Computers & Education xxx (2009) xxx–xxx

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

Controlled experiment replication in evaluation of e-learning system's educational influence

Ani Grubišić *, Slavomir Stankov, Marko Rosić, Branko Žitko

Faculty of Natural Science, Mathematics and Kinesiology, University of Split, Teslina 12, 21000 Split, Croatia

ARTICLE INFO

Article history:
Received 19 October 2007
Received in revised form 17 March 2009
Accepted 24 March 2009
Available online xxxx

Keywords:
Authoring tools and methods
Distributed learning environments
Evaluation methodologies
Intelligent tutoring systems

ABSTRACT

We believe that every effectiveness evaluation should be replicated at least in order to verify the original results and to indicate evaluated e-learning system's advantages or disadvantages. This paper presents the methodology for conducting controlled experiment replication, as well as, results of a controlled experiment and an internal replication that investigated the effectiveness of intelligent authoring shell eXtended Tutor–Expert System (xTEx-Sys). The initial and the replicated experiment were based on our approach that combines classical two-group experimental design and with factoral design. A trait that distinguishes this approach from others is the existence of arbitrary number of checkpoint-tests to determine the effectiveness in intermediate states. We call it a pre-and-post test control group experimental design with checkpoint-tests. The gained results revealed small or even negative effect sizes, which could be explained by the fact that the xTEx-Sys's domain knowledge presentation is rather novel for students and therefore difficult to grasp and apply in earlier phases of the experiment. In order to develop and improve the xTEx-Sys, further experiments must be conducted.

1. Introduction

Nowadays, many educational institutions have embraced at least one type of worldwide known phenomena called an e-learning. The e-learning presents an intersection between a world of information and communication technology and a world of education (Stankov, Grubišić, & Žitko, 2004). As the e-learning presents a wide set of applications and processes that make educational content available on different electronic media (ASTD, 2001), e-learning systems, therefore, provide access to electronically based learning resources anywhere at all times for anyone (Albert, 2001). Intelligent e-learning systems have the capability to act appropriately in uncertain situations that appear in learning and teaching process. Their main goal is to be as good as the highly successful human tutor. A special class of intelligent e-learning systems is intelligent tutoring systems (ITS).

The e-learning systems' developers have become so involved in making their system work that they have paid little attention to the process of evaluation. Since the major goal of an e-learning system is to teach, its evaluation's main test is to find out whether students learn effectively from it (Mark & Greer, 1993). It is essential to evaluate all instructional software before using it in educational process. Evaluation offers the information to make decision about using the product or not (Phillips & Gilding, 2003). So, a well-designed evaluation should provide the evidence, whether a specific approach has been successful and of potential value to the others (Dempster, 2004). One special form of an e-learning system's evaluation is the 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 the external evaluation, and as it bases itself on experiment, it is a part of an experimental research (Iqbal, Oppermann, Patel, & Kinshuk, 1999).

Experiments used in the e-learning systems effectiveness evaluation change the independent variable (tutoring strategy) while measuring the dependent variable (student's achievement), therefore, require two groups – a control and an experimental group. The control group is involved in the traditional learning and teaching process and the experimental group uses the e-learning system. Experiments are conducted to verify null-hypotheses H0: "There is no significant difference between the control group and the experimental group".

A key concern in the experimental research is related to the assumption about trueness of the conclusion from the experimental result. This concept is usually known as validity. Validity refers to whether the results of an experiment are valid, or to be exact, whether the

* Corresponding author. Tel.: +385 21 385 133.
E-mail addresses: ani.grubisic@pmfst.hr (A. Grubišić), slavomir.stankov@pmfst.hr (S. Stankov), marko.rosic@pmfst.hr (M. Rosić), branko.zitko@pmfst.hr (B. Žitko).

0360-1315/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compedu.2009.03.014

conclusions drawn from the experiment follow logically from the results of the experiment (Almqvist, 2006). The experiment validity, that is, a validity of the results, can be ensured by a replication of the same experiment. The replication is the repetition of the experiment following, as closely as possible, the original experiment.

This paper presents the results of a controlled experiment and an internal replication that investigated the effectiveness of a particular e-learning system. In the second chapter we review the age-long research and development of the Tutor–Expert System (TEx-Sys) model for building ITS (Stankov, 1997). In the third chapter we discuss some issues related to the experiment replication. In the fourth chapter we give an overview of the evaluation methodology that has been used in the initial and replicated experiment. 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) (Stankov, 2005), which is the representative of Web-based authoring shells for building ITS based on the TEx-Sys model.

2. Related work

The intelligent tutoring systems are computer systems that support and improve learning and teaching process in certain domain knowledge, respecting the individuality of a learner as in traditional "one-to-one" tutoring (Ohlsson, 1987; Wenger, 1987; Sleeman & Brown, 1982). 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 (Murray, 1996).

The first implementation of intelligent authoring shell model called the TEx-Sys 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) (Rosić, 2000) 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. The xTEx-Sys system administers questions that are not predefined, but rather generated by the system based on the predefined semantic network structures related to the domain of interest.

In the past decade, there were numerous applications of the TEx-Sys model in the learning and teaching process that involved students from primary education all the way to the academic level. Only in the period from 2001 to 2007, 1302 students took 5482 content knowledge tests in one of the TEx-Sys model versions in order to evaluate their understanding of different knowledge domains (Table 1).

Questionnaires about the students' impressions were given to the students after finishing the courses that were supported by the TEx-Sys model. The qualitative analysis of questionnaires' results revealed that most students were pleased working with the model and that they were open minded for embracing that kind of learning and teaching support. The qualitative analysis could not determine the effect of educational influence of the TEx-Sys model. That was the reason why we have conducted four experiments (the two presented in this chapter with the DTEx-Sys, and initial and replicated experiments presented in the fifth chapter with the xTEx-Sys) in order to evaluate the educational influence of the TEx-Sys model (Grubišić, Stankov, & Žitko, 2006).

