Educational Research and Reviews Vol. 6(1), pp. 110-123, January 2011
Available online at http://www.academicjournals.org/ERR
ISSN 1990-3839 ©2011 Academic Journals

Full Length Research Paper

An educational dilemma: Are educational experiments working?

Serhat Kocakaya

Department of Physics Education, Faculty of Education, Yüzüncü Yıl University, 65080 Campus/Van, Turkey.
E-mail: skocakaya@gmail.com, skocakaya@yyu.edu.tr. Tel/Fax: 0432 2251024/0432 2251038.

Accepted 23 December, 2010

The main aim of this study is to investigate the biased effects which affect scientific researches; although most of the researchers ignore them, and make criticism of the experimental works in the educational field, based on this viewpoint. For this reason, a quasi-experimental process has been carried out with five different student teachers groups, to see the John Henry effect which is one of those biased artifacts. At the end of the whole process, it was seen that the John Henry effect plays a great role on students' achievements. At the end of the study, experimental works were discussed on the basis of those artifacts and some recommendations were presented on how we can reduce the effects of those artifacts or how we can use them to get benefit for our experimental works.

Key words: Internal validity, bias effects, the John Henry effect.

INTRODUCTION

This study describes an unacknowledged artifact that may confound experimental evaluations. The present study hypothesize that, the control group members perceiving the consequences of an innovation as threatening to their achievements, may perform atypically, thereby confounding the evaluation outcomes. If researchers were to examine innovations, for example, new instructional methods such as computer assisted instruction (CAI), they would find little, if any, evidence of attempts to ascertain the "normalcy" of the control group's behavior. Under such designs, any atypical performance of those executing the control, or more appropriately the comparison treatment, would likely go undetected, confounding the result of the evaluation and thereby fundamentally misleading educational decision makers regarding the substantive worth of the innovation. This very group typically constitutes the control group in the experimental approaches to the evaluation of innovations, and lastly, sources of bias and the consequent biased responses (what the study refers to as the John Henry effect) have led to many of the non-significant difference findings that have characterized so much of the evaluation research.

We can see much of these nonsignificant differences in more studies, especially technological based instructions. Researches looking for evidence of whether group teaching and video tutorial of CAI is a more effective method will typically find answers like 'there is no difference' (Taverner et al., 2000). A recent review of 135 studies of technological interventions since 1928 identified that, each and every study showed no significant effect on the outcome of interest (Russell, 2000). Even the placebo type of studies used to teach critical appraisal skills, where the control group received no relevant instruction, typically show small and non-significant effects (Norman and Shannon, 1998). In short, abundant literature in a diversity of fields shows that, it makes little difference to student outcomes, in terms of test scores and whether teaching in higher education follows method X versus method Y. We should not be surprised at this finding, because in 1968, Dubin and Taveggia had already reported numerous published studies of comparative educational methods that reveal contradicting and disappointing results. They searched for researches that compared lectures with group discussion (56 studies), supervised independent study with face-to-face instruction (74 studies), unsupervised independent study with face-to-face instruction (25 studies) and supervised with unsupervised independent study (12 studies). Studies in favor of one method

were mostly balanced by a near equal number of studies in favor of the alternative method, if the results were standardized. The reviewers conclude “no particular method of teaching is measurably to be preferred over another when evaluated by students’ examination performances” (Dubin and Taveggia, 1968).

In the light of the literatures stated, it is therefore not surprising to find that many attempts at innovations fail. A number of factors relevant to the hypothesized artifact that is operated in the organizational innovative process, may motivate organization members, either consciously or unconsciously to obstruct, divert or defeat the proposed change. As indicated by Webb et al. (1966), classical approaches to evaluation, using experimental methodology, are often insensitive to casual factors and as such, fail to differentiate the effects resulting from such factors’ manifestations and those of the innovation or “treatment”. In identifying factors that may cause resistance, Havelock (1969) indicates that there is a need for stability within organizations; because change is disruptive, it is likely to be resisted, and may have a cause on artifacts.

There are two areas of inquiry related to the hypothesized artifact. The first is related to studies of receptivity and resistance to technological change and social innovation, while the second area is related to studies of the social psychology of the experiment (Orne, 1962) and more particularly, to the artifacts (research biasing factors) that arise therein. Subsequently, this study would like to (1) interrogate the educational experiments with some questions viewpoint of designing the experimental process, (2) briefly describe some artifacts and distinguish their effect according to the experimental process, and (3) describe the quasi-experimental of this study, which would be illustrative of instances where the John Henry effect should be considered as an alternative explanation of the hypothesized artifact.

Interrogating the educational experiments

Isaac and Michael (1981: 52) identified the purpose of experimental research as investigating "possible cause-and-effect relationships by exposing one or more experimental groups to one or more treatment conditions and comparing the results to one or more control groups not receiving the treatment". They also identified seven characteristics of the experimental research implied in their definition: (a) management of the predictor and criterion variables along with the conditions in which the investigation is conducted, (b) use of a control group, (c) attempting to control variance among the predictor and criterion variables, (d) internal validity, (e) external validity, (f) ability to manage multiple predictor, criterion and extraneous variables and (g) exercise of control which makes the experimental research powerful (but also somewhat artificial) when applied to human subjects.

Quantitative data gathered from tests, attitudinal instruments or other measurements generally may be collected before (pretest) and after (post-test) the experiment is concluded. In developing the data collection and other procedures for the experiment, the researchers should be aware of any threat to internal and external validity. Internal validity refers to the extent to which findings can be interpreted accurately, while external validity refers to the extent to which results can be generalized to larger populations. Both forms of validity are important to several education research methodologies, but are especially critical in the experimental research. Campbell and Stanley (1963) are recognized by experts (McMillan, 2004; Gay, 1992; Wiersma, 2000) in the field of experimental research, as amongst the first to identfy and categorize threats to internal and external validity. Researhers should attempt to minimize, if not eliminate, the effects of these threats on the experiment. For example, subjects should be randomly assigned whenever possible (interaction affects sellection bias). If they are not, then the researcher must be careful of making generalizations of the findings to larger populations. Subjects should not be informed that they are part of an experiment (reactive effects of subjects by Hawthorne effect) because they may behave differently, if they know they are being observed in an experiment. The researchers conluded that, once the partipicants knew that they are part of a study, they became more productive regardless of the treatment. A variation of the Hawthorn effect is the Halo effect which occurs when participants know that they are part of the experimental group and their belief that they are part of a special group pushes them to improve performance. Another variation (the opposite of Halo effect) is the John Henry effect, in which participants know that they are part of the control group and make an extra effort to improve performance. The researchers concluded that, once the partipicants knew that they are part of the control group, they feel that they are in a competition and thus become more productive against the treatment group. It is possible to increase this kind of bias effect which occurs in experimental researches and affects experiment results. Subsequently, bias effects, as stated before, will be explained and discussed.

