Experiment Replication and Meta-Analysis in Evaluation of Intelligent Tutoring System's Effectiveness

ANI GRUBIŠIĆ, SLAVOMIR STANKOV, BRANKO ŽITKO
Faculty of Natural Sciences, Mathematics and Kinesiology
Nikole Tesle 12, 21000 Split
CROATIA
ani.grubisic {slavomir.stankov, branko.zitko}@pmfst.hr

Abstract: - This paper presents the methodology for conducting controlled experiment replication, as well as, the results of a controlled experiment and an internal replication that investigated the effectiveness of an intelligent tutoring system. Since, there doesn't seem to be a common ground on guidelines for the replication of experiments in intelligent tutoring system's educational influence evaluation, this scientific method has just started to be applied to this propulsive research field. We believe that every effectiveness evaluation should be replicated at least in order to verify the original results and to indicate an evaluated e-learning system's advantages or disadvantages. On the grounds of experiment replication, a meta-analysis can be conducted in order to calculate overall intelligent tutoring system effectiveness.

Key-Words: - e-learning, intelligent tutoring systems, evaluation, effect size, effectiveness, experiment, replication, meta-analysis

1 Introduction

It is very important to evaluate all instructional software (an e-learning system) before using it in educational process. An evaluation offers information to make decision about using the product or not [24]. So, a well-designed evaluation should provide the evidence, if a specific approach has been successful and of potential value to the others [7]. One special form of evaluation is *effectiveness evaluation* designed to answer one specific research question: "What is the educational influence of an e-learning system on students?". As effectiveness evaluation concerns the whole system, it is suitable for external evaluation, and as it bases itself on experiment, it is part of an experimental research [14]. Embracing the e-learning is common among institutions, regardless of the fact that a difficult task of ensuring the quality and effectiveness still remains [2].

Experiments used in the effectiveness evaluation change the independent variable (e-learning system) while measuring the dependent variable (student's achievement) and require statistically significant groups – a control and an experimental group. The *experiment validity*, that is, validity of the results, can be ensured by a replication of the same experiment. The *replication* is the repetition of an experiment following, as closely as possible, the original experiment [1]. The main result gained through effectiveness evaluation is presented in a form of an *effect size*. The effect size measurements tell us the size of experimental effect. It is a standard way to compare the results of two experiments. Effect size is positive when the experimental group in the study outperforms the control group, and is negative when the control group is better.

The most important effect size in literature comes from one of the most stimulating studies in the field of educational psychology. That is a Bloom's [4] statement of the *2-sigma problem*. Bloom reviewed the results of published meta-analyses and conducted one experiment, which showed that students who received individualized tutoring displayed an effect size of 2.0 compared with those who received normal instruction. In other words, students who received individualized tutoring scored an average of two standard deviations above others on achievement tests. Consequently, this gain of two standard deviations became the ideal toward which normal, group oriented instruction should strive. So, the 2-sigma problem became the quest for ways that the quality of group instruction can get closer to individualized tutoring. Many researchers have taken up this search.

One possible solution for the 2-sigma problem is the usage of *intelligent tutoring systems* (ITS), which provide each student with a learning experience similar to the ideal one-to-one tutoring. Since, the latest statements regarding ITS's effectiveness are those mentioned by Fletcher in 2003 [9], it is very disputable whether or not we should rely on those results. Namely, it is not known which eleven studies did Fletcher use while calculating the effect size 0.84 of ITSs, and the effect size 1.05 for recent intelligent tutoring systems is

arguable because it has been calculated using results from only one study.

A *meta-analysis* [10] integrates the results of a set of experiments. It is usually conducted to increase the internal and the external validity of the conclusions that can be drawn from those experimental studies, giving us far more definitive statistical conclusions. It is a method that helps scientists to recognize order in something that appeared to be a disorder. One great statistician Olkin ones said (according to [13]) that meta-analysis is like a trip in a helicopter because when you are on the ground you can perfectly see individual trees, but as the helicopter rises, you begin to se patterns not visible from the ground.

There is an urgent need for conduction of a meta-analysis that would reveal an effect size of intelligent tutoring systems in general. Therefore, the researchers should "think meta-analytically" (Cumming and Finch, 2001 according to [5]), that is, be familiar with meta-analysis process and should report their results in a way that is appropriate for meta-analyzing.

A meta-analysis conducted using results gained through valid experiments based on the same methodology is a key issue in resolving a problem of making conclusions about overall intelligent tutoring system effectiveness. Therefore, a replication of experiments related to intelligent tutoring systems' effectiveness calculation, is a first step in meta-analysis.

This paper presents the results of a controlled experiment and an internal replication that investigated the effectiveness of one intelligent tutoring system. In the second chapter we review the age-long research and development of the Tutor–Expert System (TEx-Sys) model for building ITS ([29], [31]). In the third chapter we discuss some issues related to experiment replication. In the fourth chapter we give a brief overview of meta-analysis principles. Finally, in the last chapter we describe the replication of the experiment where we evaluated educational influence of the xTEx-Sys's (eXtended Tutor-Expert System) [31], which is the representative of Web-based authoring shells for building ITS based on the TEx-Sys model.

2 Background

The intelligent tutoring systems (ITS) are computer systems that support and improve learning and teaching process in certain domain knowledge, respecting the individuality of learner as in traditional "one-to-one" tutoring ([32], [20], [28]). The major problems when developing ITSs are their expensive and time consuming development process. In order to overcome those problems another approach has been chosen, namely to create particular ITSs from flexible shells acting as program generators [19].

The first implementation of an intelligent authoring shell model called the TEx-Sys [31] used in this research is the on-site TEx-Sys (1992-2001), after that followed the Web-based intelligent authoring shell (1999-2003, Distributed Tutor-Expert System, DTEx-Sys) [26] and, finally, the system based on Web services (2003-2005, xTEx-Sys).

The xTEx-Sys is a Web-based authoring shell with an environment that can be used by the following actors: an expert who designs the domain knowledge base, a teacher who designs courseware and tests for the student knowledge evaluation, a student who selects course and navigates trough the domain knowledge content using didactically prepared courseware and, finally, an administrator who supervises the system.