2.1. A variation of Bloom's experiment - DTEx-Sys

In a widely quoted research, Bloom (1984) had compared student learning under three different forms of instruction: conventional learning, mastery learning and tutoring. Using the standard deviation of a control group (attending a conventional form of instruction), he found that the average tutored student was about two standard deviations (2-σ) better than the average student who was involved in a conventional learning and teaching process. In other words, tutoring improved the achievement of 50th percentile students to that of 98th percentile students. This research had subsequently started an avalanche of research seeking ways of accomplishing this result under more practical and realistic conditions than one-to-one tutoring with human teachers.

In our variation of the Bloom's experiment replication (Stankov, Glavinić, & Grubišić, 2004), 33 students, who were taking "*Introduction to computer science*" class in academic year 2004/05, were randomly and equally divided into a control group (11 students), a tutoring group (11 students) and an experimental group (11 students). 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 2 h weekly throughout one semester.

Table 1
The number of tests performed on the TEx-Sys model and the number of students involved in every academic year from 2001 to 2007.

	TEx-Sys	DTEx-Sys	xTEx-Sys	Total
2001/2002	72 Tests 18 Students	–	–	72 Tests 18 Students
2002/2003	–	648 Tests 72 students	–	648 Tests 72 Students
2003/2004	–	591 Tests 153 Students	–	591 Tests 153 Students
2004/2005	169 Tests 119 Students	1077 Tests 165 Students	527 Tests 73 Students	1773 Tests 357 Students
2005/2006	–	–	1368 Tests 552 Students	1368 Tests 552 Students
2006/2007	–	–	1030 Tests 150 Students	1030 Tests 150 Students
Total	241 Tests 137 Students	2316 Tests 390 Students	2925 Tests 775 Students	5482 Tests 1302 Students

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 statistically significant difference between the control group and the experimental group ($t = 2.41$, $p = 0.04$), which had showed the advantage of the DTEx-Sys's over the traditional learning and teaching. Statistically insignificant difference between the experimental group and the tutoring group concerning the post-test results ($t = 0.53$, $p = 0.61$) has shown competency of the DTEx-Sys's 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.

2.2. A simple control group experiment – DTEx-Sys

In another experiment (Stankov, Grubišić, Žitko, & Krpan, 2005) that was conducted in the same academic year 2004/05, 31 students, who were taking "*Object-oriented programming*" class, were randomly divided into an experimental group (20 students) and a control group (11 students). The control group was involved in the traditional learning and teaching process and the experimental group was asked to use the DTEx-Sys for one week.

Students underwent a paper-and-pen pre-test that was distributed at the beginning of the course, which enabled us to determine that there were no statistically significant differences between the groups concerning their foreknowledge, even though, they were not numerically equivalent. After one week of learning, the students underwent a paper-and-pen post-test, which enabled us to determine that there was no statistically significant difference between the groups after the treatment ($t = 0.58$, $p = 0.56$), which had showed that the DTEx-Sys had no influence on the students' knowledge. The calculated effect size was only 0.12. Since we expected better results, encouraged by the ones gained through the previously described experiment, we decided to find out what parasitic factors might have influenced the research. Some of the reasons for such a low effect size might have been the students' lack of motivation or the length of the experiment (only one week).

3. A replication of an experiment

The replication, in the context of this paper, is the repetition of an experiment following, as closely as possible, the original experiment. The replication of controlled experiments is considered to be a critical aspect of the scientific method (Litoiu, Rolia, & Serazzi, 2000).

Pfleeger (1995) underlines that the replication means repeating an experiment under equal circumstances and not repeating measurements on the same experimental unit, which refers to literally taking several measurements of a single occurrence of the phenomenon.

At least one replication is needed if someone wants that their results are of any interest at all. Any result from an isolated study cannot show whether conclusions will hold again. The first replication shows whether or not a generalization is possible (Murray & Ehrenberg, 1993).

According to Murray and Ehrenberg (1993), there are two types of replication: close and differentiated replication. The close replication attempts to keep almost all the known conditions of the study 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.

3.1. Replication errors

While conducting the experiment, one or more of three general types of errors could arise: human error, systematic error, and random error (Farris, 2006). The *human error* (a mistake) occurs when the experimenter makes a mistake. For example, setting up experiment wrongly, misreading an instrument, or miscalculation. The *systematic error* 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 being too large or always too small. Most of the simple experiments have some systematic error. The *random errors* are variations between successive measurements made under apparently identical experimental conditions. All experiments have random error, which occurs because no measurement can be made with infinite precision. An example of a random error could be when trying to draw 100 lines on a sheet of paper, each exactly one centimeter long. Each line will be close to a centimeter, 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

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 research can be generalized. Each new replication of an experiment reduces the probability that results can be explained by human variation or experimental error (Almqvist, 2006). Replication can contribute significantly to generalizing results if replicated experiments employ probability-sampling techniques (Lucas, 2003). 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 (Deligiannis, Shepperd, Webster, & Roumeliotis, 2002). 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 (Pfahl, 2004).

To conclude, there does not seem to be a common ground on guidelines for the replication of experiments in the 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, Rodríguez, Sicilia, Cuadrado-Gallego, & Pfahl, 2006). 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 evaluated e-learning system's advantages or disadvantages.

4. Experimental design

Different evaluation methods are suitable for different purposes and the development of evaluation is a complex process. A controlled experiment enables researchers to examine relationships between teaching interferences and the students' teaching results, as well as to obtain quantitative measures of the significance of such relationships (Mark & Greer, 1993).

