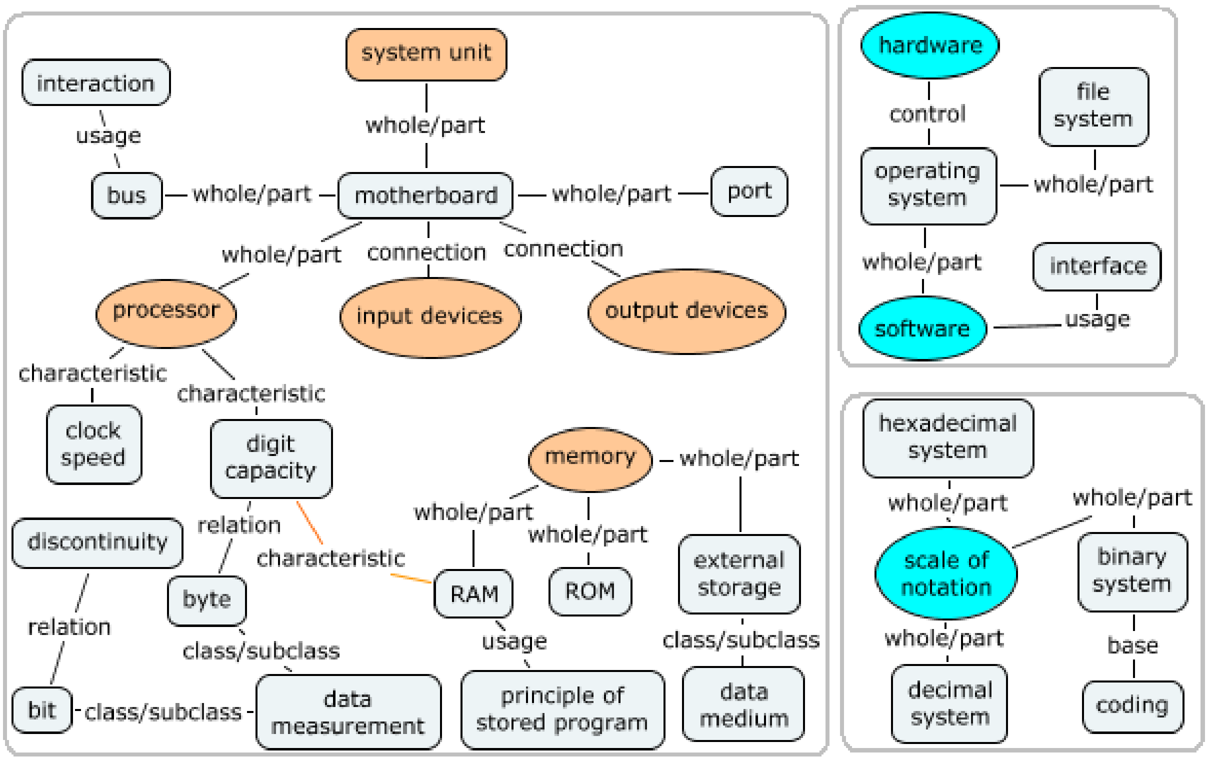

Figure 3

Fragmentation of knowledge was also evidently presented at specific “spotted” diagrams with independent groups of concepts, discussed earlier in publications [Eremin, 2009-2011].

More detailed discussion about grouping of concepts can be found in section “Test 3: stability of knowledge fragmentation”.

Analysis of numeric results in publication [Eremin, 2011] allowed selecting a simple but effective quantitative characteristic for evaluation of conceptual knowledge. It is based on resulting data about size and number of groups. If we divide *total number of concepts T* that student has named by *number of groups G*, we get *average size of a group*, which must be large when student learnt the course profoundly. As experiments showed, the results for average number of terms per group *T/G* had the most essential numeric changes among all other examined gauges. This feature prompts that such value can be considered as a reliable experimental measure [Rogosa and Willett, 1983].

