ISSN 1798-4769
Journal of Language Teaching and Research, Vol. 4, No. 5, pp. 1079-1084, September 2013

doi:10.4304/jltr.4.5.1079-1084

A Review and Analysis of the Article 'An Evaluation of the Testing Effect with Third Grade Students'

Meisam Rahimi
Department of English Language and Literature, University of Isfahan, Iran

Samira Ghanbaran
Department of English Language and Literature, University of Isfahan, Iran

Seyed Mahmood Kazemi
Department of English Language and Literature, University of Isfahan, Iran

***Abstract*—This paper reviews and analysis the article in four stages. In the first stage, major approaches to educational research are briefly reviewed and the selected article is classified in these terms and the classification is justified. In the second stage, a concise synopsis of the aims, research design and major findings of the article are presented. In the third stage, the main strengths of the selected research as a contribution to policy, practice, knowledge or understanding are explained. And in the final stage, some ways in which the reported research in the selected article could have been improved are suggested, and any difficulties which might have been encountered in following these recommendations will be analysed.**

***Index Terms*—recall, development, education, memory, testing effect, reading, comprehension**

I. Introduction

This article[1] was written by Chandra L. Brojde[2] and Barbara W. Wise[3] and published as part of the proceedings of the Cognitive Science Conference[4] in 2008 in Washington, DC.

Testing effect, according to Verkoeijen, Bouwmeester, and Camp (2012), is the robust finding that "taking one or more intervening tests after an initial encoding (study) episode produces better retention of the to-be remembered material than does restudying the same material for an equivalent amount of time."

According to Einstein, Mullet, and Harrison (2012), in recent years, psychologists have reached remarkable results indicating that learning being accompanied by testing is significantly beneficial to memory so that the benefits gained through testing outweigh those obtained through a comparable amount of time spent for additional study. It is interesting to know that the evidence shows that college students are not aware of the benefits of introducing testing into their learning. A study by Karpicke, Butler, and Roediger (2009) shows that only a small proportion of students (11%) considered self-testing as a study strategy and merely 1% of them reported that as one of their best learning strategies. On the other hand, 84% of students considered repeated reading a study strategy and almost half of them (55%) considered it to be their best strategy. The testing can be performed by students themselves as self-testing or by language educators as part of the teaching/learning process.

The article under review and analysis in this paper aims at evaluating the testing effect with young children (third grade students) where testing is performed by language educators as a means of further learning. This paper tends to look at this article in five stages. In the first stage, major approaches to educational research are briefly reviewed and the selected article is classified in these terms and the classification is justified. In the second stage, a concise synopsis of the aims, research design and major findings of the article are presented. In the third stage, the main strengths of the selected research as a contribution to policy, practice, knowledge or understanding are explained. In the fourth stage, an evaluation of how significant aspects of theory, research design, methods or researcher values influenced the analysis and conclusions of the article is presented. And in the final stage, some ways in which the reported research in the selected article could have been improved are suggested, and any difficulties which might have been encountered in following these recommendations will be analysed.

---

[1] The URL of the paper is as follows: csjarchive.cogsci.rpi.edu/proceedings/2008/pdfs/p1362.pdf

[2] Chandra L. Brojde (chandrab@colorado.edu) Department of Psychology, CB 345 University of Colorado, Boulder, Colorado 80309

[3] Barbara W. Wise (Barbara.wise@colorado.edu) CLEAR Computational Language and Education Research, CB 594 University of Colorado, Boulder, Colorado 80309

[4] The following is the URL of the archive of the Cognitive Science Conference Proceedings: http://csjarchive.cogsci.rpi.edu/Proceedings/index.html

II. Major Approaches to Educational Research

Identifying and then visualizing a research design is helpful in clarifying the suitability of the procedures carried out and will enable us to evaluate the appropriateness of any later data analysis employed (Porte 2002).

Research designs can be classified in endless ways. However, they are usually put into three major categories: experimental, quasi-experimental, and non-experimental (Marczyk, DeMatteo, & Festinger 2005). Several key questions can be asked in order to determine the classification of a particular research design: First, whether random assignment is involved or not. If the answer is positive, the design is considered randomized, or true, experimental. If answer is negative, then a second question must be asked: whether either multiple groups or multiple measurements are used or not. Positive answer leads to a quasi-experimental design and negative answer leads to a non-experimental design (Trochim 2001).

