International Education Studies; Vol. 7, No. 9; 2014
ISSN 1913-9020 E-ISSN 1913-9039
Published by Canadian Center of Science and Education

On Improving the Experiment Methodology in Pedagogical Research

Tereza Horakova[1] & Milan Houska[1]

[1] Czech University of Life Sciences Prague, Prague, Czech Republic

Correspondence: Tereza Horakova, Department of Systems Engineering, Faculty of Economics and Management, Czech University of Life Sciences Prague, Kamycka 129, 165 21 Prague 6-Suchdol, Czech Republic. Tel: 420-224-382-374. E-mail: horakovat@pef.czu.cz

Received: June 27, 2014 Accepted: July 29, 2014 Online Published: August 17, 2014

doi:10.5539/ies.v7n9p84 URL: http://dx.doi.org/10.5539/ies.v7n9p84

Abstract

The paper shows how the methodology for a pedagogical experiment can be improved through including the pre-research stage. If the experiment has the form of a test procedure, an improvement of methodology can be achieved using for example the methods of statistical and didactic analysis of tests which are traditionally used in other areas, i.e. measuring the learning outcomes. In particular, we deal with measuring the impact of different text styles of educational texts on the rate of knowledge transfer. For this purpose we are using the methods of statistical and didactic analysis. Statistical methods allow us to determine working hypotheses on statistical significance of differences in didactic characteristics of tests used for measuring the learning outcomes. The working hypotheses, whose validity have not been rejected, show us the weak points of the experiment and allow us to improve it before the experiment has been run in the full scale.

Keywords: pedagogical experiment, didactic test, pre-research, knowledge transfer, knowledge text

1. Introduction

Human learning is a non-simple process that may take place using different methods. Independently on the method, one of the most important present topics in education is the improvement of quality of the pedagogical process. This issue is of the same importance at schools (formal education) and in enterprises (informal or non-formal education) as well (EU Commission, 2014). The effort of improving the quality is also documented with the amount of public resources spent for enterprise education. This is related to the necessity of spending resources on education efficiently so that funds used for it do not become consumption but investment.

1.1 The Role of Pre-Research for Efficiency Measurement

As new approaches and innovative methods are establishing in education, at the same time new ways how to measure their impact on learning outcomes and students performance should be developed. To be systematic, any new metrics should be tested and verified before it has been deployed for the real measurement. In pedagogy, pre-research (pilot research, preliminary studies) are being used for this purpose. Nevertheless, to obtain results of satisfactory quality, even the pre-research has to be designed carefully, using appropriate qualitative or quantitative methods. In this work we focus on developing a particular methodology to carry out the pre-research to verify the properties and feasibility of the approach to measuring the efficiency of knowledge transfer through didactic texts. The following works inspired us to deal with this issue.

Garner et al. (2009) focus on scientific research in pedagogical practice which is to pass through the following three levels gradually: pilot study, pre-research, and the actual research. These stages are mentioned in Cohen (1998) as well in an extended form. Garner et al. (2009) state that the pilot study is a kind of the first probe into the patterns that we plan to examine. Many pedagogical researchers (e.g. Cobern et al., (2014); Antony, Sivanathan & Gijo (2014)) carry out pre-research which is a reduced model of the actual research itself and which serves for verifying the properties of research tools and techniques, for testing the hypotheses, and for statistical processing of results without arriving to general conclusions. Furthermore, Garner et al. (2009) distinguish between actual ex-post-facto research and experiments, and he points out that experiment is more suitable for examining the process of education because at least one independent variable which is under the control of the researcher is used. Information about particular pedagogical experiments can be found e.g. in Leutner et al. (2009), Rosales et al. (2012).

1.2 Efficiency of Teaching through Didactic Texts

Many authors (Tudor, 2012; Singer & Moscovici, 2008) deal with the problems of evaluating the education efficiency as well as with various pedagogic models. Rynne and Gaughran (2008) described the models of quality and efficiency on the pedagogic level. One of the frequently observed aspects of the efficiency of education is the efficiency of methods of teaching (Cobern et al., 2014), e.g. working with texts. The aim of experiments related to the analysis of teaching texts and textbooks is to clear up the importance of some properties and parameters of textbooks and their influence on the results of education (D. Newton & L. Newton, 2009). Duric and Song (2012) or Asaishi (2011) focused on the analysis of educational texts. The aspects that were evaluated and measured included, among others, the measure of having the textbook equipped from the didactics point of view, the measure of the difficulty of the text, the analysis of terms, the measure of information density and so on. Within the framework of interdisciplinary cooperation among related branches of science which deal with acquiring the knowledge and working with it, pieces of knowledge from the area of "knowledge engineering" commence to be used in pedagogic and didactic disciplines. Knowledge engineering provides formalized models for representing explicit knowledge which is a subject of transfer and sharing during teaching or at expert consultations. A. Glava and C. Glava (2011) emphasize the importance of the general theory of systems and system approach for this purpose. System approach has been used for deriving a new representation of knowledge, the so-called knowledge unit (Dömeová et al., 2008) as well. The advantage of this representation is the possibility of expressing the formal model of knowledge in a natural language which makes it accessible for human users. It also allows to create educational texts which contain the knowledge in an explicit form (knowledge texts). Furthermore, systems approach is suitable for the creation of metrics for evaluating the results of education which constitute one of the indications of the quality of knowledge transfer (Rosales et al., 2012). As Rottensteiner (2010) states, the key variable is the extent of understandability of text for readers and students in this case. Other researches, e.g. Igbaria (2013), Menon and Mukundan (2012) or Rahimy and Shams (2012) proved the impact of different styles of educational texts (distinguished in design, structure, narrative style, etc.) on the learning outcomes and final performance of the students.

In order to measure the efficiency of units and knowledge of texts on the education level, Houška and Beránková (2007) performed an experiment on small groups of users. It proved the feasibility of the proposed way of measuring; however, actual implementation of a follow-up study on a relevant sample of respondents and using statistic and economic methods has not been performed.

1.3 Objectives and Working Hypotheses

The objective of the paper is to demonstrate a way how to improve a particular experiment for a pedagogical research. We use statistical and didactic methods in the pre-research stage of pedagogical research for fine-tuning the imperfections of experiment methodology. Furthermore, we are aiming at proving or disproving the following working hypotheses:

H1: The distribution of the results (measured in points) achieved in the pre-test and in the post-test is normal.

H2a: The distribution of the knowledge transfer using knowledge text (measured in points) is normal.

H2b: The distribution of the knowledge transfer using common text (measured in points) is normal.

H2c: The distribution of the net knowledge transfer (measured in points) is normal.

H3: The distribution of the time of reading a leaflet (measured in seconds) is normal.

H4: Net knowledge transfer (measured in points) is influenced by the type of the text (knowledge text vs. common text).

H5: Average net knowledge transfer (measured in points) is identical in all sub-areas of the topics (chemistry, legislature, practice, economy). The distribution of results for both groups A (experimental) and B (control) is normal.