Checking experimental method for errors

Now let us enumerate the main results of our basic experiment in order to check its accuracy. So the main findings we going to check are the following.

- Experiments showed that students’ knowledge is far from ideal entire picture – learnt terms are strongly fragmented.
- Average size of groups from terms that student demonstrates may be proposed to be a gauge of learning success.
- Several kinds of students can be educed according to introduced above gauge. The most distinct result is that “weak” students demonstrate very stable growth.

Conclusions formulated in the previous research will be investigated on validity in this paper. For this purpose we shall vary several conditions and examine the influence of changes; correspondent fruits are described in the following sections.

Section “**Test 1: repetition of experiments**” depicts how stable are the obtained results when we reduplicate the same test for the same students once more. Calculating the numeric difference between these two checks, we derive a value for confrontation with growth of the evaluating parameter *T/G* after completing a course: the last gain must distinctly exceed random changes between repetitions of the same test. Such experiments also give a numeric value of method’s error.

Section “**Test 2: stability of pair selection**” checks are pairs that students construct during their testing identical or not. It is important if we want to have stable experimental results. Such results also indirectly indicate are student’s knowledge deep or superficial.

Section “**Test 3: stability of knowledge fragmentation**” continues the discussion about stability of results. Its main question may be stated like following: is fragmentation real or maybe it is some sequence of inexact method.

Section “**Test 4: evaluation of RTM effect**” examines specific mechanism of statistical error and evaluates its significance for our experiments, using numeric method from [Ostermann et al, 2008]. This effect may artificially make the lowest results higher and so its estimation is very fateful to assure us about low-performed students’ success.

Section “**Test 5: different disciplines**” argues that the results of evaluation do not depend upon the contents of learning material. All previous study was based on “*Computer architecture*” course. Another educational discipline – “*Mathematical logic*”, considered in this section, showed very close results.

Before we begin the discussion of stability problems, let us introduce some brief denotation for students groups. In order to make the description more definite, let us continue the tradition from [Eremin, 2011] to use short tokens to refer to tested groups. So whereupon group G0 is the first trial of method [Eremin, 2010], results of groups G1-G4 are the subject of analysis in [Eremin, 2011], and some aspects of student's practical assessment were investigated for groups G1-G7 in [Eremin, 2012]. The present paper that continues the previous research, considers new results for groups G8-G10.

Total number of students in all groups G0-G10 is equal to 116. So on the average experimental group consists of approximately 11 students (from 5 to 17).

Test 1: repetition of experiments

The meterage theory states that the simplest way to check an accuracy of any experimental result is to repeat a measurement several times. Although our evaluations are not so easy to reduplicate as in the case of meterage with ruler or electrical device, we may try to test students twice (no one will agree for more!) So let us consider the results of two small academic groups G8 (7 students) and G9 (5 students), which were tested *twice before* learning the control course and *once after* it.

Group G8 passed two pretests (tests before learning a course) with interval more than one month; in group G9 pretests were separated by 7 months. In both cases students could not remember their answers during the first pretest, hence they had to pass the second pretest anew. Then they learnt "Computer architecture" course and final testing as usual was implemented after completion of the course.

Let us define values *T/G* for two input tests as A_1 and A_2 correspondently. Then according to foundations of meterage theory the result of measurement must be calculated as arithmetical mean $A = (A_1 + A_2) / 2$. Absolute error is computed using formula $\Delta A = |A_1 – A_2| / 2$, and relational one – as $\Delta A / A$.

Calculations for students from group G8 give the following values for relational error: minimum – 2%, maximum – 16% and average – 8%. The results for group G9 are worse: minimum – 1%, maximum – 24% and average – 16%. Total average relational error for both groups is about 11%. This value can be taken as a total estimation of average error of method.

Obtained values of relational error we can compare with learning progress *P*, computed as $P = A_3 – A$, were A_3 is ratio *T/G* for output testing. The excess of this characteristic over relational error is equal to $(P – \Delta A) / A_3$.