Another important problem by internal and external validity is testing the significance of the experimental results findings. In fact, while doing experimental studies, if there is a significant difference in the control group based on data analysis, it is presumed as an achievement of the used method. Contrary to this, as a result of data analysis, when the differences between the control and treatment groups are non-significant, we can presume that the experimental method has no positive contribution on students’ achievement. Of course, researchers try to give more attention during the

experimental process, but it is not known if they reveal some unexpected results on students' subliminal (though they are not aware of it) during the experimental process or not.

In fact, emphasising this on some situations which iares observed in some experimental studies makes it more understandable. Nonetheless, these situations are explained in detail by the "interrogating questions" in the study's methodsection.

Describing some artifacts

Research biasing factors as alternative explanations

Calling attention to the social psychology of the experiment, Orne (1962) observed that much of the human behavioral research focuses on what is done to the subject rather than what the subject does in reaction to the cues and stimuli of an experiment. The former category, what is done to the subject, has been the focus of most inquiries into the research biasing factors in education. A brief description of these factors and comparison of their attributes and in the facet analysis (Table 1) described by Saretsky (1975) are thus displayed.

Experimenter bias effect (EBE): It refers to the experimenter's unintentional and unconscious communication of his or her expectancies upon the experimental outcome as a partial determinant of those outcomes (Rosenthal, 1963; Barber, 1973). In the social sciences point of the view, the experimenter may introduce cognitive bias into a study in several ways. First, in what is called the observer-expectancy effect, the experimenter may subtly communicate their expectations for the outcome of the study to the participants, causing them to alter their behavior to conform to those expectations. After the data are collected, bias may be introduced during data interpretation and analysis.

The Hawthorne effect (HWE): It refers to the unexpected but beneficial effects produced in experimental situations. Such effects are said to be caused by the subject's awareness that, he or she is in an experiment and the object of special attention, is an awareness that is said to have a positive effect on the subject's performance during the duration of the experimental period (Cook, 1967). Hawthorne effect was variously attributed to: (a) the novelty of a new experimental technique, (b) awareness of participations, that is, participants perceive themselves as a subject, (c) altering the social structure, and (d) knowledge of results, that is, awareness of the subject from the result outcomes and result expectations.

Demand characteristics (DC): In research and particularly psychology, demand characteristics refer to an experimental artifact, where participants form an interpretation of the experiment's purpose and unconsciously change their behavior accordingly. Pioneering research was conducted on demand characteristics by Orne (1962). Typically, they are considered as a confounding variable, exerting an effect on behavior other than that intended by the experimenter. A possible reason for demand characteristics is the expectation from the participant that, he or she will somehow be evaluated and thus, figure out a way to 'beat' the experiment to attain good scores in the alleged evaluation.

The Halo effect (HE): The Halo effect is a cognitive bias whereby the perception of one trait (that is, a characteristic of a person or object) is influenced by the perception of another trait (or several traits) of that person or object. An example would be judging a good-looking person as more intelligent. Halo effects happen especially if the perceiver does not have enough information about all traits, so that he makes assumptions based on one or two prominent traits (Medley and Mitzel, 1963; Rozenzweig, 2007).

The Placebo effect (PE): The placebo effect has its origin in biomedical, pharmacological and psycho-pharmacological research. It refers to the therapeutic effect that a chemically inert substitute (such as sugar) has upon the patient when the patient (and doctor), unaware of the substitution, believe in the efficacy of the medication.

Investigator bias effect (IBE): In discussing the investigator bias effect, Barber (1973) distinguishes between the role of the investigator, who is the conceptualizer and designer of the research activity, and the role of the experimenter, that is, the individual(s) who interact with the subject, administers the treatment, and make observations. As indicated in the procedures of the proposal, Barber contends that, the paradigm within which the investigator works determines the nature of the hypothesis, the variables selected, the data deemed relevant and the subsequent analysis and interpretation of the results.

Deutero problem (DP): It refers to the dilemma or problem that a subject is unconsciously faced with when he must choose between being a "good subject" and winning the experimenter's approval, and meeting personal needs such as the need to success and protect himself (Reicken, 1962). The effort to address these needs may be a significant determinant of the subject's performance.

Evaluation apprehension (EA): This artifact was stated by Cottrell (1972) and Rosenberg (1969). They argued that we quickly learn that social rewards and

Table 1. Facet comparison of artifacts.

Facets for comparison			
	Central aspects	**Location within the research process**	**Kinds of error contributed**
Experimenter bias effect	Expectancies held by experimenter and their effect on his behavior with subjects.	Structuring of procedures and experimenter subject interaction.	Modification of the treatment with subsequent threat to the internal validity of the test of the hypothesis.
Hawthorne effect			
Novelty	Interaction between subject and research procedures.	Initial interaction between subject and research procedures.	Modification of the treatment with subsequent threat to the internal validity of the test of the hypothesis.
Awareness of participation	Same	Throughout the research process.	Same as above.
Altered social structure	Interaction between subject, other subjects and experimenter.	Interaction between individuals.	Same as above.
Knowledge of results	Interaction between subject and a specific aspect of the research procedure.	Following the report of the subject's performance.	Same as above.
Demand characteristics	Subject's perception of his role in the experiment.	Continuous.	Modification of the subject's role with subsequent threat to the external validity of the test of the hypothesis.
Halo effect	Rater's reaction to non-relevant information in the rating process.	During measurement that involves ratings.	Measurement error not necessarily common across subjects.
Placebo effect	Control subject's interaction with research procedures.	During experimental and control procedures.	Alters performance of control subjects, resulting in an inaccurate comparison between groups.
Investigator bias effect	Paradigm under which the investigator designs, carries out and interprets the research.	Design of the experiment, generation of hypothesis, selection of variables, subjects and analysis procedures, and analysis and interpretation of outcomes.	Modification of factors with resultant threats to internal and external validity of the test of the hypothesis.
Deutero problem	Choice between being "good subject" and meeting personal needs.	Initial interaction between the subject and the experimenter.	Alters the performance of subjects with subsequent threat to external validity of the hypothesis.
Evaluation apprehension	The subject's anxiety of evaluation and subsequent behavior to avoid negative evaluation.	Initial interaction between the subject, experimenter and procedures.	Alters the performance of the subjects.
John Henry effect	The subject's perception of the innovation consequence and subsequent behavior to demonstrate the superiority of traditional methods or avoid the negative evaluation or retain the status and traditional patterns of work.	Interaction between the subject, experimenter and procedures.	Modification of the subject's performance with subsequent threats to internal validity of the hypothesis.

punishments (for example, in the form of approval and disapproval) that we receive from other people are based on their evaluations of us. On this basis, our arousal may be modulated. In other words, performance will be enhanced or impaired only in the presence of persons who can approve or disapprove our actions. For example, this behavior can be observed in the classroom when the teacher is evaluated by his/her supervisor or principal.

John Henry effect (JHE): It occurs when people in the control group view themselves as being in competition with the experimental group and so changes their behavior. Firstly, it was suggested that the Heinich's (1970) research explained the difficulty experienced by advocates of mediated forms of instruction in demonstrating the superiority of their innovations. In the educational viewpoint, Saretsky (1972a, 1972b) has delineated the John Henry effect, as the confounding influence that the undetected atypical performance (aroused by perceptions of an innovation threat), has upon an experimental evaluation of that innovation.
The artifacts were given, and during the experimental studies, the achievement of the students in the experimental process, regardless of the experimental variables that may affect the test results should be different than what was given to the cause.