In the past decade, there were numerous applications of the TEx-Sys model in learning and teaching process that involved students from primary education all the way to academic level. In the period from 2005 to 2005, there were effectiveness experiments related to one of the systems based on the Tex-Sys model (Table 1) [31].

3 A replication of an experiment

The replication, in the context of this paper, is the repetition of an experiment as closely following the original experiment as possible. The replication of controlled experiments is considered to be a critical aspect of the scientific method [16]. Pfleeger underlines that the replication means repeating an experiment under equal circumstances and not repeating measurements on the same experimental unit, which refer to literally taking several measurements of a single occurrence of a phenomenon [23].

At least one replication is needed if someone wants their results to be of any interest at all. Any result from an isolated study cannot show whether the conclusions will hold again. The first replication shows whether or not a generalization is possible [18]. According to [18], there are two types of replication: close and differentiated replication. The close replication attempts to keep almost all the known conditions of the study much the same or at least very similar as they were in the original experiment. The differentiated replication involves deliberate variations in major aspects of the study.

To conclude, there doesn't seem to be a common ground on guidelines for the replication of experiments in e-learning system's educational influence evaluation, as there are only a few replicated experiments related to the e-leaning systems' effectiveness evaluation (for example [25]). Therefore, replication has just started to be applied to this propulsive research field. We believe that every effectiveness evaluation should be replicated

Table 1. Results from effectiveness evaluation experiments

Course	Sample size after drop-off	Duration	Statistically significant difference	Original score means and standard deviations	Gain score means and standard deviations	Effect size
Academic year 2005/2006						
Introduction to computer science	Experimental group: 40 1st year students Control group: 40 1st year students	14 weeks	α =0.05, df=78			0,16
			chk test 1: t = -0,73, p=0,4676 NO	ctrl: $\overline{X}$ = 40,72, sd = 15,78 exp: $\overline{X}$ = 46,13, sd = 16,80	ctrl: $\overline{X}$ = -9,28, sd = 17,74 exp: $\overline{X}$ =-6,19 , sd =18,97	(0,17)
			chk test 2: t = 2,31, p=0,0235 YES	ctrl: $\overline{X}$ = 54,95, sd = 17,36 exp: $\overline{X}$ = 46,95, sd = 12,80	ctrl: $\overline{X}$ = 4,95, sd = 21,68 exp: $\overline{X}$ =-5,36, sd =17,86	(-0,47)
			final test: t = -3,62, p=0,0005 YES	ctrl: $\overline{X}$ = 37,48, sd = 13,44 exp: $\overline{X}$ = 51,23, sd = 12,30	ctrl: $\overline{X}$ = -12,53, sd =14,32 exp: $\overline{X}$ =-1,09, sd =13,66	(0,81)
Chemistry	Experimental group: 20 8th grade primary school pupils Control group: 21 8th grade primary school pupils	10 weeks	α =0.05, df= 39 t= -1,81, p= 0,0780 YES	ctrl: $\overline{X}$ = 24,42, sd = 8,16 exp: $\overline{X}$ = 25,70, sd = 5,28	ctrl: $\overline{X}$ = -9,83 , sd = 7,95 exp: $\overline{X}$ = -8,63 , sd = 6,71	0,60
Physics – optics	Experimental group: 40 8th grade primary school pupils Control group: 40 8th grade primary school pupils	7 weeks	α =0.05, df= 78 t=-3,67, p=0,0004 YES	ctrl: $\overline{X}$ = 44,03, sd = 22,89 exp: $\overline{X}$ = 62,28, sd = 20,91	ctrl: $\overline{X}$ = 0,16, sd = 0,24 exp: $\overline{X}$ = 0,34, sd = 0,20	0,75
Nature and society	Experimental group: 24 2nd grade primary school pupils Control group: 24 2nd grade primary school pupils	6 weeks	α =0.05, df= 46 t= -2,88, p= 0,0060 YES	ctrl: $\overline{X}$ = 79,83, sd = 10,58 exp: $\overline{X}$ = 91,00, sd = 13,10	ctrl: $\overline{X}$ = 8,17, sd = 8,30 exp: $\overline{X}$ = 14,83, sd = 7,69	0,80
	Experimental group: 24 3rd grade primary school pupils Control group: 24 3rd grade primary school pupils	6 weeks	α =0.05, df= 46 t= -3,08, p= 0,0035 YES	ctrl: $\overline{X}$ = 86,17, sd = 7,64 exp: $\overline{X}$ = 95,50, sd = 5,32	ctrl: $\overline{X}$ = 8,67 , sd = 7,24 exp: $\overline{X}$ = 15,17 , sd = 7,36	0,83
	Experimental group: 20 4th grade primary school pupils Control group: 20 4th grade primary school pupils	6 weeks	α =0.05, df= 38 t= -3,48, p= 0,0013 YES	ctrl: $\overline{X}$ = 84,00, sd =10,24 exp: $\overline{X}$ = 92,83, sd = 7,82	ctrl: $\overline{X}$ = 10,83, sd = 6,57 exp: $\overline{X}$ = 18,17 , sd = 7,98	1,11
Academic year 2006/2007						
Introduction	Experimental group: 20 1st year students	14	α =0.05, df=37			0,42
			chk test 1: t = 1,04, p=0,3051 NO	ctrl: $\overline{X}$ = 54,74, sd = 19,62 exp: $\overline{X}$ = 50,30, sd = 18,62	ctrl: $\overline{X}$ = 13,74, sd =19,62 exp: $\overline{X}$ = 7,35 , sd =18,62	(0,33)
				ctrl: $\overline{X}$ = 31,89 sd =	ctrl: $\overline{X}$ = -9,11, sd	

at least in order to verify the original results and to indicate an evaluated e-learning system's advantages or disadvantages.

3.1. Replication errors

In conducting an experiment there could happen one or more of three general types of errors: human error, systematic error, and random error [8].

Human error (a mistake) occurs when the experimenter makes a mistake. For example, setting up experiment incorrectly, misreading an instrument, or making a mistake in a calculation.