H6: All didactic characteristics (Q, P, n_H, n_L, ULI, p, q, $p*q$, see section 2 for their definitions) are identical in average for both groups A (experimental) and B (control).

H7: Standardization of tests is identical in the post-test for both groups A (experimental) and B (control) (measured by the percentile ranking of students) and, at the same time, the distribution of results (measured in points) achieved in the post-test for both groups A and B is normal.

2. Methods

2.1 Methodology of the Experiment

The proposal of the experiment came into being within the framework of a project focused on transferring knowledge in the area of processing the agricultural waste. The aim of this experiment is to find out whether knowledge transfer is influenced by the type of the text, i.e. common didactic text versus text processed using the methods of knowledge engineering.

The experiment has the form of a test procedure consisting of 4 parts: pre-test, educational text, post-test, and questionnaire. A period of 15 minutes is reserved for solving the assignments of the first section, a period of 30 minutes is reserved for reading the test, a period of 15 minutes is reserved for solving the post-test assignments, and the test participants are presented with a questionnaire at the conclusion.

Several pedagogic experiments (Ozuru et al., 2009; Tarchi, 2010) had already proven that the influence of the initial level of knowledge on further transfer in the process of education is significant; therefore, the experiment includes the pre-test, too, which is a tool for determining the initial level of knowledge of the test participants without the influence of educational text.

Text is a tool of knowledge transfer from four sub-areas (chemistry, legislature, practice, economy) within the framework of the selected problem domain corresponding to the topic of the project, i.e. biogas and biogas stations. Each person participating in the test receives the text in the form of a leaflet where the knowledge from two sub-areas is captured using common educational text, and the other two are captured using knowledge text according to the methodology proposed by Houška and Rauchová (2013). There are 4 formats of the leaflet available in total and they use a 4-digit code consisting of 0s and 1s (1100, 0011, 1010, 0101) where 0 denotes a sub-area processed using the common form, 1 denotes a sub-area processed using the knowledge form, the first position of the code represents the sub-area of chemistry, the second position of the code represents the sub-area of legislature, the third position represents the sub-area of practice, and the fourth position represents the sub-area of economy. For example, a leaflet with the 1100 code contains the sub-areas of chemistry and legislature processed using the knowledge form and the sub-areas of practice and economy processed using the form of common educational text.

Post-test is a tool for determining the effect of knowledge transfer using texts. Net knowledge transfer KT_{netto} in points is calculated using the following formula:

$$KT_{netto} = P_{POST} - P_{PRE} = KT_k + KT_o \tag{1}$$

where KT_{netto} is the net knowledge transfer in points,

P_{POST} is the number of points contained in the post-test,

P_{PRE} is the number of points contained in the pre-test,

KT_k is the net knowledge transfer in points using knowledge text,

KT_o is the net knowledge transfer in points using common educational text.

The aim of the experiment is to find out whether KT_k differs from KT_o by a statistically significant amount.

As the didactic tests (pre-test as well as post-test) constitute a tool for an experiment and they should be used to measure the knowledge transfer in a reliable way, it is necessary to verify their didactic properties within the framework of pre-research as well. Both tests contain 20 assignments with the choice of one correct answer; each of the four sub-areas is referred to by 5 questions. Garner et al. (2009) recommend to offer the choice of one out of six answers including the “I do not know” option for each question in order to prevent guessing an answer. The questions of the pre-test and the post-test survey the same information but they are formulated in a different way: where the pre-test assignment is presented in the form of completion of information, the post-test assignment contains it rephrased as a question requiring the completion of knowledge and, vice versa, where the pre-test assignment is presented in the form of a question requiring the completion of knowledge, the post-test assignment contains it rephrased as the completion of information. In each test, one half of the questions are formulated to require the completion the information, and the other half the completion the knowledge.

Respondents were separated into two groups: experimental group (A) and control group (B). The experimental group worked with pre-test, text, and post-test and the control group worked with the same materials but in a different order: post-test, text, and pre-test. This reverse order is possible, because the questions in the pre-test and the post-test are aimed at the same information, but they are formulated in a different way, as mentioned above. This separation is important for testing the methodology of prepared pre-tests and post-tests and their reciprocal correspondence.

Questionnaire is a tool for finding out the opinions of the experiment participants on each text from the point of view of their understandability, difficulty of the topics or naturalness; in the pre-research stage, the questionnaire contains questions from the point of view of time available for each part, form of testing and others as well.

2.2 Verification of Properties of a Didactic Test

Garner et al. (2009) present several formulae that can be used for the verification not only of properties of each test assignment but of the didactic test as a whole. When analyzing the difficulty of test assignments, the difficulty value (*Q*) is calculated. It represents the percentage of test participants in the sample who chose an incorrect answer for given assignment:

$$Q = 100\frac{n_n}{n} \tag{2}$$

where Q is the difficulty value,

n_n is the number of test participants in the group who chose an incorrect answer or who did not answer,

n is the total number of test participants in the sample.

Sometimes, difficulty index (*P*) is expressed as the percentage of test participants in the group who chose the correct answer for given assignment:

$$P = 100\frac{n_s}{n} \tag{3}$$

where P is the difficulty index,

n_s is the number of test participants in the group who chose the correct answer,

n is the total number of test participants in the sample.

The sensitivity of test assignments is measured using indicators that express the discrimination ability of assignments, i.e. the extent of giving the test participants with better knowledge an advantage over the participants with weaker knowledge by given assignment. Therefore, it is necessary to split the group of test participants based on the total number of acquired points into two parts, i.e. test participants with higher score and test participants with lower score of points acquired in total, before constructing the actual indicator. There are many coefficients available for determining the sensitivity of test assignments; the simplest ones include the ULI (upper-lower-index):

$$ULI = \frac{n_L - n_H}{0.5n} \tag{4}$$

where ULI is the sensitivity coefficient,

n_L is the number of people from the better group who chose the correct answer for given assignment,

n_H is the number of people from the worse group who chose the correct answer for given assignment,

n is the total number of test participants in the sample.

All sensitivity coefficients acquire values ranging from -1 to +1; the higher the value of a coefficient, the better the ability of given assignment to distinguish among people with better knowledge and people with weaker knowledge. If the coefficient is equal to 0, it means that the success rate of people from both groups is the same for given assignment and the assignment does not distinguish between these groups; if the coefficient is very low or even negative, it means that the assignments are formulated in a complicated way or the assignments with answers to choose are difficult.