Experimental design refers to the way treatments are assigned and applied to the available groups. The primary aim of experimental design is to ensure that any differences in the measured variable have resulted from the applied treatment and not from other uncontrolled variables (Gaines, 2002).

The different major types of experimental designs are classified according to the usage of random assignment to groups. If random assignment is used, we call this design a true experimental design. If random assignment is not used, and if multiple groups are used, we call it a quasi-experimental design, but if not, we call it a non-experimental design (Trochim, 2006).

Trochim (2006) mentions the following three main types of true experiments: single group, two-groups and factoral designs.

4.1. Single-group experimental design

In order to evaluate educational influence of an e-learning system with only one group of students, it is necessary to have two learning units and two learning cycles (Fig. 1). At the beginning of each cycle, initial states Si_1 and Si_2 and respectively their means Xi_1 and Xi_2, should be captured by using pre-test before introducing experimental factors. In the first cycle, a group would use an e-learning system (experimental factor F_1), and in the second cycle the same group would be involved in traditional learning and teaching process (experimental factor F_2), or vice versa. At the end of each cycle, final states Sf_1 and Sf_2 and respectively their means Xf_1 and Xf_2, should be captured by using post-test, in order to calculate effect size of an e-learning system as an experimental factor.

A single-group experimental design does not allow objective comparisons of experimental factors. Namely, in order to compare experimental factors effect sizes, a domain knowledge in both cycles has to be different, equally difficult and unrelated as much as possible (a *history threat*), but fulfilling those demands is something that is not easy to achieve. Another reason that might influence reliability of comparison between advances in each cycle is a fact that students constantly evolve and change (a *maturation threat*), so it is advisable that a single-group experimental design does not last more than one month.

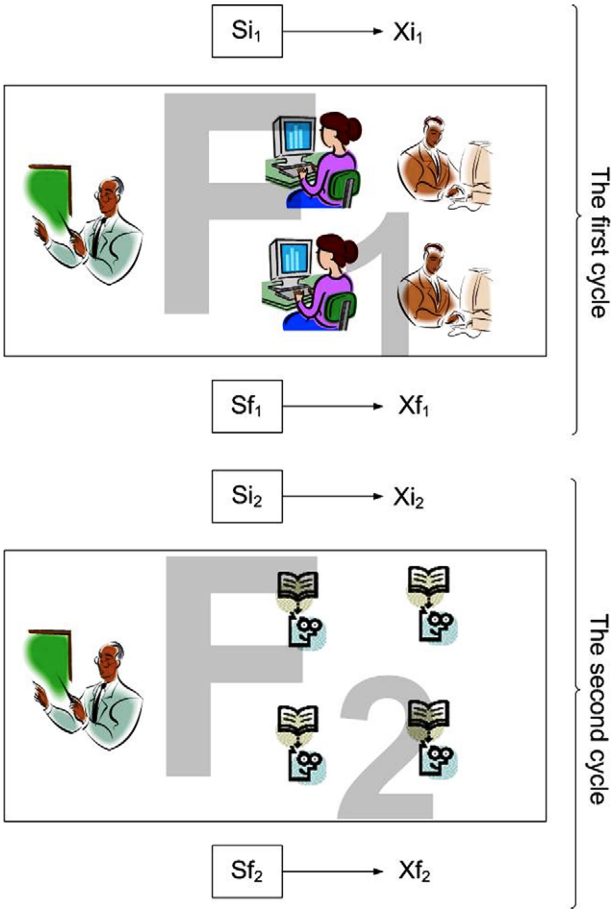


Fig. 1. A single-group experimental design.

4.2. Two-group experimental designs

At the beginning of an experiment (Fig. 2), initial states Si_{1A} and Si_{2B} and respectively their means Xi_{1A} and Xi_{2B}, should be captured using pre-test before introducing experimental factors. The group A would use an e-learning system (experimental factor F_1), and the group B would be involved in a traditional learning and teaching process (experimental factor F_2), or vice versa. At the end of an experiment, final states Sf_{1A} and Sf_{2B} and respectively their means Xf_{1A} and Xf_{2B}, should be captured using post-test, in order to calculate effect size of an e-learning system as an experimental factor.

Advantages of a two-group experimental design over single-group experimental design are: no mutual influence of factors on each other because they are introduced simultaneously; no difference in domain knowledge because they are the same; and, a test used to capture final states is the same.

The greatest problem that might occur in this experimental design is to define equal groups. When there are two groups used in an experiment, it is necessary that they are statistically equivalent. The term statistically equivalent does not mean that two groups are equal. This means that the type of equivalence is based on tests of statistical significance. In more concrete terms, statistically significant equivalence means that we are aware of the risks that there might be a difference between two groups. Statistically significant equivalence can be achieved through random assignment to groups and some matching mechanisms like the exact and caliper matching (Becker, 2000).

4.3. Factoral designs

This model presents a combination of the previous two models. Its primary purpose was to overcome some specific problems that are present in the single-group and the two-group experimental designs. In a single-group experimental design factors are introduced in a sequence, so there are two or more cycles. In a two-group experimental design, factors are introduced simultaneously, so there are two or more parallel groups. In a factoral design we have two or more parallel groups and two or more cycles. All experimental design factors are introduced simultaneously (one at a time in each group) and they are being exchanged among groups by rotation in each cycle.