Systematic error in a measurement is a consistent and repeatable prejudice or offset from the true value. This is typically the result of miscalibration of the test equipment, or problems with the experimental procedure. Systematic error is an error which causes the results to be skewed in the same direction every time, i.e., always too large or always too small. Most of the simple experiments have some systematic error.

On the other hand, variations between successive measurements made under apparently identical experimental conditions are called *random errors*. Random variations can occur in the quantity being measured or the measurement process. All experiments have random error, which occurs because no measurement can be made with infinite precision. An example of random error could be when trying to draw 100 lines on a sheet of paper, each exactly one centimetre long. Each line will be close to a centimetre, but will be longer or shorter depending on a many microscopic muscle movements. Random error can be reduced by averaging several measurements.

3.2. Validity

To study the validity of given results can be observed through three different aspects: internal validity, construct validity and external validity.

The *external validity* is the degree to which the results of the research can be generalized. Each new replication of an experiment reduces the probability that results can be explained by human variation or experimental error [1]). Replication can contribute significantly to generalizing results if replicated experiments employ probability-sampling techniques [17].

The *construct validity* is the extent to which a test may be said to measure what it has been designed to measure. A well-performed replication must also evaluate the methods used to capture data in the original experiment [6].

The *internal validity* is the degree to which conclusions can be drawn about the causal effect of the independent variable on the dependent variables. Potential threats include selection effects, non-random subject loss, instrumentation effect, and maturation effect [22].

To conclude, there doesn't seem to be a common ground on guidelines for the replication of experiments in e-learning system's educational influence evaluation, as there are only a few replicated experiments related to the e-leaning systems' effectiveness evaluation (for example, [25]). Therefore, this scientific method – replication – has just started to be applied to this propulsive research field. We believe that every effectiveness evaluation should be replicated at least in order to verify the original results and to indicate an evaluated e-learning system's advantages or disadvantages.

3.3. An example of our replication of an eminent research

In a widely quoted research Bloom [4] had compared student learning under three different forms of instruction: conventional learning, mastery learning and tutoring.

In our variation of the Bloom's experiment replication [30], 33 students, taking "Introduction to computer science" class, that were randomly and equally divided into a control group (11 students), a tutoring group (11 students) and an experimental group (11 students), participated in the first experiment in academic year 2004/05. The control group was involved in the traditional learning and teaching process, the experimental group was asked to use the DTEx-Sys and the tutoring group was tutored by human tutors (four subgroups of 2-3 students tutored by human tutors). All three different types of treatment were scheduled for two hours weekly throughout one semester.

All three groups underwent a paper-and-pen pre-test that was distributed at the beginning of the course, which enabled us to determine that there was no statistically significant difference between any two groups concerning their foreknowledge. The post-test that was applied two weeks after the end of the course, enabled us to determine that there was a statistically significant difference between the control group and the experimental group ($t=2.41$, $p=0.04$), which had showed the DTEx-Sys's advantage over the traditional learning and teaching. A statistically insignificant difference between the experimental group and the tutoring group concerning the post-test results ($t=0.53$, $p=0.61$) has shown the DTEx-Sys's competency in substituting human tutors. The calculated effect size of the DTEx-Sys was 0.82. Therefore the evaluation of the system indicated that the teaching strategy implemented by the DTEx-Sys is effective in accomplishing the task it was designed to perform.

4 A meta-analysis

In the past, scientists could not easily make a conclusion about the effectiveness of an experimental factor, because some of the studies that were taken into consideration gave positive results, some gave negative results, and some gave neutral results about effectiveness. Even if they could make that conclusion,

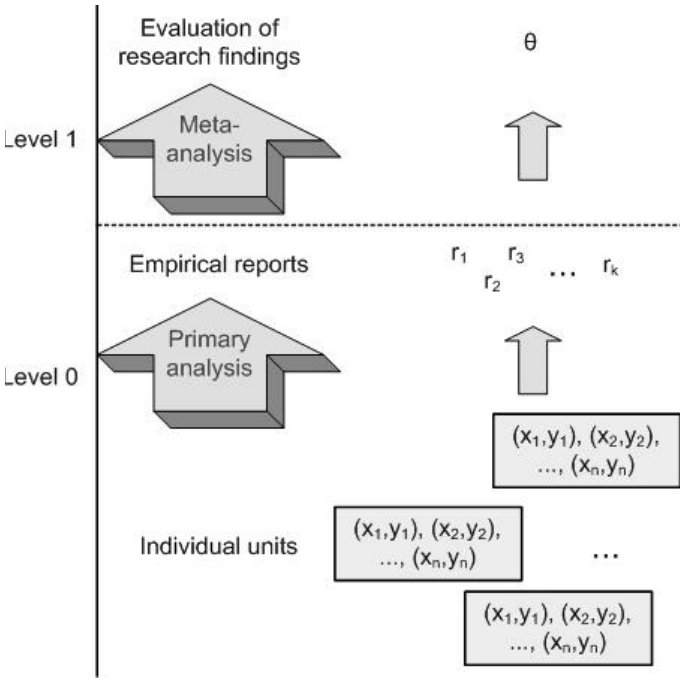

Figure 1. Levels of analysis (modified according to [27])

they were not always sure what the magnitude of the experimental factor effectiveness was. Therefore, a different approach had to be defined in order to resolve this issue.

A solution came in a form of meta-analysis. Primary analyses analyze original data from an individual study. Secondary analyses reanalyze existing data in order to answer new research question (Figure 1). A meta analysis is a relatively new methodology, developed by Glass [10] and further refined by Hedges and Olkin [12]. A meta-analysis refers to a method that goes beyond an initial set of analyses or studies (in Greek meta means "beyond" according to [21]). It is, according to Lipsey and Wilson [15], a method for reviewing and synthesizing experimental studies. It can be applied only to empirical studies that give quantitative results.

4.1. Effect size

As a result of a primary analysis, the typical outcomes are measures of the strength of the relationships that exist between the observed variables. Therefore, the basic unit of data for meta-analysis is the *effect size* (*ES)*. By examining the effect sizes from several experiments, it has to be determined whether, with all of these effect sizes combined, the impact of this experimental factor (treatment) causes the difference that is large enough to recommend changes in educational practices [15].

