Impact of Answer-Switching Behavior on Multiple-Choice Test Scores in Higher Education

Ramazan BAŞTÜRK*

Pamukkale Üniversitesi

Abstract

The multiple- choice format is one of the most popular selected-response item formats used in educational testing. Researchers have shown that Multiple-choice type test is a useful vehicle for student assessment in core university subjects that usually have large student numbers. Even though the educators, test experts and different test recourses maintain the idea that the first answer should be retained, many researchers argued that this argument is not dependent with empirical findings. The main question of this study is to examine how the answer switching behavior affects the multiple-choice test score. Additionally, gender differences and relationship between number of answer switching behavior and item parameters (item difficulty and item discrimination) were investigated. The participants in this study consisted of 207 upper-level College of Education students from mid-sized universities. A Midterm exam consisted of 20 multiple-choice questions was used. According to the result of this study, answer switching behavior statistically increase test scores. On the other hand, there is no significant gender difference in answer-switching behavior. Additionally, there is a significant negative relationship between answer switching behavior and item difficulties.

Key words: Multiple-choice test, answer switching, assessment in higher education.

Özet

Çoktan seçmeli soru formatı, eğitimle ilgili ölçme ve değerlendirme faaliyetlerinde en çok kullanılan soru formatı olarak bilinmektedir. Yapılan araştırmalar da göstermiştir ki, özellikle öğrenci sayısının çok olduğu yükseköğretimin farklı alanlarında öğrenci başarısının değerlendirilmesinde oldukça yaygın olarak kullanılan bir araç olarak görülmektedir. Eğitimcilerin, test uzmanlarının ve diğer kaynakların çoktan seçmeli sınavlarda öğrencilerin ilk verdikleri cevapların genellikle doğru cevap olduğu inancı yaygın olarak bilinmesine rağmen, bu inanç deneysel çalışmalara dayanmamaktadır. Bu araştırmamın amacı çoktan seçmeli sınavlarda cevapları değiştirme davranışının, öğrencilerin toplam puanlarına bir etkisinin olup – olmadığını ortaya çıkarmaktır. Ayrıca, cevap değiştirme davranışının cinsiyet değişkenine göre farklılık gösterip – göstermediği ve cevap değiştirme davranışları ile madde istatistikleri (madde güçlük ve ayırıcılık indeksleri) arasında bir ilişkinin olmadığı da incelenmiştir. Araştırmada çalışma grubu olarak Eğitim Fakültesine devam eden 207 öğrenci yer almaktadır. Araştırmada 20 adet çoktan seçmeli sorulardan oluşan arasınav sonuçları kullanılmıştır. Elde edilen bulgulara göre, öğrencilerin cevap değiştirme davranışlarının test puanlarını arttırdığı belirlenmiştir. Öte yandan cevap değiştirme davranışı için cinsiyet farklılığı belirlenememiştir. Ayrıca cevap değiştirme davranışı ile madde güçlük indeksi arasından ters orantılı bir ilişkinin olduğu belirlenmiştir.

Anahtar sözcükler: çoktan seçmeli soru formatı, cevap değiştirme davranışı, yükseköğretimde ölçme ve değerlendirme

The multiple- choice format is one of the most popular selected-response item formats used in educational testing (Carey, 1988; Oosterhof, 1994). It is a form of assessment in which respondents are asked to select the best possible answer/s out of the choices from a list. It is used extensively in any level of education, particularly in the middle grades through graduate. Different researchers demonstrated that Multiple-choice assessment is a useful device for student assessment in core university subjects that typically have large student numbers (Geiger, 1996; Milia, 2007). In general,

* Doç. Dr., Pamukkale Üniversitesi, Eğitim Bilimleri Bölümü, Ölçme ve Değerlendirme Ana Bilim Dalı
rbasturk@pau.edu.tr

students seem to favors multiple-choice assessment and remain one of the most commonly used assessment formats (Carey, 1988; Oosterhof, 1994; Struyen, Docht & Janssens, 2005; Wallace & Williams, 2003; Milia, 2007).