The last parameter for group G8 is always positive: it means that values of learning growth for all students are higher than random method's error and hence can be distinctly fixed. Minimal magnitude of this parameter is 3%, maximum – 86% and average – 28%. As before, G9 shows lower results: maximum – 42%, average – 14%. The most disagreeable is that one student has negative result -13%: this means that her measured growth is less than error of method and hence can not be evaluated correctly.

Thus we see that 11 students from 12 in our experiment have valid learning progress and it suggests us that our experimental method generally has valid results.

Fig. 4 illustrates these calculations by diagrams. There we can see total results for all students from both groups. Every student is represented by means of 3 columns: two pretest B1 and B2 (dark colors) and posttest A (light color). It is clear that students S2, S4 and S5 in G8 and S1 and S2 in G9 have evident learning progress. Student S4 in G9 has negative (undetermined) results: value for output testing A is lower than input one for B2. Other students have less obvious but positive educational progress.

Test 2: stability of pair selection

My previous publications concentrated an attention on statistics of experimental results. Now, studying the stability of obtained results, let us examine concrete pairs that students construct and evaluate their permanence.

To estimate this feature we have 9 doubled pretests for group G8 (note that two students do not finished the semester, so in the previous section you saw number 7 for G8!) and 5 test results for G9. The analysis is easy: a short program reads both pretest files for every student and calculates percentage of pairs that concur.