The lack of reporting effect sizes poses a problem for the meta-analysis, especially when the data reported in individual studies is not sufficient for the effect size calculations. Furthermore, experimental design characteristics also have to be taken into account when extracting an effect size in order to avoid wrong results [27]. Effect sizes can be reported in various forms. They can be classified according to the number of the observed variables and their relationships [15]: (1) one-variable relationships (proportion – p, arithmetic mean – m); (2) two-variable relationship (pre-post contrasts (unstandardized mean gain – ug, standardized mean gain – sg), group contrasts (unstandardized mean difference – um, standardized mean difference – sm, proportion difference – pd, odds-ratio – or), association between variables (two dichotomous variables (odds-ratio, phi coefficient), dichotomous and continuous variable (point-biserial coefficient, standardized mean difference), two continuous variables (product-moment correlation); (3) multivariate relationships.

Here we will briefly describe the most frequently used types of effect sizes [27]:

4.1.1 Correlation coefficient as effect size

The Pearson correlation coefficient *r* is based on n pairs of observations (x_i, y_i), i=1,..,n:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

There are several other correlation coefficients that can be used like point-biserial, biserial, rank coefficient.

4.1.2 Standardized mean difference as effect size

It is mostly used in situations where two groups (experimental and control) are examined. Both random variables are assumed to be normally distributed with common standard deviation σ but not necessarily with the same number of observations *n*. There are three popular coefficients that are based on standardized mean difference, but use different standard deviations.
The first effect size is known as Cohen's *d*:

$$d = \frac{\bar{X} - \bar{Y}}{S_{pooled}} \quad (2)$$

where S_{pooled} is standard deviation of the population.
The second is Glass's Δ:

$$d = \frac{\bar{X} - \bar{Y}}{S_{control}} \quad (3)$$

where $S_{control}$ is standard deviation of the control group.
The third is Hedges's *g*:

$$d = \frac{\bar{X} - \bar{Y}}{S_{common}} \quad (4)$$

where S_{common} is standard deviation of the sample.

4.1.3 Conversion of effect sizes

Effect sizes have to be converted because all effect sizes used in a meta-analysis have to be from the same family

[27]. If the group sizes are equal, *r* and *d* can be converted using the following formulas ([12]; [15]; Cohen, 1988; Rosenthal, 1991, according to [27]):

$$r = \sqrt{\frac{d^2}{d^2 + 4}} \tag{5}$$

$$d = \frac{2r}{\sqrt{(1 - r^2)}} \tag{6}$$

If group sizes slightly differ, we use the following formula (Aaron, Kromrey and Ferron, 1998, according to [27]):

$$r = \sqrt{\frac{d^2}{d^2 + 4 - \frac{8}{n}}} \tag{7}$$

where $n=n_1+n_2$.

If the group sizes are unequal, then we use (Aaron, Kromrey and Ferron, 1998, according to [27]):

$$r = \sqrt{\frac{d^2}{d^2 + \frac{(n_1 + n_2)^2 - 2(n_1 + n_2)}{n_1 n_2}}} \tag{8}$$

4.2. Sampling error

Besides effect sizes, there is another very important information that has to be considered in every meta-analysis. Each experiment result calculation is based on a specific sample. The size of samples used in different studies varies and different effect sizes are based on a different sample sizes. Generally speaking, larger sample sizes present the population better and therefore produce more precise effect sizes. It means that larger sample sizes have smaller *sampling error* [15].

In calculation of *mean (total) effect size* in meta-analysis, we have to use effect sizes with different sampling errors. It would be wrong just to calculate their arithmetic mean, because in that way each effect size contributes to mean effect size equally, regardless of their sampling error.

This problem can be solved by *weighting effect sizes* to determine their precision based on sample size [15]. Hedges and Olkin have proved that the best weighting measure is based on *standard error of effect size (SE)* [12]*:*

$$SE = \sqrt{\frac{N_E + N_C}{N_E \times N_C} + \frac{ES^2}{2 \times (N_E + N_C)}} \tag{9}$$

where N_E and N_C are the sample sizes of the experimental and control groups.

Less precise effect size has higher standard error, and, therefore effect size is weighted using inverse standard error square, that is, inverse variance weight. Effect size weight *w* is calculated using the following formula [15]:

$$w = \frac{1}{SE^2} \tag{10}$$

4.3. Meta-analysis phases

In order to gain valid results from meta-analysis, all meta-analysis phases have to be validly conducted [15].

4.3.1 Creating an independent set of effect sizes

In meta-analysis an observed object is a single study. Many studies present multiple effect sizes that are dependent because they are based on the same sample. Their inclusion in meta-analysis would violate assumption about data independence, what is crucial for statistical analysis [15].

For that reason, some effect size adjustments, like transformations and bias corrections for effect sizes calculations, and outliers' detection, have to be done. *Outliers* are extreme effect sizes that are contradictory to the most other effect sizes, and can easily be false. They have disproportionate influence on arithmetic mean, variance, etc., and can distort them in misleading way. Lipsey and Wilson propose that outliers should be excluded from analysis, or recoded into more moderate ones ("windsorizing" procedure) [15].

4.3.2 Calculating the mean effect size

The mean effect size is gained using weighted effect sizes from studies. It is called *weighted mean effect size* and it is calculated using this formula [15]:

$$\overline{ES} = \frac{\sum (w_i ES_i)}{\sum w_i} \tag{11}$$

where ES_i are effect sizes, w_i their weights.

4.3.3 Determining confidence intervals around mean effect size

A *confidence interval* shows a range of possible values of the mean effect size. The confidence interval around mean effect size is based on a standard error of the mean and critical z-distribution value (for example, 1.96 for $\alpha=0.05$). A formula for standard error of the mean effect size is [15]:

$$SE_{\overline{ES}} = \sqrt{\frac{1}{\sum w_i}} \tag{12}$$

where w_i are effect size weights.