According to Garner et al. (2009), when analysing the properties of a didactic test as a whole, its validity, reliability and practicality is considered. Validity is a basic property of a didactic test. Some authors (Zainal, 2012; Naqvi et al., 2010) test the validity in order to find out whether the tests really examine what should be examined. Cohen (1998) states that the reliability of a didactic test indicates its dependability, i.e. whether repeating the test under the same conditions will lead to the same or similar results. Garner et al. (2009) measure the reliability in an exact way using the reliability coefficient which acquires values ranging from 0 to 1. For tests with the number of questions higher than 10, it is desirable for the reliability coefficient to acquire at least the value of 0.8. Reliability coefficients are calculated in several ways but the one used often is the Kuder-Richardson formula; among others, Miller and Achterberg (2000) use it in their paper:

$$r_{kr} = \frac{m}{m-1}\left(1 - \frac{\sum_{i=1}^{m} p_i q_i}{s^2}\right) \quad (5)$$

where r_{kr} is the reliability coefficient,

m is the number of assignments in the test,

p is the ratio of students in the sample who chose the correct answer for the test assignment,

q is the ratio of students in the sample who chose an incorrect answer for the test assignment,

s is the standard deviation for overall results of students in given test.

By standardizing the didactic test, as stated by Jiang (2011), it is possible to compare the performance of an individual test participant to a representative sample of other test participants. A test standard is created to rank the test participant using a certain scale. The easiest method is standardizing using percentiles where percentile rank (*PR*) is determined using the following formula:

$$PR = 100\frac{n_k - 0.5n_i}{n} \quad (6)$$

where *PR* is the percentile rank,

n_i is the frequency of given result,

n_k is the cumulative frequency of given result,

n is the total number of test participants in the sample.

The value of percentage rank determines what percentage of test participants in the reference group achieved a result that is worse or equal to the result of given test participant.

2.3 Research Sample

For the pre-research we had addressed 40 students (university and high-school) in total using social marketing during December 2013, thereby creating a non-representative selection sample; according to Garner et al. (2009), this scope is sufficient for the purpose of pre-research.

The students participated in person in the whole experiment in the pre-research stage in smaller groups: there were 24 students in group A and 16 students in group B. Group A was presented with the pre-test and the post-test, group B worked with them in reverse order, i.e. its pre-test was identical to the post-test of group A and its post-test was identical to the pre-test of group A. In order to retain the equality of conditions, detailed methodology had been prepared for experiment administrators, too, who were supervising the experiment in order to ensure its proper course. By participating, the students actively contributed to the improvement of the methodology of the actual experiment itself which will take place in spring of 2014 with a representative sample of 300 test participants (people from the agriculture branch).

2.4 Statistical Analysis

In order to establish the basic idea about acquired data, indicators of descriptive statistics are used, especially the measure of central tendency and the measure of variance of acquired data which are described in detail in books (e.g. Lindsey, 2009). Furthermore, it is the calculation of the confidence interval for the average from the area of statistical induction (Basset et al., 2000), or the parametric testing of statistical hypotheses of equality of two averages (T-test) or of equality of two variances (F-test), as well as the testing of the assumption of normal distribution of data using the distribution test (Shapiro-Wilk test); these are described in detail in books (Gravetter & Wallnau, 2009). The method of single-factor analysis of variance (ANOVA) is used to determine the statistical differentness of averages for more groups. ANOVA is one of the most commonly used statistical techniques. It is a summary name for a group of very efficient methods with the common idea of decomposing the total variability into components that can be assigned to each reason of variability. The aim is to compare the level of studied quantitative variable in several groups which the basic set is decomposed into. The sorting criterion for decomposing the basic set into groups was one or more variables of the nominal type (Bassett et al., 2000). For one-dimensional analysis of variance, the influence of one factor on a certain dependent variable is observed (Lindsey, 2009).

3. Results

Pre-research took place in all stages: statistical processing of the results, testing hypotheses, and deducing the conclusions from the point of view of properties of each research tool (didactic analysis of tests). Firstly, the basic descriptive statistics were calculated; these are shown in Tab. 1. The maximum number of points acquired by test participants was 12 in the pre-test and 15 in the post-test; the average net transfer is 4.20 points which consists of the transfer using common educational texts by 52.38% in average while the remaining 47.67% results from transfer using knowledge texts. Text studying took the test participants 806.83 seconds in average, the longest time needed was 1201 seconds. This shows that the time of 30 minutes allocated for text studying is sufficient and it even allows for a margin. Furthermore, the frequencies of individual characteristics were represented graphically and the *p* values were calculated using the Shapiro-Wilk test (Tab. 2). The zero hypothesis declares that the data come from a set with normal distribution, the alternative one declares the opposite. For each observed variable, the p value is higher than the selected significance level $\alpha = 0.05$; therefore, the zero hypothesis (H0) remains valid in all cases, declaring that the data come from a set with normal distribution. The graphical representation of the distribution of data for selected variables is shown in histograms in Figures 1, 2 and 3.

Table 1. Descriptive statistics of observed variables

Variable	Range	Average	Average interval -95%	Average interval +95%	Min.	Max.	Standard deviation
P_{PRE} (points)	40.00	5.23	4.30	6.15	0.00	12.00	2.88
P_{POST} (points)	40.00	9.43	8.54	10.31	2.00	15.00	2.77
KT_{netto} (points)	40.00	4.20	3.37	5.03	-1.00	12.00	2.58
KT_o (points)	40.00	2.20	1.52	2.88	-2.00	7.00	2.11
KT_k (points)	40.00	2.00	1.46	2.54	-2.00	5.00	1.69
t^* (s)	40.00	806.83	750.13	863.52	497.00	1201.00	177.28

$^*t$: time of study.

Table 2. P values of the Shapiro-Wilk test

Variable	*p* value	$\alpha = 0.05$
P_{PRE} (points)	0.16	H0 is not rejected
P_{POST} (points)	0.33	H0 is not rejected
KT_{netto} (points)	0.15	H0 is not rejected
KT_o (points)	0.26	H0 is not rejected
KT_k (points)	0.06	H0 is not rejected
t (s)	0.17	H0 is not rejected

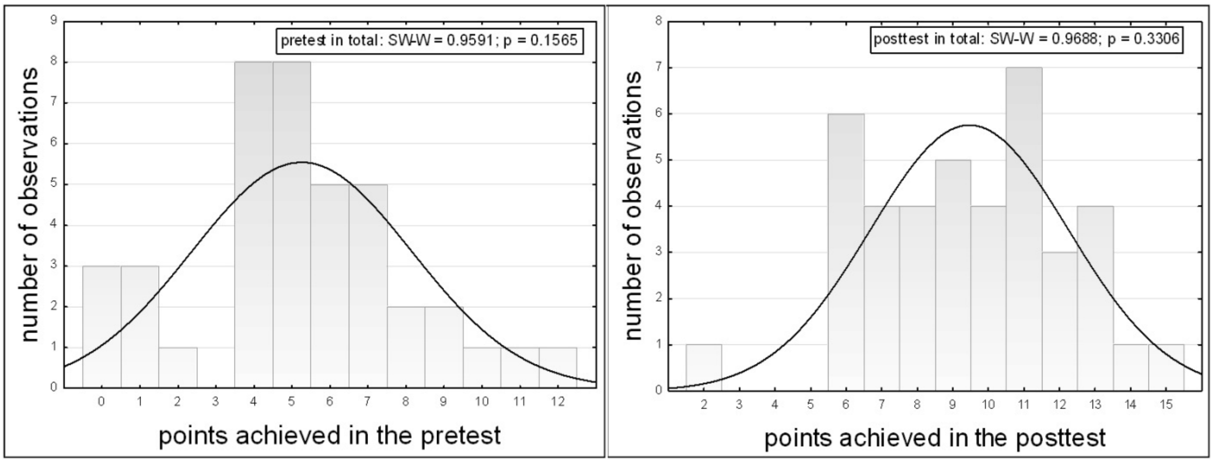

Figure 1. Histograms of variables P_{PRE} (on the left) and P_{POST} (on the right)