At the beginning of each cycle (Fig. 3), initial states Si_{1A}, Si_{2A}, Si_{2B} and Si_{1B} and respectively their means Xi_{1A}, Xi_{2A}, Xi_{2B} and Xi_{1B}, should be captured using pre-test before introducing experimental factors. In the first cycle, group A would use an e-learning system (experimental factor F_1) and the group B would be involved in traditional learning and teaching process (experimental factor F_2), and in the second cycle the group A would be involved in traditional learning and teaching process (experimental factor F_2) and the group B would be involved in traditional learning and teaching process (experimental factor F_1), or vice versa. At the end of each cycle, final states Sf_{1A}, Sf_{2A}, Sf_{2B} and Sf_{1B} and respectively their means Xf_{1A}, Xf_{2A}, Xf_{2B} and Xf_{1B}, should be captured using post-test, in order to calculate effect size of an e-learning system as an experimental factor.

In factoral design different domain knowledge is learned in each cycle. That was a big problem in a single-group experimental design, but not here. If we find that in each cycle one and the same factor is more efficient than the other, then we can conclude that a certain factor is better regardless of a difference between taught domain knowledge.

In the same way, if the same factor is more efficient in each cycle, in spite of the groups that it has been introduced in, then we can conclude that a certain factor is better regardless of a difference between the groups. This was a big problem in a two-group experimental design.

It seems that this approach has no negative sides, but it is not true. Namely, we can only be sure in effectiveness of one factor if it was found to be better in each cycle regardless of the groups' equivalence and domain knowledge (what is not highly possible). Therefore, it would be advisable to design statistically equivalent groups. Besides, conducting this kind of experiment is related to many organizational difficulties and that is the reason why it is not used very often.

4.4. Our experimental design

In a variety of different experimental designs, we have decided to combine two-group experimental design and factoral design. Our approach to evaluation of the effectiveness of an e-learning system as an experimental factor uses two statistically equivalent groups and several (n) learning cycles.

Students who participate in the experiment are randomly assigned into a control group and an experimental group. The control group is involved in the traditional learning and teaching process and the experimental group uses the e-learning system.

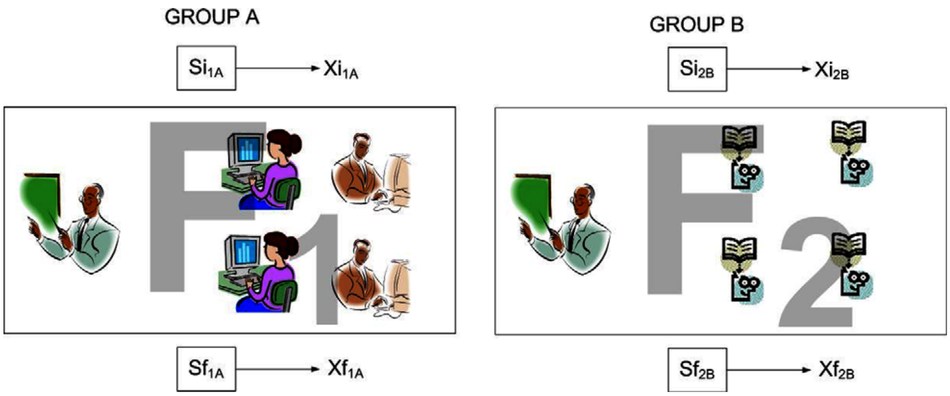


Fig. 2. A two-group experimental design.

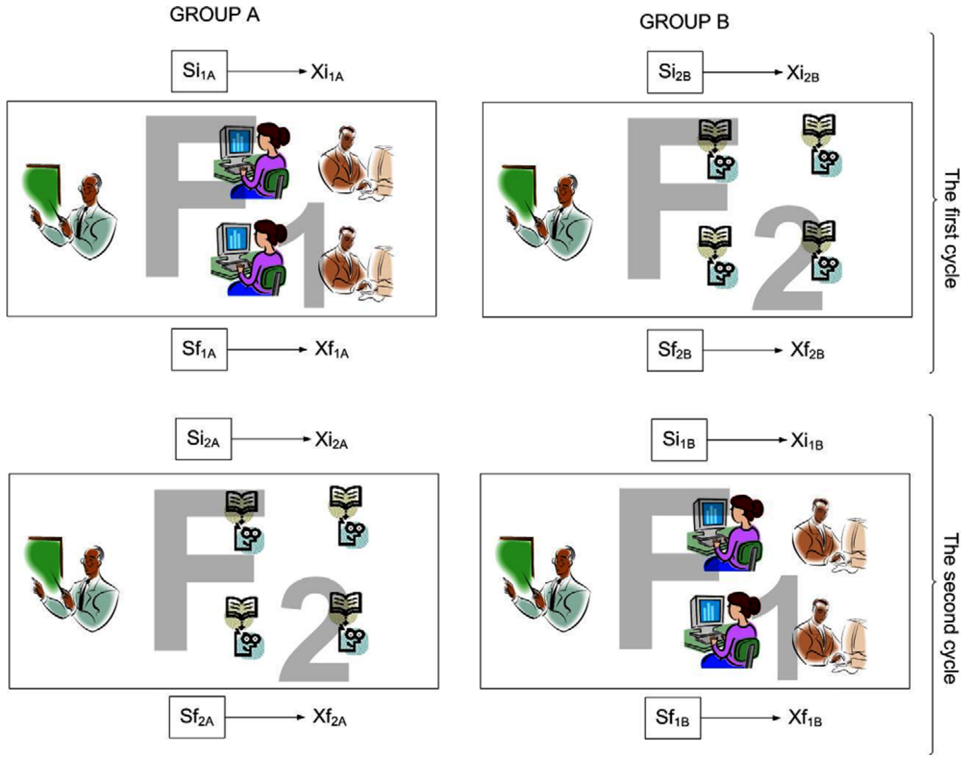


Fig. 3. A factoral design.

Our approach differs from other approaches because we add arbitrary number of checkpoint-tests in order to determine the effectiveness in intermediate states. We have named this experimental design a *pre-and-post test control group experimental design with checkpoint-tests*.