The lower and upper limits of the confidence interval are calculated using following formulas [15]:

$$\overline{ES}_L = \overline{ES} - z(SE_{\overline{ES}}),$$
$$\overline{ES}_U = \overline{ES} + z(SE_{\overline{ES}}), \tag{13}$$

where *z* is critical z-distribution value ($z=1.96$, $\alpha=0.05$ for 95% confidence interval, $z=2.58$, $\alpha=0.01$ for 99% confidence interval). If the confidence interval does not

include 0, the mean effect size is statistically significant for $p \leq \alpha$.

4.3.4 Homogeneity analysis

The homogeneity of the effect size distribution relates to the question whether the various effect sizes that are combined together into a mean effect size, all estimate the same population [15]. In homogeneous distribution each effect size differs from mean effect size only by sampling error.

A homogeneity test is based on *Q statistics*, which is distributed as a chi-square with k-1 degrees of freedom (k is the number of effect sizes) [12]. The formula for Q is:

$$Q = \sum w_i \left(ES_i - \overline{ES}\right)^2 \quad (14)$$

If Q exceeds the critical value for a chi-square with k-1 degrees of freedom, then effect sizes are homogeneous.

4.4. Meta-analysis pros and cons

These are some of the positive sides of meta-analyses seen by Lipsey and Wilson [15]. *First*, a good meta-analysis demands documenting each step and, therefore it is easy to check it's validity. It is possible because each meta-analysis includes: (i) defining the criteria that determines which type of results can be used, (ii) organized strategies for searching and identifying appropriate studies, (iii) collecting and analyzing data from the studies. *Second*, a meta-analysis presents results in a form of effect sizes, instead of qualitative summaries and statistically significance. *Third*, a systematic analysis in meta-analysis allows precise verification between results and participants' characteristics, treatment, research structure and measuring methods and instruments. *Fourth*, a meta-analysis using databases enables systematic processing of almost unlimited quantity of information gained from studies, but it is equally applicable on small number of studies.

Some of the meta-analysis shortcomings are the amount of effort and expertise it demands, and integration of different studies that could even be incomparable [15]. The latest problem, often related as problem of mixing apples and oranges, will always be present because the selection of comparable studies depends on the analyst. What is important is that the analyst explicitly elaborates the meta-analysis domain and criterions he or she uses to make distinction between comparable and incomparable studies.

5 The replicated experiments

To assess the effectiveness of the xTEx-Sys, we have conducted two experiments: the initial one in academic year 2005/06 [11] and its replication in 2006/07. Both

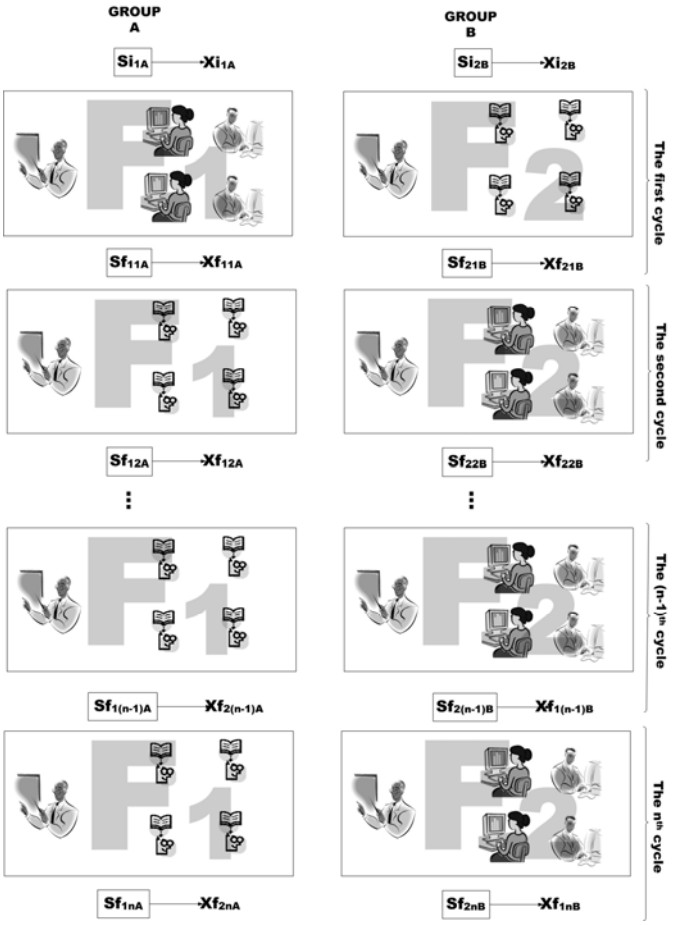

Figure 2. Pre-and-post test control group experimental design with checkpoint-tests

experiments are conducted according to design of pre-and-post test control group experimental design with checkpoint-tests (described in [11]) (Figure 2.).

5.1. Subjects

Students who participated in initial and replication experiment were undergraduate students from two Faculties from a University of Split in Croatia: the Faculty of Chemical Technology (FCT) and the Faculty of Natural Sciences, Mathematics and Kinesiology (FNSMK) that took a course called "Introduction to Computer Science".

The initial experiment started in October 2005 and lasted until the end of January 2006. At the very beginning of that experiment there were 175 students, but eventually only 120 of them completed all parts of the experiment (68%).

The replication of the initial experiment started in October 2006 and lasted until the end of January 2007. At the very beginning of that experiment there were 127 students, but only 70 of them completed all parts of the experiment (55%).

In both experiments context information about the participants was collected. Students were asked questions about personal characteristics (age, gender),

high school education, preferences and beliefs about learning styles. These questions could be answered on a voluntary basis.

Due to organizational and legality problems, we have decided, in prior, that the students from FCT would be control group students and students from FNSMK experimental group students. That prior division was later found to be proper, because the pre-test results for subgroups of defined groups in both experiments have shown that those subgroups were statistically equivalent in both experiments.

Therefore, of the 175 students that agreed to participate in the initial experiment, 86 students were assigned to a control group and 109 students to an experimental group. Of the 127 students who participated in the replication experiment, 52 students were assigned to a control group and 75 students to an experimental group.