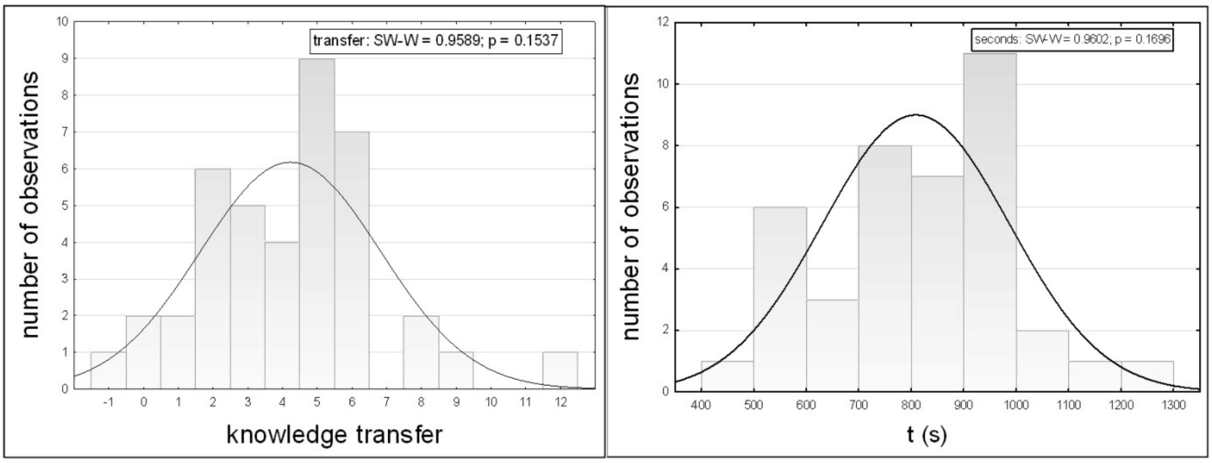

Figure 2. Histograms of variables *KTnetto* (on the left) and *t* (on the right)

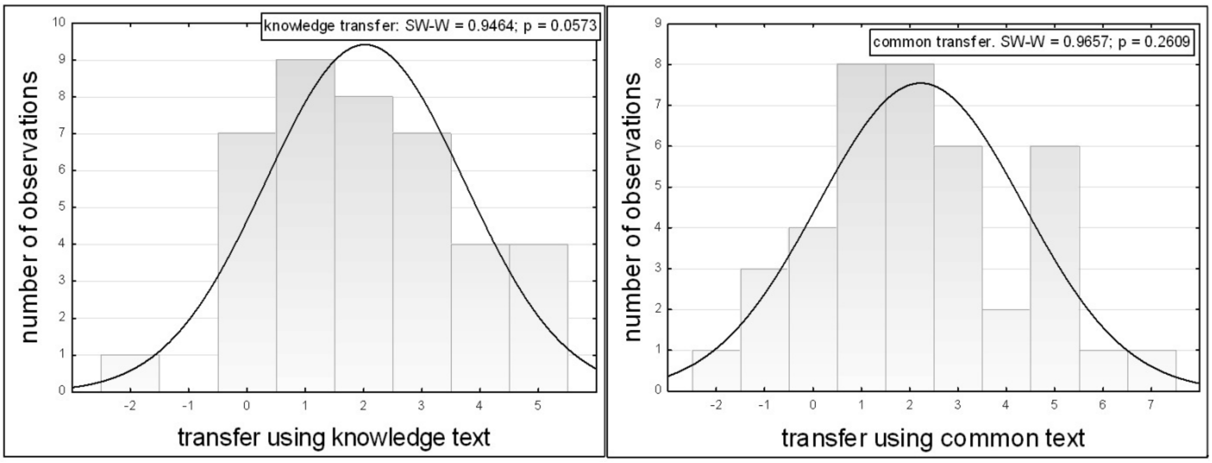

Figure 3. Histograms of variables KT_k (on the left) and KT_o (on the right)

The analysis of variance was used to determine whether the net knowledge transfer is influenced by the type of the text. The zero hypothesis H0 declares that the averages of both sets are identical; the alternative one declares that at least one set with an average that is different from the other groups exists. The *p* value of the F-test is *p* = 0.6400 which is more that the selected significance level α = 0.0500; therefore, the zero hypothesis (H0 about identical averages) remains valid in all cases. The graphical representation is shown in a box-plot in Figure 4.

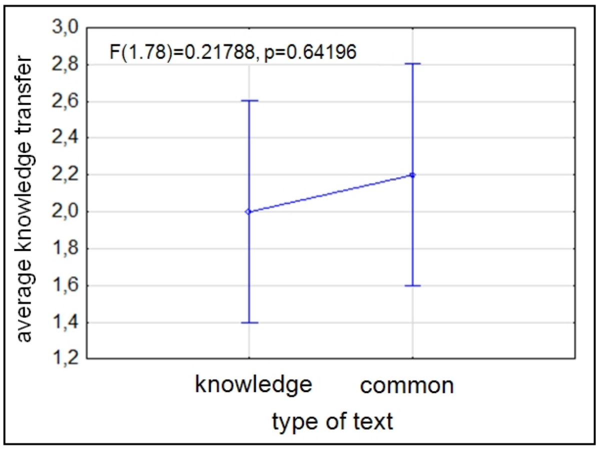

Figure 4. Box-plot showing the influence of the differentness of the type of text on the average knowledge transfer

Furthermore, the analysis of variance was used to determine whether the topic of the sub-area of the text is influenced by the type of the text. The *p* value of the F-test is *p* = 0.0014 which is less than the selected significance level *α* = 0.0500; therefore, the zero hypothesis (H0 about identical averages) is not valid any more but there is at least one area with an average that is different from the others. More detailed evaluation (Tukey test) has shown that statistically significant difference is found between the topics of legislature and chemistry (*p* = 0.000751), and between the topics of legislature and economy (*p* = 0.022417) which can be seen in Figure 5 as well. For future, it is necessary to focus on the legislature area in particular which contributed to the transfer by the smallest amount in average because this topic has also achieved the highest frequency from the point of view of the highest difficulty as compared to the other topics in the answers in the questionnaire.