At the beginning of an experiment (Fig. 4), initial states $\mathrm{Si}_{1\mathrm{A}}$ and $\mathrm{Si}_{2\mathrm{B}}$ and respectively their means $\mathrm{Xi}_{1\mathrm{A}}$ and $\mathrm{Xi}_{2\mathrm{B}}$, should be captured using pre-test before introducing experimental factors. The group A would use an e-learning system (experimental factor F_1) and the group B would be involved in traditional learning and teaching process (experimental factor F_2) in each cycle. At the end of each cycle, intermediate states $\mathrm{Sf}_{11\mathrm{A}}$, $\mathrm{Sf}_{12\mathrm{A}}$,. . ., $\mathrm{Sf}_{1(n-1)\mathrm{A}}$, $\mathrm{Sf}_{21\mathrm{B}}$, $\mathrm{Sf}_{22\mathrm{B}}$. . ., $\mathrm{Sf}_{1(n-1)\mathrm{B}}$ and respectively their means $\mathrm{Xf}_{11\mathrm{A}}$, $\mathrm{Xf}_{12\mathrm{A}}$,. . ., $\mathrm{Xf}_{1(n-1)\mathrm{A}}$, $\mathrm{Xf}_{21\mathrm{B}}$, $\mathrm{Xf}_{22\mathrm{B}}$. . ., $\mathrm{Xf}_{1(n-1)\mathrm{B}}$, should be captured using $n-1$ checkpoint tests, in order to calculate partial effect size of an e-learning system. At the end of an experiment, final states $\mathrm{Sf}_{1n\mathrm{A}}$ and $\mathrm{Sf}_{2n\mathrm{B}}$ and respectively their means $\mathrm{Xf}_{1n\mathrm{A}}$ and $\mathrm{Xf}_{2n\mathrm{B}}$, should be captured using post-test, in order to calculate partial effect size of an e-learning system as an experimental factor.

So, both groups take a 45-min pre-test at the very beginning of the experiment to determine some individual starting level of knowledge or understanding. At a later point, approximately every four weeks, they should take the exact comparable 45-min checkpoint tests and 45-min post-test at the end of the experiment, to determine the extent to which knowledge and understanding has been improved by the educational intervention.

The effectiveness of an e-learning system is then evaluated by comparing within-groups pre-test to checkpoint tests and post-test scores.

5. Description of the experiments

To assess the effectiveness of the xTEx-Sys, we have conducted two experiments: the initial one in academic year 2005/06 (Grubišić et al., 2006) and its replication in 2006/07. Both experiments are carried out according to the pre-and-post test control group experimental design with checkpoint-tests described earlier in the fourth chapter.

5.1. Subjects

Students who participated in the initial and the replication experiment were undergraduate students from two faculties: the Faculty of Chemical Technology and the Faculty of Science, both at University of Split in Croatia, 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), education, preferences and beliefs about learning styles. These questions could be answered on a voluntary basis. The results are presented in Table 2.

