Engineering Educators Bringing the World Together
2025 ASEE Annual Conference & Exposition
Palais des congrès de Montréal, Montréal, QC • June 22–25, 2025
ASEE

Paper ID #48769

Do True-False Question Variants Create Bias?

Dr. Jacqueline Jenkins, Cleveland State University

Do True-False Question Variants Create Bias?

In an introductory transportation engineering course, four versions of four tests were developed as a measure to deter cheating. These tests included a total of 45 true-false questions. Of those, five questions were written as true statements (true only), and seven questions were written as false statements (false only). The remaining 33 questions were written as a true statement (true variant) and a corresponding false statement (false variant) and assigned to the different test versions. This approach raised three concerns: 1) that the student performance on the different versions of a test would differ; 2) that true-false statements that were written as a true statement and corresponding false statement would provide some hint to students; and 3) that student performance on the true variant and false variant questions would differ. The test responses from 53 students were analyzed. For each of the four tests, the performance on the true-false questions on the four versions was found to be comparable. In addition, the performance on the true variant questions and true only questions was found comparable as was the performance on the false variant and false only questions. Lastly, the performance on the true variant questions was found to be equal to or greater than the performance on false variant questions. The conclusion is that assigning true and false variants across versions of a test does not introduce a bias when the proportion of true to false questions is consistent.

Introduction

For more than a decade, true-false questions have been included in the midterm and final tests in the introductory course, CVE 446/546 Transportation Engineering at Cleveland State University (CSU). The true-false section of a test would typically include ten to fifteen questions and contribute no more than 25% of the total test score. These questions were based on facts, pieces of knowledge students had previously applied to solve homework problems, some of which students needed to recall when solving the computational problems on the test. The odd trivia question was also included to see who had been paying attention during the lectures. Anecdotally, student performance on the true-false questions was for the most part similar with the performance on the analysis and design problems.

Over the past few years, the size of the class has more than tripled, such that during tests there could no longer be an empty seat between students. To deter cheating, and potentially recognize the occurrence of cheating, four versions of each test were developed. The quantitative problems were written with different sets of input values. When grading such problems, it was evident if a student retrieved answers from a neighbor. For the true-false questions, two approaches were tried. The first approach was to use the same true-false questions on each version of a test but switch their order. The reordering of questions changed the pattern of the true and false statements; however, students could still potentially discern which questions were the same across test versions based on the apparent length of the questions. The second approach was to keep the questions in the same order but rewrite some of the questions, such that some versions of the test would include a true statement (true variant), and others would include the corresponding false statement (false variant). Turns out that this approach did not help to identify cheaters, but hopefully did serve to dissuade such behavior. This approach also provided an opportunity to examine whether the design of true-false questions impacts student performance.

Through a post course analysis of student responses to the true and false test questions, three concerns were addressed. The first concern was that the different versions of a test impacted student performance, that some versions of a test were easier than other versions because they included the true variant or false variant questions. The second concern was that the process of converting true statements into false statements and false statements into true statements resulted in some sort of hint to students, thus the student performance on the true variant and false variant questions differed from questions that were developed only as true statements or only as false statements. The third concern was that the student performance on the true variant and false variant questions differed.

The third concern is critical. If student performance on true questions differs from that on false questions, assuming both questions are written by the same rules governing language, content, format, etc. then the overall performance on a true-false test can be influenced by the balance between true and false questions. It follows then that versions of tests with different proportions of true and false questions would introduce a performance bias. Conversely, switching out true statements for false statements or false statements for true statements would not be an issue if student performance on each type of question is similar.

Literature Review

The key article guiding this study was that of Frisbie and Becker [1]. They reviewed 17 educational measurement textbooks published between 1980 and 1991, presented a list of rules for writing true-false statements, outlined the perceived advantages and disadvantages of using true-false tests, and presented a handful of researchable questions. The advantages included the ability to cover a large amount of content, the ease of constructing questions, and the ability to quickly and accurately grade responses, as well as the ability to obtain a lot of information about achievement while minimizing the amount of reading required by the test takers. The disadvantages largely describe validity and reliability issues, stemming from the content and style of the questions, that guessing is a viable response strategy, and that responses can be given based on knowing what is wrong without having to show what is right.

