RedFame

Journal of Education and Training Studies
Vol. 4, No. 10; October 2016
ISSN 2324-805X E-ISSN 2324-8068
Published by Redfame Publishing
URL: http://jets.redfame.com

Do the Guideline Violations Influence Test Difficulty of High-stake Test?: An Investigation on University Entrance Examination in Turkey

Erkan Hasan Atalmis

Correspondence: Erkan Hasan Atalmis, School of Education, Kahramanmaras Sutcu Imam Univeristy, Kahramanmaras, Turkey.

Received: July 1, 2016 Accepted: July 4, 2016 Online Published: July 14, 2016

doi:10.11114/jets.v4i10.1738 URL: http://dx.doi.org/10.11114/jets.v4i10.1738

Abstract

Multiple-choice (MC) items are commonly used in high-stake tests. Thus, each item of such tests should be meticulously constructed to increase the accuracy of decisions based on test results. Haladyna and his colleagues (2002) addressed the valid item-writing guidelines to construct high quality MC items in order to increase test reliability and validity. However, violating these guidelines is very common in high-stake tests. This study addressed two of these guidelines:, "AVOID the complex MC (Type K) format" and "Word the stem positively, avoid negatives such as NOT or EXCEPT", respectively. After reviewing a total of 2336 MC items extracted from university entrance examination (UEE) in Turkey administered over the past 15 years, we investigated impact of the violations of item-writing guidelines on test difficulty using multiple regression analysis. The findings showed that test difficulty was not statistically changed when MC items with negative stem were used in a test. They, however, indicated that the use of complex MC items has a statistically negative influence on the test difficulty. The paper concludes with possible results of the cases whereby items constructed by violating item-writing gudielines are eliminated from the test, and directions for future studies.

Keywords: Item-writing guidelines, complex multiple-choice-items, high-stake tests, test difficulty

1. Introduction

MC items are widely used in classroom assessments and high-stake tests since they are considered to provide accurate and objective scores as well as time-efficiency in administration and scoring. Test takers' skills and abilities are measured accurately and then critical decisions about individuals and groups are made in accordance with the results (Kingston & Kramer, 2013; Madaus, 1988). However, constructing reliable and valid MC items might be time-consuming and challenging because it requires clearly defined problem statements, plausible incorrect responses, and correct answers, making it essential for these items to be produced by highly-experienced item writers.

Previous studies addressed the valid item-writing guidelines to construct high quality MC items in order to increase test reliability and validity (Haladyna and Downing, 1989a; Haladyna and Downing, 1989b; Haladyna et al., 2002). Haladyna and his colleagues (2002) classified more than 30 guidelines in five categories as follows: content, formatting, style, writing stem, and choices concerns.

Constructing a test of MC items conforming to these guidelines increases the effectiveness of the test in terms of measurement of intended skills and evaluation of students' success. However, existing literature reported that many MC items used in classroom assessment and test banks are in poor quality in terms of conforming to these guidelines (Downing, 2005; Masters et al., 2001; Mehrens & Lehmann, 1991; Tarrant et al., 2006). Violating such guidelines makes the items flawed, decreasing the overall quality of test. Consequently, students' scores are adversely affected, which increases the difference of students' expected scores (true score) and observed scores (Downing, 2005; Haladyna & Rodriquez, 2013).

The related literature also shows that writing MC items by violating item-writing guidelines is very common in high-stake tests, test banks and classroom assessment from different disciplines (Hansen & Dexter, 1997; Masters et al., 2001; Downing, 2005; Tarrant et al., 2006; Tarrant &Ware, 2008). For example, Hansen and Dexter (1997) examine 440 items posed in ten test banks on auditing, and found that 75% of them were constructed ignoring item-writing guidelines. Masters and his colleagues (2001) reviewed 2913 items in 17 test banks on nursing and found around 76% of the items contain item-writing guidelines violations. Downing (2005) evaluated a total of 219 items extracted from four tests

administered to medical students, and revealed that approximately 46% of them contain item-writing violations. Tarrant and her colleagues (2006) evaluated a total of 2770 MC items collecting from test-banks and examinations of clinical nursing courses, and found 46% of the items have at least one item-writing guideline violations. Tarrant and Ware (2008) examined 10 high-stakes tests in a nursing school and found that 47% of a total of 664 items have item-writing guideline violations. Pate (2014) examined a total of 187 MC items from four classroom examinations as to discipline of medicine and found that 52% of them written with violations. The most common violations these studies identified were the use of "All of the Above" (AOTA) and "None of the Above" (NOTA), negative wording stem, unfocused stem, options' homogeneity, complex MC items and implausible distractors.