Aim

The aim of this study was to (a) determine whether there are any statistical differences in the level of student teachers' achievements when the control group feel itself in a competition (John Henry effect) and (b) make criticism of the educational research experiments viewpoint on whether or not we do put the experiment process into practice properly to improve students' achievement.

METHOD

A quasi-experimental design was used in this study to measure how the biased effects, especially John Henry effect, change the students' achievements. For that purpose, five groups were investigated in the process, but none of the groups were aware of the experiment process.

Sample

The study was conducted on five groups and it consists of 116 fourth year students of teachers training program in the Education Faculty of Yüzüncü Yıl University, Van/Turkey. Students were chosen from history, Turkish language and literature, biology, physics and mathematics education departments due to take same course. These five groups were divided into two categories as, social group and science group. History and Turkish language and literature students were marked as the social group, while biology, physics and mathematics students were marked as the science group.

Interrogating questions

The point we must first consider in the experimental studies is the experimental process used for the research group directed by whom and how. Let us try to answer questions to this issue (Q: Question).

Q1: Does the experimental study cover the whole term?
Q2: Are the researchers and formal teachers who taught the course the same?
Q3: Does the research method used in the study differ from the usual process which students are used to (so that, in general, experimental studies are like that)?
Q4: Is there, at least, a control group to compare the experimental results? And are the first two questions valid for this control group?
Q5: Are the experimental and control group students aware of the execution of the experimental study? If teaching of the control group is not managed by the researcher, does the teacher have an awareness of the experimental process?
Q6: Were students in the experimental group informed about result expectations of the experimental process?
Q7: Is there any saying to students whether the performance of students during the experimental process will affect them in their course achievement or not?

The artifacts, which were given in Figure 1, arise according to the seven questions and answers (Yes or No) given for the questions of the experimental process.

Figure 1 shows the articfacts which may arise during the experimental process according to answers given to questions ("Yes" or "No" answers, rather than report the positive or negative, show direction of the experimental process that can be affected by which artifacts).

Data collection instruments

For data collection, students' achievement were not only evaluated with the achievement test, but also with their performance during the whole term used for comparing their improvements. At the beginning of comparing their initial cognitive situation, their scores obtained from OSS (student selection exam) central exam were taken into consideration. To compare their post achievement at the end of the experiment, a test that consist of ten questions and four performance studies which they present during the entire term at the "Planning and evaluation in education" course was used. All students in the five groups were divided into sub-groups to perform their performance studies. Contents of the performance studies are listed as follows:

1a) Performing one target behavior according to each cognitive levels of the Bloom's taxonomy and one multiple-chosen and one written-question which is coherent with that behavior. They have to prepare these questions according to their teacher professions.
1b) Evaluating the other groups' target behaviors and questions based on quality and accuracy of the questions according to Blooms' taxonomy.
2a) Performing a lesson plan for two hours related to their profession (Every sub-group will be prepare a lesson plan).
2b) Performing an annual plan which contains the 2010 to 2011 academic year's content related to their profession (only one annual plan will be prepared by the whole class).
3) Performing an achievement test related to their profession and

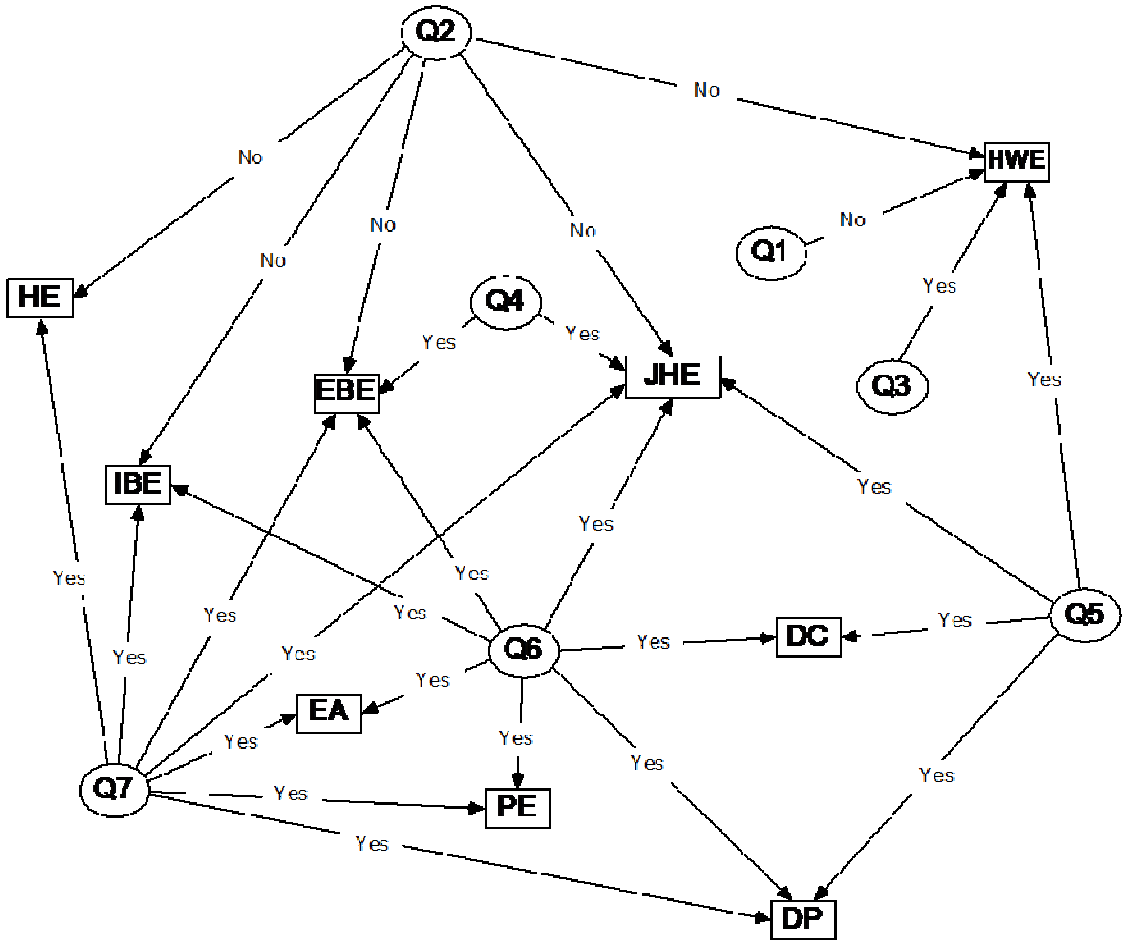

Figure 1. Artifacts likely to be encountered according to the answers given to the questions.

containing at least 25 multiple-chosen questions. They have to ensure reliability and validity of the test (only one achievement test will be prepared by the whole class).
4) Using teacher abilities during 2 h. At the first hour, they have to acquaint five different techniques selected in thirty-five techniques which are used by educators (question-answer, argument, cooperative learning, six thinking hats, simulation, computer aided instruction, etc), while at the second hour, they have to carry out a course related to their profession, but during the course they have to use at least one technique which they have presented at the first hour.