One of the questions raised by Frisbie and Becker [1] was about whether the proportion of false statements on a test, or knowledge of that proportion, impacts performance. At the time, Sax [2] recommended that approximately half the statements should be false because test takers who did not know an answer were more likely to select the true response. Similarly, Gronlund and Linn [3] recommended having approximately the same number of true and false questions. They suggested that this approach would not favor those test takers who adopted a strategy of consistently selecting true or consistently selecting false for questions they were not certain of. However, Gronlund and Linn [3] also recommended that the exact same number of true and false questions should not be used as it would provide a hint to those test takers who guessed. However, this rationale implies that the test taker has some knowledge of the proportion of true and false questions.

Guessing the answers to true-false questions has been studied, mainly in terms of test reliability. For blind guessing, where the test taker has no knowledge of the answer, the test taker has a 50% chance of selecting the correct true-false answer, however test takers often have partial knowledge about questions, and can therefore improve their chances of guessing correctly. Burton [4] and Burton and Miller [5] developed statistical models of test reliability illustrating the impact of blind guessing. They point out that partial knowledge would tend to yield higher scores but would also increase the variability of scores and therefore reduce test reliability. However, they point out that test reliability can be improved by increasing the number of questions, or discouraging guessing through negative marking, where marks are deducted for incorrect answers. Wang [6] suggested that correcting for guessing is not always needed and derived critical values of passing scores indicating when the influence of blind guessing is negligible, based on the binomial distribution and the number of test questions. Reid [7] proposed a complex scoring formula accounting for the number of questions, the number of correct responses and the number of incorrect responses. Interestingly, Lackey and Lackey [8] did not use negative marking when studying the impact of students' first language on test scores in an undergraduate mechanical engineering course and found a good correlation between student performance on true-false questions and open-ended questions on tests. The correlation also illustrated the value of true-false questions to obtain a measure of achievement.

Methodology

Recall, the second approach taken to curb cheating in the introductory transportation engineering at CSU was to have multiple versions of each test and to keep the true-false questions in the same order on each version but rewrite some of the questions, such that some versions of the test included the true variant of a statement, and others included the false variant. Therefore, the main tasks of this study were to craft true and false questions for four versions of four tests, administer the tests, and collect and analyze the student responses.

The true-false questions were written in simple language, without the use of determiners (e.g. all, never, always), and without negatives or double negatives. Each question addressed a single, meaningful idea and were clearly true or false. These characteristics were consistent with the rules for constructing true-false questions reported by Frisbie and Becker [1].

A total of 45 true-false questions were written for four tests. The Operations, Design I, and Design II tests each had ten true-false questions while the Planning test had fifteen. The number of true-false questions included on each version of each test is detailed on Table 1. The numbers shown in parentheses represent true only and false only questions. A true only question was only written as a true statement, and a false only question was only written as a false statement. The true only and false only questions appeared on each version of a test. The numbers outside of the parentheses represent the true and false variants. These were the questions that were written as a true statement and as a corresponding false statement and assigned to different versions of a test. For instance, the true variant "A simple horizontal curve is a circular arc with a consistent radius." was assigned to versions A and D of the Design I test and the corresponding false variant "A simple horizontal curve is an elliptical arc with a consistent radius." was assigned to versions B and C.

Table 1. Number of True and False Questions

Version	Operations		Design I		Design II		Planning	
	True	False	True	False	True	False	True	False
A	3 (1)	4 (2)	3 (2)	3 (2)	4 (1)	4 (1)	6 (1)	7 (2)
B	4 (1)	3 (2)	3 (2)	3 (2)	4 (1)	4 (1)	6 (1)	7 (2)
C	3 (1)	4 (2)	3 (2)	3 (2)	4 (1)	4 (1)	6 (1)	7 (2)
D	3 (1)	4 (2)	3 (2)	3 (2)	4 (1)	4 (1)	6 (1)	7 (2)

Note: the inconsistency in the number of true and false questions on the Operations test was the result of a typo

To improve readability, the true-false questions have been numbered sequentially. There were five true only questions (6, 11, 17, 25, 34), seven false only questions (2, 7, 16, 20, 30, 40, 42), and the remaining 33 questions had both a true variant and a false variant.