This study focuses on the specific guidelines such as "negative wording stem" and "complex MC items" (see Table 1) because they are very widespread in high-stake tests such disciplines as medicine and nursing (Tarrant et al., 2006; Downing, 2005; Pate et al., 2014; Baig et al., 2014). However, violating these guidelines raises issues with test reliability and validity by affecting items making them more difficult or easier (Downing, 2005; Albenese, Kent & Whitney, 1976). Frey et al. (2005) suggest that violating these guidelines causes potentially confusing wording issues for some students. Moreover, complex MC items decrease test validity as indicated by Liu & Wilson (2009) who reported gender differences in PISA performance on complex MC items. A recent study investigated by Haladyna and Rodriquez (2013) suggests that negative phrases are more challenging to understand for some students while reading complex MC items requires more time. Thus, students perform worse while handling the items containing violation of such guidelines than conventional multiple choice items (Downing, 2005; Tarrant and Ware, 2008).

Table 1. Examples of Item Types

Conventional MC item	Complex MC item	MC item with negative stem
Which is a city in Europe?	I. Berlin	Which city is NOT in Europe?
A. Berlin*	II. Chicago	**A.** Athens
B. Chicago	III. Paris	**B.** Berlin
C. Mexico City	Which city (or cities) is/ are in Europe?	**C.** London
D. New Delphi	**A.** only I **B.** only III	**D.** Paris
E. Sydney	**C.** I and II **D.** I and III*	**E.** Chicago*
	E. II and III	

*Key of the items

Specifically, there is a limited number of empirical studies addressing the impact of the using complex MC items and negative stem on item characteristics. Studies related to complex MC items reported consistent results. Namely, complex MC items were found statistically more difficult than conventional MC items although item discrimination did not statistically differ across conventional and complex MC items in anatomy and pharmacology (Nnodim, 1992; Tripp & Tollefson, 1985). However, inconsistent results were found concerning item difficulty when items with negative stem were employed. Downing, Dawson-Saunders, Case, and Powell (1991) found that using negative stem did not statistically change item difficulty in medicine while Tamir (1993) concluded that items with negative stem were found statistically more difficult than conventional MC items in biology.

The limitation of previous studies was that they mostly focused on the items from particular disciplines, such as biology, medicine, anatomy, and pharmacology. However, this study aims to investigate whether the violation of such guidelines has an impact on test difficulty after controlling different disciplines, such as mathematics, science and social sciences. Data were obtained from university entrance examinations (UEE) which are high stake tests held in Turkey from 1999 to 2013. More details about the examination is provided in the following section. In order to attain the objectives of the study, two specific research questions were posed:

(1) Does the use of negative stems influence test difficulty across different disciplines?

(2) Does the use of complex MC items influence test difficulty across different disciplines?

The organization of this paper is as follows: Section 2 provides the content and significance of UEE in Turkey. Section 3 describes data and data collection as well as method of data analysis. The results are presented in Section 4, and the last section offers related discussion on findings and conclusion.

1.1 University Entrance Examination in Turkey

The University Entrance Examination (UEE) in Turkey has been implemented as a high-stake standard test administered to senior and graduate high school students who want to further their education in institutions of higher education since late 1960s. It was organized by Measuring, Selection and Placement Center (OSYM). It is a paper and pencil test consisting of only MC items with five options. The examination covers a wide range of topics from different disciplines, including Turkish language, social sciences (philosophy, history, and geography), math (algebra and geometry), science (physics, chemistry, and biology), and foreign language (optional).