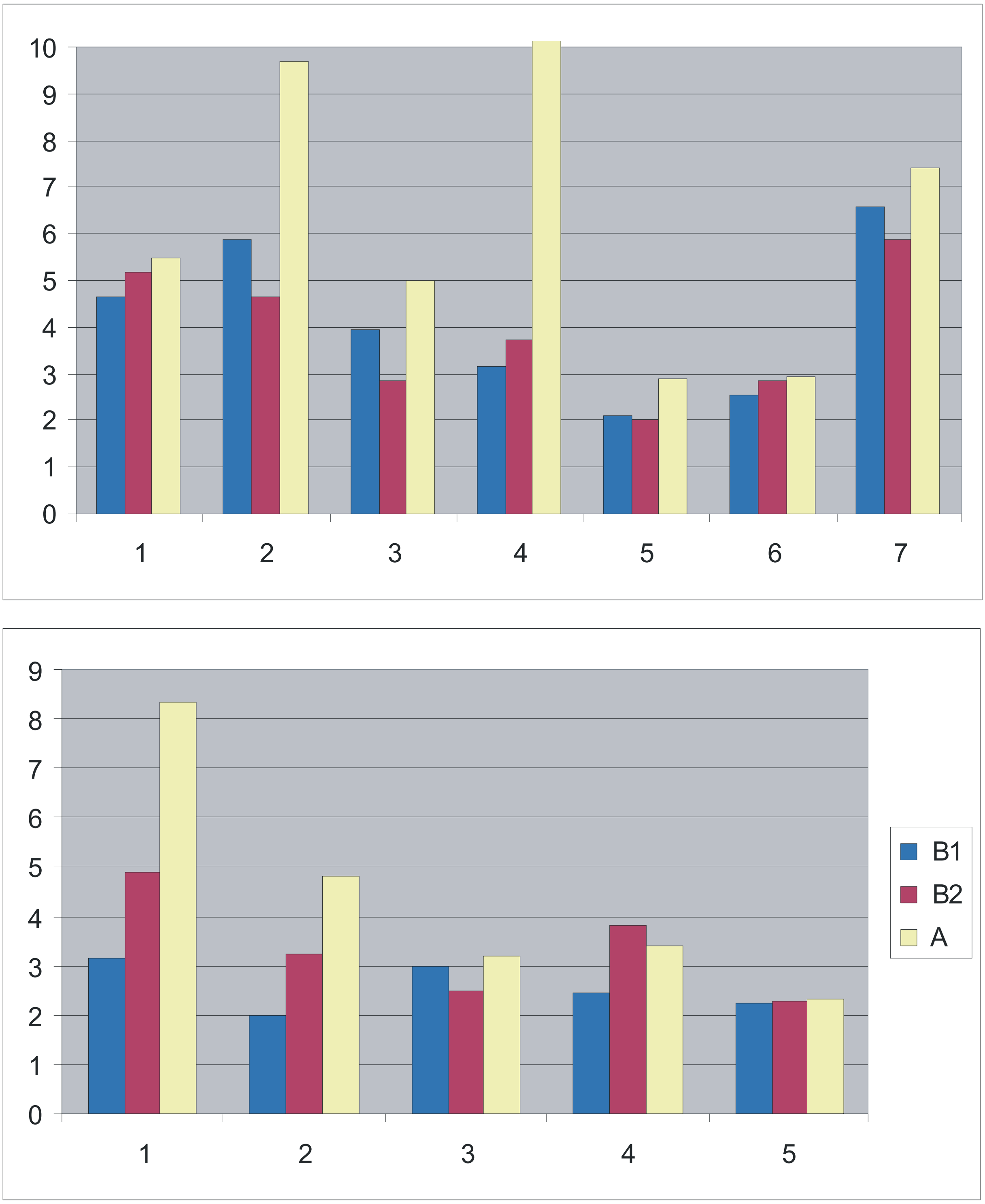

Figure 4

The results of such processing are the following. Students from group G8 demonstrate minimum 13% and maximum 48% congruence; average value is 30%. “Weaker” students from small group G9 gave maximum 24% and average value only 11%. To understand such difference we must have in view that groups has different specialties and G8 students named much more pairs: for G8 average number of pairs was 33 and they included on the average 41 topics, while G9 had only 14 links from 21 terms! On my opinion all results are too low, but this is what my students really showed.

The statistical analysis also showed a very interesting fact: a plot for number of overlap pairs looks very similar to correspondent plot for our gauge *T/G*, selected in present research as a measure of learning success. The linear correlation between these variables is very high and equals to 0.92 for G8; value 0.69 was formally calculated for 5 students of G9 and total magnitude for all 14 students is 0.88. I suppose such accordance is a certain indirect confirmation that we are really evaluating some objective ability of a student to work with conceptual knowledge. Unfortunately the number of students in G8 and G9 is too small to make a confident conclusion.

Test 3: stability of knowledge fragmentation

Results of our study indicate that students can not build (or maybe can’t show that they are able to built) an entire picture of the discipline they learnt. It is a serious blame, so we must check this situation more carefully.

The first easy to do control is to give students a direct task to build an entire picture. We must emphasize that during the basic experiment students did not know its aim [Eremin, 2010]. What if they will specially try to build one interrelated picture?

Such estimation experiment was implemented in 2009 for several students. They were asked to build one connected picture, but nevertheless the results were fragmented. Although student specially tried to construct non-fragmented system of terms and total number of groups decreased 2-5 times, nobody reached only one group of concepts. Furthermore the best result among these students was 4(!) groups.

Usually when students do not specially worry about minimization of group’s number, the result 4-5 groups is good. Only two from all 116 students managed to form 2 groups. In the first case student formed 2 groups in pretest, but her posttest contains 6 groups. In the second case in two pretests results were 12 and 7, and in posttest he showed 2 groups. So it seems to be rather chance than consistent pattern.

Hence we can see from all these facts that it is very long distance from students’ answers (remember fig. 3) to ideal teacher’s picture (fig. 1). But let us check this thesis more careful yet.

For this purpose in 2013 another method was trialed for building of groups from topics. The list of terms was absolutely the same as earlier, but the task sounds in other way. Students were proposed to link to every key concept as many related concepts as possible. To make the task clearer it was an example, demonstrating that to category *fruit* we can link several *subclasses* like *apple*, *orange* and so on. Applying this method to computer architecture, we may specify for instance that concept *functional unit* has subclasses *processor*, *memory*, *input devices* and *output devices* (remember fig. 2).