The purpose of using multiple different versions of a test was to curb cheating. Therefore, when tests were distributed in class, two versions of the test would be given to students in every other row. Within a single row, the two versions would be distributed in an alternating pattern. Thus, a student would not have a neighbor with the same version of the test. Students were aware that there were multiple test versions.

For all four tests, students received one mark for each correct answer and zero marks for incorrect answers. Hence, negative marking was not implemented. No correction for guessing was applied.

During the tests, students were not aware of the research study. After handing in the last test, students received a consent form. By signing the form, they agreed to have their true-false answers included in the analysis, as per the approved protocol (IRB-FY2025-66 True False Questioning).

Results

There were 53 students registered in CVE 446/546 Transportation Engineering during the Fall 2024 semester. All the students were present for each of the four tests. All students gave their consent to have their true-false responses included in the analysis for this study. One student did not complete any questions on the Operation, Design 1 and Planning tests. These tests were excluded. Two students left a true false question unanswered on the Operations test, three students left a true false question unanswered on the Design II test, and four students left a true false question unanswered on the Planning test. These nine non-responses were taken as incorrect answers. The resulting sample sizes for each of the versions of the four tests are given in Table 2.

Table 2. Number of Completed Tests

Version	Operation	Design I	Design II	Planning
A	16	12	13	12
B	14	14	14	13
C	12	12	14	14
D	10	14	12	13
Total	52	52	53	52

The frequencies of true (correct) and false (incorrect) responses for 38 true questions are shown on Figure 1. Note that questions 6, 11, 17, 25 and 34 were asked using only a true statement. Therefore, the number of responses for these true only questions was approximately double that of the true variant questions.

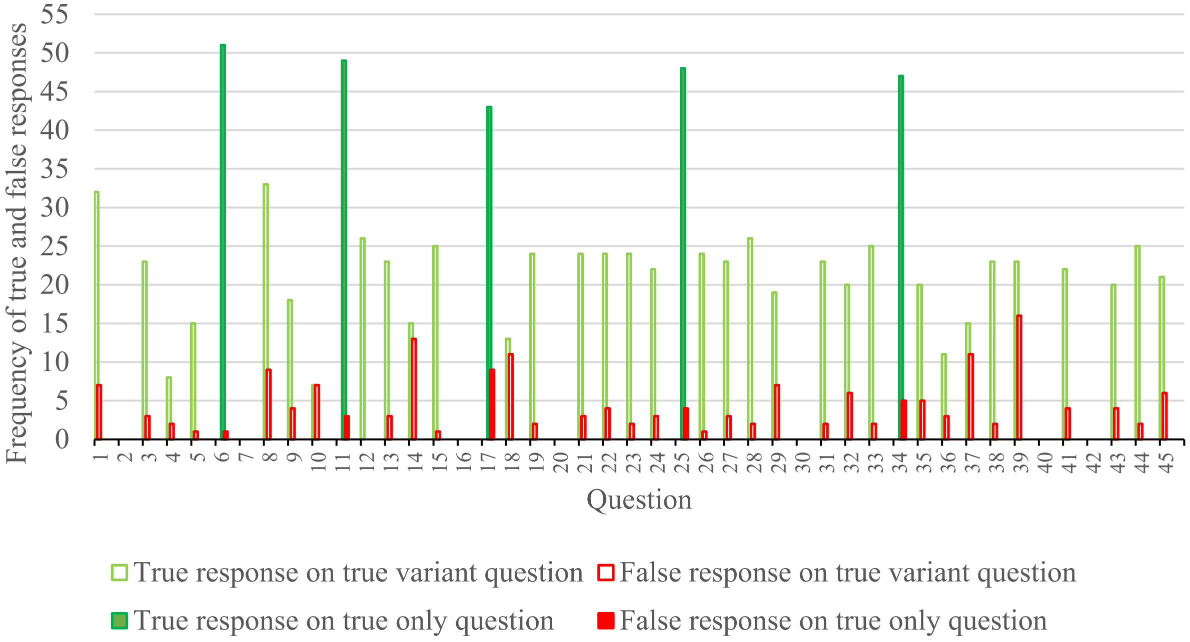

Figure 1. Frequency of true and false responses on true questions

The frequencies of true (incorrect) and false (correct) responses for 40 false questions are shown on Figure 2. Note that questions 2, 7, 16, 20, 30, 40, and 42 were asked using only a false statement. Therefore, the number of responses on these false only questions was approximately twice that of the false variant questions.