All performances have been presented by the whole participants during one term and at the end of the term, an achievement test consisting of ten written-questions was applied to all the students. The score they got from four performances and their last achievement test was taken into consideration to evaluate the total achievement which they achieved during the term.

Application process

In this study, all artifacts' detailed investigation could not be specified. The main reason of examining the John Henry effect in this study specifically is the lack of researches which examined the existence of this effect. Although, the John Henry effect is shown as the reason of non-significant differences in some studies, there is very few studies which directly examine the existence of this effect. For that reason, influencing the John Henry effect only, within these artifacts, has been taken into account to examine student achievement. The minimal impact of other mentioned artifacts on students at the process of questioning, in which questions will be kept in mind, the following paths were followed.
The experiment process was conducted during the entire period, not covered for partial process (Q1: Yes). The study tried to organize a circumstance in which they could not realize that they are in a different process. Thus, to reveal that some artifacts may arise when the case of "No" are told is prevented. With the same logic, the course was conducted by researchers throughout the entire period, and between each group of five students, researches did not take place before any interaction at the start of the course. Both researchers and students were met for the first time during this course (Q2: Yes). Subjects are not previously known by researchers and, for the subjects, the researcher do not have any prejudgments on students. Likewise, students did not recognize the researcher previously, although the researcher was an official member of the faculty, so it was assumed that the students did not have the inspiration of participating in an experimental environment. However, a direct experimental method has not been tested in the present study. In the course, throughout the semester, different

teaching methods and techniques were used, but this process was identical to each group and the possible effects on each group was in an equal level (Q3: No, Q4: No and Q5: No).

If we already give direct answers to Q6 and Q7, we must say “No” for both questions as the present study is not experimental, but in terms of the logic of the study, it will be examined if the John Henry effect will occur or not in this study, and if both the history teaching students (in the social group) and the biology teaching students (in the science group) were informed about the expectations of success at the end of the study using motivation text. In short, partially "Yes" and "No" can be said for Q6 in this study. The students already have the intention of creating a sense of competition in their subliminal with the motivation text, while the John Henry effect that will take place here, already constitute the main objective of this study, which is whether the examination of the artifact will be appear or not. For Q7, an explanation by the researcher is as follows: the students taught were in a formal course in which they have to be successful for graduation. In this context, students are aware that their achievement given from both 10-written questions and their performance studies will determine their success in the course, but they are not aware that this process is an experimental environment, and their success in this experimental process will determine their term’s achievements. In this context, it is thought that the results of artifacts which stemmed from Q7 are disposed of in this experiment process.

The study mentioned earlier that, the John Henry effect is only an artifact that it wants to occur in the study process, and as such, it tried to minimize and reduce the effects of other artifacts, except that of John Henry, which affect all groups equally in this process.

The experiment and analysis of data

Actually, there is no comparison in the experimental design in this study. All courses carried out during the term are performed similarly for all the five groups. There is only one difference in detecting how biased effects, especially “John Henry effect”, influence the students achievement, in giving some motivation (does the motivation aim makes them feel that they are in the control group?) to history and biology student teachers at the beginning of the term. In the Turkish education system, a central exam (OSS) is used by high school students to locate a university and this location depend on the scores they get in OSS. According to the scores they get, the OSS (in the social group) and the history teaching students have lesser scores than the Turkish language and literacy teaching students when they have already located a university, while the biology teaching students (in the science group), as well as the history teaching students, have lesser scores than the physics and mathematics teaching students. For motivating the history and biology teaching students, the “motivation text” stated below was presented verbally to them at the beginning of the term and it was presented only once.

“Dear friends, you will learn the ‘two-tailed normal distribution (curve)’ in this lesson. The two-tailed test is a statistical test used in inference, in which a given statistical hypothesis, H_0 (null hypothesis), will be rejected when the value of the statistic is either sufficiently small or large. The test is named after the "tail" of the data under the far left and far right of a bell-shaped normal data distribution or bell curve. However, the terminology is extended to tests relating to distributions other than normal.

In general, a test is called two-sided or two-tailed if the null hypothesis is rejected for values of the test statistic falling into the tail of its sampling distribution, while it is called one-sided or one-tailed if the null hypothesis is rejected only for values of the test statistic falling into one specified tail of its sampling distribution. For example, when we look at the ‘normal distribution graphic (Figure 2)’, there is a ‘critical region’ used to recognize how the data

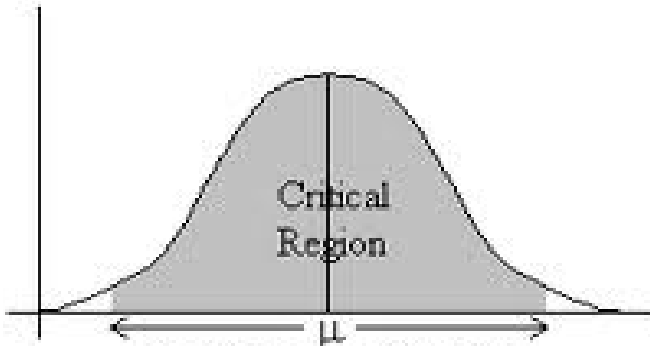

Figure 2. Normal distribution (curve).

are distributed normally. If we interpret the graphic according to the achievement, it will be easily understood that the graphic already shows the scores which one can get from an examination. When the scores obtained from the test results were analyzed, the district named ‘critical region’ showed boldly the stacking points field of the students (95%). About 5% of the class can get too high and too low scores, but 95% of the class in this part of the region will be a measure of the class achievement. This graphic shows the similarities in shape for each class, but values in the class according to the achievement ‘mode’ value can prolapse to the left or right. Let us explain this situation as follows. Turkish language and literature teaching students are more succesfull than students taught history according to the OSS exam which represents the students’ achievement. This means that if the students and other groups are tested with the same quiz, ‘mode’ value will be higher than the students’ class, even if the graph will emerge in similar shape. In a similar way, at the end of this term, we will also see that although some of the students will take very high and low points, while the students’ graphic will look like this normal situation and it will be valid for the other groups. The only difference is that they will show more success than the control students and the ‘mode’ value. In other words, the average of their class achievement, will be higher than that of the control (here, the speeches of students taught History were given and mentioned as the other group is the students taught Turkish Language and Literature)”.

The difference on speech for the other control group, Biology teaching students, is only for physics and mathematics teaching students which are mentioned as the other group.

The aforementioned motivation text presented only history and biology teaching students, while other processes are the same with other group students. The main aim of the text is to revive a competition upon their deep subliminal to create a side effect (John Henry effect). To see how the John Henry effect influences the students’ achievement, students’ scores which they got in OSS were assumed and used as pre-test to compare students’ initial achievements and then, the scores which they got both in the written exam and the four performance studies were assumed and used as post-test to compare the last situation. When the data were analysed, history teaching students’ achievement was compared with Turkish language and literacy teaching students’ achievement, while biology teaching students’ achievement was compared with compared with physics and mathematics teaching students.