The experimental task looks quite similar, but upon a close view it stimulates grouping of concepts: it directly prompts a student that he must link to selected concept as many terms as possible – pair strategy has no such prompt. Besides new modification of method makes irrational to construct the smallest groups from two terms: it means introducing generalized concept for one term – such generalization is useless!

Let us mark the basic method as **MCP** (*Method of Constructing Pairs*) and proposed modification as **MGT** (*Method of Grouping Terms*). The results for group G10, tested both ways, are the following (see tab. 1).

Table 1

	method	1	2	3	4	5	6	7	8	9	10	11	12	mean
terms	MCP	57	49	50	54	40	36	41	58	60	61	71	50	52.3
	MGT	44	41	57	47	25	34	42	39	34	62	49	20	41.2
groups	MCP	15	14	8	15	13	9	10	12	12	16	14	14	12.7
	MGT	6	8	10	7	5	6	7	7	7	14	10	5	7.7
T/G	MCP	3.8	3.5	6.3	3.6	3.1	4.0	4.1	4.8	5.0	3.8	5.1	3.6	4.2
	MGT	7.3	5.1	5.7	6.7	5.0	5.7	6.0	5.6	4.9	4.4	4.9	4.0	5.4

The most essential changes were observed in the number of groups – it decreased approximately 1.7 times in average. At once total number of terms also decreased 1.3 times, so finally ratio T/G grew about 1.3 times.

We see that although the number of group noticeably decreased, even its minimal value 5 is far from the ideal. Hence essential change of method did not lead to disappearance of fragmentation.

Comparing data for MCP and MGT variants of method, I also determined common pairs for both experiments. Average ratio of common pairs to total number of links is approximately 34% for MCP and 40% for MGT (as it follows from tab. 1, total number of terms for MCP is greater, so fraction for it is smaller). Comparing with values 30% and 24% for experiments in groups G8 and G9, described in the previous section, we can discover some accordance of these results.

Test 4: evaluation of RTM effect

Beginning to learn the course “Computer architecture”, students already know many facts about computer. So after the final control we must compare the results before and after the course digest, expecting to seclude student’s real growth. Analysis of testing results before (pretest) and after teaching (posttest) is wide-spread practice in educational research.

So natural to general logic of education, pretest/posttest method is in a center of an active theoretical discussion. It traces back to 1956 when F.M. Lord published his work “The measurement of growth” [Lord, 1956], in which concluded that “differences between scores tend to be much more unreliable than the scores themselves”. Later Rogosa and Willett [Rogosa and Willett, 1983] investigated the problem wider and proved that limitations of difference measurement are not universal: in many important situations the difference scores are reliable. The authors formulated a very clear practical guideline: “the difference score cannot detect individual differences in change that do not exist, but it will show good reliability when individual differences in true change are appreciable”. We already used this principle of maximum pretest/posttest difference while selecting a suitable experimental gauge.

As careful investigation shows, difference scores have limitations of various natures. One of them is **regression towards the mean** (RTM; e.g. [Barnett et al, 2005]). As RTM may cause noticeable experimental error, it is very useful to explore this phenomenon.

RTM was first described in 1886 by F. Galton, who studied the dependence of adult children heights from specially calculated middle height of parents. He discovered that extremely tall parents had shorter children and, conversely, children of short parents tend to be taller than their parents. Later on such effects were found in many various areas, especially in medicine and education [Bland and Altman, 1994].

The substance of RTM may be explained by means of the following example. Suppose a normal student who felt himself bad or was strongly disconcerted with something while doing pretest and therefore demonstrated extremely low result. Posttest was long after and our imaginary student naturally got his average score (much closer to the mean than the primary one). Such situation has two abnormalities: first, our examinee will be mistakenly classified according to pretest to low level of primary knowledge and, second, his numeric growth will be undeservedly high. Conjunction of these circumstances may lead to wrong conclusion that students with low initial knowledge show more essential growth than others: unfortunately this effect is indistinguishable from really high learning outcome. Trying to avoid such fault we ought to do primary measurement more carefully, i.e., like theory usually recommends, several times; but is it real to oblige students to pass through the same pretest 3-4 times? To reduce RTM effect practically, Davis [Davis, 1976] suggested measuring pretest baseline twice, using one data set to calculate differences and the other one for baseline definition. In my experiments the idea of independent evaluation of students' growth and their baseline was put to use even more strictly: students were ranked according to rating, absolutely independent from pretest.