A. True Experimental Designs

In an experimental design, study participants are randomly selected and assigned to experimental and control groups. In such designs, all sources of internal validity are completely controlled. Random numbers tables are often used for assigning research participants to the groups. Random assignment decreases the likelihood that the obtained results are due to extraneous factors or nuisance variables rather than the independent variables.

Generally, three major designs are used for conducting a true experiment: (1) a randomized two-group posttest only or pretest-posttest design, (2) a Solomon four-group design, or (3) a factorial design (Marczyk, DeMatteo, & Festinger 2005).

The notation used in the following description is as follows:

X = experimental manipulation (independent variable)
Y = experimental manipulation (independent variable) other than X
O = observation
R = random assignment
NR = non-random assignment

B. Randomized Two-group Design

This design is composed of two groups or two levels of an independent variable. The primary purpose of this design is to determine whether a particular independent variable causes an effect (causality). Two basic types of this design are used: the posttest only and the pretest-posttest design.

1. Randomized Two-Group Posttest Only Design

The design is as follows:

R ➔ X1 ➔ O
R ➔ X2 ➔ O

Source: Marczyk, DeMatteo & Festinger (2005)

This simple design incorporates all required elements of a true experimental design: (1) random assignment (2) experimental and control groups, and (3) observations following the treatment.

2. Randomized Two-Group Pretest-Posttest Design

This design is typically utilized for randomized experiments:

R ➔ O ➔ X1 ➔ O
R ➔ O ➔ X2 ➔ O

Source: Marczyk, DeMatteo & Festinger (2005)

The pretest included is beneficial in several respects: First, the researcher can ensure that the groups are truly equivalent. Second, it provides information which enables researchers to compare the participants who completed the posttest to those who did not.

3. Solomon Four-Group Design

We can consider this design as a combination of the randomized posttest only and pretest posttest two-group designs, as depicted below:

R ➔ O ➔ X1 ➔ O
R ➔ O ➔ X2 ➔ O
R ➔➔➔➔ X1 ➔ O
R ➔➔➔➔ X2 ➔ O

Source: Marczyk, DeMatteo & Festinger (2005)

The main advantage of this design is that it controls for the possible effects of pretest on posttest results. This design can also be considered as a very basic example of a factorial design.

C. Factorial Design

This design empirically examines the effects of more than one independent variable, both individually and in combination, on the dependent variable.

R ➔ X_1 ➔ Y_1 ➔ O
R ➔ X_1 ➔ Y_2 ➔ O

R ➔ X_2 ➔ Y_1 ➔ O
R ➔ X_2 ➔ Y_2 ➔ O
Source: Marczyk, DeMatteo & Festinger (2005)

This design is beneficial in several respects: First, more than one independent variable can be examined simultaneously. Second, several hypotheses can be tested in a single research study. Finally, the interaction between independent variables can also be examined.

D. Quasi-experimental Designs

Due to the feasibility of random assignment in real-world environments, researchers must often use quasi-experimental designs. Studies based on these designs occur in real-life settings as opposed to a laboratory. Quasi-experimental designs make use of both control and experimental groups; however, subjects are not normally randomly selected nor randomly assigned to these groups. These designs, therefore, do not manage satisfactorily to control for extraneous variables as intervening factors in research outcomes. The major attraction of such designs is the fact that they do not disrupt the research environment (i.e. the school system and/or re-assignment of subjects to other classes) and make use of the available groups.

A variety of quasi-experimental designs are presented which can be divided into two main categories: interrupted time-series designs and nonequivalent comparison-group designs (Cook & Campbell 1979).

E. Nonequivalent Comparison-group Designs

These designs do not employ random assignment. In these designs, groups as similar as possible are selected. Unfortunately, the resulting groups might be nonequivalent. However, careful analysis and cautious interpretations may still lead to some valid conclusions (Graziano & Raulin 2004). Two major types of this design are: 1. Nonequivalent Groups Posttest-Only (Two or More Groups) 2. Nonequivalent Groups Pretest-Posttest (Two or More Groups)

1. Nonequivalent Groups Posttest-Only (Two or More Groups)

In this design the experimental group receives the treatment while the control group does not:
NR ➔ X1 ➔ O
NR ➔ X2 ➔ O
Source: Marczyk, DeMatteo & Festinger (2005)

The results of a study employing this design may be considered largely uninterpretable due to the fact that there is a low probability that the obtained results could be attributed to the intervention.