5.2. Procedure

The initial experiment and its replication were conducted following the same plan. After a short introduction during which the purpose of the experiment and general organizational issues were explained, data on personal characteristics and background knowledge was collected by means of a questionnaire. Then the pre-test was conducted. Following the pre-test, a brief introduction into organizational issues related to the treatments was given.

During the experiments, there were three treatment-test cycles. The tests were used to measure the dependent variable – student knowledge. After completing first treatment, the both groups performed the first checkpoint test (CHK1), after second treatment they performed the second checkpoint test (CHK2), and, finally, at the end of the experiments they performed the post-test (END). All tests in both experiments were respectively identical. During the whole procedure, the time slots reserved for completing a certain step of the schedule were identical for the experimental and control groups.

To be able to analyze results, it was important to find out the size of the student drop-off from each group. At the end of initial experiment, of 86 control group students only 40 completed all parts of the experiment and of 109 experimental group students only 80 completed all parts of the experiment. At the end of replication experiment, of 52 control group students only 19 completed all parts of the experiment and of 75 experimental group students only 51 completed all parts of the experiment.

Therefore, we had to statistically equalize the control and the experimental groups in both experiments using the caliper matching [3]. In the initial experiment, there were, at the end, 40 control group students and 40 experimental group students. In the replicated experiment, there were, at the end, 19 control group students and 20 experimental group students.

5.3. Data analysis

Standard significance testing was used to investigate the effect of the treatments on the dependent variable. First, it has to be checked whether groups' initial competencies were equivalent before comparing the gains of the groups. That means calculating the means of pre-test score for both groups and their standard error of mean.

Now, a null-hypothesis H0 has to be stated for every checkpoint-test and post-test: "There is no significant difference between the control and the experimental group" ($H0_{CHK1}$, $H0_{CHK2}$, …, $H0_{END}$).

Next, the gain scores from the pre-test to every checkpoint-test and the post-test for both groups have to be calculated. The means of gains for every test and for both groups, as well as, their standard means of error have to be calculated. A prerequisite for applying the t-test is the assumption of normal distribution of the variables in the test samples. A test to check this assumption was conducted.

Then the t-values of means of gain scores have to be computed to determine if there is a reliable difference between the control and the experimental group for every testing point (the checkpoints and at the end of the course). If there is statistically significant difference at every testing point (same or slightly rising), it implies that the e-learning system has had a positive effect on the students' understanding of the domain knowledge. In other words, the null-hypothesis is rejected.

5.4. Results

Table 2 contains the descriptive statistics for the initial experiment and the replication. The columns "Pre-test", "CHK1", "CHK2" and "END" show the calculated values for mean, median, and standard deviation of the raw data collected during the pre-test, first checkpoint test, second checkpoint test and post-test, respectively, of the initial experiment (E) and the replication (R) for both experimental groups and control groups.

The columns of Table 2 that start with "Gain" show the calculated values for mean, median, and standard deviation of the differences between post-test, first checkpoint test, second checkpoint test and pre-test scores of the initial experiment (E) and replication (R).

The zero or negative difference between average first checkpoint test scores and average pre-test scores occurred twice during the initial experiment and not even once during the replication. The same phenomenon, relating second checkpoint test, occurred once during the initial experiment and twice during the replication, and relating post-test, it occurred twice during the initial experiment and once during the replication.

Table 2

	Pre-test	CHK1	CHK2	END	Gain CHK1 and Pre-test	Gain CHK2 and Pre-test	Gain END and Pre-test
E: initial experiment							
Control group (40 students)							
Mean	50,00	40,72	54,95	37,48	*-9,28*	4,95	*-12,53*
Median	51,49	42,50	58,00	37,00	*-7,87*	6,78	*-13,54*
Stdev.	18,01	15,78	17,36	13,44	*17,74*	21,68	*14,32*
Experimental group (40 students)							
Mean	52,31	46,13	46,95	51,23	*-6,19*	*-5,36*	*-1,09*
Median	52,98	49,38	45,50	51,50	*-8,59*	*-4,24*	*-2,01*
Stdev.	14,76	16,80	12,80	12,30	*18,97*	*17,86*	*13,66*
R: replication experiment							
Control group (19 students)							
Mean	41,00	54,73	31,89	40,79	13,74	*-9,11*	*-0,21*
Median	35,00	55,00	27,00	37,00	14,00	*-9,00*	*3,00*
Stdev.	14,97	17,88	22,04	17,37	19,62	*23,30*	*11,79*
Experimental group (20 students)							
Mean	42,95	50,30	42,05	57,20	7,35	*-0,90*	14,25
Median	39,50	48,00	38,00	56,00	5,50	*-6,00*	13,00
Stdev.	13,48	21,32	24,21	11,27	18,62	*22,78*	12,14

In the following, the results of statistical hypotheses testing are presented for each hypothesis ($H0_{CHK1}$, $H0_{CHK2}$, …, $H0_{END}$) individually. Table 3 shows the results of testing hypothesis H0 using a two-tailed t-test for dependent groups. Column one specifies the test and the related study, i.e. initial experiment (E) and replication (R). Column two represents the effect size, column three the degrees of freedom, column four the t-value of the study, column five the critical value (the commonly accepted practice is to set $\alpha = 0.05$) that the t-value has to exceed to be statistically significant, and column six provides the associated p-value.

By examining columns four and five of Table 3, it can be seen that the experimental groups achieved a statistically significant result for dependent variable twice in the initial experiment, and once in the replication experiment. It should be noted, though, that in both experiments the post-test values support the direction of the expected positive learning effect.

5.5. Interpretation of results and discussion

At the end, we summarize the results of the initial experiment and its replication with regards to null hypothesis H0 in Table 4. Statistical significance (stat. sig.), mentioned in that table means that null hypothesis could be rejected at significance level $\alpha = 0.05$. Practical significance (pract. sig.) means that null hypothesis could not be rejected, but effect size is $\Delta \geq 0.5$. If statistical significance is achieved, practical significance is not mentioned. Positive effect (+) means that no practical significance could be observed, but effect size is $\Delta>0$. No effect or negative effect (-) means that effect size is $\Delta\leq 0$. On the second checkpoint test the control group performed better than the experimental way in statistically significant sense.