A special mathematical methods were developed to evaluate is this kind of experimental inaccuracy significant or not. The most suitable estimation was proposed in [Ostermann et al, 2008]. To apply it we must know only general statistical characteristics of pre- and posttest such as the mean and the standard deviation; calculating formula also contains correlation between the results of two measures.

The authors of the cited paper not only described the mathematical basics of evaluation technique, but kindly attached an Excel file to realize all calculations (http://www.biomedcentral.com/1471-2288/8/52). It remains just to substitute our values into definite cells of a spreadsheet and get immediate outcome. My experimental parameters for groups G1-G6 were the following: pretest – the mean 4.87 with deviation 2.66 and posttest – 5.58 with 2.76; correlation between scores 0.46; total number of students – 67. The main estimation result for specified values is that if the "true mean" (term "true" in publication represent value in "real nature", but not the statistical one!) of test is lower than 5.1, RTM effect is not essential with p-value < 0.025. It is reasonable to assign average value 5.23 to this "true" variable in our case. Then, as Excel diagram shows, p-value will not exceed 0.05. According to such small magnitude, the RTM effect on our experiments must not be significant.

Test 5: different disciplines

This section shows that the results of experiments do not depend upon learning discipline. All previous results were obtained for "*Computer architecture*". What will change if we take another educational discipline?

In order to answer this question some additional experiments were fulfilled during learning "*Mathematical logic*" course. Method and software for testing were identical to previous experiments, I just changed the text file with the list of terms (new list contained 88 concepts).

The only difference in processing of the experimental data was the following. During "Computer architecture" control, students were rank-ordered according to some rating: the criterion of such arrangement was time of finishing all the tasks, given by the teacher. Students with small numbers finished the course earlier; hence they are supposed to demonstrate better results in learning the course content. In opposite, columns for «the slowest» students form the right part of the picture (see fig. 4).

The course of mathematical logic is organized in slightly different manner: instead of control tasks in computer class, students must solve several tasks in a written form. The marks for this work also allowed arranging students according to their results. It is important to note, that in both cases students' rating was defined *independently* from the results of experiments: this requirement is aimed at reducing RTM statistical effect, which was discussed in the previous section.

It is evident that the results of experiments for both educational courses can not be compared directly – we previously need to normalize them, using common scale. The normalization procedure must be executed for both coordinates – arguments and results.

Let us begin from abscess – student's rating *n*. It is integer dimensionless value, but the problem is that groups have different size. Hence we must reduce values of this rating according to total number of students *N* in the group. To solve this problem we introduce the formal variable *X*, which has a meaning of improved non-dimension student's rating. We assign value X = 0 to the first student with n = 1, and X = 1 – to the last student with n = N. For any other student with rating *n* we can calculate variable X_n, using formula

$$X_n = (n-1) / (N-1), \text{ where } n = 1, 2, \ldots, N$$

We see that *X* is a real variable, possesses the value in the interval from 0 to 1.

Now we consider the ordinate – our learning progress gauge *T/G*. As semantic nets for computer architecture and mathematical logic are quite different, it is incorrect to put them together. Following [Eremin, 2011] let us introduce the ratio

$$K_{tg} = (T_2/G_2) / (T_1/G_1),$$

where index 2 relates to posttest and 1 – to pretest. Non-dimension value K_{tg} describes the growth of conceptual knowledge and we can suppose is suitable for comparison of results for different disciplines.