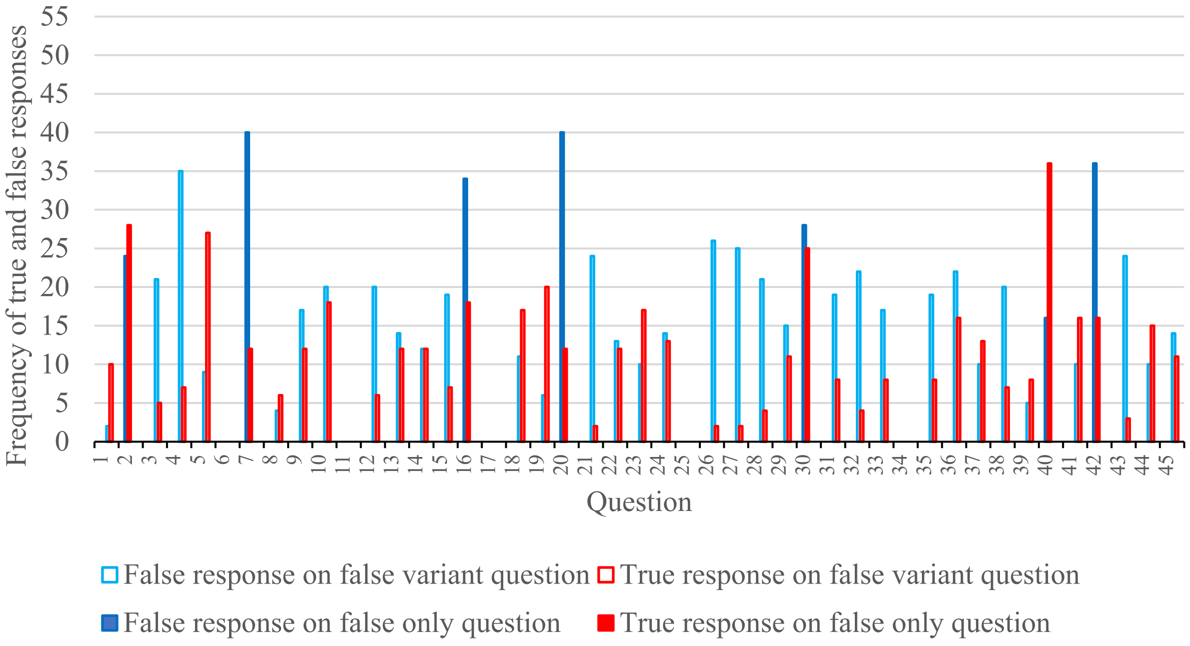

Figure 2. Frequency of true and false responses on false questions

Analysis

The analysis began by calculating the correct response rates. True and false response frequencies were first aggregated for each version of each test, then aggregated for each test. The summary statistics for each version and each test are given on Table 3. Lastly, the true and false response frequencies were aggregated for true only questions (mean=0.912, variance=0.003), false only questions (mean=0.597, variance=0.030), true variant questions (mean=0.819, variance=0.018) and false variant questions (mean=0.593, variance=0.049).

Table 3. Correct Response Rate Summary Statistics for Four Versions of Four Tests

Version	Operation Test		Design I Test		Design II Test		Planning Test	
	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
A	0.744	0.032	0.675	0.095	0.715	0.072	0.672	0.054
B	0.686	0.050	0.793	0.056	0.871	0.019	0.754	0.030
C	0.675	0.081	0.783	0.051	0.721	0.045	0.719	0.037
D	0.636	0.083	0.629	0.110	0.783	0.039	0.687	0.069
All	0.690	0.060	0.719	0.051	0.774	0.037	0.696	0.037

To test whether there were differences in the correct response rate between different versions of a test, a single factor ANOVA was conducted for each of the four tests. The null hypothesis was that the mean correct response rates for each version of a test were the same. The alternative hypothesis was that the mean correct response rate for at least one version of a test was different. The four versions of the Operations test were found to compare (p=0.81), the four versions of the Design I test were found to compare (p=0.48), the four versions of the Design II test were found to compare (p=0.32), and the four versions of the Planning test were found to compare (p=0.74), at a 0.05 level of significance.

