

The Impact of the “all-of-the-above” Option and Student Ability on Multiple Choice Tests

Yi-Min Huang

University of Washington
Seattle, Washington, USA
chym@engr.washington.edu

Michael Trevisan

Washington State University
Pullman, Washington, USA
Trevisan@mail.wsu.edu

Andrew Storfer

Washington State University
Pullman, Washington, USA
astorfer@wsu.edu

Abstract

Despite the prevalence of multiple choice items in educational testing, there is a dearth of empirical evidence for multiple choice item writing rules. The purpose of this study was to expand the base of empirical evidence by examining the use of the “all-of-the-above” option in a multiple choice examination in order to assess how different student ability groups would respond to this particular alternative. Ten experimentally manipulated items were generated with “all-of-the-above” as one of the options and were incorporated into three different test formats. Test formats were randomly distributed to university students in the study. The test scores in these test formats were compared as well as the experimentally manipulated items. Results showed that when “all-of-the-above” is used as the correct answer, the item is more difficult for all students, despite the literature assumption that it provides a cueing effect to students. Research findings corroborate literature assumptions that high ability students score significantly higher than other ability students in this type of option.

Introduction

Multiple-choice items (MCQ) remain the most widely and commonly used item format. The reasons are straightforward. In comparison to other item formats, the lower cost and efficiency in using and storing MCQ items is simply too compelling to disregard. In addition, more MCQ items can be administered in a given time frame than any other item format. As a consequence, the reliability of the test data can be increased, better content sampling obtained, and validity improved (Haladyna & Downing, 1989; Trevisan, Sax, and Michael, 1991; 1994). Today, with the importance and significance policy makers and educators place on multiple-choice items in large-scale K-12 achievement tests, college entrance examinations, and certification tests for example, the demand for well-written items will remain high.

Thorndike (1967) stated that constructing good test items is perhaps the most demanding type of creative writing. Haladyna and Downing (1989a) argue that the essence to a good multiple-choice item lies within good item-writing skills. Haladyna (1999) added that the

process of creating good test items also requires a deep understanding of test material content, the type of mental behavior intended, the choice of an item format and the skill in actually writing the item.

Scholars in the educational measurement profession have argued for empirical research to form the foundation for sound item writing. For decades however, authors have noted the lack of scholarly work in this field (Cronbach, 1970; Ebel, 1951; Haladyna, 1999; McDonald, 2002; Nitko, 1984; Thorndike, 1967; Wesman, 1971; Wood, 1977). The lack of empirical research involving item writing is due in part to recognition of the difficulties involved in such research (Ebel, 1951; Rodriguez, 1997; Wood, 1977). In lieu of empirically based item writing rules, rules of MCQ item writing have been merely passed down by experts in the field to novice test writers based on opinion, experience, and knowledge. This practice continues today.

Haladyna and Downing (1989) examined popular measurement textbooks to establish a taxonomy of item writing rules. The authors also conducted a literature search to identify empirical studies that examined the validity of these rules. The taxonomy allows for the evaluation of the state of evidence for item writing rules, and makes apparent, gaps in the literature. Further empirical work to bolster the evidentiary base for item writing rules is also offered.

In 2002, Haladyna, Downing and Rodriguez offered a taxonomy of MCQ item writing rules focused on classroom assessment. The Haladyna and Downing (1989) taxonomy was revised to account for empirical work conducted since 1989 and tailored to account for factors pertinent for the classroom. The authors discuss ramifications of the taxonomy for large-scale testing.

All-of-the-above

One item writing rule found in the literature is the all-of-the-above (AOTA) option. Sometimes referred to as a complex item type, measurement experts offer conflicting recommendations for use of AOTA. One line of reasoning for use of AOTA is that the item format tends to be more difficult than standard MCQ items (Dudycha and Carpenter, 1973) and can therefore, better discriminate between low and high achievers. In direct contrast to the aforementioned line of reasoning, some argue that the AOTA format tends to be easier for test-wise students (Harasym, Leong, Violato, Brant, and Lorscheider, 1998; Haladyna, Downing and Rodriguez,, 2002). This is thought to occur in two ways. First, students who can identify at least one option that is incorrect and with this knowledge, logically eliminate the AOTA option, will find this item format easier than others who cannot. And second, students who can identify at least two options as correct, and then wager that the AOTA option is likely the correct answer, will also find the item format easier than others who cannot.