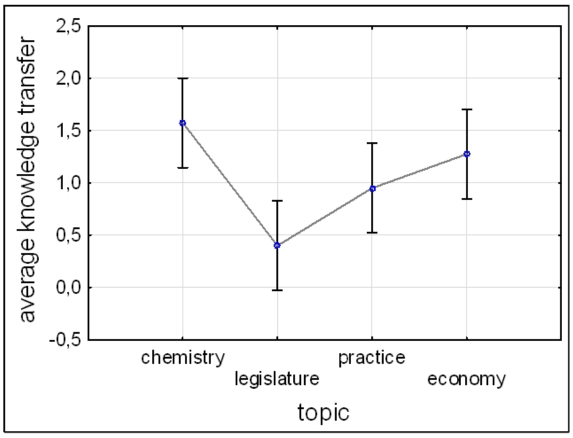

Figure 5. Box-plot showing the influence of the topic of sub-areas of the problem domain on the average knowledge transfer

Furthermore, didactic analysis of the tests was performed in order to determine whether the research tools that had been used measured the knowledge in a reliable way as expected. The test participants were assigned to groups A and B. 24 students were assigned to group A, 16 students to group B. Group A was presented with the pre-test and the post-test, group B worked with them in reverse order; for further information, see the chapter Material and Methods. The aim of didactic analysis was to calculate the indicators of difficulty of each test assignment, their sensitivity, test reliability and percentile scale of the post-test results for group A and group B, and to compare these tests to each other statistically.

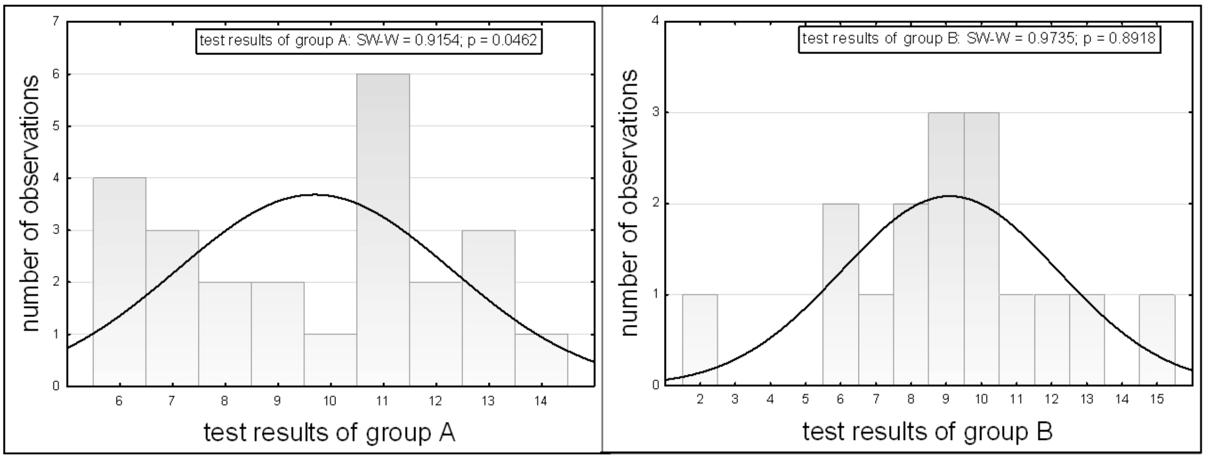

Figure 6. Histograms for the post-test results of groups A (on the left) and B (on the right)

Table 3. Didactic analysis of the post-test for group A

Question No.	Q	P	n_H	n_L	ULI	$p= n_s/n$	$q= 1-p$	$p*q$
1	29.17	70.83	6.00	11.00	0.42	0.71	0.29	0.21
2	37.50	62.50	4.00	11.00	0.58	0.63	0.38	0.23
3	29.17	70.83	5.00	12.00	0.58	0.71	0.29	0.21
4	62.50	37.50	6.00	3.00	-0.25	0.38	0.63	0.23
5	20.83	79.17	9.00	10.00	0.08	0.79	0.21	0.16
6	66.67	33.33	3.00	5.00	0.17	0.33	0.67	0.22
7	66.67	33.33	2.00	6.00	0.33	0.33	0.67	0.22
8	29.17	70.83	10.00	7.00	-0.25	0.71	0.29	0.21
9	12.50	87.50	11.00	10.00	-0.08	0.88	0.13	0.11
10	70.83	29.17	1.00	6.00	0.42	0.29	0.71	0.21
11	62.50	37.50	4.00	5.00	0.08	0.38	0.63	0.23
12	29.17	70.83	6.00	11.00	0.42	0.71	0.29	0.21
13	29.17	70.83	10.00	7.00	-0.25	0.71	0.29	0.21
14	83.33	16.67	1.00	3.00	0.17	0.17	0.83	0.14
15	75.00	25.00	2.00	4.00	0.17	0.25	0.75	0.19
16	79.17	20.83	1.00	4.00	0.25	0.21	0.79	0.16
17	58.33	41.67	2.00	8.00	0.50	0.42	0.58	0.24
18	66.67	33.33	0.00	8.00	0.67	0.33	0.67	0.22
19	75.00	25.00	1.00	5.00	0.33	0.25	0.75	0.19
20	50.00	50.00	5.00	7.00	0.17	0.50	0.50	0.25

Table 4. Didactic analysis of the post-test for group B

Question No.	*Q*	*P*	n_H	n_L	*ULI*	$p = n_s/n$	$q = 1\text{-}p$	$p*q$
1	12.50	87.50	8.00	6.00	-0.25	0.88	0.13	0.11
2	31.25	68.75	5.00	6.00	0.13	0.69	0.31	0.21
3	43.75	56.25	4.00	5.00	0.13	0.56	0.44	0.25
4	87.50	12.50	1.00	1.00	0.00	0.13	0.88	0.11
5	43.75	56.25	5.00	4.00	-0.13	0.56	0.44	0.25
6	43.75	56.25	3.00	6.00	0.38	0.56	0.44	0.25
7	43.75	56.25	4.00	5.00	0.13	0.56	0.44	0.25
8	37.50	62.50	6.00	4.00	-0.25	0.63	0.38	0.23
9	93.75	6.25	0.00	1.00	0.13	0.06	0.94	0.06
10	56.25	43.75	3.00	4.00	0.13	0.44	0.56	0.25
11	31.25	68.75	4.00	7.00	0.38	0.69	0.31	0.21
12	62.50	37.50	4.00	2.00	-0.25	0.38	0.63	0.23
13	56.25	43.75	2.00	5.00	0.38	0.44	0.56	0.25
14	50.00	50.00	3.00	5.00	0.25	0.50	0.50	0.25
15	31.25	68.75	6.00	5.00	-0.13	0.69	0.31	0.21
16	68.75	31.25	3.00	2.00	-0.13	0.31	0.69	0.21
17	75.00	25.00	1.00	3.00	0.25	0.25	0.75	0.19
18	75.00	25.00	0.00	4.00	0.50	0.25	0.75	0.19
19	87.50	12.50	1.00	1.00	0.00	0.13	0.88	0.11
20	62.50	37.50	1.00	5.00	0.50	0.38	0.63	0.23

Figure 6 shows the histograms of the post-test results separately for group A and for group B. According to the amount of the *p* value (higher than the selected significance level $\alpha = 0.0500$) for the Shapiro-Wilk test with group B, the data come from a group with normal distribution ($p = 0.8919$). For group A, the p value is $p = 0.0462$, i.e. with the selected significance level $\alpha = 0.0500$, the data do not come from a set with normal distribution; however, when the significance level is reduced to $\alpha = 0.0100$, the zero hypothesis cannot be dismissed.