To test whether there were differences between the correct response rates of the true only questions and true variant questions, the response rates were first plotted on Figure 3 and then analyzed using a single factor ANOVA. The correct response rates for the true only questions ranged from 0.83 to 0.98 while the range for the true variant questions was from 0.50 to 1.00. The means were found to compare at the 0.05 level of significance (p=0.14).

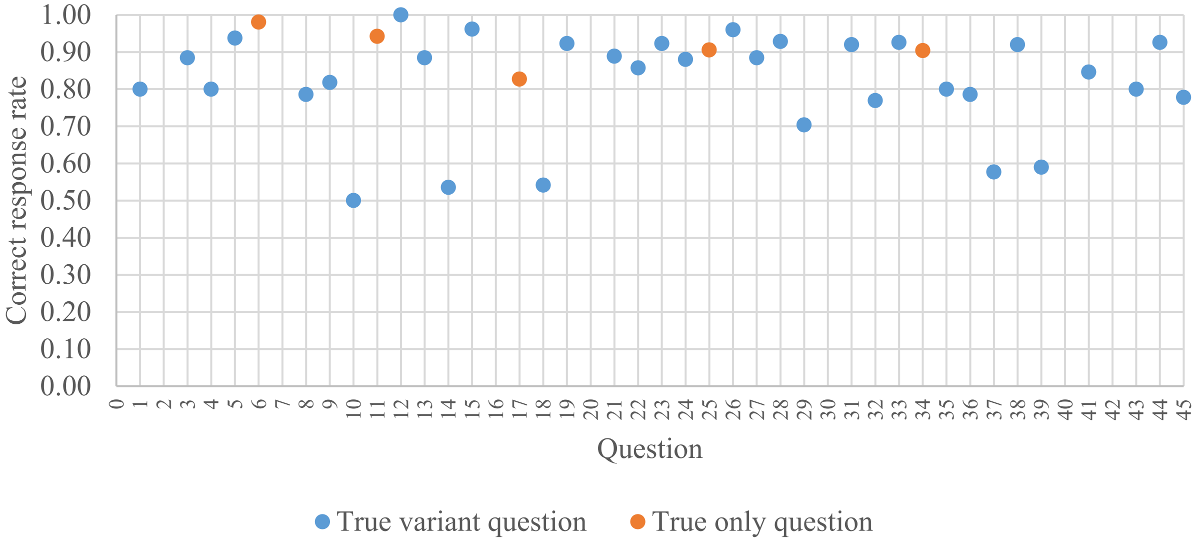

Figure 3. Correct response rates on true only and true variant questions

Similarly, to test whether there were differences between the correct response rates for the false only questions and the false variant questions, the response rates were first plotted on Figure 4 and then analyzed using a single factor ANOVA. The correct response rates for the false only questions ranged from 0.31 to 0.77 while the range for the false variant questions was from 0.17 to 0.93. The means were found to compare at the 0.05 level of significance (p=0.96).