Using the introduced above scales, we can put dots for growth coefficient K_{tg}, preliminary translating abscissa X_n for every student. The result scatter plot is shown in fig. 5. Every round black point conforms a student from groups G1-G6; not filled dots were got for mathematical logic: round ones describe group G9 and square dots – G8. Note that mathematical logic course was in the first semester, when 7 students studied in G9.

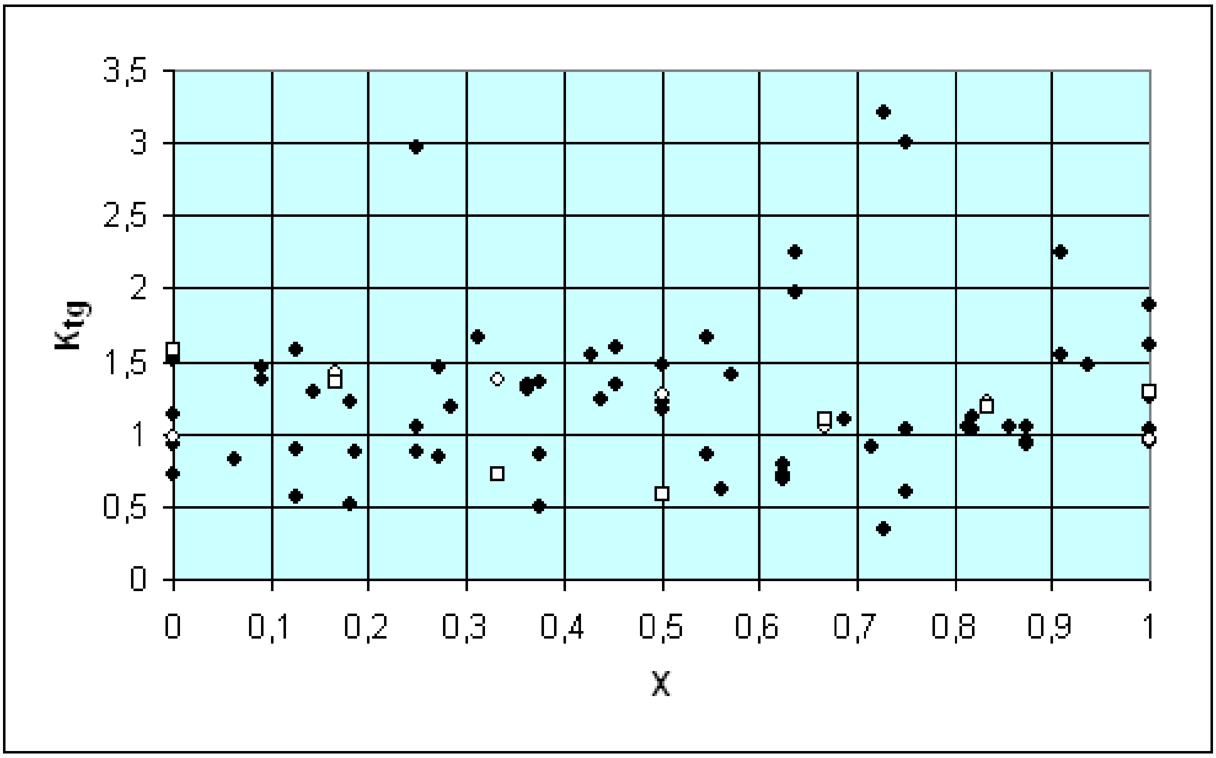

Figure 5

Diagram shows that dots for mathematical logic lie among the dots for computer architecture and variation of their positions do not exceed the dispersion of the last ones. Hence no dependence from learning material can be noticed.

Contrariwise, the values of *T/G* for mathematical logic are noticeably lower than for computer architecture. So as it was expected, these variables for different disciplines can not be compared without conversion.

And one more interesting detail. By occasional concatenation of circumstances one student made his output test twice. Two resulting values, processed as it was described in the section "Test 1: repetition of experiments", gave for posttest relational error about 11%, i.e. very close to average pretest error.

