Distractor Similarity and Item-Stem Structure: Effects on Item Difficulty

M. Evelina Ascalon; Lawrence S. Meyers; Bruce W. Davis; Niels Smits

* Institute of Psychology, Erasmus University Rotterdam, † Department of Psychology, California State University, Sacramento, USA  ‡ PSI, Burbank, CA, USA  § Faculty of Psychology and Education, Free University of Amsterdam,

To cite this Article

To link to this Article: DOI: 10.1080/08957340701301272
URL: http://dx.doi.org/10.1080/08957340701301272
Distractor Similarity and Item-Stem Structure: Effects on Item Difficulty

M. Evelina Ascalon  
_Institute of Psychology  
_Erasmus University Rotterdam_

Lawrence S. Meyers  
_Department of Psychology  
_California State University, Sacramento_

Bruce W. Davis  
_PS, Burbank CA_

Niels Smits  
_Faculty of Psychology and Education  
_Free University of Amsterdam_

This article examined two item-writing guidelines: the format of the item stem and homogeneity of the answer set. Answering the call of Haladyna, Downing, and Rodriguez (2002) for empirical tests of item writing guidelines and extending the work of Smith and Smith (1988) on differential use of item characteristics, a mock multiple-choice driver’s license examination was administered to high school students with items having item stems that were either open-ended or in question form and with distractors structured to be either similar or dissimilar to the correct answer. Analyses at the test level indicated that the similarly structured distractors raised the mean difficulty level by .12. No effect was found for item-stem format. Differential item function analyses on each of the test items further supported the effect of distractor similarity on test performance. Implications of this study for item writing and standard setting, as well as implications for future research, are discussed.
This study is an investigation of a component of Haladyna, Downing, and Rodriguez’s (2002) taxonomy of item-writing guidelines and an extension of Smith and Smith’s (1988) exploratory work into the differential use and impact of item characteristics information on standard setting. We further investigated two of the item characteristics that Smith and Smith studied and further assessed the relationship of these characteristics to examinee scores. More specifically, our purpose was to investigate the effect of selected item characteristics on item difficulty. The two characteristics we chose to manipulate were similarity of distractors to the correct answer (Smith & Smith, 1988), and whether the premise or stem of the item was formatted in an open-ended (i.e., sentence-completion) format or as a question (Crehan & Haladyna, 1991). We chose these item characteristics because they have been highlighted by item developers for use in item writing (e.g., Haladyna et al., 2002) and examinees have been prone to attend to these characteristics when selecting their answers (e.g., Smith & Smith, 1988).

Past research on the constructs of item-stem format and distractor similarity indicate that more research appears to be warranted. Haladyna et al.’s (2002) taxonomy recommends writing answer options that are homogeneous in content and grammar structure (see Rule 23); however, they were unable to find sufficient empirical studies to support this recommendation. Smith and Smith (1988) found that the similarity of answer options was associated with lower examinee performance (i.e., greater difficulty). However, their study did not adequately isolate distractor similarity and therefore it remained questionable as to whether distractor similarity truly affected item difficulty. Haladyna and Downing (1989) reported findings from Green (1984) that more heterogeneous options (less similar answer options) produced more difficult items. These findings are contrary to Smith and Smith’s findings. Given these mixed results, we chose to further investigate the construct of distractor similarity.

Haladyna et al. (2002) also recommended that item stems should be written in question format, even though the literature they had reviewed did not provide a clear indication of how this would affect item difficulty. There are many types of open-ended statements, for example, “The president of the United States__.” “With regard to the military, the role of the president of the United states is__.” This study focused on the latter types of statements that lead to a narrow and specific response. Haladyna et al. reported on three studies (Crehan & Haladyna, 1991; Eisley; Rachor & Gray) that empirically tested the effect of a question-stem versus. a statement-stem format. The findings from the three studies are mixed. None of the studies found a difference in discrimination between the question and sentence-completion formats, one found a difference in difficulty (i.e., items written in question format were more difficult; Eisley), and one found a difference in reliability (Eisley). Because the findings concerning this format appeared to be inconclusive, we decided to investigate the item-stem format further.
Based on the previous research, we hypothesized that exam characteristics would correlate with item responses. More specifically,

1. Items that contain distractors that are more similar to the correct answer will be more difficult (i.e., have lower $p$ values).
2. Items that contain item stems written as questions will be less difficult (i.e., have higher $p$ values) than those written as open-ended statements.