In Table 3, it is obvious that for 10 questions, more than 50% of test participants in group A chose an incorrect answer. For four questions, the sensitivity coefficient is negative; this has to be improved in the research itself by rephrasing the respective questions. In Table 4, it is obvious that for 12 questions, more than 50% of test participants in group B chose an incorrect answer. For six questions, the sensitivity coefficient is negative, and for one question, it equals zero; this has to be improved in the research itself by rephrasing the respective questions. A negative *ULI* coefficient appears in both groups for question 8 only which refers to specific investment costs for building a biogas station (this question is from the sub-area “economy”).

Furthermore, reliability coefficients of tests were calculated, achieving $r_{kr} = 0.97228$ for group A and $r_{kr} = 0.59880$ for group B. This significant difference was caused by differing values of the sum of indicators $p*q$ (last columns in Table 3 and Table 4) and the standard deviation for the results of testing. For group A, the test reliability is very good; for group B, it is necessary to get the reliability coefficient close to the threshold of 0.8, as stated by Garner et al. (2009).

For each indicator, basic descriptive characteristics of the post-test were calculated as well; see Table 5 which shows the group that the indicator refers to, i.e. (A) or (B), for each variable. For example, Table 5 shows that the left-hand side of the confidence interval for the average for the sensitivity coefficient acquires negative values, which implies that the respective questions should be rephrased. Furthermore, each average value of selected didactic indicator of the post-test for group A and group B was tested using the T-test resp. the F-test, i.e. a check was performed in order to determine whether their average values resp. variances are statistically different, or not.

Each result of testing including the *p* values of F-tests and T-tests is shown in Table 6. The indicators do not differ from each other by a statistically significant amount, with the exception of the *nL* indicator (number of people in the better group who chose the correct answer for given assignment); the *p* value of the T-test ($p = 0.0000$) is lower than the selected significance level $\alpha = 0.0500$ here. From the point of view of standardization of didactic texts, percentile scales were constructed and also tested from the point of view of their statistical differentness; the *p* value of the T-test is higher than the selected significance level ($\alpha = 0.0500$, $p = 0.8286$) so the zero hypothesis about the equality of averages cannot be dismissed and the percentile ranking does not differ for group A and B by a statistically significant amount. Figure 7 shows the percentile ranking of each group.

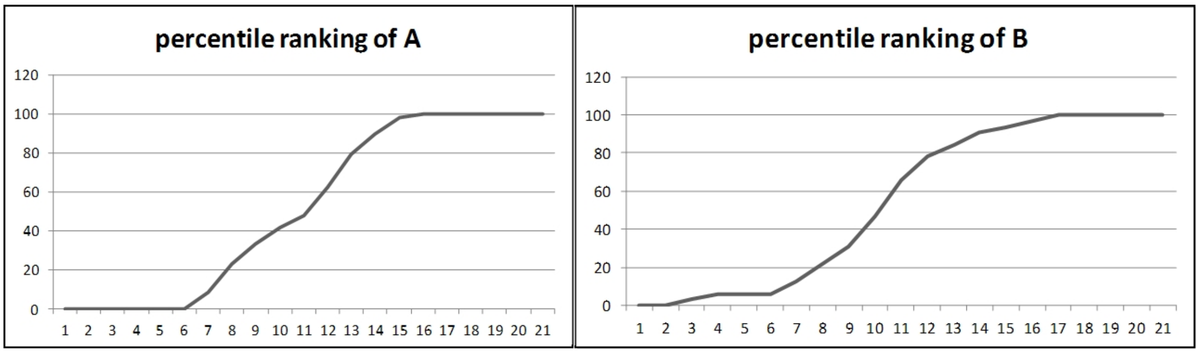

Figure 7. Graphical representation of percentile ranking of groups A and B

Table 5. Descriptive statistics of didactic characteristics of tests

Variable	Range of questions	Average	Average interval -95 %	Average interval +95 %	Min.	Max.	Standard deviation
Q(B)	20.000	54.688	44.472	64.903	12.500	93.750	21.828
P(B)	20.000	45.313	35.097	55.528	6.250	87.500	21.828
n_H(B)	20.000	3.200	2.186	4.214	0.000	8.000	2.167
n_L(B)	20.000	4.050	3.198	4.902	1.000	7.000	1.820
ULI(B)	20.000	0.106	-0.008	0.221	-0.250	0.500	0.244
p(B)	20.000	0.453	0.351	0.555	0.063	0.875	0.218
q(B)	20.000	0.547	0.445	0.649	0.125	0.938	0.218
$p*q$(B)	20.000	0.203	0.175	0.230	0.059	0.250	0.058
Q(A)	20.000	51.667	41.263	62.070	12.500	83.333	22.230
P(A)	20.000	48.333	37.930	58.737	16.667	87.500	22.230
n_H(A)	20.000	4.450	2.854	6.046	0.000	11.000	3.410
n_L(A)	20.000	7.150	5.807	8.493	3.000	12.000	2.870
ULI(A)	20.000	0.225	0.094	0.356	-0.250	0.667	0.280
p(A)	20.000	0.483	0.379	0.587	0.167	0.875	0.222
q(A)	20.000	0.517	0.413	0.621	0.125	0.833	0.222
$p*q$(A)	20.000	0.203	0.186	0.219	0.109	0.250	0.036

Table 6. T-test and F-test for selected didactic characteristics

T-test	Average 1	Average 2	T value	p value (T-test)	Standard deviation 1	Standard deviation 2	F value	p value (F-test)
Q(B) vs. Q(A)	54.688	51.667	0.434	0.667	21.828	22.230	1.037	0.937
P(B) vs. P(A)	45.313	48.333	-0.434	0.667	21.828	22.230	1.037	0.937
n_H(B) vs. n_H(A)	3.200	4.450	-1.384	0.175	2.167	3.410	2.477	0.055
n_L(B) vs. n_L(A)	4.050	7.150	-4.079	0.000	1.820	2.870	2.487	0.054
ULI(B) vs. ULI(A)	0.106	0.225	-1.430	0.161	0.244	0.280	1.312	0.560
p(B) vs. p(A)	0.453	0.483	-0.434	0.667	0.218	0.222	1.037	0.937
q(B) vs. q(A)	0.547	0.517	0.434	0.667	0.218	0.222	1.037	0.937
$p*q$(B) vs. $p*q$(A)	0.203	0.203	-0.016	0.988	0.058	0.036	2.664	0.039

4. Discussion

Measuring the difficulty of test tasks and analyzing the tested person is offered by variety of software or this possibility is implemented in e-learning systems. Prextová (2012) shows the possibilities of chosen software for evaluation of test tasks and person. Software could include not just reliability of test as we can see in our contribution but also reliability of each test item. The software have already provided visualization of distribution of individual items (as we have counted in Table 3 and Table 4) by bubble chart or item characteristics curves that rank the items according the difficulty. This visualization of extreme data items is very clear and useful for teacher decision about reformulation of each task.