Conclusion

Additional experiments were implemented to check the stability of results, obtained in previous research of students' conceptual knowledge. Variation of some method's conditions gave possibility to estimate how these changes influence on observed picture. Several different tests were organized for these purposes. First of all calculation of the difference between two pretests in the same group evaluates the magnitude of relational error of testing procedure (its average value was near 11%). Then the stability of grouping concepts process was investigated. In spite of some difference in numeric results, the main conclusions were found to be stable. Special evaluation of RTM effect, which can cause incorrect results for low-performed students, showed that we may not worry about this effect in our study. And at last after confrontation of results for two different educational disciplines we come to a decision that scaled statistical characteristics do not depend upon learning material.

So all control experiments described in the paper assured that we can rely on previous study of conceptual knowledge.

Acknowledgement

The paper is published with partial support by the project ITHEA XXI of the ITHEA ISS (www.ithea.org) and the ADUIS (www.aduis.com.ua).

Bibliography

[Anderson and Krathwahl, 2001] L.W.Anderson, D.R.Krathwahl. A Taxonomy for Learning, Teaching and Assessing. Longman, New York, 2001.

[Barnett et al, 2005] A.G.Barnett, J.C.van der Pols, A.J.Dobson. Regression to the mean: what it is and how to deal with it. International Journal of Epidemiology, 34(1), 2005.

[Bland and Altman, 1994] J.M.Bland, D.G.Altman. Some examples of regression towards the mean. British Medical Journal, 309, 1994.

[Davis, 1976] C.E.Davis. The effect of regression to the mean in epidemiologic and clinical studies. American Journal of Epidemiology, 104(5), 1976.

[Eremin, 2009] E.A.Eremin. About computer method to study the entirety of system from basic concepts, which students form in the result of digestion a course. In: Human Aspects of Artificial Intelligence: IBS "Information Science and Computing", 12(3), 2009. (in Russian)

[Eremin, 2010] E.A.Eremin. A Computer Method to Study the Entirety of Students' Knowledge Acquired During an Educational Course. Information Theories and Applications, 17(4), 2010.

[Eremin, 2011] E.A.Eremin. Experimental study of students' knowledge entirety. Information Technologies and Knowledge, 5(3), 2011. (in Russian)

[Eremin, 2012] E.A.Eremin. In quest of effective gauge for learning course digestion. In: Multidisciplinary academic research. Conference Proceedings. Prague, 2012.

[Koponen et al, 2002] I.T.Koponen, T.Mäntylä, J. Lavonen. Challenges of Web-based education in physic teachers' training. In: Proceedings of International Conference on Information and Communication Technologies in Education. Badajoz, Spain, 2002.

[Lee et al, 2011] H.S.Lee, O.L.Liu, M.C.Linn. Validating measurement of knowledge integration in science using multiple-choice and explanation items. Applied Measurement in Education, 24(2), 2011.

[Lord, 1956] F.M.Lord. The measurement of growth. Educational and Psychological Measurement, 16(4), 1956.

[Loubser et al, 2001] C.P.Loubser, C.H.Swanepoel, C.P.C.Chacko. Concept formulation for environmental literacy. South African Journal of Education, 21(4), 2001.

[Ostermann et al, 2008] T.Ostermann, S.N.Willich, R.Lüdtke. Regression toward the mean – a detection method for unknown population mean based on Mee and Chua's algorithm. BMC Medical Research Methodology, 8, 52, 2008.

[Rogosa and Willett, 1983] D.R.Rogosa, J.B.Willett. Demonstrating the reliability of the difference score in the measurement of change. Journal of Educational Measurement, 20(4), 1983.

Author's Information

***Evgeny A. Eremin** – Senior lecturer, Perm State Pedagogical University, 614990, Sibirskaya st., 24, Perm, Russia; e-mail: eremin@pspu.ru*

Major Fields of Scientific Research: Knowledge representation, Computer learning, Computer architecture, Programming