2. Nonequivalent Groups Pretest-Posttest (Two or More Groups)

In this design, the dependent variable is measured before and after the intervention:
NR ➔ O ➔ X1 ➔ O
NR ➔ O ➔ X2 ➔ O
Source: Marczyk, DeMatteo & Festinger (2005)

This design has two advantages over the previous one: First, the researcher will be more confident that the obtained results are due to the independent variable. Second, the between-group differences can be measured before exposure to the treatment.

F. Interrupted Time-series Designs

This design employs periodic measurements on a group prior to the intervention to establish a stable baseline. This will help the researcher to interpret the impact of the independent variable more accurately. After the intervention, the researcher will make several more periodic measurements.

G. Non-experimental Designs

These designs do not involve any control over the variables and the environments under investigation and consequently they will not be able to rule out extraneous variables as the cause of the observed outcome. Here, three of the most widely used designs in this category are reviewed briefly.

H. Case Studies

In this design, a single person or a few people are examined in-depth. A study based on this design provides an accurate and complete description of the case under investigation.

I. Naturalistic Observation

This design involves observing organisms in their natural settings. The main advantage of this approach is that participants are observed in a natural setting where they do not realize that they are studied.

J. Survey Studies

In this type of study, information about behaviors, attitudes, and opinions of a large number people is obtained through asking questions. Some surveys merely describe what people express as their opinions and activities. Others try to find relationships between the respondents' reported behaviors and opinions and their characteristics. When surveys are used to determine relationships, they are called *correlational studies.*

In the study reviewed, the participants are neither randomly selected nor randomly assigned. Therefore, the design in this study is not true experimental design. On the other hand, there is control over the following variables: 1. testing 2. age and 3. level of proficiency. As a result, the design of the study is not non-experimental. Consequently, the design is quasi-experimental. As other characteristics of the design which make it quasi-experimental are: 1. This study occurs in real-life settings (classroom) as opposed to a laboratory. 2. This study makes use of both control and experimental situations (instead of groups).

III. Aims, Research Design and Major Findings

A. Aims

According to Chandra and Barbara (2008), the testing effect has been investigated comprehensively with regard to undergraduate students through various materials. However, the effect has not been dealt with comprehensively in lower levels of education, specially using educationally related material in real classroom settings.

The reviewed article is a first attempt to deal with this issue from two perspectives: first, whether young children also demonstrate an general testing effect regarding educationally related material, specially whether initial multiple choice tests will lead to better retention of the material or not. Second, whether the overall testing effect will be sensitive to the content (type) of the initial questions.

B. Research Design

The participants of the study were ten female and eight male children with an average age of nine years and one month. Met in their classrooms, children were asked to complete a reading comprehension and a word reading test to make sure that students' reading level was in line with their level of education.

Three stories were selected from the Houston Museum of Science Horizon Plus Science Stories Series. These stories had been written at the third grade level. Three conditions were presumed: 1. A story with fact questions 2. A story with main idea and inference questions 3. A story without any questions used as a control. Considering the three conditions, three versions of each story were prepared and assigned to different conditions in a counterbalanced way. Each story included graphics depicting plot points and was five pages long.

Each student was required to read the three texts, one week apart for three weeks. In the first two conditions and for each story, children required to answer three multiple-choice questions as they read the text and two multiple-choice questions right after reading the text. As for the first condition, all the questions dealt with details of the text. Regarding the questions in the second condition, the first three were inference questions and the last two were about the main ideas. In the third condition, children did not answer any questions after reading the story. Children did not have any access to the stories as they were answering the multiple-choice questions.

As the participants read the stories and answered the multiple choice questions (except for the reading with no question), they gave an oral summary of that story. Children who were not willing to answer questions or those who gave a short answer up to two sentence summaries were required to give more elaborate answers. Having finished their summaries, they immediately answered 10 open-ended questions which asked for short answers of only one or two words. The questions were prepared based on guidelines which were developed for a related study. The questions consisted of five fact questions and five inference or main idea questions. Each question was composed of a stem and four choices, three distractors and one correct answer. The most prominent themes in the texts as well as the content of the multiple-choice questions from the two experimental procedures were used as the basis for question development.