Table 3 shows that null hypothesis $H0_{CHK1}$ could not have been rejected in any experiment. Regarding the first checkpoint test, the expected positive learning effect could be observed only in the initial experiment, but it was statistically insignificant. In other words, in the initial experiment, the experimental group performed better than the control group, but it was not statistically significant. In the replication experiment, the control group performed better than the experimental group, but it also was not statistically significant.

The null hypothesis $H0_{CHK2}$ has been rejected only in the replication experiment (Table 3). Regarding the second checkpoint test, the expected positive learning effect could be observed only in the replication experiment, but it was statistically insignificant. In other words, in the initial experiment, the control group was statistically significantly better than the experimental group. In the replication experiment, the experimental group performed better than the control group, but it also was not statistically significant.

The null hypothesis $H0_{END}$ has been rejected in both

Table 3

	Effect size Δ	df	t-value	Crit. t α = 0.05	p-value
First checkpoint test					
E	0,17	78	-0,73	1,99	0,4676
R	-0,33	37	1,04	1,68	0.3051
Second checkpoint test					
E	*-0,47*	*78*	*2,31*	*1,99*	*0,0235*
R	0,35	37	-1,11	1,68	0,2742
Post test					
E	*0,79*	*78*	*-3,62*	*1,99*	*0,0005*
R	*1,23*	*37*	*-3,77*	*1,68*	*0,0006*

Table 4

Experimental group vs. Control group	Dependent variable – student knowledge	
	Statistical significance / Practical significance	Positive effect size / Negative effect size
Initial experiment		
First checkpoint test	none	+
Second checkpoint test	Stat. sig.	-
Post-test	Stat. sig.	+
Replication experiment		
First checkpoint test	none	-
Second checkpoint test	none	+
Post-test	Stat. sig.	+

experiments (Table 3). Regarding the post-test, the expected positive learning effect has been observed in both experiments, and it was statistically significant. In other words, in the initial experiment, the experimental group was statistically significantly better than the control group. In the replication experiment, the experimental group was also statistically significantly better than the control group.

After the initial experiment results' analysis, we have calculated that the xTEx-Sys's educational influence has the average effect size of 0,16 sigma (standard error of effect size SE_E = 0,224, effect size weight w_E = 19,93). After the replication experiment results' analysis, we have calculated that the xTEx-Sys's educational influence has the average effect size of 0,42 sigma (standard error of effect size SE_R = 0, 323, effect size weight w_R = 9,59).

Starting out from the results presented in the previous section, interpretations and possible explanations of the outcomes of the experiments will be given below, followed by a discussion of the validity of the results.

The strong effect observed for post-test when comparing the performance of experimental to control groups in both experiments can probably be attributed to the inclusion of the xTEx-Sys in the treatments of the experimental groups.

Also, the positive impact of working with the xTEx-Sys calculated using first checkpoint test which was found in the initial experiment, was not confirmed by the replication. The good thing is that the negative statistically significant impact of working with the xTEx-Sys calculated using second checkpoint test which was found in the initial experiment, was not confirmed by the replication. That negative impact had happened due to organizational problems related to scheduling of the experiment, when the experimental group has taken the second checkpoint-test before the control group.

6 Conclusion

The empirical studies presented in this paper investigated the effect of the intelligent authoring shell xTEx-Sys. The system's educational effectiveness was analyzed by comparing the test results of students who used the xTEx-Sys to the test results of students who were traditionally tutored in the initial and the replicated experiment.

Although the results of the two studies are promising, we expected to get larger average effect sizes. A reasonable explanation for the small, or even negative partial effect sizes, could be that the xTEx-Sys's domain knowledge presentation is rather novel for students and therefore difficult to grasp and apply in earlier phases of experiment. When students get familiarized with the system's knowledge presentation, the system itself is very efficient (large post-test partial effect sizes for both experiments). As a consequence, in future experiments, the presentation of the xTEx-Sys should be improved.

As mentioned before, in order to develop and improve the xTEx-Sys, further experiments must be conducted. The following questions should be addressed by future experiments: What is the main reason why the initial experiment yielded positive effect for the first checkpoint test while the replication did not? Is this due to high pre-test scores or other unknown factors? Why were the pre-test scores in the replication much lover than in the initial experiment? Are the similar average effect sizes of two experiments with same students, but different domain knowledge, influenced by subjects more than the system itself? Or is the system evenly effective regardless of domain knowledge? Could the xTEx-Sys it be further improved in order to produce a more positive impact in every stage of the experiment?

It should be emphasized that the presented exploratory research is just the first step of a series of experiments, which – after modification of the treatments and inclusion of subjects with different backgrounds – might yield more generalisable results in the future. Results gained through the conducted experiments have shown a need for adding some extended functions for courseware development and learning management in the xTEx-Sys.

In this paper we have presented an approach to replication of an experiment for intelligent tutoring system's effectiveness evaluation. Our intention is to continue replicating experiments using the same experimental design in order to calculate the xTEx-Sys's total effect size using meta-analysis's methods.

The results of experimental studies are often reported in ways that make it is difficult to locate the calculated effect sizes or even. Therefore, the researchers should be familiar with meta-analysis process and statistics basis, and should start to report their results in a way that is appropriate for meta-analyzing. We believe that a meta-analysis, conducted using results from experiments that are based on the same methodology, is a key issue in resolving a problem of making conclusions about overall intelligent tutoring system effectiveness. Therefore, a replication of experiments is a first step in meta-analysis related to intelligent tutoring systems' effectiveness calculation, because the replications ease the problem of studies comparability.

7 Acknowledgments

This work has been carried out within scientific project 177-0361994-1996 „Design and evaluation of intelligent e-learning systems“, funded by the Ministry of Science, Education and Sports of the Republic of Croatia.

References:

[1] Almqvist J. P. F. (2006) Replication of controlled Experiments in Empirical Software Engineering - A Survey. MS thesis, Department of Computer Science, Faculty of Science, Lund University.

[2] Baruque, L.B., Melo, N.R. (2006). "Towards Metrics for the Assessment of Web-Based Education", in International Journal WSEAS Transactions on Computers, Vol. 5, Issue 11, November 2006, ISSN: 1109-2750, pp.2668-2673.

