Experiment Replication in Evaluation of E-Learning System's Effectiveness

ANI GRUBIŠIĆ, SLAVOMIR STANKOV, BRANKO ŽITKO
Faculty of Natural Sciences, Mathematics and Kinesiology
Nikole Tesle 12, 21000 Split
CROATIA

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 e-learning system. 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. Therefore, 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.

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

1 Introduction

It is very important to evaluate all instructional software before using it in educational process. An evaluation offers information to make decision about using the product or not [14]. So, a well-designed evaluation should provide the evidence, if a specific approach has been successful and of potential value to the others [6]. One special form of e-learning system's 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 [8].

Experiments used in the e-learning systems effectiveness evaluation change the independent variable (tutoring strategy) while measuring the dependent variable (student's achievement) and require statistically significant 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 and the experimental group".

A key issue in research is the question of whether the conclusion from the experimental result is "true". This concept is usually referred to as validity. Validity refers to whether the results of an experiment are valid, to be exact, whether the conclusions drawn from the experiment follow logically from the experiment's results [2]. 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.

This paper presents the results of a controlled experiment and an internal replication that investigated the effectiveness of an 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 [18]. In the third chapter we discuss some issues related to experiment replication. 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) [20], which is the representative of Web-based authoring shells for building ITS based on the TEx-Sys model.

2 Background

The e-learning presents intersection between a world of information and communication technology and a world of education [19]. As the e-learning presents a wide set of applications and processes that make educational content available on different electronic media [3], e-learning systems, therefore, provide access to electronically based learning resources anywhere at anytime for anyone [1]. The intelligent e-learning systems have capability to act appropriately in uncertain situations that appear in learning and teaching process. Special class of intelligent e-learning systems are the intelligent tutoring systems (ITS).

The intelligent tutoring systems 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 ([21], [12], [17]). 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 [11].

The first implementation of an 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) [15] 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.

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 [9].

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 [13].

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 [10].

According to [10], 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 [16]). 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.

4 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 [7] and its replication in 2006/07. Both experiments are conducted according to design of pre-and-post test control group experimental design with checkpoint-tests (described in [7]).

3.1. 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.

3.2. 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). As a final point, the subjects got the chance to evaluate the xTEx-Sys by filling in another questionnaire, providing data on subjective judgment of teaching quality (Fig 1). 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 [4]. 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.

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

Figure 1

3.3. 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.

3.4. Results

Table 1 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 1 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

Table 1

	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

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

In the following, the results of statistical hypotheses testing are presented for each hypothesis ($H0_{CHK1}$, $H0_{CHK2}$, …, $H0_{END}$) individually. Table 2 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 2, 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.

In addition to filling in the questionnaires about personal characteristics and subjective perceptions, 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.

3.5. 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 3. 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 2 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

Table 2

	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 3

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.	+

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 2). 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 experiments (Table 2). 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 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.

5 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.

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 xTEx-Sys's educational influence has the average effect size of 0.16 sigma.

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 xTEx-Sys's educational influence has the average effect size of 0.42 sigma.

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.

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.

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 order to get it as close as possible to the Bloom's 2-sigma target [4].

6 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] Albert, D. (2001) E-learning Future – The Contribution of Psychology. In: Roth, R., Lowenstein, L., Trent, D. (eds.): Catching the Future: Women and Men in Global Psychology, Proceedings of the 59th Annual Convention, International Council of Psychologists, Winchester, England, pp. 30-53.

[2] 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.

[3] ASTD (2001) A Vision of E-Learning for America's Workforce. Report of the Commission on Technology and Adult Learning.

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

[5] 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.

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

[7] Grubišić, A., Stankov, S., Žitko, B. (2006). "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.

[8] 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.

[9] 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.

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

[11] 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

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

[13] 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.

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

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

[16] 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.

[17] 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.

[18] 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)

[19] Stankov, S., Grubišić, A., Žitko B. (2004) E-learning paradigm & Intelligent tutoring systems. In: Kniewald, Z. (ed.): Annual 2004 of the Croatian Academy of Engineering. Croatian Academy of Engineering, Zagreb, pp 21-31.

[20] 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.

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