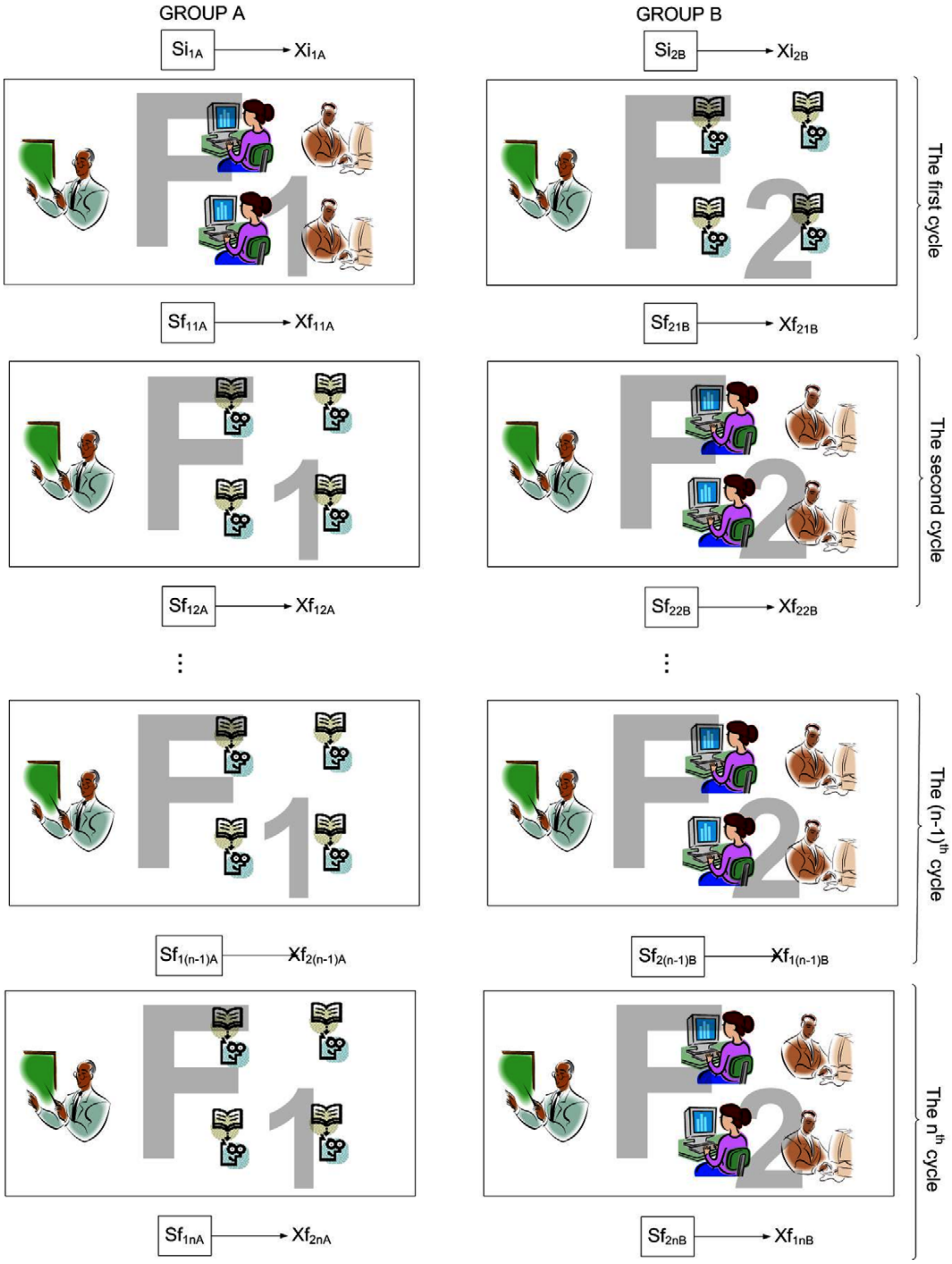


Fig. 4. Pre-and-post test control group experimental design with checkpoint-tests.

Due to organizational and legality problems, we have decided, in prior, that the students from the Faculty of Chemical Technology would make control group students and students from the Faculty of Science would be 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, 175 students agreed to participate in the initial experiment: 86 were assigned to the control group and 109 were assigned to the experimental group. Accordingly, 127 students agreed to participate in the replication experiment: 52 students were assigned to the control group and 75 students were assigned to the experimental group.

A teacher, who has been teaching the control group students in the initial experiment, has been teaching them in the replicated experiment as well. Also, a teacher, who has been teaching the experimental group students in the initial experiment, has been teaching them in the replicated experiment as well. The control group teacher and the experimental group teacher were not the same person. The course structure and the teaching materials were the same for both groups in both experiments.

5.2. Procedure

The initial experiment and its replication were conducted following the plan presented in Table 3. After a short introduction, during which the purpose of the experiment and general organizational issues were explained, data on personal characteristics and background

Table 2
Personal characteristics.

	Initial experiment	Replication
Average age	19	19
Share of women (%)	69	64
Share of subjects majoring in		
Computer science (%)	34	52
Other (non-software related) (%)	66	48
Learning computer science using e-learning system		
Interesting (%)	60	89
Boring (%)	40	11
Learning computer science using e-learning system		
Helps (%)	51	71
Does not help (%)	2	1
I do not know (%)	47	28

Table 3
Initial and replication experiment phases.

	Initial experiment	Replication
Session 1	2 h	2 h
Introduction to experiment	5 min	5 min
Personal characteristics	15 min	15 min
Pre-test	45 min	45 min
Introduction to treatments	55 min	55 min
Session 2 – learning cycle 1	30 days	30 days
Learning and teaching process – part 1	2 h/week	2 h/week
Checkpoint test 1	45 min	45 min
Session 3 – learning cycle 2		
Learning and teaching process – part 2	2 h/week	2 h/week
Checkpoint test 2	45 min	45 min
Session 4 – learning cycle 3		
Learning and teaching process – part 3	2 h/week	2 h/week
Post-test	45 min	45 min
Questionnaire	15 min	15 min
Total	3 months	3 months

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 learning cycles. Tests were used to measure the dependent variable – student knowledge. After completing the first treatment, both groups performed the first checkpoint test (CHK1), after the second treatment they performed the second checkpoint test (CHK2), and, finally, at the end of the experiments they performed the post-test (END). As a final point, subjects got the chance to evaluate the xTEx-Sys by filling in a questionnaire, providing data on the subjective judgment of a teaching quality (Fig. 5). All tests in both experiments were respectively identical and their results were scored on a 0–100 points scale. During the whole procedure, time slots, reserved for completing a certain step of the schedule, were identical for the experimental and the control group.

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, only 40 of 86 control group students and only 80 of 109 experimental group students completed all parts of the experiment. At the end of the replication experiment, only 19 of 52 control group students and only 51 of 75 experimental group students completed all parts of the experiment.

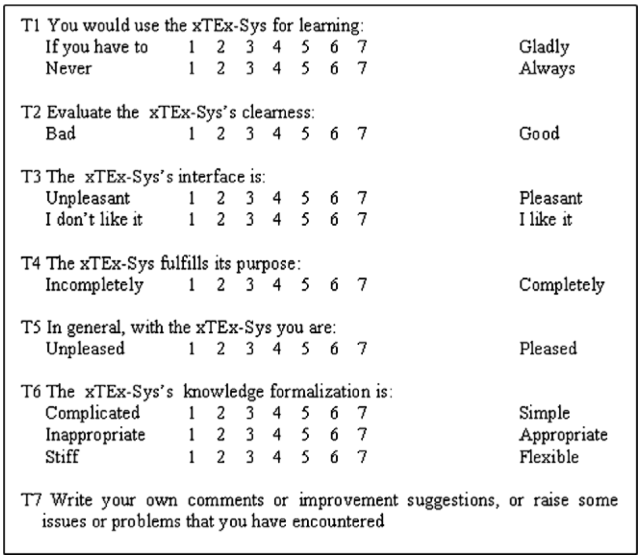
Therefore, we had to statistically equalize the control and the experimental subgroups in both experiments using the caliper matching with ±5 points range (Becker, 2000). Therefore, we randomly matched the students from the control group with the students from the experimental group who had the same pre-test results or their results varied in 5 points. For example, if we had a student from the control group that gained 57 points on pre-test, then we could match him or her with a student from the experimental group (randomly chosen) whose pre-test result was between 52 and 62 points.