Authors that argue the AOTA format provides cueing effect for test-wise students recommend against the use of AOTA. Others however, argue for limited use, suggesting that when the correct answer is AOTA, its use is warranted. In addition, some measurement experts do not differentiate between the use of none of the above (NOTA) and AOTA, while others make the distinction. Also, no study has differentiated between AOTA as the correct answer and AOTA as a distractor.

There are only four empirical studies found in the literature that investigated the impact this item writing rule has on various psychometric properties of items, such as difficulty and discrimination. Each study maintains a different rationale for investigation and assumptions about the use of AOTA. Each is described below.

The first study was done by Hughes and Trimble (1965). The authors' line of reasoning for the study was that use of what they referred to as complex item types are more difficult for students. The authors experimentally tested three types of complex items, items that included one of the following options: (1) both 1 and 2 are correct, (2) none of the above is correct, and (3) all of the above are correct. Statistically significant differences were found compared to the control group, with slightly lower mean test scores for the experimental tests. Item difficulty indices were slightly lower for all three experimental tests. No impact on item discrimination was found. The authors tentatively suggest that complex item formats, including use of AOTA, increase the difficulty of the item. The authors suggest that student knowledge may have influenced the findings.

There are design limitations to the study that diminish the validity of its findings. In particular, small sample sizes and small numbers of items were used in this study. The authors did not mention how group assignment was made. Thus, randomization is in doubt. In addition, no experimental comparison of complex items when AOTA, for example, was used as the correct alternative or as a distractor was made. The design confounded AOTA and NOTA and therefore, the authors cannot offer definitive statements about impact attributed to use of AOTA.

Dudycha and Carpenter (1973) investigated the effect on item difficulty and discrimination of what they refer to as "inclusive items" (p. 116). Inclusive items have AOTA or none of the above (NOTA) as an option. Using a repeated measures design they found statistically significant differences between items with an inclusive option and those that do not have inclusive options on item difficulty indices. Items with an inclusive option tended to be more difficult. In addition, statistically significant differences were also found between these two types of items on item discrimination indices (point biserials). The authors conjectured that inclusive items require more cognitively from students and as a consequence, the difficulty indices are smaller in magnitude. However, they did not formally test this idea. Given the findings, the authors recommend against the use of AOTA (or NOTA). The authors could not explain the impact on discrimination.

Two issues with this study make the findings problematic regarding the AOTA item writing rule. One, the authors did not differentiate between AOTA or NOTA. As a consequence, it is not possible to determine the unique impact of AOTA items on item difficulty or discrimination. Two, the authors did not differentiate between using the option as the correct response and using the option as a distractor.

Mueller (1975) compared item difficulty and discrimination indices for the complex item types investigated by Hughes and Trimble (1965), as well as with items with substantive responses only. Mueller (1975) found AOTA items to be slightly less difficult than the other item types, particularly when AOTA was keyed as the correct answer. The author qualified the findings by stating that the AOTA option was over represented in the study. No impact on discrimination was found. This was a descriptive study, rather than experimental. Thus,

statements of effect are not possible. The authors did not make a definitive recommendation concerning use of AOTA.

Perhaps the most compelling study to date is that done by Harasym, Leong, Violato, Brant, and Lorscheider (1998). The researchers investigated the impact of AOTA on student performance, item discrimination, and test score reliability and validity. A major focus of the study was to compare AOTA to identical multiple true-false (MTF) items. The items were scored with an innovative software package that allows for more than one correct answer within an item. The author found large differences in test performance, favoring the test format that included AOTA items. By examining the frequency that students chose various distractors and comparing test performance between AOTA items and non-AOTA items, the authors argued that cueing was in part a cause for differences in test performance. As a consequence, the authors recommend against the use of AOTA.