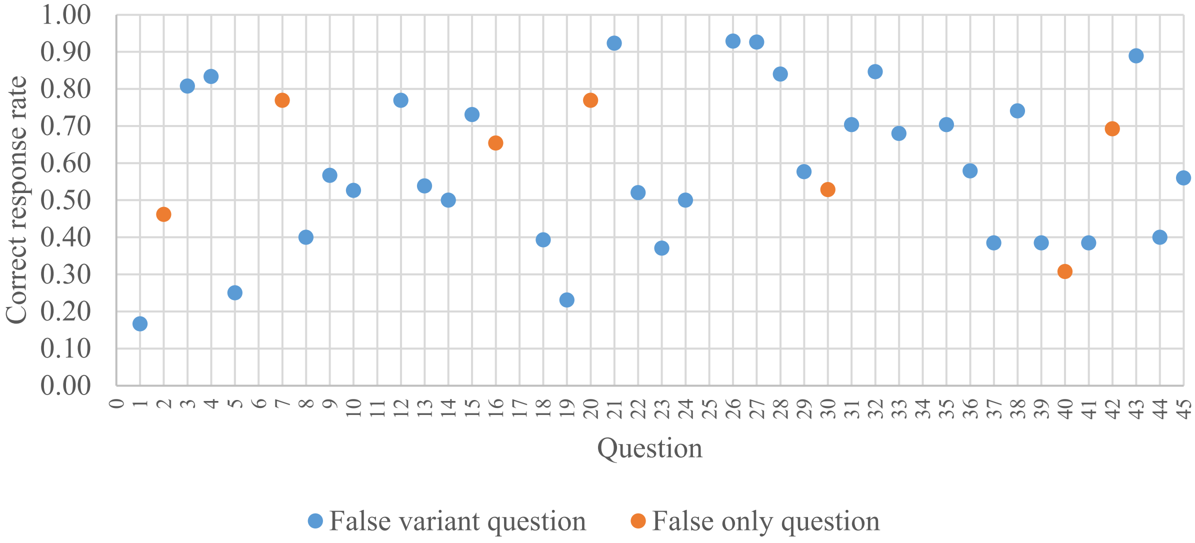

Figure 4. Correct response rates on false only and false variant questions

To test whether there were differences between the correct response rates of the true variant questions and the false variant questions, the Chi-squared test for differences in probabilities between two random samples was applied to each of the 33 questions that had a true variant and false variant. For each question, the number of correct responses and incorrect responses for the true variant and the false variant were tabulated as a 2 x 2 contingency table as shown in Table 4. Using the two-tail test, the null hypothesis was that the response rates for the true variant and the false variant were the same and the alternate hypothesis was that they were statistically different. For each test, the test statistic was calculated as

$$T_1 = \frac{\sqrt{N}(O_{11}O_{22} - O_{12}O_{21})}{\sqrt{n_1 n_2 C_1 C_2}}$$

and then compared to the standard normal distribution with α=0.05. The null hypothesis was rejected when the calculated T_1 value was less than -1.96 or greater than 1.96.

Table 4. 2 x 2 Contingency Table

	Correct response	Incorrect response	Totals
True statement	O_{11}	O_{12}	n_1
False statement	O_{21}	O_{22}	n_2
Totals	C_1	C_2	N

The results of the 33 Chi-squared tests are shown on Figure 5. H_o was rejected for questions 1, 5, 8, 12, 13, 15, 19, 22, 23, 24, 31, 33, 41 and 44. The test statistic values for these questions indicated that the correct response rates were greater for the true variant than the false variant.