FINDINGS AND RESULTS

The data collected has been presented here in

Table 2. Analysis of the pre-test results between the social groups.

		N	$\overline{X}$	Std.Dev.	df	t	P
OSS score	HT	21	314.292	1.597	48	-13.688	0.001*
	TLLT	29	327.664	4.255			

*p<,001; HT: History teaching; TLLT: Turkish language and literacy teaching.

Table 3. Analysis of the pre-test results between the science groups.

OSS		Sum of squares	df	Mean square	F	Sig.
Science group	Between groups	32012.930	2	16006.465	818.758	0.001*
	Within groups	1231.631	63	19.550		
	Total	33244.560	65			

*<0.001.

Table 4. Analysis of the post hoc (Scheffe) multiple comparison test results between science groups.

	(I) Department	(J) Department	Mean difference (I-J)	Std. error	Sig.	95% Confidence ınterval	
						Lower bound	Upper bound
Science group	Biology	Mathematics	-51.136540*	1.405246	0.000	-54.65965	-47.61343
		Physics	-3.295972*	1.280833	0.043	-6.50717	-0.08478
	Mathematics	Biology	51.136540*	1.405246	0.000	47.61343	54.65965
		Physics	47.840568*	1.355731	0.000	44.44160	51.23954
	Physics	Biology	3.295972*	1.280833	0.043	0.08478	6.50717
		Mathematics	-47.840568*	1.355731	0.000	-51.23954	-44.44160

*The mean difference is significant at the 0.05 level.

accordance with the aim of this study. To determine whether there are differences in the social groups [history students (HT) and Turkish language and literacy teaching students (TLLT)], in regard to achievements, the data were subjected to t-test analysis. The result of the analysis is shown in Table 2.

According to the data in Table 2, there is significant difference between two students groups regarding OSS scores. As a result, it can be concluded that, based on the OSS score, the initial level of the TLLT students was higher than the HT students. To determine whether there are differences between the science groups based on the OSS score, the collected data were subjected to one-way anova analysis. However, the results of the analysis are shown in Table 3.

According to the data in Table 3, there is a significant difference between three students groups regarding OSS scores. Determining which groups significantly differ from others, Scheffe test was performed and the results of the Scheffe test have been shown in Table 4.

According to the data in Table 4, there are significant differences between three students groups regarding OSS scores. As a result, it can be concluded that, based on the OSS score, the initial level of the Biology students is lower than the physics and mathematics students.

When Tables 2, 3 and 4 were evaluated, history and biology teaching students' initial achievement was significantly lower in the related groups. They are also selected for seeing the John Henry effect intentionally if the motivation text is revised again.

After the experimental process, to determine whether there are any differences between social groups, based on the John Henry effect's results upon the final achievement, the collected data were subjected to t-test analysis. However, the results of the analysis of the final tests have been shown in Table 5.

According to the data in Table 5, there is no significant difference between two students groups regarding post-test scores. As a result, it can be concluded that, based on the post-test score, the last achievement level of the social group students' achievement level is equal to each one. When Tables 2 and 5 are examined together, it is obvious that while history teaching students have significantly lower achievement level as per Turkish

Table 5. Analysis of the post-test results between social groups.

		N	$\overline{X}$	Std. Dev.	df	t	P
Final achievement (post-test)	HT	21	73.76	8.949	48	-0.206	0.838
	TLLT	29	74.24	7.496			

HT:History teaching; TLLT: Turkish language and literacy teaching.

Table 6. Analysis of the post-test results between science groups.

Post-test		Sum of squares	df	Mean square	F	Sig.
Science group	Between groups	3443.959	2	1721.979	33.524	0.001*
	Within groups	3235.996	63	51.365		
	Total	6679.955	65			

*<0.001

Table 7. Analysis of the post hoc (Scheffe) multiple comparison test results between science groups.

	(I) Department	(J) Department	Mean difference (I-J)	Std. error	Sig.	95% confidence interval	
						Lower bound	Upper bound
Science group	Biology	Mathematics	12.278*	2.278	0.000	-54.65965	-47.61343
		Physics	16.654*	2.076	0.000	-6.50717	-0.08478
	Mathematics	Biology	-12.278*	2.278	0.000	47.61343	54.65965
		Physics	4.376	2.198	0.146	44.44160	51.23954
	Physics	Biology	-16.654*	2.076	0.000	0.08478	6.50717
		Mathematics	-4.376	2.198	0.146	-51.23954	-44.44160

*The mean difference is significant at the 0.05 level.

language and literacy teaching students' achievement level, at the end of the term, they finished with the same scores.

To determine whether there are any differences between science groups based on the post-test score, the collected data were subjected to one-way anova analysis and the results were shown in Table 6.

According to the data in Table 6, there is significant difference between three students groups regarding post-test scores. Determining which groups significantly differ from others, scheffe test was performed and the results were shown in Table 7.

According to the data in Table 4, there is significant difference between three students groups regarding post-test scores. As a result, it can be concluded that, based on the post-test score, the last achievement level of the Biology students is significantly higher than the physics and mathematics students. When Table 3, 4, 6 and 7 were examined together, it was obvious that while biology teaching students have significantly lower achievement level as per physics and mathematics teaching students' achievement level, at the end of the term, they finished with higher achievement. Figure 3 shows the comparison of the pre-test and post-test results as bar graphic presentations. Likewise, in the tables, we can clearly see differences between pre-test and post-test.

DISCUSSION

In this study, artifacts which can influence achievement of the students in educational studies have been defined and the existence and effect of the John Henry effect have been investigated. The John Henry effect refers to when the students or teachers know that they are in the control group (students or control group teachers), they see themselves in a competition against the treatment group and show extraordinary performance to gain more success. This effect usually requires at least one experimental and one control group, but in this study, no