For this study, the authors did not contrast AOTA with a standard or conventional item type. The contrast was with the MTF format. The MTF format, though promising, has seen little work in the literature. Thus, integrating these findings with the other studies on AOTA is problematic. The authors investigated the use of AOTA as a correct response only, employing AOTA as an incorrect response only to mask the presence of the experimentally manipulated use of AOTA as the correct response. While the argument was made that cueing was a likely cause for differences in test performance, the authors also stated that student knowledge likely played a part in this difference. However, student knowledge was held constant in the study.

Collectively, these studies provided mixed results and recommendations concerning the use of AOTA. The different designs, study limitations, and contrasts do not allow for definitive or careful statements about the use of AOTA. What has emerged from these studies is that student knowledge or ability is a factor in the use of AOTA, yet no study systematically investigated this possibility. In addition, there may be differences in item performance when AOTA is the correct answer versus when it is used as a distractor. To date, no study has systematically investigated this possibility either.

Haladyna and Downing (1989) suggested that the use of AOTA was controversial and that empirical work to date, did not allow for definitive statements about its use. The authors recommended further work. Haladyna, Downing and Rodrigues (2002), after reviewing textbooks that focused more on classroom assessment (rather than large-scale testing) and considering the Harasym et al. (1998) study, stated that "we continue to support this guideline to avoid AOTA" (p. 319).

We argue, that given the lack of consensus concerning the use of AOTA among a small number of previous studies, and the consistent call over three decades to build empirical support for item writing rules, further work on AOTA is warranted. The purpose of this study is to examine the impact of AOTA on student performance and item and test characteristics.

This study is an improvement over empirical work represented in the literature by employing student ability as an independent variable, a factor directly mentioned or alluded to in previous studies. In addition, this study is an improvement over previous studies by examining the use of AOTA as both a distractor and the correct answer. The research hypotheses ($p < 0.05$) of this study are stated below:

1. Statistically significant differences exist between ability groups in test format A, the test score order will be High, Average, Low.
2. Statistically significant differences exist between ability groups in test format B, the test score order will be High, Average, Low.
3. Statistically significant differences exist between ability groups in test format C, the test score order will be High, Average, Low.
4. Statistically significant differences exist for low ability students, with the order of test formats A, C, B being favored.
5. Statistically significant differences exist for high ability students, with the order of test formats A, C, B being favored.

Method

A total of 624 college students from a large land-grant university in the Pacific Northwest (USA) gave consent to participate in the study. After the data were sorted and compiled according to the research design, data from 457 students were utilized. The first midterm exam of this introductory class was administered. This test is a 5-option multiple-choice examination. The course instructor developed the test items to form three different test formats based on his professional judgment and lecture materials. These formats are:

1. Test format A – 10 questions with AOTA as the correct response.
2. Test format B – 10 questions with AOTA as the incorrect response (distractor).
3. Test format C – 10 questions with AOTA as either the correct or incorrect response.

By using the class lecture materials and the course professor's professional judgment, three additional correct answers were added to the questions to make the AOTA option the correct alternative for format A. In format B, the AOTA option was used as a distractor. In order to control for cuing effects and test-wiseness, half of the items in format C employed AOTA as the correct answer and the other half were used as a distractor. Test items were further screened for readability, grammar, syntax, and connection to the class content.

The three forms of the multiple-choice test were randomly assigned to individual students; each student received only one form of the test. Students were asked to estimate and mark their semester Grade Point Average (GPA) on their scoring sheet. The GPA estimate was used as a proxy for an ability measure as employed by Green, Sax, and Michael (1982) and Trevisan, Sax, and Michael (1991). Student scores were later categorized into high, average and low ability groups by using the following GPA cutoffs: High (3.7 – 4.0), average (3.0 – 3.4) and low (0.0 – 2.7). The purpose of using this noncontiguous design is to increase power by controlling within group variability and maximizing the spread between groups

(Cronbach & Snow, 1977; Trevisan, Sax & Michael, 1991). Students with GPAs between these defined ability group cutoffs were eliminated and not used in the data analysis.

