International Education Studies; Vol. 10, No. 2; 2017
ISSN 1913-9020 E-ISSN 1913-9039
Published by Canadian Center of Science and Education

The Psychological Effect of Errors in Standardized Language Test Items on EFL Students' Responses to the Following Item

Saman Khaksefidi[1]

[1] Higher Educational Complex of Saravan, Iran

Correspondence: Saman Khaksefidi, Higher Educational Complex of Saravan, Iran. E-mail: khaksefidi@gmail.com

Received: July 24, 2016 Accepted: August 30, 2016 Online Published: January 30, 2017

doi:10.5539/ies.v10n2p84 URL: http://dx.doi.org/10.5539/ies.v10n2p84

Abstract

This study investigates the psychological effect of a wrong question with wrong items on answering to the next question in a test of structure. Forty students selected through stratified random sampling are given 15 questions of a standardized test namely a TOEFL structure test in which questions number 7 and number 11are wrong and their answers to the next question is being analyzed. Three way ANOVA determines whether the total scores of the students, also their age, and gender affect the students' responses to the upcoming question or it is just a psychological factor affecting students' responses. The results showed that only 20 percent provided true answers; among them were only very proficient and proficient students. Age and gender were not considered significant factors in this regard. The correct answers were attributed to their full mastery over the structure and their personality type rather than the psychological factor. Most of the students were psychologically affected by a wrongly given item. Thus, it is suggested that test givers and teachers ignore the wrongly given question and also the upcoming one in order to prevent the psychological factor that makes students provide unsure and false answers to the question.

Keywords: psychological effect, errors, EFL student, following item, language test

1. Introduction

Test developers intend to make appropriate tests, including a good stem and discriminating distractors in a multiple choice test. Burton, Sudweeks, Merrill, and Wood (1990) contend that the purpose of the distractors is to appear as plausible solutions to the problem for those students who have not achieved the objective being measured by the test item. Conversely, the distracters must appear as *im*plausible solutions for those students who *have* achieved the objective. Only the answer should appear plausible to these students, but we are dealing here in this study with the case of totally wrong items, even the correct answer has been brought by mistake wrongly. This can affect the students because the more they try to find a suitable answer to the question, the more frustrated they will be, and a result of this ambiguity, it is very probable that they might not be able to concentrate enough, and answer correctly to the next upcoming question.

There are some guidelines on how the stem and the items including the correct item and the distracters should be. For the student who does not possess the ability being measured by the item, the distractors should look as plausible as the answer. Unrealistic or humorous distractors are nonfunctional and increase the student's chance of guessing the correct answer. When more than one of the alternatives can be successfully defended as being the answer, responding to an item becomes a frustrating game of determining what the teacher had in mind when he or she wrote the item. Such ambiguity is particularly a problem with items of the best answer variety where more than one alternative may be correct, but only one alternative should be clearly the best. If competent authorities cannot agree on which alternative is clearly the best, the item should either be revised or discarded (Burton, Sudweeks, Merrill, & Wood, 1990).

Item analysis is an excellent way to periodically check the effectiveness of your test items. It identifies items that are not functioning well, thus enabling you to revise the items, remove them from your test, or revise your instruction, whichever is appropriate. In fact, revising plays a very significant role in test developing. It has happened in many situations, even in TOEFL tests and widely nationwide tests that due to the lack of enough attention to revision, there have been some wrong items that have negatively affected the students.

Performance in language test tasks can be influenced by a wide range of features, which can interact unpredictably with characteristics of individual test-takers (O'Sullivan, 2000). Collectively, these influences can be considered as contributing to task difficulty, a topic that has attracted a lot of interest. There remains an assumption in many language testing contexts that test tasks are interchangeable for all sections of the test population (Lumley, & O'Sullivan, 2005).The students who are more tolerant of ambiguity, can handle the case much better dealing with the wrong items in a test, and as a result of that they might probably not have any problems with answering the next question correctly in case they have the necessary knowledge to answer.