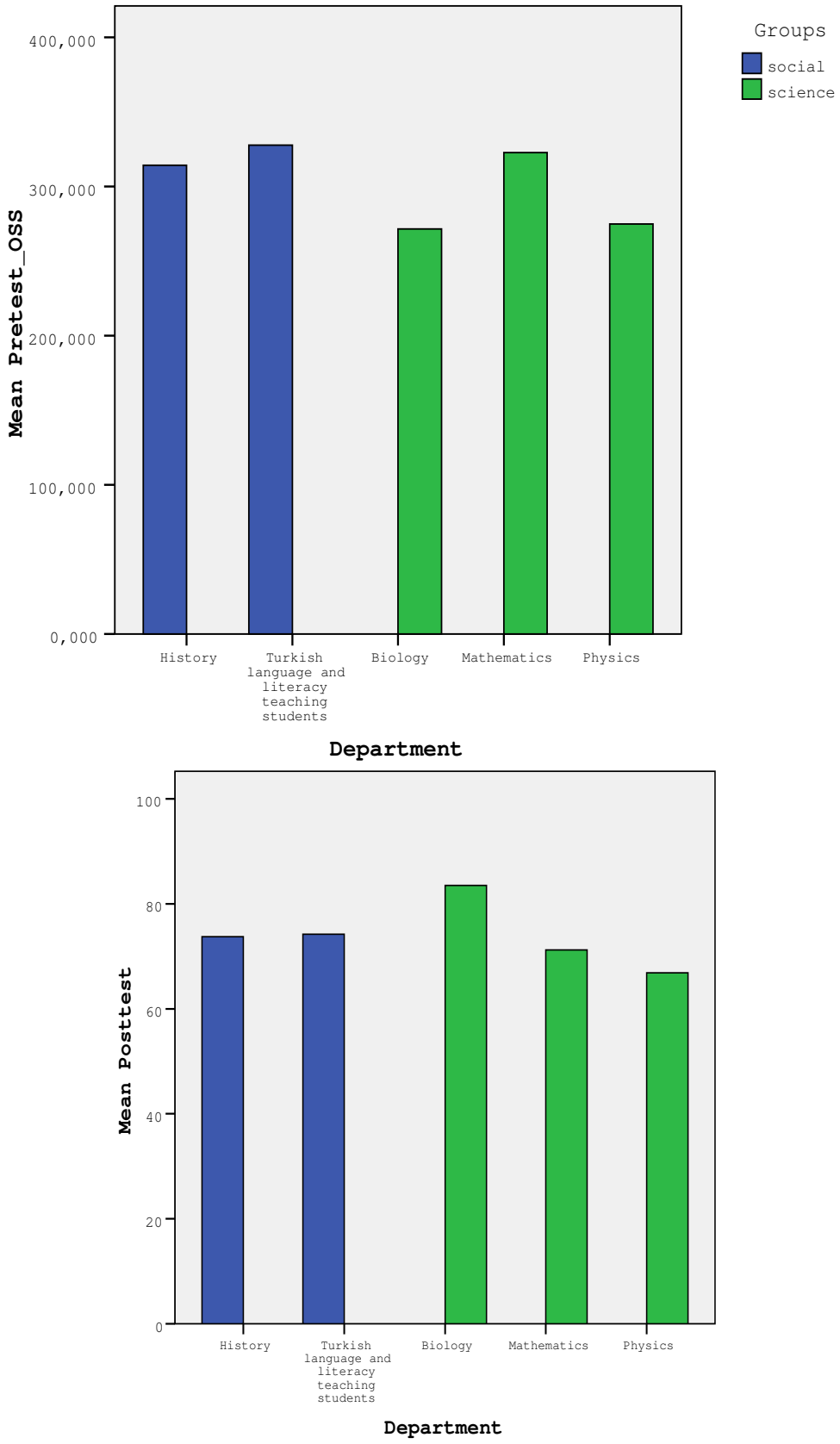

Figure 3. Presentation of the pre and post test results as bar graphic.

experimental or control group is formed. Due to the fact that the scores they got when they located a university are lower than other group members (social and science), history and biology teaching students (known as the control group) were selected and thought as the ideal group to examine the influence of the John Henry effect on them (Tables 2, 3 and 4). It was investigated that when a sense of competition has been created in their subliminal with the help of the “motivation text”, there will be an achievement increase from other group members.

Although the teaching process of study is as sameas for all five classes, at the end of the term, achievement of the history and biology teaching students (known as the control group) has shown more achievement increase than other members of their own groups. The artifacts that were mentioned in this study and which the researcher tried to control carefully, increase the control groups’ achievements, which were concluded in the result of the John Henry effect. Even though it could be thought that EBE and IBE biases may occur, it can still be sincerely said that the “motivation text” is the only difference, and as such, there is no situation which can influence the result of the study process.

The Hawthorn effect is the most effective artifact which may occur in this study process due to the fact that the changing learning environment stems from using various methods and techniques (constructivist learning, problem based learning, computer assisted learning, mastery learning, etc.). However, when courses were proceeded by the researcher and applied to students, the entire process was performed in the same manner to all classes. For that reason, it is assumed that the Hawthorn effect which could occur in this study has affected all classes in the same way and it has contolled it if it has to. In addition to the answer given to the question mentioned in the application process, it can be said that there is no answer that came out of other artifacts which did not come out of the four artifacts previously mentioned.

If we summarize the entire work, it can be seen that the main purpose of this study is to investigate the effects of artifacts on the educational experimental process and under which circumtances they may arise. Furthermore, the existence of the John Henry effect has been investigated and it is concluded that the increase of the control groups’ achievement has stemmed from this effect. Finally, designing the experimental process that interprets the significancy of the statistical finding obtained from the educational studies was criticized based on how it was controlled accurately.

CONCLUSION AND RECOMMENDATIONS

"John Henry" effect, described by Heinich (1970) to explain that there are no significant differences in the findings of his study, was firstly investigated by Saretsky (1972b, 1975) and discussed at some length by Cook and Campbell (1976). According to the educational viewpoint: If subjects in a control group find out that they are in a competition with those in an experimental treatment, they tend to work harder. When this occurs, differences between control and treatment groups are decreased, thereby minimizing the perceived treatment effect. Martella et al. (1999) refer to this effect as a compensatory rivalry by the control group and a compensatory equalization of treatments. Levenstein (1970) has recently attempted to handle this problem with an anti-John Henry design, in which a general screening service was offered and, among those who accepted, the experimental and control groups were established by random assignment. The control group was unaware of the selection, but continued to receive the screening service, in that the design appears to be proving effective. An important issue is to distinguish between a generalized compensatory rivalry and one in which the rivals have extended contact with the experimental group. The main reason for examining the John Henry effect in this study is, specifically, the lack of researches on the existence of this effect. Although the John Henry effect is shown as the reason for the non-significant differences in some studies, very few studies directly examine the existence of this effect. As such, the study only observed the Saretsky’s (1975) study which directly examined the existence of this effect and which showed similarities with the results of the present study. Saretsky concluded that the John Henry effect is likely to occur when an innovation is introduced in such a manner as to be perceived as threatening to jobs, status, salary, or traditional work patterns. Saretsky provided convincing evidence that the John Henry effect resulted in a marked increase in the control group’s achievement in labs, when those labs were compared with labs in which performance-contracting was employed. He obtained data on performance of the control subjects for two years prior to the experimental year. These data showed that, during the experimental year, the control group gains in laboratory skills, as measured by standardized tests, were much higher than its gains in the two preceding years. Due to the fact that performance-contracting was very threatening to managers, it became obvious that the managers made a very strong effort during the year of the experiment.

In this study, unlike Saretsky’s work, students were used as participants and their achievement was measured depending on the process assesment and not with the standardized tests. Moreover, at the end of the Saretsky’s experimental work, it was reported that some decreases were observed on achievement of the control group in a shorter time after finishing the experimental process. That the present study is based on an evaluation process, repeating the study or re-measuring the students’ achievement after a particular time period is not possible, although the control group's achievement

could not be examined after the end of the retention times. Moreover, in terms of the control group's acvhievement, this study showed similar results with Saretsky's work.