Results

Table 1 presents the means of total test scores, GPA, p-value and sample sizes for each test format and ability group. The descriptive statistics are presented for the low, average, high and combined ability groups. Table 2 presents the mean total score, GPA, p-value and sample sizes for the ten experimentally manipulated test items in each test format and ability group.

Table 3 presents KR-20s for the 55 item different formats of the test. These internal-consistency estimates are presented as the unadjusted and adjusted KR-20s respectively. Correlations between test scores and GPAs were also calculated and presented (validity coefficient) for each format of the test and ability group.

Table 1

Means (test score, GPA and p-values), number of items, sample sizes, and standard deviations for each test form and ability group.

Test Formats	Mean Score	Mean GPA	Mean p	Number of Items	Sample Size	S
A (AOTA = all correct)						
L	34.46	2.35	0.62	55	57	6.54
A	36.74	3.19	0.67	55	65	6.61
H	42.03	3.84	0.76	55	30	5.87
C	36.93	3.00	0.67	55	152	6.96
B (AOTA = all incorrect)						
L	36.06	2.32	0.65	55	49	7.14
A	37.47	3.17	0.68	55	60	6.45
H	44.54	3.83	0.81	55	41	6.40
C	38.94	3.07	0.71	55	150	7.49
C (AOTA = half correct, half incorrect)						
L	34.40	2.33	0.62	55	53	6.18
A	37.06	3.19	0.67	55	72	5.99
H	43.93	3.82	0.80	55	30	5.03
C	37.48	3.02	0.68	55	155	6.76

Note. L = Low Ability, A = Average Ability, H = High Ability, C = Combined Ability Groups

Table 2

Means (test score for AOTA items, GPA and p-values), number of items, sample sizes, and standard deviations for each test form and ability group.

Test Formats	Mean Score	Mean GPA	Mean p	Number of Items	Sample Size	S
A (AOTA = all correct)						
L	5.70	2.35	0.57	10	57	1.87
A	6.40	3.19	0.64	10	65	1.67
H	7.20	3.84	0.72	10	30	1.37
C	6.30	3.00	0.67	10	152	1.77
B (AOTA = all incorrect)						
L	6.33	2.32	0.63	10	49	1.99
A	7.15	3.17	0.71	10	60	1.74
H	8.12	3.83	0.81	10	41	1.63
C	7.15	3.07	0.71	10	150	1.92
C (AOTA = half correct, half incorrect)						
L	6.21	2.33	0.62	10	53	1.47
A	6.14	3.19	0.61	10	72	1.44
H	7.30	3.82	0.73	10	30	1.51
C	6.39	3.02	0.64	10	155	1.52

Note. L = Low Ability, A = Average Ability, H = High Ability, C = Combined Ability Groups

Table 3

The unadjusted and adjusted KR-20s, and validity coefficients for each test form (all 55 items) and ability group

Test Formats	Adjusted KR-20	Unadjusted KR-20	Validity Coefficient
A (AOTA = correct)			
L	0.73	0.72	0.15
A	0.78	0.78	0.16
H	0.78	0.77	0.23
C	0.81	0.80	0.39
B (AOTA = incorrect)			
L	0.80	0.79	0.00
A	0.75	0.77	0.19
H	0.81	0.81	0.44
C	0.84	0.84	0.41
C (AOTA = half correct, half incorrect)			
L	0.72	0.72	0.03
A	0.74	0.74	0.27
H	0.74	0.73	0.15
C	0.81	0.80	0.45

Note. L = Low Ability, A = Average Ability, H = High Ability, C = Combined Ability Groups

The results of the study showed that statistically significant differences existed between ability groups in Test Format A with $F(2, 149) = 13.62, p < .05$. The results favored high, average, and low ability groups, respectively. This corroborated hypothesis 1. Significant differences were found between ability groups in Test Format B with $F(2, 147) = 20.45, p < .05$, favoring high, average, and low ability students – an outcome that was predicted. The results also showed that significant differences existed between ability groups in Test Format C with $F(2, 152) = 25.47, p < .05$, favoring high, average, and low ability groups, as predicted.

No significant differences were found for low ability students across different test formats. The test score trend in test formats was B, A, C. No significant differences were found for high ability students across different test formats. The test score trend in test formats also favored B, A, C. Both trends were not the ones hypothesized.