The system of UEE has changed a lot over the past several decades. For example, the exam which basically covered 9th grade subjects in all disciplines was held in a single stage from 1999 and 2005. The purpose of the exam was to select and place students in universities. In 2006, the system introduced a two-stage exam which was similar to the one implemented prior to 1999. Students achieving scores higher than cut score on the first examination are allowed to take the second stage of the exam. In other words, not all students can take the second stage of the exam. Thus, the main purpose of the first stage is to select students for higher education whereas that of the second one is to place those who getting the passing score in the institutions of higher education taking the score they achieved on the second stage of the exam into consideration. Another distinction between the stages is that the first stage covers 9th grade materials, like the system implemented prior to 2006, while the second stage covers more advanced materials offered from 10th to 12th grades.

UEE is so competitive that only students achieving sufficient scores on the exam could further their education at a college or university. To illustrate, around 1.9 million students took this exam in 2015 and less than half of them were able to be placed in a two or four year programs offered at colleges or universities, respectively (OSYM, 2015). Moreover, high-quality universities require relatively high scores on the test in concern. Therefore, some students are likely to retake this exam in subsequent years to enroll at better universities. In short, each individual correct answer plays a significant role for students in shaping their future life. Therefore, each item on this examination should be effectively and meticulously constructed to increase the accuracy of decision based on the test results.

2. Method

2.1 Sample of Items

A total of 2336 MC items taken from UEE implemented over the past 15 years (items in single-stage exams from 1999 to 2005, and items in the first stage of the exams from 2006 to 2013) analyzed in this study. The examination contains four main different disciplines with sub-disciplines, exactly the same proportion of items from each: 25% math, 25% science (physics, chemistry, and biology), 25% Turkish language, and 25% social science (history, geography, and philosophy). For example, the 180 items posed in 2001 consisting of 45 items prepared for each discipline; 120 items in 2008 with 30 items in each discipline; 160 items in 2012 with 40 items in each discipline.

Although OSYM publicly announces the results of students' scores and ranking, and the mean of test for each discipline every year, the item statistics are not announced due to security issues. Therefore, in this study, test difficulty for each discipline is calculated dividing the mean of test by the number of items. For example, the difficulty of the mathematics test of 40 items, of which mean was measured 7.98 in 2013, was calculated as 7.98/40=0.20.

2.2 Data Analysis

After a total of 2336 items from different disciplines at UEE over the past 15 years were evaluated based on item-writing guidelines, the frequency distributions, descriptive statistics, and test difficulty of disciplines were individually calculated. Subsequently, multiple regression method was conducted by using Stata Version 13 (StataCorp, 2013) in order to examine the impact of item-writing guidelines violations on test difficulty. Dependent variable was set as test difficulty across different disciplines while types of guideline violations, complex MC items and negative stem, were accepted as independent variables. Eight different disciplines were used as control variables as follows: Turkish language, philosophy, history, geography, math, physics, chemistry, and biology.

3. Results

3.1 Item-writing Guidelines Violations across Different Disciplines over 1999-2013

Table 2 summarizes the number and frequency of items that contain item-writing guideline violations in UEE over the past 15 years. A total of 768 items revealed to contain guideline violations: 451 with negative stem (guideline #9) and 315 complex MC items (guideline #17). The average percentage of the items containing such violations was found 32% over the past 15 years: 19% of items with negative stem (guideline #17) and 13% of the items which are complex MC items (guideline #9).

Table 2 depicts the number of guideline violations across the disciplines of mathematics, Turkish language, science (physics, chemistry, biology), and social science (history, geography, and philosophy). Only 1% of mathematics items, all consisting of items with negative stems, seem to contain guideline violations,. Approximately 35% of Turkish language items contained guideline violations, 33% of which were those with negative stems, and 2% of which were complex MC items. Science items were the ones containing the most guideline violations at UEE over the past 15 years. Around 49% of science items contain such kind of violations (18% with negative stem and 30% complex MC items). Specifically, 64% of biology items (42% complex MC items and 22% with negative stem), 59% of chemistry items (25% comlex MC items and 34% negative stem), and 30% of physics items (%25 complex MC items, 4% negative stem) were formed by violating item writing guidelines.