In the initial experiment, in the end, there were 40 control group students randomly matched with 40 (out of 80) experimental group students using ±5 points range caliper matching. Also in the replicated experiment, in the end, there were 19 control group students randomly matched with 20 (out of 51) experimental group students using ±5 points range caliper matching.

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 are equivalent before comparing the gains of the groups. This implies calculating the means of pretest score for both groups and their standard error of mean. Now, a null-hypothesis H0 has to be stated for every checkpoint-test and the 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



T1 You would use the xTEx-Sys for learning:
If you have to 1 2 3 4 5 6 7 Gladly
Never 1 2 3 4 5 6 7 Always

T2 Evaluate the xTEx-Sys's clearness:
Bad 1 2 3 4 5 6 7 Good

T3 The xTEx-Sys's interface is:
Unpleasant 1 2 3 4 5 6 7 Pleasant
I don't like it 1 2 3 4 5 6 7 I like it

T4 The xTEx-Sys fulfills its purpose:
Incompletely 1 2 3 4 5 6 7 Completely

T5 In general, with the xTEx-Sys you are:
Unpleased 1 2 3 4 5 6 7 Pleased

T6 The xTEx-Sys's knowledge formalization is:
Complicated 1 2 3 4 5 6 7 Simple
Inappropriate 1 2 3 4 5 6 7 Appropriate
Stiff 1 2 3 4 5 6 7 Flexible

T7 Write your own comments or improvement suggestions, or raise some issues or problems that you have encountered

Fig. 5. Questionnaire about subjective judgment of teaching quality.

(Kolmogorov–Smirnof test). 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, in favor of the experimental group, 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

The descriptive statistics for the initial experiment and the replication is presented in Table 4. Columns "Pre-test", "CHK1", "CHK2" and "END" show calculated values for mean, median, and standard deviation of raw data collected during the pre-test, the first checkpoint test, the second checkpoint test and the post-test, respectively, of the initial experiment (E) and the replication (R) for both experimental and control groups.

Columns of Table 4 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 during relating post-test, it occurred twice during the initial experiment and once during the replication.

The results of statistical hypotheses testing are presented for each hypothesis ($H0_{CHK1}$, $H0_{CHK2}$,. . .,$H0_{END}$) individually. Table 5 shows the results of testing hypothesis H0 using a two-tailed *t*-test for independent groups. Column one specifies the test and the related study, i.e. the initial experiment (E) and the replication (R). Column two represents the effect size, column three the degrees of freedom, column four

Table 4
Descriptive statistics for the initial experiment and replication.

	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

Table 5
Results of testing hypothesis H0.

	Effect size Δ	d*f*	*t*-value	Crit. $t\ \alpha = 0.05$	*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 6
Summary of the results of the experiments.

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	Statistical significance	–
Post-test	Statistical significance	+
Replication experiment		
First checkpoint test	None	–
Second checkpoint test	None	+
Post-test	Statistical significance	+

the *t*-value of the study, column five the critical value (the commonly accepted practice is to use $\alpha = 0.05$) that the *t*-value has to exceed to be statistically significant, and column six provides the associated *p*-value.

In addition to filling in the questionnaires about personal characteristics and subjective perceptions, at the end of experiments, participants in the experimental groups had the chance to make comments or improvement suggestions, and could raise issues or problems that they encountered during the treatments. Apart from some improvement suggestions related to technical aspects of the system usage, comments mainly supported the findings of the quantitative analyses. Negative comments mainly addressed the difficulty of understanding the structure of the domain knowledge that is based on semantic network with frames.

5.5. Interpretation and discussion of the results

At the end, we summarize the results of the initial experiment and its replication with regards to null hypothesis H0 in Table 6. 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 $\Delta \geqslant 0.5$. If statistical significance is achieved, practical significance is not mentioned (Cohen, 1988). Positive effect (+) means that effect size is $\Delta > 0$. No effect (zero) or negative effect (–) means that effect size is $\Delta \leqslant 0$. For example, on the second checkpoint test in the first experiment, the control group performed better than the experimental group, in statistically significant sense.

By examining columns three, four and five of Table 5, it can be seen that both the experimental and the control group achieved a statistically and practically significant result for dependent variable once in the initial experiment, and only the experimental group 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.

Table 6 shows that null hypothesis $H0_{CHK1}$ has been accepted for both experiments (no statistical significance). 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 also it was not statistically significant.

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

The null hypothesis $H0_{END}$ has been rejected for both experiments (Stat. sig.). 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.

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 the post-test, when comparing the performance of the experimental to the control groups in both experiments, can probably be attributed to the inclusion of the xTEx-Sys in the treatments of the experimental groups. Cohen (1988) proposes a practical interpretation of effect sizes and considers that effect sizes below 0.2 are small, effect sizes between 0.2 and 0.7 are moderate, and effect sizes above 0.7 are large.

After the initial experiment results' analysis, we have calculated that the first checkpoint-test had a small partial effect size of 0.17 (there was no statistically significant difference between the groups), the second check-point-test had a moderate partial effect size of −0.49 (there was a statistically significant difference between the groups in a favor of the control group) and finally the post-test had a large partial effect size of 0.79 (there was a statistically significant difference between the groups in favor of the experimental group). The educational influence of the xTEx-Sys, in the initial experiment, has the average effect size of 0.16σ.

After the replication experiment results analysis, we have calculated that the first checkpoint-test had a small partial effect size of −0.33 (there was no statistically significant difference between the groups), the second checkpoint-test had a moderate partial effect size of 0.35 (there was no statistically significant difference between the groups) and finally the post-test had a large partial effect size of 1.23 (there was a statistically significant difference between the groups in favor of the experimental group). The educational influence of the xTEx-Sys, in the replicated experiment, has an average effect size of 0.42σ.