Discussion

The present study yielded significant differences among ability groups in Test Format A, B and C. Although the Student-Newman-Keuls multiple comparison procedure confirmed the significant differences between low and high ability groups, and between average and high ability groups for Test Formats A and B results from this procedure showed statistically significant differences among the three ability groups for Test Format C.

The findings of hypotheses 1, 2 and 3 corroborated the assumption found in the literature that high ability students are more likely to succeed in test items that consist of complex alternatives such as the “all-of-the-above” option (i.e. Ebel & Frisbie, 1991; Frary, 1991; Osterlind, 1989).

No statistically significant differences were found among the three test formats for low ability students ($p = .33$). However, the overall test scores for the low ability group showed the order of test formats B, A, C being favored. For the 10 experimentally manipulated items that consisted of AOTA as an alternative, the test score again favored the trend of test formats B, C, A, this trend was not the one hypothesized. Once again, there is no significant difference for the low ability group ($p = .16$). Therefore, research expectations were not confirmed.

There were no statistically significant differences found among the three test formats for high ability students ($p = .20$). The test scores showed the order of test formats B, C, A being favored; this was not the trend hypothesized. For the 10 experimentally manipulated items that included “all-of-the-above” option, the test score favored the order of test formats B, C, A. Consequently, this was not the trend hypothesized. Despite the significant differences between the test formats ($p = .02$) for high ability students, research expectation was not confirmed.

Perhaps partial explanation for the non-significance in the hypotheses can be attributed to the low number of experimentally manipulated items. There was a total of ten items in each test format that included a manipulated "all-of-the-above" option. The low number of the experimental items tends to decrease test reliability. In turn, low reliability decreases power in the design. Considering previous empirical studies, the number of experimentally

manipulated AOTA option items ranged from a low of nine to a high of twenty-six items (e.g. Harasym, et al., 1998; Hughes & Trimble, 1965; Mueller, 1975), although none of these studies statistically compared the findings across ability groups. In addition, these studies did not consider the student ability-item format relations. One recommendation is to increase the number of experimentally manipulated items in similar studies employing this design.

Further explanation of the findings may be obtained by considering the student sample size. Given the GPA cut-offs, student test data were eliminated. Future study should include larger sample sizes. Additional research studies might also examine the optimum GPA cut-offs to increase the power for this type of research design.

Hypotheses 4 and 5 did not corroborate the literature assumption that "all-of-the-above" when used as the correct option would be the easiest alternative. Researchers such as Ebel and Frisbie (1991), Harasym, et al. (1998), Mueller (1975), Osterlind (1989) have all commented on how items containing an "all-of-the-above" alternative would be the least difficult among other complex alternatives. The reason is that there would be unwarranted cues provided to test wise students who recognize that at least two of the options are correct, thereby deducing the correct alternative (Ebel & Frisbie, 1991). In the present study, the results were somewhat contrary. The statistical findings, although not significant, favored Test Format A (where all of the "all-of-the-above" items were correct) as the most difficult for both low and high ability groups. Test Format B (where all of the "all-of-the-above" items were distractors) was regarded as the least difficult. Although statistical significance was not found for the means of the overall test scores, significant differences were found for the 10 manipulated "all-of-the-above" items. It was found that these 10 "all-of-the-above" items were significantly easier in test format B for the high ability group. Perhaps partial explanation for this finding is that there is a greater difference between the test formats for these 10 manipulated items than the overall test items.

Additional item analyses including item difficulty, item discrimination and distractor analysis were also calculated and presented. For the purpose of the current research, the items consisting of "all-of-the-above" option were examined in each test formats. The least difficulty, on the average, was Format B ($p = .71$), where items containing "all-of-the-above" were used as a distractor. The difficulty level between Format A and Format C is equivalent (both $p = .64$). When the scores are broken down between ability groups, test format A and test format C both have comparable item difficulty index. For all ability groups, test format B seemed to be the easiest, with the highest item difficulty indices. For low ability group, the favored trend of test formats appeared to be Format B ($p = .63$), Format C ($p = .61$) and Format A ($p = .57$). For average ability group, the favored trend of test formats are Format B ($p = .70$), Format A ($p = .64$) and Format C ($p = .61$). Finally, for high ability students, the favored trend of test formats are Format B ($p = .81$), Format C ($p = .73$) and Format A ($p = .72$). These results did not follow the hypothesized trends,

however, they do corroborate with Frary (1991) that high ability students are more likely to succeed in complex alternatives (such as “all-of-the-above”, “none-of-the-above” options) versus low ability students.