2. Literature Review

Considering errors seems inevitable in language testing. "Because error will always be present in the use of any test, it is essential that we build the expectation of error in to decision-making procedures" (Fulcher & Davidson, 2007, p. 114).

Fulcher (2013) contends that all the effects of testing should be taken into consideration in order to be able to evaluate the overall impact that our actions have (Fulcher, 2013, p. 22). This can also be generalized to errors that occur in tests and their effects that can affect the test-taker's performance.

According to Backman (1990), "in any language situation, as with any non-test situation in which language use is involved, the performance of an individual will be affected by a large number of factors, such as the testing context, the type of tasks required, and the time of the day, as well as her mental alertness at the time of the test, and her cognitive and personality characteristics" (Backman, 1990, p. 31). Sometimes disorders occur in mental alertness due to some errors in tests.

Rhoades and Madaus (2003) contend that errors occur when test questions are ambiguous, poorly designed, or miskeyed and on a high-stakes exam, one poorly written question can determine a student's classification as "passing" or "failing," and reversal is difficult." (Rhoades & Madaus, 2003, pp. 16-17)

When those responsible for constructing the questions, assembling the trial test, and reviewing it, are satisfied that each question meets the criteria for relevant, reasonable, valid and fair items, the test is ready for trial. Only items which have survived this review should be subjected to trial with candidates like those who will eventually attempt the final version of the test (Izard, 1998).

Buck (1990) attributes the negative result to lack of attention to the questions or to a low level of interest in the content of the text, concluding that, 'Question preview may not motivate listening as much as test developers hope'.

Cohen (1998) and Nevo (1989) both found test takers in MCQ reading comprehension tests tended to adopt a strategy of matching the options with the text and selecting an option because (a) it had a word/words that also appeared in the text; (b) it had words similar in sound, or meaning, to words in the text; (c) it had a word which belonged to the same word family as a word appearing in the text; or (d) it just seemed to be related in some other way to word(s) in the text**.**

A quality teacher-made test should follow valid item-writing rules, but as many researchers point out, empirical studies establishing the validity of item-writing rules are in short supply and often inconclusive, and, ''item writing-rules are based primarily on common sense and the conventional wisdom of test experts'' (Millman & Greene, 1993, p. 353).

Frey, Edwards, Petersen, Pedrotti, and Peyton (2005) maintain that teachers of classroom assessment must rely on advice, opinion, experience, and common sense to direct their students in constructing classroom tests that produce reliable and valid scores.

Moreover, each of the four moderators of accuracy (judge, target, trait, and information) may combine with other moderators to form interaction effects. For example, the interaction of trait characteristics and rating targets is referred to as "palpability." Palpability describes a situation in which "traits...stand out in certain targets, relative…to the same trait in other targets" (Funder, 1995, p. 664). In other words, this effect could occur when a trait possesses certain characteristics when rating oneself, but different characteristics when rating others (Russell & Zickar, 2005).

This study aims at answering to the following questions:

1) Do wrong items in a question have an impact on the answer to the next question?
2) Does this affect all the students?
3) What is the role of age and gender in answering to this question?

The significance of this study lays in the fact that answering to totally wrong items can affect the students' motivation and mood for answering to the rest of the questions effectively, especially one or two questions that come afterward. Not only students' preparation for the test should be considered, but also the effects of a good or bad test stem and choice should be taken into consideration. A minor mistake by a test developer can lead to serious problems in the mind of test takers.

3. Method

3.1 Participants

Forty upper-intermediate students, including 20 male and 20 female students, participated in the test. The students were selected through stratified random sampling from among two institutes in Shiraz, Iran. Four upper-intermediate classes including 20 students were randomly chosen. In each class only 10 students were chosen randomly to answer the questions.

3.2 Data Collection

Fifteen multiple choice TOEFL structure questions were selected out of the total number of 40 questions. The students were required to answer them carefully. The correct item choices in question number 7 and number 11 have been deliberately changed into wrong ones to test the students' responses to the next question. Choice (b) the word “until” has been swapped with the word “for” and also “widely was used” was replaced with “widely used” in choice(c) of question number 11; therefore none of the answers could be correct.

