The present study involved the testing of two common multiple-choice item writing rules. A recent review of research revealed that much of the advice given for writing multiple-choice test items is based on experience and wisdom rather than on empirical research. The rules assessed in this study include: (1) the phrasing of the stem in the form of a question versus a partial sentence; and (2) the use of the inclusive "none of the above" option instead of a specific content option. Limited empirical research suggests that using the partial sentence format and the inclusive "none of the above" option may lead to undesirable item and test characteristics, while textbook authors essentially are divided on their opinions about the validity of each rule. The items used in this study were from the instructor's manual for D. Myer's (1986) text entitled "Psychology." Items were randomly assigned to be rewritten to reflect the experimental conditions under investigation. Two instructors of an introductory psychology course selected 32 multiple-choice items for the study. The rewritten tests were administered to 228 students enrolled in two sections of an introductory psychology class. About half of the students in each section received Form A and the other half received Form B, resulting in 115 Form A and 113 Form B responses. The same manipulated items were combined with 18 different non-manipulated items in a third section of the class to comprise Forms C and D, whose administration resulted in 59 Form C and 59 Form D responses. Results offer no evidence to support the use of either type of stem and limited evidence to caution against use of the "none of the above" option. Two data tables and examples of the four item formats used are provided. (TJH)
The Validity of Two Item-Writing Rules

Kevin Crehan
University of Nevada, Las Vegas
and
Thomas M. Haladyna
Arizona State University West Campus
The Validity of Two Item-Writing Rules

ABSTRACT

A recent review of research revealed that much of the advice given for writing multiple-choice test items is based on experience and wisdom rather than empirical research. The present study involved the testing of two common item writing rules: (1) the phrasing of the stem in the form of a question versus a partial sentence and (2) the use of the inclusive "none of the above" option instead of a specific content option. Limited empirical research suggests that using the partial sentence format and the inclusive 'none of these' option may lead to undesirable item and test characteristics, while textbook authors essentially are divided on their opinions about the validity of each rule. Results of this experimental study offer no evidence to support the use of either type of stem and limited evidence to caution against use the option "none of the above."
The Validity of Two Item-Writing Rules

A number of writers in the field of educational measurement have commented that multiple-choice (MC) item writing, despite its widespread popularity and use, has received little scholarly attention in the past (Cronbach, 1970; Ebel, 1951; Millman & Green, in press; Nitko, 1984; Roid and Haladyna, 1982; Wesman, 1971; Wood, 1977). In a review of empirical research on item writing, Haladyna and Downing (1989a) reported finding 96 empirical studies of which 53 dealt with only two item-writing practices, the optimal number of options and the desirability of key balancing. Most item-writing rules have been studied fewer than 10 times. Thus, the empirical foundation for the validity of many item-writing rules is weak, and the basis for many rules is often authoritative wisdom passed on through textbooks and other professional publications and presentations.

The study reported here addresses two item-writing rules which are popularly prescribed in treatments on MC item writing in textbooks and other sources in the educational measurement literature (Haladyna & Downing, 1989b). The first rule is: "Don't use 'none of the above' as an option"; the second rule is "Use either the question format or the completion format when phrasing the stem."

None of the Above

In a review of 46 references dealing with the topic MC
item writing, Haladyna & Downing (1989b) found that 34 (73%) of these references stated support or lack of support for the "Don't use 'none of the above' as an option" rule. This was the tenth most often mentioned rule, and this survey was taken as evidence of the importance of the rule for item writers. However, authors were divided on their support for this rule, with 19 for and 15 against. Obviously some controversy exists in the validity of the rule.

Empirical research on this item writing rule has been limited to only ten studies (Boynette, 1950; Dudycha & Carpenter, 1973; Forsyth & Spratt, 1980; Hughes & Trimble, 1965; Mueller, 1975; Oosterhof & Coats, 1984; Rimland, 1960; Schmeiser & Whitney, 1975; Wesman & Bennett, 1946; Williamson & Hopkins, 1967). All of these studies involved the item characteristic of difficulty, but only five studied item discrimination and reliability, and only two validity. In all instances, the use of "none of the above" option made items more difficult, the mean effect across nine studies where results were aggregable was 4.8%. With discrimination, avoiding the inclusive "none of the above" option made items slightly more discriminating, .03, while reliability was improved by a factor of .04.