Martin, Montgomery, and Saphian (2006) use percentile rank and test scores achieved on high school level to predict academic performance across four years of course-work. Even if the authors chose different methods (regression analysis) based on personal characteristics and properties, they could predict the results with satisfactory accuracy. As we were able to describe the source of differences in students' performance among domains of the educational texts (see Figure 5), they identified the same among the study years and the key attributes of the students.

Razak, Khairani, and Thien (2012) examine the quality of mathematics test items using Rasch Masurement Model. Although the authors use the didactic test for verification of knowledge in mathematics, we used a similar design creating tests, based on different types of test task that could have a form of knowledge or information, but in the biogas issue. In the study of Razak, Khairani, and Thien (2012) items were sorted according the difficulty level (low, intermediate and high). And items were divided into six categories (knowledge, comprehension, application, analysis, synthesis, evaluation). These analyses enable observe the ability of each student answering each question. In addition, the difficulty level of each question designed can be evaluated and continuously improved for future examinations. Each tested student has a general ability in the measured dimension as well as strengths and weaknesses. Rash Measurement Model is also connected with similar scaling as you can see in Figure 7.

Rash Measurement Model can be used also to measure validity and reliability of the test items that we can see in paper of Ariffin, Omar, Isa, and Sharif (2010). On the other hand the validity and reliability of test can be also measure by other metrics as shown in Miller & Achtenberg (2000). They have used Kuder-Richardson formula, the same formula as we used in our paper. These authors also worked with similar design of the pedagogical experiment, because they have worked with some pre-tests and post-tests and also they have two groups (experimental and control) as you can see in our paper too (Table 3 and Table 4).

Application of the analysis of properties of test assignments and didactic tests is commonly used as an unbiased tool for systematic survey of results of teaching as one can observe in works of many authors (Uljens, 1997; Draves & Coates, 2011); on the other hand, using statistics only does not always have to be relevant (Akyüz, 2013). The paper shows how this type of analysis can be used in another area as well; it is the area of pre-research of a pedagogic experiment, being of the nature of a test procedure. Using this analysis, the tools of actual research can be improved.

From the point of view of new modern methods of testing and evaluating, as stated by Dvorak (2012), robust methods of creating and evaluating the tests can also be used for obtaining more information from the tests, e.g.

Item Response Theory (IRT), Measurement Decision Theory (MDT) or Knowledge Spaces. The latter methods mentioned are being found suitable for the 21st-century testing where the so-called knowledge spaces can be used for modelling knowledge and for constructing tests which can provide extensive information about the characteristics of the test participant. Moreover, knowledge spaces can provide a detailed map of the learning progress (development continuum, progression map) so they can be used not only for testing in pedagogic practice but quite for measuring the net knowledge transfer which is made clean of the influence of original, previous knowledge within the framework of the experiment.

5. Conclusion

Some pieces of knowledge that can help improve the actual research have resulted from the pre-research:

1) Time needed to study the texts is sufficient because the average time is 806.83 seconds (13.43 minutes); test participants have 30 minutes available during the test so it is possible to reduce this time to 25 minutes, and still allow for a sufficient margin. The remaining times for the pre-test (15 minutes) and the post-test (15 minutes) are also sufficient, as confirmed by the participants in discussion or in the closing questionnaire.

2) To focus on the texts from the area of legislature because the knowledge transfer in these areas was lower than in the areas of economy or chemistry by a statistically significant amount. At the same time, the test participants designated this area as the most difficult one. For future, it is necessary to check the topic of legislature for texts from the point of view of their readability and semantic as well as syntactic difficulty, too, and to compare these indicators statistically to corresponding indicators from other topics (chemistry, economy, practice) as well.

3) To focus on the test questions where the *ULI* sensitivity coefficients were negative or close to zero and to rephrase them or to change possible answers to choose from.

4) After rephrasing the answers, to re-check for both tests whether the n_L parameter does not differ for both groups by a statistically significant amount.

5) To re-calculate the reliability coefficients and to focus on getting the coefficient of group B closer to the threshold of 0.8.

6) Statistical analysis of data from a non-representative sample showed that the knowledge transfer is not influenced by the type of the text, i.e. the average knowledge transfer using classical educational text is equal to the one of knowledge text (more general conclusions cannot be inferred, as they are a subject of separate research and work with a representative sample); however, it is desirable for future to focus also on other benefits of knowledge texts in the process of education that are measurable in an exact way.

There were seven working hypotheses formulated at the beginning; for six of them, the validity was proven (H1, H2a, H2b, H2c, H3, H7). For H4, statistically non-significant influence of the text on the net knowledge transfer was proven; for H5, two sub-areas of topics (legislature and chemistry) differ from each other in the average net knowledge transfer by a statistically significant amount; for H6, the equality was not shown for the n_L characteristics.

Other data was also acquired from the questionnaires as well as from the actual tests; however, further surveys and possible analyses (influence of texts on answers of the information or knowledge type, comparison of texts from the point of view of their understandability, naturalness and suitability for education and other) exceed the scope of this paper.

Acknowledgments

The paper is supported by the grant project of the Internal Grant Agency of the FEM CULS Prague "Determining the quantitative characteristics of knowledge texts", No. 20131020, and by the grant project of the Internal Grant Agency of the CULS Prague "Measuring the efficiency of knowledge transfer in the agricultural waste processing sector", No. 20131001.

References

Akyüz, G. (2013). Reflections from research methods education program: Effect on pre-service teachers' attitudes and anxieties. *New Horizons in Education*, *61*(2), 1-12.

Antony, J., Sivanathan, L., & Gijo, E. V. (2014). Design of Experiments in a higher education setting. *International Journal of Productivity and Performance Management*, *63*(4), 513-521. http://dx.doi.org/10.1108/IJPPM-07-2013-0130

Ariffin, S. R., Omar, B., Isa, A., & Sharif, S. (2010). Validity and Reliability Multiple Intelligent Item using Rasch Measurement Model. *Procedia–Social and Behavioral Sciences*, *9*(0), 729-733.

Asaishi, T. (2011). An analysis of the terminological structure of index terms in textbooks. *Procedia–Social and Behavioral Sciences*, *27*, 209-217. http://dx.doi.org/10.1016/j.sbspro.2011.10.600

Bassett, E. E., Bremner, J. M., & Morgan, B. J. T. (2000). *Statistics: Problems and solution*. Singapore: World Scientific Publishing Co. Pte. Ltd.