3.3 Data Analysis

Descriptive statistics has been run through SPSS 16.0 to demonstrate the total score of each student and also his/her answers to question number 8 (the question immediately after question number 7) and question number 12(the question immediately after question number 11). Mean, mode, median, and standard deviation of the cores out of 15 have been calculated. The frequency of correct and wrong responses is also considered here in this study. Item difficulty is measured for questions number 8 and 12 to make sure it is not the difficulty of the item affecting the answers of students to the questions. One way ANOVA was applied in order to see the effect of the students’ total scores, age, and gender on their answers to Questions 8 and 12that immediately follow the wrongly given questions number 7 and number 11.

4. Results

The frequency of seven for the score of 7 and the frequency of 6 for the score of 9 illustrated in the table below signals that most of the students have gained scores close to the mean. Eighty percent of the students gave a wrong answer to question number 8 which means that there might be some factors affecting their answers.

Table1. Students’ scores out of 15

	Score	Frequency	Percent	Valid Percent	Cumulative Percent
Valid	3	1	2.5	2.5	2.5
	4	3	7.5	7.5	10.0
	5	4	10.0	10.0	20.0
	6	7	17.5	17.5	37.5
	7	7	17.5	17.5	55.0
	8	4	10.0	10.0	65.0
	9	7	17.5	17.5	82.5
	10	3	7.5	7.5	90.0
	11	1	2.5	2.5	92.5
	12	1	2.5	2.5	95.0
	13	1	2.5	2.5	97.5
	14	1	2.5	2.5	100.0
	Total	40	100.0	100.0	

Table 1 indicates the frequency of the scores among the students. Very few students were able to answer all the questions correctly.

Table 2. Answers given to question number 8

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Correct	8	20.0	20.0	20.0
	Wrong	32	80.0	80.0	100.0
	Total	40	100.0	100.0	

According to Table 2, twenty percent of the test takers responded correctly to question number 8, on the other hand, 80 percent chose a wrong response to the question.

Table 3. Answers given to question number 12

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Correct	7	17.50	17.50	17.50
	Wrong	33	82.50	82.50	82.50
	Total	40	100.0	100.0	

Table 3 depicts that 82.50 percent of the students responded wrongly to item 12 which indicates that a large number of the students we unable to provide correct answers to this item.

Item difficulty was measured for both Items 8 and 12 in another test removing the wrong question. The item difficulty for Item 8 was equal 0.5 and the item difficulty for Item 12 was equal 0.6 showing that the items were not difficult and had an acceptable level of item difficulty.

Table 4. Students' proficiency level

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	proficient and very prof.	8	19.5	20.0	20.0
	weak and not very proficient	32	78.0	80.0	100.0
	Total	40	97.6	100.0	
Missing	System	1	2.4		
Total		41	100.0		

According to Table 4, proficient and very proficient students who got more than 12 out of 15 are just 7.5% that is a small minority of the whole students. This means that there were few proficient and very proficient students.

Table 5. ANOVA result for Item 8

		Sum of Squares	df	Mean Square	F	Sig.
Age	Between Groups	.440	1	.440	1.756	.193
	Within Groups	9.535	38	.251		
	Total	9.975	39			
Gender	Between Groups	.246	1	.246	1.924	.173
	Within Groups	4.854	38	.128		
	Total	5.100	39			
Proficiency level	Between Groups	.262	1	.262	1.621	.211
	Within Groups	6.138	38	.162		
	Total	6.400	39			

Table 5 shows three-way ANOVA results for Item 8. The results reveal that age, gender, and proficiency level did not significantly affect the answers to Item 8 with the significance level ($p<0.05$) which was0.19, 0.17, and 0.21,

respectively.