A possible explanation for the observed item difficulty trend (test formats B, C, and A) can be attributed to the different alternatives presented in each of the different test formats. For the purpose of this present study, the 10 experimental items for each test formats were examined. A distractor analysis was conducted and the proportion of students choosing a particular option in each item was reported. It was found that for test format A, two items have a highly selected distractor; these distractors should be further examined to eliminate the discrepancy between the option selections. Also, an item with negative stem in the question should be avoided (Haladyna, Downing & Rodriguez, 2002). For test format B, one specific item needed to be further examined, because two of the distractors were not selected at all by any students. These distractors were not good competitors for the correct answer and should be reexamined. Finally for test format C, further similar improvements described above in test format A and B should be reconsidered.

For the present study, the items consisting of “all-of-the-above” option were examined in each test format. The highest mean discrimination index occurred when “all-of-the-above” option is used as a distractor in Format B ($r_{pbis} = .32$). There were little differences between Format A where “all-of-the-above” is always the correct answer and Format C where “all-of-the-above” is half correct answer, half the distractor. Therefore, Format A ($r_{pbis} = .21$) discriminates as well as Format C ($r_{pbis} = .24$). Although all the mean item discrimination indices have at least a “fair” standing according to the discrimination scheme developed by University of Washington, some indicated “good” discrimination indices. These were: (a) Format B for low ability group; (b) Format B in high ability group; (C) Format C in high ability group.

The purpose of this study was to examine the use of the “all-of-the-above” option and student ability in multiple choice testing. The present study also represents one of the five empirical verifications of the option “all-of-the-above”. The first three results in this study confirmed research hypotheses which indicated that high ability students will do better than average and low ability students in complex alternative questions. However, this study did not corroborate previous literature’s assumption that indicated when “all-of-the-above” option is used as the correct answer, it should be easier than when it is used as a distractor (i.e. Ebel & Frisbie, 1991; Osterlind, 1989). Consequently, hypotheses four and five were not supported by the current research findings. Our study findings warrant the following recommendations:

1. Conduct additional studies that compare the use of the “all-of-the-above” option as the correct answer and the distractor.
2. Increase the experimentally manipulated item size, which incorporated the “all-of-the-above” options.
3. Conduct additional studies with officially recorded student ability measures.
4. The GPA cutoffs were based on previous empirical studies (i.e. Trevisan, Sax and Michael, 1991). Additional studies are recommended to establish the optimum GPA

cutoffs on student ability measure in order to increase the power of this type of noncontiguous research design.

5. Increase the sample size in future studies.
6. Broaden populations sampled to include pre-college students.

Although the results of this study did not confirm all research hypotheses, the study supported the existing literature assumption that the high ability students were more likely to score an item with “all-of-the-above” option correctly (e.g. Ebel & Frisbie, 1991; Frary, 1991; Osterlind, 1989). However, items with “all-of-the-above” as the correct answer appear to be more difficult than when it is used as a distractor. These findings contradicted the assumption in the literature that when “all-of-the-above” is used as a correct answer; it is apparently easy for students to select this option if they identify two or more correct options (e.g. Ebel & Frisbie, 1991; Harasym, et al., 1998; Muller, 1975; Osterlind, 1989).

Educational scholars continue to emphasize the need to align content standards, classroom instructions, classroom assessment and high-stakes testing in order to promote student learning (Haladyna, Downing, and Rodriguez, 2002; Pellegrino, Baxter, and Glaser, 1999; Snow and Mandinach, 1991). Current educators who are engaged in writing MCQ will need to consider the implications of using this AOTA writing rule in order to provide quality assessment for students. For those who are interested in examining these item writing guidelines, these recommendations should be taken into considerations.