Table 2. Guideline Violations across Different Disciplines over 1999-2013

Disciplines		# of items	Type of Flawed Items						Item Statistics	
			Neg. Stem		Complex MC Items		All		Test difficulty mean	Standard deviation
			# of items	% of items	# of items	% of items	# of items	% of items		
Mathematics		594	8	0.01	0	0.00	8	0.01	0.22	0.06
Turkish Language		594	198	0.33	9	0.02	207	0.35	0.48	0.07
Science	Physics	241	9	0.04	61	0.25	72	0.30	0.12	0.04
	Chemistry	185	62	0.34	48	0.25	110	0.59	0.13	0.05
	Biology	168	37	0.22	71	0.42	108	0.64	0.12	0.06
	Combination	594	108	0.18	180	0.30	290	0.49	0.12	0.05
Social Science	History	253	51	0.20	105	0.42	156	0.62	0.33	0.07
	Geography	207	75	0.36	21	0.10	96	0.46	0.28	0.07
	Philosophy	134	11	0.08	0	0.00	11	0.08	0.35	0.08
	Combination	594	137	0.23	126	0.21	263	0.44	0.32	0.08
Total		2376	451	0.19	315	0.13	768	0.32	0.29	0.15

A total of 44% of social science items guideline violations: nearly half of them were those with negative stems and other half was complex MC items. 62% of history items (42% complex MC items and 20% with negative stem), 46% of geography items (10% complex MC items and 36% with negative stem), and 11% of philosophy items, all of which were constituted by those with negative stem.

Table 2 also indicated test difficulty means across different disciplines over the past 15 years. The test difficulty mean for UEE between the years1999 and 2013 was counted 0.29, which signifys that 29% of them were correctly responded by examinees. Science is the most difficult part of the test with the lowest test difficulty mean, which was calculated 0.12. The test difficulty mean across physics, chemistry, and biology varies in the range between 0.12 and 0.13. Test difficulty mean of social science was 0.32. The related mean across history, geography, and philosophy varies in the range between 0.28 and 0.35. Test difficulty mean of mathematics was found 0.22 while that of Turkish language items was 0.48, making it the easiest part of UEE.

3.2 The Impact of Item-writing Guidelines Violations on Test Difficulty

This section reports the impact of item-writing guidelines violations, the use of negative stems and complex MC items on test difficulty by conducting multiple regression analysis. Test difficulty was set as a dependent variable, which was obtained from each discipline of UEE from between 1999 and 2013. Independent variables were the percentage of use of negative stems and complex MC items in each discipline from 1999 to 2013. The disciplines were accepted as the control variable in this model.

Table 3 shows the results of multiple regression analysis of Model 1 and Model 2. The difference between the models is that impact of the use of negative stems and complex MC items on test difficulty was examined in Model 1without using control variables which were employed in Model 2. Two of the findings indicated that both models were statistically significant ($F_{model\ 1}(2, 99) = 12.36$, $p<0.001$; $F_{model\ 2}(9, 92)=48.52$, $p<0.001$). %20 and 83% of the total amount of variance in test difficulty was explained by Model 1 and Model 2, respectively.

Table 3. Unstandardized Coefficients for Regressions Predicting Test Difficulty

	Model 1	Model 2
Item-writing guidelines violations Measures (independent variables)	Test Difficulty	Test Difficulty
Use of negative stem	-0.001	-0.001
	[0.001]	[0.001]
Use of complex MC items	-0.003**	-0.002**
	[0.001]	[0.001]
Disciplines (control variables)		
Mathematics (cons.)	-	0.223**
		[0.016]
Physics	-	-0.045
		[0.031]
Chemistry	-	-0.020
		[0.037]
Biology	-	-0.005
		[0.043]
Turkish language	-	0.284**
		[0.029]
History	-	0.215**
		[0.043]
Geography	-	0.106**
		[0.033]
Philosophy	-	0.129**
		[0.024]
R-Square	0.20**	0.83**

*p<.05 **p<.01

The findings of Model 1 showed that test difficulty did not statically change when negative stem was used in the construction of test items ($\beta_{\text{negative stem}} = 0.001$, $SE_{\text{negative stem}} = 0.001$, $p = 0.180$). However, the use of complex MC items negatively and statistically influenced the test difficulty ($\beta_{\text{complex}} = 0.002$, $SE_{\text{complex}} = 0.001$, $p < 0.01$).