Table 6. ANOVA result for Item 12

		Sum of Squares	df	Mean Square	F	Sig.
Age	Between Groups	.704	1	.704	2.886	.098
	Within Groups	9.271	38	.244		
	Total	9.975	39			
Gender	Between Groups	.017	1	.017	.125	.726
	Within Groups	5.083	38	.134		
	Total	5.100	39			
Proficiency level	Between Groups	.150	1	.150	.912	.346
	Within Groups	6.250	38	.164		
	Total	6.400	39			

Table 6 in a similar way displays three-way ANOVA results for Item 12. None of the variables such as age, gender, and proficiency level significantly affect the students' answers to this item. The results show that there must be some other factor affecting the students' answers to Items 8 and 12 since the items are neither very difficult nor very simple.

5. Discussion and Conclusion

Due to 80 percent wrong responses to question number 8, some possibilities emerge. The first possibility is that the level of difficulty of the question has been high, therefore results in providing the wrong answer to this question, but this should be rejected because a considerable number of students who got higher scores and were more proficient also answered the question wrongly and the level of difficulty of Item 8 was equal to 0.5 and Item 12 also was equal to 0.6. This means that either the stem or the items have been misleading or some other factors have influenced the students' response. Looking at the statistics for the answers to this item in some other situations without manipulating the previous item, the results show that proficient learners have been answering correctly to the item. There remains one possibility, and the possibility is that there should be another factor involved in the inaccurate response by the testees. The most probable possibility can be due to the previous question's ambiguity in its items which led the students into a state of perplexity and had negative emotional and psychological effects on the students. As a result of this, the students were unable to concentrate enough on the question, and therefore couldn't answer the question correctly. In reply to the first research question, it can be said that wrong items on a test may affect the response to the next question. The second question seems to have a more complicated answer. According to the results, it can be concluded that a wrong question may affect both proficient and weak students, but not all the students. But one point to add here is that very proficient learners who got 13 or 14 out of 15 were not affected by the wrong question, and also some students with more than medium level of proficiency scoring 9 or 10 or 11 answered the question correctly which might be due to their personality (being tolerant of ambiguity or learning the structure well or guessing. By the way, guessing seems less probable because they're not weak students rather they're to some extent proficient learners, so we can finalize the discussion to say that only weak and proficient students are affected by wrong questions in order to answer the next question, but very proficient students' capabilities remain intact by the influence of a wrong question. So, proficiency also cannot be considered a determining factor in the ability to answer a question immediately following a wrong question since some proficient students were not able to answer the question correctly. It can be concluded that the psychological effect of the wrong question has had a detrimental but temporary effect which affects the students' knowledge of answering to the upcoming question and therefore led to a false answer although the students were quite dominant to answer the question correctly. Therefore, wiping out the question immediately following the wrong question seems an appropriate measure to be taken by test givers.

Searching for a strategy to control this negative feature can be areas for further research to be carried out.

5.1 Implications

Test givers and teachers are advised to consider the effect of a wrongly given item on upcoming question. It is suggested that they omit the wrong question and also the question immediately following that one because psychological effects resist for a while, say some seconds or minute in one's mind and they can affect one's reply to a question. This can be done in norm-referenced questions such as national entrance exams because a true or

false response to a question can play a very significant role in being accepted or rejected in those exams.

References

Angoff, W. H., & Sharon, A. T. (1974).The evaluation of differences in test performance of two or more groups. *Educational and Psychological Measurement, 34*, 7-16.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Buck, G. (1990). *Testing Second Language Listening Comprehension* (Unpublished doctoral dissertation, University of Lancaster).

Burton, S. J., Sudweeks, R. R., Merrill, P. F., & Wood, B. (1990). *How to prepare multiple-choice test items: Guidelines for university faculty*. Brigham: Brigham Young University Testing Services and The Department of Instructional Science.

Cohen, A. D. (1998). Strategies and processes in test taking and SLA. In L. F. Bachman, & A. D. Cohen (Eds.), *Interfaces between Second Language Acquisition and Language Testing Research* (pp. 90-111). Cambridge University Press, Cambridge.

Frey, B. B., Edwards, L. M., Petersen, S., Pedrotti, J. T., & Peyton, V. (2005). Item-writing rules: Collective wisdom. *Teaching and Teacher Education, 21*, 357-364.