On the other hand, traditional objective tests, particularly multiple-choice type, are being criticized for a variety of different reasons including: test preparation practices may increase the test scores in high-stakes situation (Haney & Madaus, 1989); objective tests can lead to a narrowing of the educational curriculum (Shepard, 1989); multiple-choice assessment fosters surface or deep approaches to learning (Milia, 2007), the number of response items to include, the appropriate positioning of correct item options, etc. However, literatures demonstrate that less attention has been given to what should a student do when faced with the dilemma of choosing the correct answer (Geiger, 1996; Vispoel, 2000; Milia, 2007).

Researchers have found that the theory that a student should trust their first intuition and stay with their original answer on a multiple choice test is a myth and it is not dependent with empirical findings (Geiger, 1996; Nieswiadomy, Arnold & Garza, 2001; Milia, 2007). Benjamin, Cavell & Shallenberger, (1984) surveyed of teaching academics and found that 55 % of the participants believed that changing the original answer would lead to an incorrect choice, compared to 16 % who felt answer changing would benefit the overall score.

Milia, (2007) found that the educators, students, books offering exam preparation and university student support services support the idea that the first answer should be retained. In addition, Milia (2007) randomly selected 19 Australian web sites to assess their advice to students and found that they generally suggest retaining the original answer. Additionally, prior studies have shown that even though many students change answers in test, the widespread "conventional wisdom" has been that the first multiple-choice answer selected is usually believed to be the best answer (Geiger, 1996; Pressley & Ghatala, 1988; Zakay & Glicksohn, 1992).

The main purpose of this study was to investigate how the answer switching behavior affects the multiple-choice test score in higher education. Additionally, three sub-purposes were investigated: The first one was to investigate the practice of answer switching for all participants. Second sub-purpose was to examine any differences between Male and Female students' answer switching behavior in multiple-choice type test in higher education. And third sub-purpose of this study was to explore any significant relationship between item parameters (item difficulty and item discrimination) and number of answer switching behavior.

Method

Participants

The participants in this study consisted of 207 upper-level College of Education students from mid-sized university. Of these, 37.2 % (77) of the participants were Female and 63.8 % (130) of the participants were Male students. The students were enrolled in the upper-level courses of "Educational Measurement and Evaluation." The courses used in this study were taught by the same instructor for 7 years as part of his normal teaching loads.

Course

"Measurement and Evaluation in Education" course is one of the mandatory undergraduate level courses in College of Education and it is 5 ECTS credit hours. Aims and objectives of the course are to teach the role of measurement and assessment in teaching and learning, instructional goals and objectives, planning classroom tests and assessment, constructing objective test items: *Short answer items, true / false, matching, multiple-choice type items*. Measuring complex achievement: *The interpretive exercise essay type questions, performance – based assessment, alternative assessment techniques and observational techniques* and validity and reliability of testing devices and interpreting tests scores.

Procedures

In order to assess the students' achievement of the course, two exams, Midterm and Final, were used. Midterm exam was a multiple-choice type and Final exam was essay and calculation type exam. Since Final exam was different from the multiple-choice type, it is not included in this study. The Midterm exam was completed in pencil on a card read by an optical marker machine. Students were instructed that they must answer every question. Two different judges individually reviewed the Midterm exam cards to identify any answer changes by noting eraser marks. The changes were categorized as: (1) False to True (F–T); (2) True to False (T–F); and (3) False to False (F–F). In cases of multiple changes to an item that included an erased answer that was correct, the item was coded as True to False (T–F) (Milia, 2007).

Research Instruments

Midterm exam contained 20 multiple-choice type questions with five distracters and was administered in the 7th week of the course (middle of the semester). Table 1 demonstrated the item parameters (item difficulty and item discrimination) of the test. It can be seen from the Table 1 that the item difficulties, the proportion of students in the analysis group who answer the item correctly, ranged from 0.30 to 0.97 and item discrimination, which indicates the difference in the performances of the upper and lower groups, ranged from 0.05 to 0.64. Reliability analysis of the test calculated with Kuder-Richardson (KR - 20) type analysis and it was found 0.56. Item 15 has minimum (1) answer switches and Item 17 has the maximum (30) answer switches.