Another main viewpoint of this study is: Are there methodological faults that we are overlooking? There is this thought that the experiment process was not controlled well. Maybe, because we misunderstand the soul of education and for that reason we assume that students will react to our actions the way we expect them to, but we overlook the fact that there are some psychological artifacts, and as such, they affected our experimental design undesirably. In providing solution to these problems, we must think of better research designs. The major question which is not resolved in a higher educational outcome research is: what happens to the student? We should know how education affects his or her behavior before we decide that there is no effect. Is constructivist learning, problem based learning, computer assisted learning or another instruction method we used in our experimental design different from that of the curricula? There is no easy solution to the problem. The issue is that we want to make sure that student mastery reflected in knowledge gains and performance changes, and we know that this outcome is only reached if students study (Cate, 2001).

Cate (2001) suggest that if we adopt more elaborate designs for the educational outcome research, including analyses of the effects on students, we might be able to find effects that otherwise remain hidden. Such a design could be envisioned as some sort of three stage path design, with (A) an independent variable (the experimental teaching condition, with or without student characteristics) affecting (B) an intermediate variable (what happens to the students?), which in turn affects (C) a dependent variable (what are the outcome effects?). The intermediate variable can first be considered a dependent and then an independent variable. Two separate hypotheses can then be tested. The what-happens-to-the-students question requires hypotheses of student behavior. Importantly, we will have to predict whether a supposedly 'better' educational method should lead to more or less student activity. As long as we have no hypothesis of this intermediate effect, hypothesizing about the outcome effect seems premature. We might argue that, in an ideal world, all educational designs and researches would be hungry for reliable information about what works to improve learning, quick to eschew untested fads and keen to work with researchers to identify and address pressing research questions. In the real world, we would not be eager to collaborate in developing a magical solution for education. Is there a magical solution to solve our problems? Are the new methods used in our experimental researches really working? Or if there are no significant differences, is it not working? Can we see clearly the main outcome reason for our experimental designs? Can we really isolate all the unobserved effects which can influence the study?

This study has no intentions of reviling at experimental researches and new research designs. For that reason, it did not illustrate the findings of any researcher, whether they are "significant or non-significant differences"; but if the goal of the educational experimental designs is to provide recommendations on methods of education with high likelihood of success, we should, minimally, be able to predict what happens to students. Perhaps we should be more modest in our outcome expectations. What do we expect from effective education, that is, does teaching really serves its purpose? A better world? Better students? Better teachers? Better student grades? Or should we start to expect better learning behaviors from students? If the purpose of the educational research is to create and disseminate knowledge and tools that can be used to improve learning, all educational research ultimately must be judged in terms of its success in creating knowledge or tools that can be used to improve learning. The improvement of learning is the objective that should drive all educational research. Research into teacher practices, curriculum materials, managing educational institutions, teacher professional development, assessment and reporting practices, successful transition from school to work has, as its ultimate purpose, the improvement of learning. Educational researchers are encouraged to design their research to address significant issues or questions in education that will benefit students, teachers, schools, colleges and others involved in the educational practice or policy. Methodologies are chosen to match the requirements of the research problem. After the ones chosen to match the requirements of the research problem, we need better ties between theory and practice. Perhaps a key to improving the usefulness of the much educational research would be a sharper focus on the research questions being addressed. Many research studies in education proceed without an explicit specification of the questions they are attempting to answer. Instead, they are conducted as data gathering exercises in some area of educational activity, presumably in the hope that the data, when carefully analyzed, will yield useful insights, and if sufficient data are collected, they frequently do; but too often the reported 'findings' of this form of research are one or two of the most interesting bits of information caught in the research net.

For the main purpose of the present study, the most important point to consider (after the experimental design) to optimize our goals is, how we can terminate the artifacts which may ocur in experiments or how we can minimize their effect even if they are terminated, or how we reverse their effect for our benefits. Isolating the control and treatment group from each other is the only one way to eliminate the impact of the John Henry effect on the control group. Otherwise, recognizing the experimental process by the control group students' or their teachers can reduce validity of the data gotten. In

this study, although somewhat unusual, the John Henry effect was used as the only variable to increase student achievement, this effect can not be used in favor of our benefit. Already, Saretsky (1972b, 1975) states in his studies that this effect increases the control group students' achievement only during the competition and this increase is not permanent, due to the fact that retention of this achievement reduces after a period. Briefly, creating the control and treatment groups in different schools is the best way to eliminate the John Henry's effect.

Howthorn effect is another important confounding artifact that occurred in educational studies. Originally, Hawthorne effect research was a series of studies on the productivity of workers manipulated by various conditions (pay, light levels, rest breaks, etc.), but each change resulted on average over time in productivity rising, including a return to the original conditions, eventually. A proposal can be suggested when this issue is examined from the educational perspective. The Hawthorn effect refers to a situation, where if we change anything in the enviroment, productivity rises. We can turn this stituation in our favor by changing the methods and techniques used in the educational and instructional environments consistently. Although changing something in the enviroments raises the productivity, some researchers reported that after a few times, the amount of production turns back to the old situation (Mayo, 1933; Roethlisberger and Dickson, 1939; Gillespie, 1991). For that reason, we should create continuous dynamics and new methods and techniques to build a productive educational environment instead of stable enviroments. To achieve this: (a) the curriculum has to be recovered from stability to a dynamic structure and (b) teachers who put this curricullum into practice need to be trained both in their educational life and in their profession with in-service training to implement the mentioned dynamic enviroments. For this, the educational institutions interact closely with the government and all necessary funds should be set to contruct the infrastructures for the creation.

Although not of central importance here, the huge importance in the educational research in general is the issue of teacher effects. When one of the variables was the teacher, the effect of different teachers was always bigger than the effect of different treatments (usually what was meant to be was studied). Basically, teachers have a huge effect, but we did not understand it at all. To prevent this bias in our experiment, we have to use the same teacher both in the control and treatment group as questioned in my research (Q2). So given the importance of teacher effects, what is the evidence? Rosenthal and Jacobson (1992) also mention briefly that the research showed that 10 secs of video without the sound of a teacher allows students to predict the ratings they will get as a teacher. Similarly, hearing the sound without vision and content (rhythm and tone of the voice only) were enough too. This is a powerful evidence that teachers differ in ways they cannot easily or normally control, but which are very quickly perceptible, and which at least in students' minds, determine their value as a teacher [Marsh's (1987) work shows that student ratings of teachers do relate to their learning outcomes]. Subsequent research done by Chaiken and Derlega (1974) involved using videotape to capture teachers' interaction with students that had been identified as bright students. These interactions revealed that teachers smile and make more eye contact with bright students, while other students are treated in a generalized standard manner. As a result, those students that their teachers have higher expectations of them generally do better, which proves the correlation between expectations and performance.

Solving this problem derived from the teacher can be solved by applying the same manner of representation on all students, and the student groups, used in the experimental studies, should be taught by the same teacher. Similar problems can occur when students in the schools, used for experimental studies, were taught by different instructors. When the students realize that they are in an experimental environment, they show a trend towards the experimental expectations. This can occur not only as a result of these experiments, but also as the expectations of the experimenter in the process. When this occurs, students have the willingness to participate more actively in performing the expectations of the experimenter and the experimental process, so that it can cause an unpredicted achievement increase on experiment results. To solve these mentioned problems, "application schools" should be opened within each university, and researchers should be formal instructors in these shools. In this way, the HWE, HE, DP, DC and EA effects that may ocur when the researcher and formal teacher is different, can be mininimized.