Model 2 was used to find answers to the research questions in this study because disciplines were controlled in this study. The results indicated that the effect of the use of negative stem on test difficulty was not statistically significant ($\beta_{\text{negative stem}} = 0.001$, $SE_{\text{negative stem}} = 0.001$, $p = 0.194$) while that of the use of complex MC items on test difficulty was negative and statistically significant ($\beta_{\text{complex}} = 0.002$, $SE_{\text{complex}} = 0.001$, $p < 0.01$). This result shows that when the use of complex MC items in a test increases by 1%, test difficult value decreases by 0.002. This means that complex MC items make the test more difficult. For instance, when a complex MC items is used in a test of 15 items, percentage of the use of complex MC items in this test is 1/15 = 0.07 (7%), and the test difficult value decreases by 0.014 (0.002*7).

4. Discussion and Conclusion

The purpose of this study was to examine how the two most frequented item-writing guideline violations, the usage of negative stem and complex MC items, influence test difficulty. We examined items containing item-writing guideline violations across different disciplines at university entrance examination (UEE) in Turkey from 1999 to 2013. The findings demonstrated that %13 of the items across different disciplines were complex MC items while %19 of them were those with negative stems. Mathematics and science items contained the least and the most guideline violations, respectively. Social science items contained slightly less violations than science items (around 45%). Thus, approximately half of the science items and social items were not well-constructed, approving the results of the previous studies (Downing, 2005; Tarrant et al., 2006; Tarrant & Ware, 2008; Baig et al., 2014).

Subsequently, regression analysis was conducted to examine the impact of the above-mentioned violations, the use of negative stem and complex MC items, on test difficulty across disciplines. The findings displayed that the test difficulty did not statistically change when items with negative stem were used. However, complex MC items statistically decreased the test difficulty, which means that they were found more difficult than conventional MC items. Another finding of the study was that the test difficulty decreased by 0.002 when the use of complex MC items in a test increased by 1%. In this respect, taking consideration into biology items with 42% complex MC items, test difficulty decreased by around 0.08, confirming the results of earlier studies carried out by various scholars (Albenese Kent, & Whitney, 1976; Nnodim, 1992; Tripp & Tollefson, 1985). These results were expected since students are required to comprehend every single statement in the root of complex MC items to earn full credit, which might decreases the probability of guessing correct answer (Liu & Wilson, 2003). Moreover, such type of items increases test anxiety of the students since they spend more time to read them when compared to the conventional MC items.

The findings also showed that negative stem does not have a significant impact on test difficulty, which is consistent with the results of the study conducted by Downing et al. (1991). This specific finding might be attributed to that students become very familiar with these items since MC items with negative stem are widely used not only in high-stakes test but also in classroom assessments. Particularly, in their educational career, they might have been provided with some strategies to cope with mistakes on MC items with negative stem. They might also how to transfer negative statements into positive ones. Therefore, these item types might be challenging for the students in elementary schools rather than those attending high schools.

To sum up, items with negative stems and complex MC items are commonly used across diffferent disciplines at UEE. Particularly, complex MC items could decrease the test difficuly, which might influence the realibility and validity of the test since such types of items might be tricky for the students. In other words, if the test were administered once again after the items containing guideline violations were eliminated, the test results could be relatively more accurate. When a high-stake test with MC items is well-constructed without violating item-writing guidelines, students' level of test anxiety is likely to decrease as they are expected to easily understand the items in a shorther period of time (Haladyna & Rodriguez, 2013). This could also reduce their preparation time for such kind of high-stake tests. For example, some students can start to manage their preperation period themselves more effectively instead of trying to elicit them from tutors and institutions, which is likely to help them understand the structure of the items of tests, and use manage time-effectively during the tests.

5. Limitation and Future Studies

A limitation of this study is that the items were obtained from the first stage of UEE in Turkey from 1999 to 2013, which were given only to senior and graduate high school students. To generalize the results, an analysis of more items posed in different high stake tests taken by students at different grades is needed. In future studies, two test forms might be designed to evaluate students' progress in particular subject matters, one containing only conventional MC items and one containing complex MC items, and compared in terms of various aspects. This design is believed to enable the researchers to measure more accurate results as to how different types of items influence item difficulty. In addition, it might provide an opportunity to see whether other psychometric properties, such as item discrimination and test reliability, change across the use of conventional MC items and complex MC items.