Table 1. Item parameters and number of answer switches on the test

Item No	Item Difficulties	Item Discrimination	Number of Answer Switches
I17	0.30	0.48	30
I7	0.39	0.43	19
I18	0.61	0.64	8
I19	0.66	0.50	10
I11	0.67	0.48	14
I13	0.67	0.02	13
I16	0.71	0.45	4
I20	0.73	0.34	7
I2	0.82	0.27	12
I10	0.83	0.38	6
I8	0.84	0.27	10
I3	0.85	0.32	7
I15	0.86	0.34	1
I1	0.88	0.21	13
I12	0.90	0.21	6
I14	0.92	0.18	3
I4	0.96	0.11	3
I5	0.97	0.11	2
I6	0.97	0.07	2
I9	0.98	0.05	2

Results

Overall Changes

In this exam, a total of 171 (4.1 %) answer changes behavior were recorded. Of these, 101 (59 %) were F–T, 42 (25 %) were T–F, and 28 (16 %) were F – F. It can be seen from the Table 2 that there were more F – T changes than T – F changes by a ratio of 2.36:1. In addition, 53 % of the test-takers switched ≥ 1 answer. The mode (70) was a single change and the maximum was 4.

Table 2. Answer switching behavior by gender

	Gender		
Switch	**Male**	**Female**	**Total**
False to True	28	73	**101**
True to False	14	28	**42**
False to False	11	17	**28**
Total	**53**	**118**	**171**

Individual Student Changes

While the aggregate analysis supports answer-changing behavior, it is also important to analyze the results from an individual student's perspective. Therefore, number of answer switching was also analyzed on an individual basis to identify how many students actually gained or lost points for the exam due to their own answer-changing behavior. In addition, gender differences on answer-changing behavior and the outcome of that behavior are analyzed.

According to Table 3, 38 out of 77 Male (% 49) and 70 out of 130 Female students (% 53) demonstrated answer switching behavior. This study demonstrated that there is no significant gender difference in answer-switching behavior in this midterm exam (χ^2 = 0.39; p > 0.05). Even though Female students exhibited higher change activity than Male students, this is not enough to be significant.

Table 3. Answer switching behavior by gender

	Gender				
Answer Switching	**Male**	**Female**	**Total**	χ^2	**p**
Yes	38	70	108	0.39	0.53
No	39	60	99		
Total	**77**	**130**	**207**		

Table 4 demonstrated that changing answers from False to True helped 68 (63%) students to increase their exam score, 12 (11%) decreased their exam score by switching from True to False and 28 (26%) were not impacted since their switch was False to False.

Table 4. Score differences by gender

	Gender		
Score	**Male**	**Female**	**Total**
Increase	19	49	68
Same	14	14	28
Decrease	5	7	12
Total	**38**	**70**	**108**

Changing answers F – T helped 19 (50 %) Male participants to increase their exam score, 5 (13 %) decreased their exam score by switching T – F, and 14 (37 %) were not impacted since their switch was F–F.

For Female students, changing answers F – T helped 49 (70 %) Female students to increase their exam score, 7 (10 %) decreased their exam score by switching T – F, and 14 (20%) were not impacted since their switch was F – F.

As summarized in Table 5, a paired - sample *t* test indicated that all participants' scores were changed from 14.99 to 15.66 and this difference is statistically significant ($t_{[1,107]}$ = 7.24, *p* < .01, η^2 = 0.33). In addition, results indicated that both Male and Female groups increased their score with answer switching behavior and these differences were statistically significant. It can be seen from the Table 5 that analysis suggested that Males increased their test score from 14.66 to 15.13 and this difference is statistically significant ($t_{[1,37]}$ = 3.27, p < .01, η^2 = 0.23). Like Male students, Female students also were increased their test scores from 15.17 to 15.94 and this difference is statistically significant ($t_{[1,69]}$ = 6.58, p < .01, η^2 = 0.39). Levene's test of homogeneity of variance indicated that variances are equal across both groups.