**Question Format Versus Completion Format**

One of the most fundamental requirements in MC item writing is that one states the item in a question format or a completion format. On the surface there appears to be no reason
to challenge either format. According to Haladyna & Downing (1989a), the rule is one of the most common given in treatments on MC item writing, 41 of 46 references mentioned it, and all 41 support the use of either format. Paradoxically, the small body of empirical research leads to the opposite conclusion.

Studies of this item writing rule include: Board and Whitney (1972), Dudycha & Carpenter (1973), Dunn & Goldstein (1959), Schmeiser & Whitney (1975a; 1975b), and Schrock & Mueller (1982). These six studies observed effects on item difficulty in each instance, discrimination in three cases, reliability four times, and validity twice. In general, the question format appears to have an advantage over the sentence completion format with respect to making items slightly easier, having little or no effect on item discrimination, and making test scores based on such items more reliable and valid. For reliability, the improvement was a median .065, which is a reduction of 6.5% error variance in test scores. Validity was improved by .06 in two studies (Board & Whitney, 1972; Schmeiser & Whitney, 1975b). Based on these few studies, it appears the evidence favors the use of the question format over the completion format in phrasing the MC stem.

The present study further investigates these two item-writing rules.
METHOD

The items used in this study were from the instructor's manual for Myer's (1986) text entitled *Psychology*. Two instructors of an introductory psychology course selected 32 MC items for the study. Each item was keyed to the objectives of the course and met the standard requirements for MC item writing. Each item also had adequate performance characteristics as judged from previous uses. Items were randomly assigned to be rewritten to reflect the experimental manipulations as outlined below:

<table>
<thead>
<tr>
<th>No. of Items</th>
<th>Version 1</th>
<th>Version 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>8 completion</td>
<td>completion</td>
<td>completion</td>
</tr>
<tr>
<td>option 'e' (CE)</td>
<td>none of these (CN)</td>
<td></td>
</tr>
<tr>
<td>8 question</td>
<td>completion</td>
<td>completion</td>
</tr>
<tr>
<td>option 'e' (QE)</td>
<td>option 'e' (CE)</td>
<td></td>
</tr>
<tr>
<td>8 question</td>
<td>question</td>
<td>question</td>
</tr>
<tr>
<td>none of these (QN)</td>
<td>option 'e' (QE)</td>
<td></td>
</tr>
<tr>
<td>8 completion</td>
<td>question</td>
<td>question</td>
</tr>
<tr>
<td>none of these (CN)</td>
<td>none of these (QN)</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1 provides an example of one item written in all four variations.

-------------------------

Insert Figure Inabout here

-------------------------
The manipulations were balanced both within and between the two versions. Version 1 items were combined with eighteen non-manipulated items to comprise Form A of the final exam for two sections of an introductory psychology class while Version 2 items were combined with the same eighteen items to comprise Form B. Test forms were key balanced with the option 'none of these' being keyed three times in sixteen appearances or approximately one-fifth of the time.

The tests were administered to two sections of the class with approximately one-half the students in each section receiving Form A and the other half receiving Form B resulting in 115 Form A and 113 Form B responses. In addition, the same manipulated items were combined with eighteen different non-manipulated items in a third section of the class to comprise Forms C and D. Forms were key balanced as above and test administration in this class resulted in 59 Form C and 59 Form D responses.

This design was chosen to allow comparison of item format manipulations controlling for examinee ability. That is, when Version 1 CE items are combined with Version 2 QE items, we have sixteen items not employing the option 'none of these'. When Version 1 QM items are combined with Version 2 CN items we have these same sixteen items employing the option 'none of these'. Item characteristics can be compared between these sixteen item
sets since all subjects in the study responded to one or the other of the eight item subscales under each condition. Since, at best, small effect sizes were anticipated hypothesis testing was conducted with alpha set at the .10 level for each statistical test.