References

Albanese, M. A., Kent, T. H., & Whitney, D. R. (1976). A comparison of the difficulty, reliability and validity of complex multiple choice, multiple response and multiple true-false items. In *Annual Conference on Research in Medical Education. Conference on Research in Medical Education*, 16, 105-110.

Baig, M., Ali, S. K., Ali, S., & Huda, N. (2014). Evaluation of Multiple Choice and Short Essay Question Items in Basic Medical Sciences. *Pakistan journal of medical science, 30*(1), 3.

Downing, S., Dawson-Saunders, B., Case, S. M., & Powell, R. D. (1991). The psychometric effects of negative stems, unfocused questions, and heterogeneous options on NBME Part I and Part II characteristics. A paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL

Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education, 10,* 133–143. http://dx.doi.org/10.1007/s10459-004-4019-5

Frey, B. B., Petersen, S., Edwards, L. M., Pedrotti, J. T., & Peyton, V. (2005). Item-writing rules: Collective wisdom. *Teaching and Teacher Education, 21*(4), 357-364. http://dx.doi.org/10.1016/j.tate.2005.01.008

Haladyna, T. M., & Downing, S. M. (1989a). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2*(1), 37-50. http://dx.doi.org/10.1207/s15324818ame0201_3

Haladyna, T. M., & Downing, S. M. (l989b). Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2*(1), 51-78. http://dx.doi.org/10.1207/s15324818ame0201_4

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. New York, NY: Routledge. Routledge.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item writing guidelines for classroom assessment. *Applied Measurement in Education, 15,* 309–334. http://dx.doi.org/10.1207/S15324818AME1503_5

Kingston, N. M., & Kramer, L. M. B. (2013). High-stakes test construction and test use. In T.D. Little (Ed.), *The Oxford Handbook of Quantitative Methods: Statistical Analysis* (pp. 189-205). Oxford University Press. http://dx.doi.org/10.1093/oxfordhb/9780199934874.013.0010

Liu, O. L., & Wilson, M. (2009). Gender differences in large-scale math assessments: PISA trend 2000 and 2003. *Applied Measurement in Education*, *22*(2), 164-184. http://dx.doi.org/10.1080/08957340902754635

Madaus, G. F. (1988). The distortion of teaching and testing: High ‐ stakes testing and instruction. *Peabody Journal of Education*, *65*(3), 29-46. http://dx.doi.org/10.1080/01619568809538611

Nnodim, J. O. (1992). Multiple-choice testing in anatomy. Medical Education, 26, 301–309. http://dx.doi.org/10.1111/j.1365-2923.1992.tb00173.x

OSYM (2015). The system of student selection and placement in higher education institutions in Turkey. Ankara. Higher Education Council. Student Selection and Placement Center.

Pate, A., & Caldwell, D. J. (2014). Effects of multiple-choice item-writing guideline utilization item and student performance. *Currents in Pharmacy Teaching and Learning*, *6*(1), 130-134. http://dx.doi.org/10.1016/j.cptl.2013.09.003

Stata, C. (2013). *Stata Statistical Software: Release 13.* College Station, TX: Stata-Corp LP.

Tamir, P. (1993). Positive and negative multiple choice items: How difficult are they? *Studies in Educational Evaluation 19*, 311–332. http://dx.doi.org/10.1016/S0191-491X(05)80013-6

Tarrant, M., & Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical Education*, *42*(2), 198-206. http://dx.doi.org/10.1111/j.1365-2923.2007.02957.x

Tarrant, M., Knierim, A., Hayes, S. K., & Ware, J. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse education in practice*, *6*(6), 354-363. http://dx.doi.org/10.1016/j.nepr.2006.07.002

Tripp, A., & Tollefson, N. (1985). Are complex multiple-choice options more difficult and discriminating than conventional multiple-choice options?. *Journal of Nursing Education*, *24*(3), 92-98.