Table 5. Comparison of exam score by answer switching and gender

		Answer Switching						
		Before		After				
Gender	**N**	**Mean**	**SD**	**Mean**	**SD**	**D**	**t**	**p**
Male	38	14.66	2.22	15.13	2.28	0.47	3.27	0.00
Female	70	15.17	2.43	15.94	2.49	0.77	6.58	0.00
Total	**108**	**14.99**	**2.36**	**15.66**	**2.44**	**0.67**	**7.24**	**0.00**

Relationship between item parameters and number of answer switching

Item Difficulty

The item difficulty index, p, is the proportion or percentage of students in the analysis group who answer the item correctly (Carey, 1988; Linn & Gronlund, 1995; Varma, 2010). The range of item difficulty is from 0.00, indicating no student taking the exam answered the item correctly, to 1.00, indicating all students answered the item correctly (Oosterhof, 1994). Pearson correlation demonstrated that there is a statistically significant relationship between item difficulties and number of answer switching behavior (r = - .847 p < .05). It can be seen from the Figure 1 that the negative significant correlation result indicated that when the item is getting difficult, number of answer switching behavior increase.

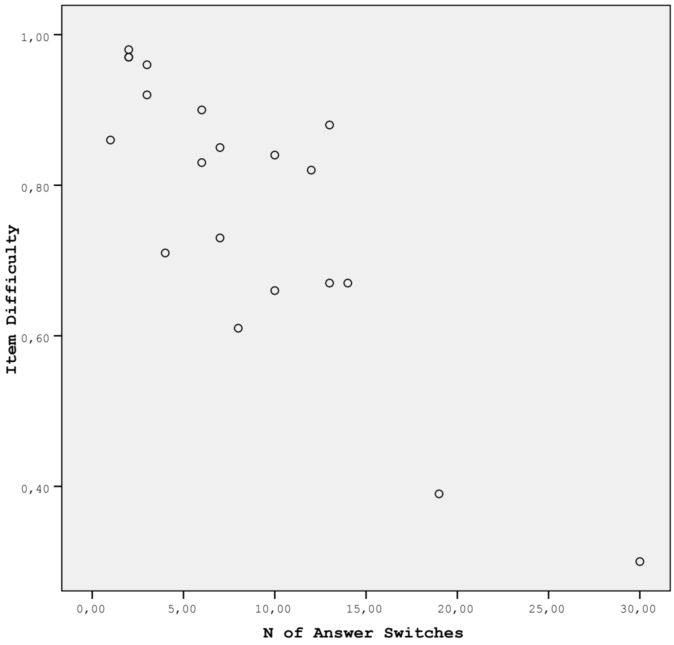

Figure 1. Relationship between item difficulty and number of answer switching behavior

Item Discrimination

The item discrimination analysis is based on the assumption that students who receive high scores on the overall test should score better on an item-by-item basis than students who receive low scores on the overall test (Carey, 1988; Varma, 2010). Another words, the discrimination of an item refers to its ability to distinguish between more and less knowledgeable students (Oosterhof, 1994). There are different item discrimination indices calculated by different measurement perspective (Linn, & Gronlund, 1995; Kelley, Ebel & Linacre, 2002). For example, The *Discrimination Index* (D) is computed from equal-sized high and low scoring groups on the test. *The Point-biserial Correlation*, is the Pearson correlation between responses to a particular item and scores on the total test and Rasch person measures and their responses to the item, the *point-measure correlation (*Kelley, Ebel & Linacre, 2002). Item discrimination index can range from - 1.00 to + 1.00. In this research, item discrimination indexes are calculated by point-biserial correlation techniques. Pearson correlation result indicated that there is no statistically significant relationship between item discrimination and number of answer switching behavior (r = .414; p > .05).

Discussion and Conclusion

The results obtained from this study support the view that students have very low answer switching behavior and prefer to stay with their first answers. In this research, only 4.1 % of answers were switched. This result is above in line with Kruger et al. (2005) and Milia, (2007) but below than Geiger's (1996) finding of 6%. Despite the fact that the absolute level of switching is low, in this research indicated that just over half the participants had at least one or more answer switching behavior.

Consistent with the literature (Geiger, 1996; Kruger et al., 2005; Milia, 2007), the results from this research suggest that answer switching led to improved test score for approximately half the undergraduate students. This study demonstrated that there is no significant gender difference in answer switching behavior in multiple-choice type exam. Female students had more answer-switching behavior than male students but this is not enough to be significant. This result is consistent with some literature (Kruger et al., 2005; Milia, 2007). However, the literature on answer-switching behavior has reported fairly mixed results for slight behavioral differences due to gender. Therefore more research needed to continue to investigate this potential determinant in order to ascertain whether research results on gender, including this study, are sample driven or influenced by other factors (Geiger, 1996).