An interactive computer program designed to improve children's reading was utilized to present the stories and multiple-choice questions. The program recorded the children's voice as they read the texts orally. Children could also click on the words and listen to their pronunciation.

During each of the three sessions, children's voice was recorded as they read the story through a headset accompanied with a microphone. In the first session, children first went through the reading comprehension and word reading tests. Moreover, the administrator presented an sample story to make the students familiar with the structure of the texts and the computer environment. Children could listen to the pronunciation of the unfamiliar words by clicking on them.

The transcription of the oral summaries of the students was provided. A 10-point checklist based on similar material to the questions posed at the end of stories was used to analyze the transcriptions. Based on the number of items of the checklist included in the summary, a score was assigned to each summary, ranging from zero to ten. Similarly, children were assigned a score from zero to ten based on the number of questions answered correctly at the end of each story.

C. Major Findings

Regarding summaries, a one-way repeated-measure ANOVA with planned contrasts was conducted. The result showed that there was a significant difference between the condition with questions and conditions without questions, $F(1,15) = 4.54$, $p = 0.05$. Nevertheless, no significant difference was found between inference/main idea question condition and the fact condition, $F(1, 15) = 0.11$, $p = 0.75$

Similarly, a second one-way repeated-measures ANOVA with planned contrasts was conducted for open-ended questions. The result of the analysis indicated that there existed a significant difference between the condition with

multiple-choice questions and conditions without multiple-choice questions, F (1,15) = 5.802, p = 0.03. Moreover, no significant difference was found between inference/main idea question condition and the fact condition, F (1,15) = 0.195, p = 0.67.

Generally, the results clearly show that 1) no testing effect can be found with regard to lower level education students (third-grade in this case), 2) the testing effect is strong in both inference/ main idea questions condition and fact questions condition and 3) the testing effect is strong in both summary or short-answer questions conditions at the final test.

IV. The Main Strengths

As mentioned in the article, the interactive program used for presenting the material had two particular features among others. First, the users had the opportunity to answer the questions more than once as far as they did not get the correct answer. Second, the participant received feedback on their performance in answering the questions. In case that a student answered a questions correctly, the program would provide the feedback and the answer would be fixated in his mind. On the other hand, if they provided the wrong answer, the program would indicate it making them think more deeply. Consequently, these features provided conditions which perhaps led to a more established and elaborative mental representations of the material which in turn decreases the extent of forgetting and consequently would lead to a greater testing effect.

A headset is something which usually comes with a computer. Therefore, using a headset for recording the oral summaries of the children wouldn't distract their attention as much as in the case when they might have used the traditional paper and pencil method. For the reason, using computers for presenting material and conducting tests improved the validity of the test.

The main focus of schools is certainly learning. On the other hand, tests are usually used in these systems mainly for evaluation and to a lesser extent for enhancing learning. So, it can be mentioned as a strength in the design of the study that it deals with real classroom settings with educationally relevant material.

V. Possible Improvements

Two reasons are presented in the article for the fact that the testing effect for inference/main idea questions was not greater than fact questions. There might be another explanation for this phenomenon as follows. As expressed in the article, the stories presented as the material of the experiment encompassed some graphics showing plot points in the story. On the other hand, the final multiple-choice questions asked about the most prominent themes in the story. Consequently, the observed testing effect might have been partly due to the children's visual memory. This problem could be easily eliminated by removing the pictures included.

According to the definition presented in the introduction testing effect is the finding that "taking one or more intervening tests after an initial encoding (study) episode produces better retention of the to-be remembered material than does restudying the same material for an equivalent amount of time." A second issue is that in the reviewed study is that no comparison group is available where the participants restudy the text instead of taking the test. In the reviewed study, the comparison is between a group with an initial test and a group without an initial test, whereas the comparison must be between a group with an initial test and a group with the same amount of time as spent on testing with restudying. This matter could be solved through having the second group which did not receive the test restudy the material in the same amount of time spent on testing in the first group.

According to Chandra and Barbara (2008), research on testing material as a means of improving reading comprehension suggest that the design, format, and content of test questions might influence the comprehension of the material. In this regard, two suggestions can be made to improve the study. First, since multiple-choice questions restrict the premise of the material covered in testing, especially in this case that there were only five questions for each condition, using oral summaries as the initial and final testing would lead to a more valid investigation of the testing effect. Second, using tests during reading and immediately after reading the texts might lead to different results with respect to testing effect. So, two separate studies can be conducted to investigate the two conditions separately. In the first suggestion, the scoring procedure might be problematic. In order to make the task easier, a checklist consisting of various points in the text can be provided according to which the oral summaries will be scored. Obviously, the more elaborate the checklist is, the more valid the final results will be.