[3] Becker, L.A. (2000) Online syllabus - Basic and Applied Research Methods. Retrieved 14/09/2007 from web.uccs.edu/lbecker/Psy590/default.html

[4] Bloom, B.S. (1984) The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. Educational Researcher, 13, pp. 4-16.

[5] Cumming, G. (2006). Meta-analysis: Pictures that explain how experimental findings can be integrated. In A. Rossman and B. Chance (Eds.), Proceedings of the Seventh International Conference on Teaching Statistics, Salvador, Brazil. Voorburg: The Netherlands: International Statistical Institute.

[6] Deligiannis I. S., Shepperd M., Webster S., Roumeliotis M. (2002) A review of experimental investigations into object-oriented technology. Empirical Software Engineering, 7(3), pp. 193–231.

[7] Dempster, J. (2004) Evaluating e-learning developments: An overview. Retrieved 14/09/2007 from warwick.ac.uk/go/cap/resources/eguides

[8] Farris T. (2006) Experimental error. Pearson Custom Publishing, Retrieved 14/09/2007 from www2.volstate.edu/tfarris/PHYS2110-2120/experimental_error.htm

[9] Fletcher, J.D. (2003) Evidence for Learning from Technology-Assisted Instruction. In: Technology applications in education: a learning view, (H.F. O'Neal, R.S. Perez (Ed.)), Mahwah, NJ: Lawrence Erlbaum Associates, pp. 79-99.

[10] Glass, G. V (1976). Primary, secondary, and meta-analysis of research. Educational Researcher, 5, pp. 3-8.

[11] Grubišić, A., Stankov, S., Žitko, B. (2007). "EVEDIN: A System for Automatic Evaluation of Educational Influence", in International Journal WSEAS Transactions on Computers, Vol. 6, Issue 1, January 2007, ISSN: 1109-2750, pp.95-102.

[12] Hedges, L.V., Olkin, I. (1985) Statistical methods for meta-analysis. New York: Academic Press.

[13] Hunt, M. (1997). How Science Takes Stock. The Story of Meta-Analysis. New York: Sage.

[14] Iqbal, A., Oppermann, R., Patel, A., Kinshuk. (1999) A Classification of Evaluation Methods for Intelligent Tutoring Systems., In: Arend, U., Eberleh, E., Pitschke, K. (eds.): Software Ergonomie '99 , B. G. Teubner, Stuttgart, Leipzig, pp. 169-181.

[15] Lipsey, M.W., Wilson, D.B. (2001) Practical meta-analysis. Sage Publications, Inc., Thousand Oaks, California.

[16] Litoiu M., Rolia J., Serazzi G. (2000) Designing process replication and activation: A quantitative approach. IEEE Transactions on Software Engineering, 26(12), pp. 1168–1178.

[17] Lucas J. W. (2003) Theory-testing, generalization and the problem of external validity. Sociological Theory, 21(3), pp. 236–253.

[18] Murray, L. R., Ehrenberg, A. S. C. (1993) The design of replicated studies. American Statistician, 47(3): pp. 217–228

[19] Murray, T. (1996) Having It All, Maybe: Design Tradeoffs in ITS Authoring Tools. In Proceedings of the Third International Conference on Intelligent Tutoring Systems, Montreal

[20] Ohlsson, S. (1987) Some Principles of Intelligent Tutoring. In Lawler & Yazdani (Eds.), Artificial Intelligence and Education, Volume 1. Ablex: Norwood, NJ, pp. 203-238.

[21] Perseus Digital Library Project. Ed. Gregory R. Crane. Tufts University. Accessed on 21.02.2008. http://www.perseus.tufts.edu

[22] Pfahl, D. (2004) Evaluating the learning effectiveness of using simulations in software project management education: Results from a twice replicated experiment. Information and Software Technology (Elsevier), v46, pp. 127-147.

[23] Pfleeger S.L. (1995) Experimental design and analysis in software engineering, part 2: How to set up an experiment. ACM SIGSOFT Software Engineering Notes, 20(1), pp. 22–26.

[24] Phillips, R., Gilding, T. (2003) Approaches to evaluating the effect of ICT on student learning. ALT Starter Guide 8.

[25] Rodríguez, D., Sicilia, M. A., Cuadrado-Gallego, J. J., Pfahl, D. (2006) e-Learning in Project Management Using Simulation Models: A Case Study Based on the Replication of an Experiment. IEEE Transactions on Education 49(4), pp. 451-463.

[26] Rosić, M. (2000) Establishing of Distance Education Systems within the Information Infrastructure. Faculty of Electrical Engineering and Computing, Zagreb, Croatia, MS Thesis (in Croatian)

[27] Schulze, R. (2004). Meta-analysis: A comparison of approaches. Cambridge, MA: Hogrefe & Huber.

[28] Sleeman, D., Brown, J. S. (1982) Introduction: Intelligent Tutoring Systems. In D. Sleeman & J. S.

Brown (Eds.), Intelligent Tutoring Systems. New York: Academic Press. pp. 1-11.

[29] Stankov, S. (1997): Isomorphic Model of the System as the Basis of Teaching control Principles in an Intelligent Tutoring System. Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture, Croatia, PhD Thesis (in Croatian)

[30] Stankov, S., Glavinić, V., Grubišić, A. (2004) What is Our Effect Size: Evaluating the Educational Influence of a Web-Based Intelligent Authoring Shell?, Proceedings INES 2004 8th International Conference on Intelligent Engineering Systems, Nedevschi, Sergiu; Rudas, Imre J. (eds.). Cluj-Napoca: Faculty of Automation and Computer Science, Technical University of Cluj-Napoca, 2004., pp. 545-550

[31] Stankov, S., Rosić, M., Žitko, B., Grubišić, A. (in press): TEx-Sys model for building intelligent tutoring systems, Computers & Education (2007), doi:10.1016/j.compedu.2007.10.002

[32] Wenger, E. (1987) Artificial Intelligence and Tutoring Systems. Los Altos, California: Morgan Kaufmann Publishers.