According to Cohen (1988), the average effect size 0.16 of the initial experiment implies that the xTEx-Sys has a very small educational influence on students' learning and teaching process. Furthermore, the average effect size 0.42 of the replicated experiment implies that the xTEx-Sys has a moderate educational influence on students' learning and teaching process. If we observed only post-test results (such as in classical two-group experimental design with pre-and-post tests), effect sizes (0.79 in the initial experiment and 1.23 in the replicated experiment) are greatly higher and could be classified as large effect sizes.

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, and the students from the experimental group informed the students from the control group about the questions that were in the test.

Threat to internal validity was preceding decision about control and experimental groups. Namely, due to organizational and legality problems, we had to decide what faculty would make the extended control group students, and what would make the extended experimental group students. That prior division was later found to be proper, because the pre-test results for subgroups of defined groups have shown that those subgroups were statistically equivalent in both experiments. Threat to external validity relates participating subjects' expertise level in computer science or related.

6. Conclusion

The empirical studies presented in this paper investigated the effect of using one 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 the 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? Is the system evenly effective regardless of domain knowledge? Could the xTEx-Sys 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 order to get it as close as possible to the Bloom's 2-σ target.

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

Albert, D. (2001). E-learning future – The contribution of psychology. In R. Roth, L. Lowenstein, & D. Trent (Eds.), *Catching the future: Women and men in global psychology, proceedings of the 59th annual convention, international council of psychologists* (pp. 30–53). England: Winchester.

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.

ASTD. (2001). *A vision of e-learning for America's workforce*. Report of the commission on technology and adult learning.

Becker, L. A. (2000). *Online syllabus – Basic and applied research methods*. Retrieved 14/09/2007 from <web.uccs.edu/lbecker/Psy590/default.html>.

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*, 4–16.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Ass.

Deligiannis, I. S., Shepperd, M., Webster, S., & Roumeliotis, M. (2002). A review of experimental investigations into objectoriented technology. *Empirical Software Engineering, 7*(3), 193–231.

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

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

Gaines, D. R. (2002). The role of ancillary variables in the design, analysis, and interpretation of animal experiments. *ILAR Journal, V43*(4), 214–222.

Grubišić, A., Stankov, S., & Žitko, B. (2006). An approach to automatic evaluation of educational influence. In *Proceedings of the 6th WSEAS international conference on distance learning and web engineering (DIWEB 06)* (pp. 20–25). Lisabon, Portugal.

Please cite this article in press as: Grubišić, A., et al. Controlled experiment replication in evaluation of e-learning system's educational influence. *Computers & Education* (2009), doi:10.1016/j.compedu.2009.03.014

Iqbal, A., Oppermann, R., Patel, A., & Kinshuk (1999). A classification of evaluation methods for intelligent tutoring systems. In U. Arend, E. Eberleh, & K. Pitschke (Eds.), *Software Ergonomie 99* (pp. 169–181). Stuttgart, Leipzig: B.G. Teubner.

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

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

Mark, M. A., & Greer, J. E. (1993). Evaluation methodologies for intelligent tutoring systems. *Journal of Artificial Intelligence and Education, 4*(2/3), 129–153.

Murray, R. L., & Ehrenberg, A. S. C. (1993). The design of replicated studies. *American Statistician, 47*(3), 217–228.

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.

Ohlsson, S. (1987). Some principles of intelligent tutoring. In Lawler & Yazdani (Eds.). *Artificial intelligence and education* (Vol. 1, pp. 203–238). Norwood, NJ: Ablex.

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* (Vol. 46). Elsevier. pp. 127–147.

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), 22–26.

Phillips, R., Gilding, T. (2003). Approaches to evaluating the effect of ICT on student learning. *ALT starter guide 8*.

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), 451–463.

Rosić, M. (2000) *Establishing of distance education systems within the information infrastructure*. Faculty of electrical engineering and computing, Zagreb, Croatia, MS Thesis (in Croatian).

Sleeman, D., & Brown, J. S. (1982). Introduction: Intelligent tutoring systems. In D. Sleeman & J. S. Brown (Eds.), *Intelligent tutoring systems* (pp. 1–11). New York: Academic Press.

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).

Stankov, S., Grubišić, A., & Žitko, B. (2004). E-learning paradigm and Intelligent tutoring systems. In Z. Kniewald (Ed.), *Annual 2004 of the Croatian academy of engineering* (pp. 21–31). Zagreb: Croatian Academy of Engineering.

Stankov, S., Glavinić, V., & Grubišić, A. (2004). What is our effect size: Evaluating the educational influence of a web-based intelligent authoring shell? In *IEEE international conference on intelligent engineering systems 2004 – INES 2004*, Cluj-Napoca, Romania.

Stankov, S., Grubišić, A., Žitko, B., Krpan, D. (2005). Vrednovanje učinkovitosti procesa učenja i poučavanja u sustavima za e-učenje, Školski Vjesnik – časopis za pedagoška i školska pitanja, 54 (2005), 1–2; 21–31 (in Croatian).

Stankov, S. (2005). *Principal investigating project TP-02/0177-01 web oriented intelligent hypermedial authoring shell*. Ministry of Science and Technology of the Republic of Croatia, 2003–2005.

Trochim, W. M. (2006). *The research methods knowledge base* (2nd ed.). Retrieved 14/09/2007 from <http://www.socialresearchmethods.net/kb/>.

Wenger, E. (1987). *Artificial intelligence and tutoring systems*. Los Altos, California: Morgan Kaufmann Publishers.

Please cite this article in press as: Grubišić, A., et al. Controlled experiment replication in evaluation of e-learning system's educational influence. *Computers & Education* (2009), doi:10.1016/j.compedu.2009.03.014