Fulcher, G. (2013). *Practical language testing*. Routledge.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*. London, England & New York, NY: Routledge.

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review, 102*, 652-670.

Green, A., & Yanagawa, K. (2008). To show or not to show: The effects of item stems and answer options on performance on a multiple-choice listening comprehension test. *System, 36*, 107-122.

Izard, J. (1998). Module 7, *Trial testing and item analysis in test construction*. UNESCO International Institute for Educational Planning.

Lumley, T., & O'Sullivan, B. (2005). Performance in tape-mediated assessment of speaking The effect of test-taker gender, audience and topic on task. *Language Testing, 22*, 415-437.

Millman, J., & Greene, J. (1993). The specifications and development of tests of achievement and ability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). Phoenix, AZ: American Council on Education.

Nevo, N. (1989). Test-taking strategies on a multiple-choice test of reading comprehension. *Language Testing, 6*, 199-215.

O'Sullivan, B. (2000). *Towards a model of performance in oral language testing* (Unpublished doctoral dissertation, University of Reading).

Rhoades, K., & Madaus, G. (2003). *Errors in standardized tests: A systemic problem* [National Board on Educational Testing and Public Policy Monograph]. *Boston: Boston College*.

Russell, S. S., & Zickar, M. J. (2005). An examination of differential item and test functioning across personality judgments. *Journal of Research in Personality, 39*, 354-368.

Appendix

1. From 1949 onward, the artist Georgia O'Keeffe made New Mexico _____.

(A) her permanent residence was

(B) where her permanent residence

(C) permanent residence for her

(D) her permanent residence

2. Just as remote-controlled satellites can be employed to explore outer space, _____employed to investigate the deep sea.

(A) can be robots

(B) robots can be

(C) can robots

(D) can robots that are

3. In _____ people, the areas of the brain that control speech are located in the left hemisphere.

(A) mostly of

(B) most

(C) almost the

(D) the most of

4. Stars shine because of _____ produced by the nuclear reactions taking place within them.

(A) the amount of light and heat is

(B) which the amount of light and heat

(C) the amount of light and heat that it is

(D) the amount of light and heat

5. _____ is not clear to researchers.

(A) Why dinosaurs having become extinct

(B) Why dinosaurs became extinct

(C) Did dinosaurs become extinct

(D) Dinosaurs became extinct

6. Although many people use the word “milk” to refer cow’s milk, _____ to milk from any animal, including human milk and goat’s milk.

(A) applying it also

(B) applies also

(C) it also applies

(D) but it also applies

*7. The first transatlantic telephone cable system was not established _____ 1956.

(A) while

(B) for

(C) on

(D) when

8. _____ no two people think exactly alike, there will always be disagreement, but disagreement should not always be avoided; it can be healthy if handled creatively.

(A) There are

(B) Why

(C) That

(D) Because

9. Drinking water _____ excessive amounts of fluorides may leave a stained or mottled effect on the enamel of teeth.

(A) containing

(B) in which containing

(C) contains

(D) that contain

10. In the 1820’s physical education became _____ of the curriculum of Harvard and Yale Universities.

(A) to be part

(B) which was part

(C) was part

(D) part

*11. Pewter, _____ for eating and drinking utensils in colonial America, is about ninety percent tin, which copper or bismuth added for hardness.

(A) was widely used

(B) widely used it

(C) widely wasused

(D) which widely used

12. A moth possesses two pairs of wings _____as a single pair and are covered with dustlike scales.

(A) function

(B) are functioning

(C) that function

(D) but functions

13. Soap operas, a type of television drama series, are so called because at first, they were _____.

(A) often which soap manufacturers sponsored

(B) sponsored often soap manufacturers

(C) often sponsored by soap manufacturers

(D) soap manufacturers often sponsored them

14. The Woolworth Building in New York was the highest in America when _____ in 1943 and was famous for its use of Gothic decorative detail.

(A) built (B) it built (C) was built (D) built it

15. Humans, _____, interact through communicative behavior by means of signs or symbols used conventionally.

(A) like other animals (B) how other animals

(C) other animals that (D) do other animals