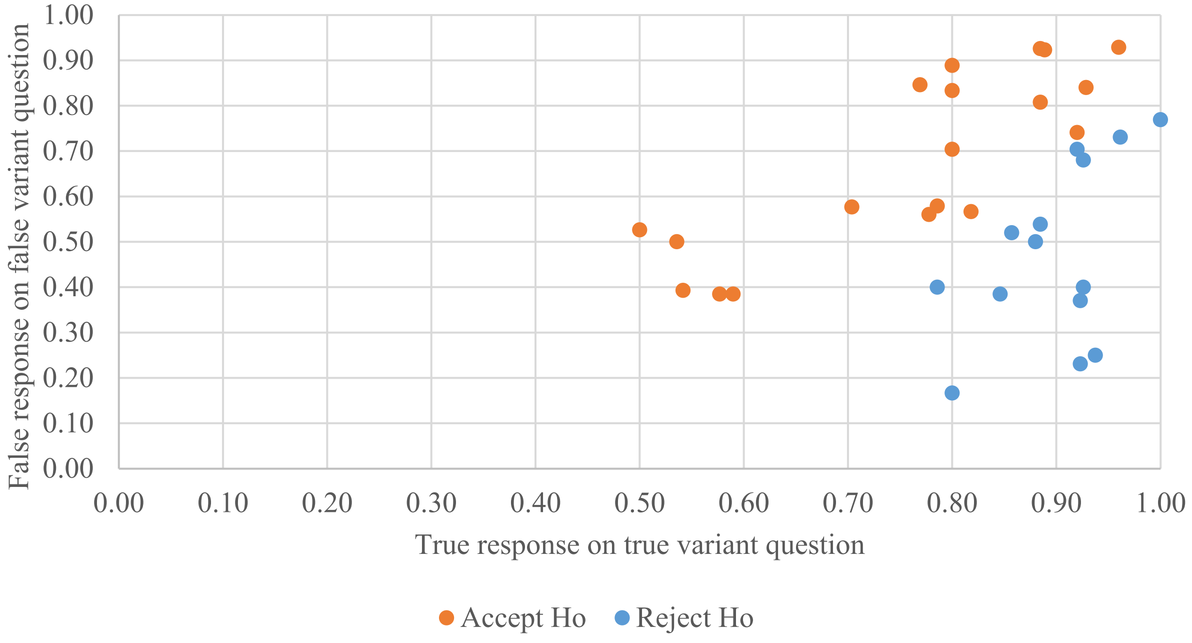

Figure 5. Comparison of correct response rates

Overall, the analysis results indicate that: 1) for each of the four tests, the mean correct response rate for the four versions were comparable; 2) the mean correct response rate of the true only questions was comparable to that of the true variant questions; 3) the mean correct response rate of the false only questions was comparable to that of the false variant questions; and 4) the correct response rates for the true variant questions were either comparable to or greater than the correct response rates for the false variant questions.

Conclusion and Discussion

The analysis indicated that students performed similarly on true only and true variant questions and that students performed similarly on false only and false variant questions. These results are not surprising. When approaching a true variant question, students were not aware of the existence of an associated false variant question, and vice versa. Hence, the conclusion to be drawn is that the true variant questions appeared no different than the true only questions, and the false variant questions appeared no different than the false only questions. Perhaps this result also provides some face validity regarding the rules used to write the true and false questions.

The analysis also indicated that students performed somewhat better on true variant questions than on false variant questions. Given that the performance on true only and true variant questions was comparable and that the performance on false only and false variant questions were comparable, the difference between the performance on true variant and false variant questions was not the writing of the questions. The difference is likely the level of knowledge and/or the strategy taken when guessing an answer. The result aligns with the sentiment from Sax [2], that test takers are more likely to choose true for questions they do not know the answer to. In other words, test takers are more likely to respond true to a true statement than false to a false statement.

The most important finding from the analysis was that for each of the tests, the performance on all four versions was comparable. This means that true variant and false variant questions can be switched between versions of a test but with one caveat; the proportion of true questions to false questions needs to be consistent. If the proportion of true to false questions changes, then the version with the greater proportion of true variants may experience improved performance.

References

[1] D. A. Frisbie and D. F Becker, "An analysis of textbook advice about true-false tests." *Applied Measurement in Education*, 4 (1), 1991, p. 67-83.

[2] G. Sax, *Principles of educational and psychological measurement and evaluation* (3rd ed.) Belmont, CA, Wadsworth, 1989, 678 p.

[3] N. E. Gronlund and R. L. Linn, *Measurement and evaluation in teaching* (6th ed.) New York, Macmillin, 1990, 530 p.

[4] R. F. Burton, Quantifying the effects of chance in multiple choice and true/false tests: question selection and guessing of answers, *Assessment & Evaluation in Higher Education*, 26 (1), 2001, p. 41-50.

[5] R. F. Burton and D. J. Miller, Statistical modelling of multiple-choice and true/false tests: ways of considering, and of reducing, the uncertainties attributable to guessing. *Assessment & Evaluation in Higher Education*, 24(4), 1999, p. 399-411.

[6] J. Wang, Critical values of guessing on true-false and multiple-choice tests, *Education* 116 (1) 2001, p. 153-158.

[7] F. Reid, An alternative scoring formula for multiple-choice and true-false tests, *The Journal of Educational Research*, , 2001, p. 335-339.

[8] L.W. Lackey and W. J. Lackey, Influence of true/false tests and first language on engineering students' test scores. *Journal of Engineering Education*, January, 2002, p. 25-32.