RESULTS

Table 1 presents the means and standard deviations of item difficulties, mean point-biserials and the Kuder-Richardson 20 reliability estimates of each subscale for the four forms of the test.

Insert table 1 about here

In order to test for differences in difficulty and discrimination for the question versus completion format item statistics for the Form A-QE items were combined with item statistics for the Form B-QN items and were compared to the Form A-CN items combined with the Form B-CE items. Similarly item statistics for the same item types on Forms C and D were combined. In order to test for differences in difficulty and discrimination for the inclusive versus specific option hypothesis item statistics for Form A-CE items were combined with Form B-QE and were compared to the Form A-QN items combined with
Form B-CN items. Similarly, item statistics for the same item types were combined on Forms C and D. Summary statistics for the combined items are presented in Table 2.

---------------------
Insert Table 2 about here
---------------------

DIFFICULTY

The observed difference in difficulty was .02 higher for the question format. A correlated one-tailed t-test showed non-significance at the .10 level (t = .56, df = 15, r = .70, p = .29). The t-test for the same comparison on Forms C and D showed similar results with a mean difference of .003 and a non-significant t statistic (t = .10, df = 15, r = .76, p = .46). Differences between using and not using the option 'none of these' was tested by combining Form A CE with Form B QE item difficulties and comparing these with Form A Q and Form B C item difficulties. The difference in mean difficulty was .027 with use of 'none of these' being lower. The dependent t-test was significant at the .1 level (t = 1.44, df = 15, r = .916, p = .085). The same test for Forms C and D had similar results with a mean difference of .043 (t = 1.59, df = 15, r = .67, p = .065).

DISCRIMINATION

Differences in mean point-biserials between the question and
completion formats were non-significant for both replications. Differences in mean point-biserials between using and not using the inclusive 'none of these' option were .034 and .033 for Form A vs Form B and Form C vs Form D respectively and favored not using the inclusive option in both instances. The observed differences, however, failed to reach significance at the .10 level. The correlated t-tests for Form A versus Form B and Form C versus Form D had p values of .18 and .20 respectively.

DISCUSSION

While this study fails to offer support to a recommendation regarding use of either the question or completion format over the other, observed results regarding use of the "none of these" option are consistent with previous findings in direction and magnitude. Differences in difficulty were statistically significant and in 3 to 4% range favoring the specific option over the inclusive option format. Item discriminations were also observed to be slightly over .033 higher for the specific option format. This result, while not statistically significant, is at the same level as observed in previous research. Lack of statistical significance may be attributable to the low power to detect a difference of this magnitude with sixteen subjects (items) and the low correlations between the item discriminations between forms (.183, .488). It is noted that differences in item
discrimination observed in this study are estimated to result in
differences in reliability of about .04 favoring use of the
specific option over use of "none of these". Future research on
this should use the knowledge of this effect size to determine
the sample size necessary to detect a .03 or greater effect with
reasonable power.
Two Item Writing Rules
Page 12

REFERENCES


of these' as an option in test construction. *Journal of Educational Psychology*, 27, 541-549.