Cobern, W. W., Schuster, D., Adams, B., Skjold, B. A., Mugaloglu, E. Z., Bentz, A., & Sparks, K. (2014). Pedagogy of Science Teaching Tests: Formative assessments of science teaching orientations. *International Journal of Science Education*, *36*(13), 2265-2288. http://dx.doi.org/10.1080/09500693.2014.918672

Cohen, B. (1998). *Developing Sociological Knowledge*. Hampshire, UK: Cengage Learning.

Dömeová, L., Houška, M., & Beránková, M. (2008). *Systems Approach to Knowledge Modelling*. Hradec Králové: Graphical Studio Olga Čermáková.

Draves, W. A., & Coates, J. (2011). *The Pedagogy of the 21st Century*. River Falls, WI, US: LERN Books.

Duric, A., & Song, F. (2012). Feature Selection for Sentiment Analysis Based on Content and Syntax Models. *Decision Support Systems*, 53(4), 704-711. http://dx.doi.org/1016/j.dss.2012.05.023

Dvorak, J. (2012). On practical issues of measurement decision theory: An experimental study. *CSEDU 2012–Proceedings of the 4th International Conference on Computer Supported Education*, *2*, 94-99.

EU Commission. (2014). *European Qualification Framework*. Retrieved from http://ec.europa.eu/eqf/home_en.htm

Garner, M., Wagner, C., & Kawulich, B. (2009). *Teaching Research Methods in the Social Sciences*. Farnham, UK: Ashgate.

Glava, A., & Glava, C. (2011). The Model and the Didactic Modelling: An Analytic Point of View, *Procedia–Social & Behavioral Sciences*, *15*, 2228-2231. http://dx.doi.org/10.1016/j.sbspro.2011.04.084

Gravetter, F. J., & Wallnau, L. B. (2009). *Statistics for the Behavioral Sciences*. Wadsworth: Cengage Learning.

Houška, M., & Beránková, M. (2007). Individual Learning Based on Elementary Knowledge Concept: Experiments and Results. *Proceedings of International Conference–Interactive Computer Aires Blended Learning*, 48-53

Houška, M., & Rauchová, T. (2013). Methodology of creating the knowledge text. *Proceedings of the 10th International Conference on Efficiency and Responsibility in Education (ERIE 2010)*, 197-203.

Igbaria, A. K. (2013). A content analysis of the WH-questions in the EFL textbook of horizons. *International Education Studies*, *6*(7), 200-224. http://dx.doi.org/10.5539/ies.v6n7p200

Jiang, F. (2011). Explore ways of using computer technology to construct test database of "pedagogy" course. *Proceedings: 2011 IEEE International Symposium on IT in Medicine and Education*, 289-291. http://dx.doi.org/10.1109/ITiME.2011.6130835

Leutner, D., Leopold, C., & Sumfleth, E. (2009). Cognitive load and science text comprehension: Effects of drawing and mentally imagining text content. *Computers in Human Behavior*, *25*(2), 284-289. http://dx.doi.org/10.1016/j.chb.2008.12.010

Lindsey, J. K. (2009). *Introduction to Applied Statistics: A modelling approach*. New York: Oxford University Press.

Martin, J. H., Montgomery, R. L., & Saphian, D. (2006). Personality, Achievement Test Scores, and High School Percentile as Predictors of Academic Performance across Four Years of Coursework. *Journal of Research in Personality*, *40*(4), 424-431.

Menon, S., & Mukundan, J. (2012). Collocations of high frequency noun keywords in prescribed science textbooks. *International Education Studies*, *5*(6), 149-160. http://dx.doi.org/10.5539/ies.v5n6p149

Miller, C. K., & Achterberg, C. L. (2000). Reliability and Validity of a Nutrition and Food-Label Knowledge Test for Women with Type 2 Diabetes Mellitus. *Journal of Nutrition Education*, *32*(1), 43-48.

Naqvi, S. I. H., Hashmi, M. A., & Hussain, A. (2010). Validation of objective-type test in biology at secondary school level. *Procedia–Social and Behavioral Sciences*, *2*(2), 3909-3913. http://dx.doi.org/10.1016/j.sbspro.2010.03.615

Newton, D. P., & Newton, L. D. (2009). A procedure for assessing textbook support for reasoned thinking. *Asia-Pacific Education Researcher*, *18*(1), 109-115.

Ozuru, Y., Dempsey, K., & McNamara, D. S. (2009). Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. *Learning and Instruction*, *19*(3), 228-242. http://dx.doi.org/10.1016/j.learninstruc.2008.04.003

Prextová, T. (2012). Use of the Program Winsteps for Analyzing Test Tasks and Test Persons. *Procedia–Social and Behavioral Sciences*, *47*(0), 1077-1082.

Rahimy, R., & Shams, K. (2012). An investigation of the effectiveness of vocabulary learning strategies on Iranian EFL learners' vocabulary test score. *International Education Studies*, *5*(5), 141-152. http://dx.doi.org/10.5539/ies.v5n5p141

Razak, A. N., Khairani, A. Z., & Thien, L. M. (2012). Examining Quality of Mathemtics Test Items using Rasch Model: Preliminarily Analysis. *Procedia-Social and Behavioral Sciences*, *69*(0), 2205-2214.

Rosales, J., Vicente, S., Chamoso, J. M., Muñez, D., & Orrantia, J. (2012). Teacher–student interaction in joint word problem solving. The role of situational and mathematical knowledge in mainstream classrooms. *Teaching and Teacher Education*, *28*(8), 1185-1195. http://dx.doi.org/10.1016/j.tate.2012.07.007

Rottensteiner, S. (2010). Structure, Function and Readability of New Textbooks in Relation to Comprehension. *Procedia–Social and Behavioral Sciences*, *2*(2), 3892-3898. http://dx.doi.org/10.1016/j.sbspro.2010.03.611

Rynne, A., & Gaughran, W. (2008). Cognitive modeling strategies for optimum design intent in parametric modeling (PM). *Computers in Education Journal*, *18*(3), 55-68.

Singer, F. M., & Moscovici, H. (2008). Teaching and Learning Cycles in a Constructivist Approach to Instruction. *Teaching and Teacher Education*, *24*(6), 1613-1634. http://dx.doi.org/10.1016/j.tate.2007.12.002

Tarchi, C. (2010). Reading comprehension of informative texts in secondary school: A focus on direct and indirect effects of reader's prior knowledge. *Learning and Individual Differences*, *20*(5), 415-420. http://dx.doi.org/10.1016/j.lindif.2010.04.002

Tudor, S. L. (2012). A study on the efficiency of using combined modern and traditional didactic strategies. *Procedia–Social and Behavioral Sciences*, *33*, 989-993. http://dx.doi.org/10.1016/j.sbspro.2012.01.270

Uljens, M. (1997). *School Didactics and Learning*. Oxford: Psychology Press.

Zainal, A. (2012). Validation of an ESL writing test in a Malaysian secondary school context. *Assessing Writing*, *17*(1), 1-17. http://dx.doi.org/10.1016/j.asw.2011.08.002

Copyrights