This study demonstrated that there is a statistically negative significant relationship between item difficulties and number of answer-switching behavior. Negative significant correlation result indicated that when the item is getting difficult, number of answer switching behavior is increase. On the other hand, result indicated that there is no statistically significant relationship between item discrimination and number of answer-switching behavior. Therefore, more research needed to continue to investigate the relationship between answer-switching behavior and item statistics. In addition, the effects of item distracters on answer- switching behavior could be important topic to be investigated.

According to the result of this study, answer switching behavior statistically increase multiple-choice test scores. Findings suggest that educators should advise students to prepare soundly for multiple-choice exams and to make their choice after carefully reviewing each answer option. Benjamin et al., (1984) pointed out that the switching answers from "right to wrong" may be more painful and therefore more memorable. On the other hand, they advised that it is probably a good idea to change an answer if there is an additional reflection indicates that a better choice could be made.

Even though the finding that answer switching behavior statistically increase test scores and it is consistent with the literature, there are some limitations to this study. First limitation is that the data were collected with the undergraduate students in college of education. Other students registered with other faculties for example college of engineering or art and sciences may different behavior than the students in college of education. The second limitation is that relying on visible eraser marks results in an underestimate of answer switching. Milia (2007) pointed out that the visible eraser marks are produced when students use greater effort to mark the item and concluded that other cases of answer switching were not identified.

References

Benjamin, L. T., Cavell, T. A. & Shallenberger, W. R. (1984). Staying with the initial answers on objective tests: Is it a myth? *Teaching of Psychology, 11*, 133-141.

Carey, L. M. (1988). *Measuring and evaluating school learning*. Allyn and Bacon, Inc. Newton, Massachusetts.

Geiger, M. (1996). On the benefit of changing multiple-choice answers: Student perception and performance. *Education, 117,* 108–117.

Haney, W. & Madaus, G. (1989). Searching for alternatives to standardized tests; whys, whats, and whithers. *Phi Delta Kappan, 70,* 683 – 687.

Kelley, T., Ebel, R. & Linacre, J. M. (2002). Item discrimination Indices. *Rasch Measurement Transactions*, 16:3, p.883-4.

Kruger, J., Wirtz, D., & Miller, D. T. (2005). Counterfactual thinking and the first instinct fallacy. *Journal of Personality and Social Psychology*, *88*, 725–735.

Linn, R. L. & Gronlund, N. E. (1995). *Measurement and assessment in teaching*. 7th ed. Prentice-Hall, Inc. Columbus, OH.

Milia, L. D. (2007) Benefiting from multiple-choice exams: the positive impact of answer switching. *Educational Psychology*, *27(5),* 607 — 615

Nieswiadomy, R. N., Arnold, W. K. & Garza, C. (2001). Changing answers on multiple-choice examinations taken by baccalaureate nursing students. *Journal of Nursing Education, 40,*142–144.

Oosterhof, A. (1994). *Classroom applications of educational measurement (2nd Ed.)*. New York, NY: Macmillan College Publishing Company.

Pressley, M. & Ghatala, E.S. (1988). Delusions and about performances on multiple-choice comprehension test items. *Reading Research Quarterly*, *23*, 454-464.

Shepard, L. A. (1989). Why we need better assessment. *Educational Leadership*, *46(7)* 4 – 9.

Struyen, K., Docht, P. & Janssens, S. (2005). Students' perceptions about evaluation and assessment in higher education. *Assessment and Evaluation in Higher Education*, *30*, 325–341.

Varma, S. (2010). *Preliminary item statistics using point-biserial correlation and p values*. Retrieved April 14, 2010, from, http://www.eddata.com.

Vispoel, W. (2000). Reviewing and changing answers on computerized fixed-item vocabulary tests. *Educational and Psychological Measurement*, *60*, 371–384.

Wallace, M. A. & Williams, R. L. (2003). Multiple-choice exams: explanations for student choices. *Teaching of Psychology, 30,* 136–138.

Zakay, D. & Glicksohn, J. (1992). Overconfidence in a multiple-choice test and its relationship to achievement. *The Psychological Record*, *42*, 519-524.