References

- Cronbach, L. J. (1970). [Review of the book *On the theory of achievement test items*]. *Psychometrika*, 35, 509-511.
- Dudycha, A.L., & Carpenter, J. B. (1973). Effects of item formats on item discrimination and difficulty. *Journal of Applied Psychology*, 58, 116-121.
- Ebel, R. L. (1951). Writing the test item. In E. F. Lindquist (Ed.), *Educational Measurement*, American Council on Education, Washington, DC.
- Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement*. (5th ed.). Englewood, NJ: Prentice Hall.
- Frary, R. B. (1991). The none-of-the-above option: An empirical study. *Applied Measurement in Education*, 4, 115-124.
- Green, K., Sax, G., & Michael, W. B. (1982). Validity and reliability of tests having different numbers of options for students of differing levels of ability. *Educational and Psychological Measurement*, 42, 239-245.

International Journal for the Scholarship of Teaching and Learning
<http://www.georgiasouthern.edu/ijstol>
 Vol. 1, No. 2 (July 2007)
 ISSN 1931-4744 © Georgia Southern University

Haladyna, T. M. (1999). *Developing and Validating Multiple-Choice Test Items* (2nd ed.). New Jersey: Lawrence Erlbaum Associates, Inc.

Haladyna, T. M., & Downing, S. M. (1989a). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, 2 (1), 37-50.

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 (3), 309-334.

Harasym, P. H., Leong, E. J., Violato, C., Brant, R., & Lorscheider, F. L. (1998). Cuing effect of "all of the above" on the reliability and validity of multiple-choice test items. *Evaluation and the Health Professions*, 21, 120-133.

Hughes, H. H., & Trimble, W. E. (1965). The use of complex alternatives in multiple-choice items. *Educational and Psychological Measurement*, 25, 117-126.

McDonald, M. E. (2002). *Systematic assessment of learning outcomes: Developing multiple-choice exams*. Boston: Jones and Bartlett Publishers.

Mueller, D. J. (1975). An assessment of the effectiveness of complex alternatives in multiple choice achievement test items. *Educational and Psychological Measurement*, 35, 135-141.

Nitko, A. J. (1984). [Review of the book *A technology for test item writing*]. *Journal of Educational Measurement*, 21, 201-204.

Osterlind, S. J. (1989). *Constructing test items*. Norwell, MA: Kluwer Academic.

Pellegrino, J. W., Baxter, G. P., and Glaser, R. (1999). Addressing the "Two Disciplines" problems: Linking theories of cognition and learning with assessment and instructional practice. *Review of Research in Education*, 24, 307-353.

Rodriguez, M. C. (1997, April). *The art and science of item-writing: A meta-analysis of multiple-choice item format effects*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Rudner, L. M., & Schafer, W. D. (2002). *What teachers need to know about assessment*. Washington D.C.: National Education Association.

Snow, R. E., and Mandinach, E. B. (1991). *Integrating assessment and instruction: A research development agenda*. Princeton: Educational Testing Services.

Thorndike, R. L. (1967). The analysis and selection of test items. In S. Messick & D. Jackson (Eds.), *Problems in Human Assessment*. New York: McGraw-Hill.

Trevisan, M. S., Sax, G., & Michael, W. B. (1991). The effects of the number of options per item and student ability on test validity and reliability. *Educational and Psychological Measurement*, 51, 829- 837.

International Journal for the Scholarship of Teaching and Learning

<http://www.georgiasouthern.edu/ijstol>

Vol. 1, No. 2 (July 2007)

ISSN 1931-4744 © Georgia Southern University

University of Washington (2002). *Scorepak: Item analysis*. Retrieved April 13th, 2004 from <http://www.washington.edu/oea/item.htm>

Wesman, A. G. (1971). Writing the test item. In Thorndike, R. L. (Ed.), *Educational Measurement*, American Council on Education, Washington, DC.

Wood, R. (1977). Multiple choice: A state of the art report. *Evaluation in Education: International Progress*, 1, 191-280.