The next issue is related to the reliability and validity of the 10 initial and final multiple-choice questions. These questions had not been piloted. So, their reliability and validity could not be established.

The next two issues are related to recording the voice of children as they read the story and the interactive program. As for recording the children, there has been no need to record the children's voice as they read the story. This perhaps acted as a distractor which, in this case, reduced the comprehension of the texts. As there was no need for recording, the researcher could simply eliminate this section of the research procedure. As for the second issue, the interactive program allowed the child to listen to the pronunciation of the unknown words by clicking on them. This again could be a distractor which could reduce the comprehension of the texts.

The next issue is related to the three versions of the stories. As expressed in the article, considering the three conditions, three versions of each story were prepared and assigned to different conditions in a counterbalanced way. Now, the question is why three versions of each story. The researcher could simply assign the three stories to the three conditions in a counterbalanced fashion.

Some more illuminating results might have been obtained if the gender variable was controlled. The participants, as mentioned, were 8 males and 10 females. The testing effect might be different for different genders.

REFERENCES

[1] Cook, T. D. & Campbell, D. T. (1979). Quasi-experimentation: Design and analysis issues for field settings. Chicago: Rand McNally.
[2] Einstein, G. O., Mullet, H. G. & Harrison T. L. (2012). The Testing Effect: Illustrating a Fundamental Concept and Changing Study Strategies. *Teaching of Psychology* 39/3, 190-193.
[3] Graziano, A. M. & Raulin, M. L. (2004). Research methods: A process of inquiry (5th edn.). Boston: Allyn & Bacon.
[4] Karpicke, J. D., Butler, A. C. & Roediger, H. L., III. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory* 17, 471–479.
[5] Marczyk, G., DeMatteo, D. & Festinger, D. (2005). Essentials of Research Design and Methodology. Hoboken, New Jersey: John Wiley & Sons.
[6] Porte, G. K. (2002). Appraising Research in Second Language Learning: A practical approach to critical analysis of quantitative research. Philadelphia, PA: John Benjamins.
[7] Trochim, W. M. K. (2001). The research methods knowledge base (2nd edn.). Cincinnati, OH: Atomic Dog Publishing.
[8] Vanderstoep, S. W. & Johnston, D. D. (2009). Research methods for everyday life: Blending qualitative and quantitative approaches. San Francisco, CA: Jossey-Bass.
[9] Verkoeijen, P. P. J. L., Bouwmeester, S. & Camp, G. (2012). A Short-Term Testing Effect in Cross-Language Recognition. *Psychological Science* 23/6, 567– 571.

Meisam Rahimi was born in Isfahan, Iran on March 14, 1983. He received his BA degree in English language and Literature from the University of Isfahan, Isfahan, Iran in 2008. He is currently an MA student in Teaching English as a Foreign Language (TEFL) at the University of Isfahan. His areas of interest include computational linguistics, computer-assisted language learning (CALL) and language acquisition.

Samira Ghanbaran was born in Iran, Esfahan on September 21st ,1986.I earned my diploma in Physics and Mathematics at Esfahan University of Technology Highschool, Esfahan, Iran on June 19th , 2004.

I continued my studies in BA in English Literature, English department of the University of Esfahan, Esfahan, Iran. Degree was earned on September 21st, 2008.

I'm a MA student in English Teaching, English department of the University of Esfahan, Esfahan, Iran.

She has worked for four years as a LANGUAGE TEACHER. Still she is a Teacher in Pooyesh Language School, Esfahan, Iran. She is interested in Discourse Analysis, Testing Theories, Sociolinguistics, Pragmatics and all related courses in second language learning.

Seyed Mahmood Kazemi was born in Isfahan, Iran in 1975. He got his BA in TEFL, from the University of Isfahan, Isfahan, Iran, in 1999 and is now an MA student in tefl, University of Isfahan, Isfahan, Iran.

He has been an English Teacher in Isfahan, Iran from 1999. His research interests are language testing, psycholinguistics, and discourse analysis.