### TABLE I

Mean (P) and standard deviation (S) of difficulty indices, mean point-biserials (D) and KR20 reliability (r) for each 8 item subscale across test forms.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Item Type</th>
<th>Form A</th>
<th>Item Type</th>
<th>Form B</th>
<th>Item Type</th>
<th>Form C</th>
<th>Item Type</th>
<th>Form D</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>CE</td>
<td>.705</td>
<td>CN</td>
<td>.636</td>
<td>CE</td>
<td>.733</td>
<td>CN</td>
<td>.634</td>
</tr>
<tr>
<td>S</td>
<td>.180</td>
<td>.209</td>
<td>.136</td>
<td>.132</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>.379</td>
<td>.401</td>
<td>.450</td>
<td>.386</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>r</td>
<td>.535</td>
<td>.591</td>
<td>.619</td>
<td>.627</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>QE</td>
<td>.792</td>
<td>CE</td>
<td>.798</td>
<td>QE</td>
<td>.798</td>
<td>CE</td>
<td>.778</td>
</tr>
<tr>
<td>S</td>
<td>.140</td>
<td>.130</td>
<td>.149</td>
<td>.143</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>.352</td>
<td>.415</td>
<td>.355</td>
<td>.405</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>r</td>
<td>.469</td>
<td>.568</td>
<td>.528</td>
<td>.566</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>QN</td>
<td>.806</td>
<td>QE</td>
<td>.790</td>
<td>QN</td>
<td>.731</td>
<td>QE</td>
<td>.718</td>
</tr>
<tr>
<td>S</td>
<td>.092</td>
<td>.097</td>
<td>.134</td>
<td>.131</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>.319</td>
<td>.409</td>
<td>.401</td>
<td>.402</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>r</td>
<td>.388</td>
<td>.489</td>
<td>.611</td>
<td>.572</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>CN</td>
<td>.622</td>
<td>QN</td>
<td>.666</td>
<td>CN</td>
<td>.667</td>
<td>QN</td>
<td>.653</td>
</tr>
<tr>
<td>S</td>
<td>.217</td>
<td>.158</td>
<td>.202</td>
<td>.192</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>.328</td>
<td>.419</td>
<td>.395</td>
<td>.415</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>r</td>
<td>.385</td>
<td>.560</td>
<td>.549</td>
<td>.539</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Means and standard deviations for item difficulties and discriminations, and with estimated reliability on the combined sixteen item scales for each item type:

<table>
<thead>
<tr>
<th>Forms Type</th>
<th>Item Type</th>
<th>Mean Diff.</th>
<th>Standard Deviation</th>
<th>Mean Disc.</th>
<th>Standard Deviation</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>A&amp;B</td>
<td>Q</td>
<td>.729</td>
<td>.159</td>
<td>.386</td>
<td>.114</td>
<td>.74</td>
</tr>
<tr>
<td>A&amp;B</td>
<td>C</td>
<td>.710</td>
<td>.195</td>
<td>.371</td>
<td>.106</td>
<td>.72</td>
</tr>
<tr>
<td>C&amp;D</td>
<td>Q</td>
<td>.725</td>
<td>.181</td>
<td>.384</td>
<td>.124</td>
<td>.74</td>
</tr>
<tr>
<td>C&amp;D</td>
<td>C</td>
<td>.722</td>
<td>.178</td>
<td>.400</td>
<td>.142</td>
<td>.75</td>
</tr>
<tr>
<td>A&amp;B</td>
<td>E</td>
<td>.743</td>
<td>.144</td>
<td>.394</td>
<td>.107</td>
<td>.75</td>
</tr>
<tr>
<td>A&amp;B</td>
<td>N</td>
<td>.720</td>
<td>.179</td>
<td>.360</td>
<td>.117</td>
<td>.70</td>
</tr>
<tr>
<td>C&amp;D</td>
<td>E</td>
<td>.726</td>
<td>.129</td>
<td>.426</td>
<td>.168</td>
<td>.70</td>
</tr>
<tr>
<td>C&amp;D</td>
<td>N</td>
<td>.682</td>
<td>.138</td>
<td>.393</td>
<td>.125</td>
<td>.75</td>
</tr>
</tbody>
</table>

*Reliability estimate based on average point-biserials for sixteen items after Guilford (1965).*
In their classic nine-year study, Friedman and Roseman found that competitive, hard-driving, impatient, and easily angered individuals are especially susceptible to:

a. stomach ulcers.
b. cancer.
c. heart attacks.
d. accidents.
e. none of these

In their classic nine-year study, Friedman and Roseman found that competitive, hard-driving, impatient, and easily angered individuals are especially susceptible to which of the following?

- a. stomach ulcers
- b. cancer
- c. heart attacks
- d. accidents
- e. none of these

In their classic nine-year study, Friedman and Roseman found that competitive, hard-driving, impatient, and easily angered individuals are especially susceptible to:

- a. stomach ulcers
- b. cancer
- c. heart attacks
- d. accidents
- e. strokes

In their classic nine-year study, Friedman and Roseman found that competitive, hard-driving, impatient, and easily angered individuals are especially susceptible to which of the following?

- a. stomach ulcers
- b. cancer
- c. heart attacks
- d. accidents
- e. strokes