EBE and IBE is the another bias which is not mentioned here to reduce their effects on educational studies. These effects emphasize the possible occurence that is revealed when planning and conducting experiments and the expectations of the researchers in the experimental environment. As it happens like the situation previously mentioned, some unacceptable effects which can sharpen the students may occur as a result of the direction of the experimenters and investigators' conscious or unconscious expectations. When this happened, results of the data analysis may show parallelism with expectancies of the experiment. The only way to avoid that kind of situation is that the experimenter and investigator should consider this kind of situation and ensure that the experimental process is free from such errors.

After all those explanations, we know that all the artifacts can have important and unexpected effects. So we cannot trust results that do not at least try to control them. Currently, we do not understand how any of these

effects work. This could probably be done, but would require some concentrated research, for example, on uncovering how expectancies are communicated unconsciously or anyway implicitly, and what expectancies are in fact generated. Moreover, if we want to know that our experiment results are affected by those artifacts or not, we should interview students at the end of the experimental process to learn students' awareness status in the study. We should support our results both quantitatively and qualitatively to ensure the validity of our design. If we do only quantitative analysis to recognize our experimental process and its effectiveness, we probably can encounter unexpected outcomes which stemmed from those confounding artifacts.

Finally, we have to construct a warm enviroment for the benefit of students to achieve more educational goals. We should not look at students as subjects, but should prepare students so that they could be ready for a better learning enviroment. Also, we should look at students from the psychological point of view too. We should not look at them as hamsters which are used only as subject. Besides the new instructional methods, we should motivate the students to believe that they can achieve other methods. However, in education learning depends almost entirely on the learner's actions, so if the learner believes they cannot learn, they are just as unlikely to learn as a walker is to be found at the top of a mountain which he believed he could not climb. In education, if a student does not believe that he can improve at something, then he will not try, but an experiment might make him change his assumption and start making an effort to learn. Conclusively, it should be noted "students are our goal and not subject".

REFERENCES

Barber TX (1973). Pitfalls in research: Nine investigator and experimenter effects. In R. Travers (Ed.). Second handbook of research on teaching. Rand Mcnally and Co. Chicago, Illinois.

Campbell DT, Stanley JC (1963). Experimental and quasi-experimental design for research on teaching. Rand McNallyand Co. Chicago, Illinois.

Cate OT (2001). What happens to the student? The neglected variable in educational outcome research. Adv. Health Sci. Educ., 6: 81-88.

Chaikin AL, Derlega VJ (1974). Variables affecting the appropriateness of self disclosure. J. Consult. Clin. Psychol., 42: 588-593.

Cook DL (1967). The impact of the Hawthorn Effects in experimental designs in educational research. United States Office of Education, Cooperative Research Project, Washington, D.C., No. 1757.

Cook TC, Campbell DT (1976). The design and conduct of quasi-experiments and true experiments in field settings. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology. Chicago: Rand McNally.

Cottrell NB (1972). Social facilitation. In C. McClintock (ed.), Experimental Social Psychology, New York: Holt, Rinehart & Winston, pp. 185-236.

Dubin R, Taveggia TC (1968). The Teaching Learning Paradox. A Comparative Analysis of College Teaching Methods. Eugene, OR: Center of Advanced Study of Educational Administration, University of Oregon.

Gay LR (1992). Educational research (4th Ed.), New York: Merrill.

Gillespie R (1991). Manufacturing knowledge: A history of the Hawthorne experiments (Cambridge : Cambridge University Press).

Havelock RG (1969). Planning for innovation through dissemination and utilization of knowledge. Center for Research on Utilization of Scientific Knowledge, University of Michigan, Ann Arbor.

Heinich R (1970). Technology and the management of instruction. Department of Audio-Visual Instruction, Inc. Washington, D.C.: Associations for Educational Communications and Technology.

Isaac S, Michael WB (1981). Handbook in research and evaluation (2nd ed.). San Diego, CA: EdITS.

Levenstein P (1970). Cognitive growth in preschoolers through verbal interaction with mothers. Am. J. Orthopsychiatr., 40: 426-432.

Marsh HW (1987) Student's evaluations of university teaching: Research findings, methodological issues, and directions for future research. Int. J. Educ. Res., 11(3): 253-388.

Martella RC, Nelson R, Marchand-Martella NE (1999). Research methods: Learning to become a critical research consumer. Boston: Allyn & Bacon.

Mayo E (1933). The human problems of an industrial civilization (New York: MacMillan) Ch. 3.

McMillan JH (2004). Educational research: Fundamentals for the consumer (4th ed.). Boston: Person Education.

Medley DM, Mitzel HE (1963). Measuring classroom behavior by systematic observation. In N.L. Gage (Ed.), Handbook of Research on Teaching. Rand McNally and Co. Chicago.

Norman GR, Shannon S (1998). Effectiveness of instruction in evidence based medicine: A critical appraisal. Can. Med. Assoc. J., 158: 177–181.

Orne MT (1962). On the social psychology of the psychological experiment: with particular reference to demand characteristics and their implications. Am. Psychol., 17: 776-793.

Reicken H (1962). A program for research or experiments in social psychology, Decisions, Values and Groups. Washburne NF(Ed.) Pergamon Press, New York.

Roethlisberger FJ, Dickson WJ (1939). Management and the worker (Cambridge, Mass.: Harvard University Press).

Rosenberg MJ (1969). The conditions and consequences of evaluation and apprehension. Artifacts in Behavioral Research, Rosenthal, R. and Rosnow, R. (Eds.) Academic Press, New York, pp. 79-114.

Rosenthal R, Jacobson L (1992). Pygmalion in the classroom: Teacher expectation and pupils' intellectual development. Irvington publishers: New York.

Rosenthal R (1963). On the social psychology of the psychological experiment: The experimenter's hypothesis as unintended determinant of experimental results. Am. Sci., 51: 268-283.

Rozenzweig P (2007). Halo effect. Free Press, Simon & Schuster Inc. New York, NY 10020.

Russell TL (2000). The No Significant Difference Phenomenon. North Carolina State University.

Saretsky G (1972a). The consequences of an innovation as determinants of control group behavior: an exploratory study. Indiana University, Mimeo.

Saretsky G (1972b). The OEO P.C. experiment and the John Henry effect. Phi Delta Kappan, 53: 579-581.

Saretsky G (1975). The John Henry Effect: Potential confounder of experimental vs. control group approaches to the evaluation of educational innovations. The Annual Meeting of the American Educational Research Association, Washington, D.C., USA. (Eric document, ED: 106309).

Taverner D, Dodding CJ, White JM (2000). Comparison of methods for teaching clinical skills in assessing and managing drug-seeking patients. Med. Educ., 34(4): 285–291.

Webb EJ, Campbell DT, Schwartz RD, Sechrest L (1966). Unobtrusive measures: Nonreactive research in the social sciences. Rand McNally & Company, Chicago.

Wiersma W (2000). Research Methods in Education (7th ed.). Boston: Allyn & Bacon.