This study was conducted in five stages. First, about four dozen multiple-choice items on driving regulations were developed. For each hypothesis to be tested, items were written in question or open-ended format with distractors that were either very similarly or not similarly structured (this issue is explained further in the Method section). Second, the degree of similarity/dissimilarity of these distractors was evaluated by a group of raters who were knowledgeable about test construction (graduate students and consultants). Third, the most suitable items in revised form were compiled into four different versions of a written driver’s license exam (this exam choice was selected based on Hudson & Campion’s [1994] study on standard setting, which used a written driver’s license exam) that balanced the levels of the two independent variables of similarity and item-stem format. Fourth, the exams were administered to a large sample of high school students. Finally, we assessed the effect of our manipulations on the test and item scores.

METHOD

Participants

High school students who were attending summer school at three northern California high schools were asked to voluntarily participate in the exam administration by their principals or teachers. Eight hundred exams were given to the three high schools, of which 520 exams were returned (84 exams were returned from the first high school, 77 were returned from the second, and 359 were returned from the third). The participants included 241 males and 246 females (33 students did not report their gender), ranging in age from 13 to 19 years ($M=16.03$ years, $SD=1.07$).

Item Development

Stage 1

The test was developed over a series of stages. First, 43 items were written about important concepts from the 1999 California Driver Handbook (State
of California Department of Motor Vehicles, 1999). Items were written for each section of the handbook to ensure that each area was represented in the exam.

Stage 2

In the second stage of item development, items were written in question or open-ended format with distractors that were either very similarly or dissimilarly structured. A distractor was operationalized as similar when its content (e.g., theme, words, sentence length) was comparable to the correct answer. For example, the distractors in Example 1 are all considered similar to the correct answer.

Example 1

7. Driving in the center left turn lane is allowable only for
   a. driving a distance of 150 feet.
   b. driving a distance of 200 feet.
   c. driving a distance of 250 feet.
   d. driving a distance of 300 feet.

Each of the distractors in Example 1 conveys driving a distance of a number of feet. This set of distractors is considered to be highly similar to the correct answer, Option b. All items were constructed so that all distractors were plausible.

A distractor was operationalized as dissimilar when its content (and therefore its structure) was not comparable to the correct answer. For example, the following options in Example 2 are all considered dissimilar to the correct answer.

Example 2

7. When is driving in the center left turn lane allowable?
   a. It is allowable only for motorcycles.
   b. It is allowable only for driving a distance of 200 feet.
   c. It is allowable only for emergency vehicles.
   d. It is allowable only for traffic congested situations.

Each of the distractors (a, c, and d) in Example 2 conveys information different from the correct answer in its responses. Options a and c concern vehicles and Option d concerns a situation. The placement of the correct answer was held constant for each item across each of the versions. For illustration, in Examples 1 and 2, the placement of the correct answer (Option b: driving a distance of

---

1This study was conducted in 1999.
200 feet) is identical in each of the versions. In addition, as can be seen in the two above examples, the item stem of Example 2 is similar to that of Example 1, but is in question form\(^2\). Moreover, as recommended by Haladyna et al. (2002), all item stems were written to be clear and to contain the central idea (see Rules 14 and 15). Twenty-seven items remained after this stage.

**Stage 3**

In the third stage of item development, the item characteristics (item importance and distractor similarity) of the 27 items were evaluated using a questionnaire. As a preliminary assessment, four consultants in a personnel test development department of a government consulting agency reviewed the initial 27-item questionnaire for general feedback. The consultants were each given a copy of the questionnaire and asked to provide contextual and grammatical feedback on the questionnaire, which we incorporated prior to the administration of the item characteristics review.

Next, 13 psychology graduate students from a California university, who were knowledgeable about test construction, evaluated the items using ratings of importance, similarity, and dissimilarity. The graduate students used a 5-point rating scale to assess the importance of the item in relation to driving in the State of California using anchors of 1 (*not very important*) to 5 (*very important*). Second, they used a 5-point rating scale to assess the similarity between the statement and the question using anchors of 1 (*not very similar*) to 5 (*very similar*). Finally, they used a 5-point rating scale to assess the extent to which the distractors were similar (or dissimilar) to the keyed answer using anchors of 1 (*not very similar*) to 5 (*very similar*). A target of 16 items was planned for the driver’s license exam in order to have four items representing each combination of the two independent variables. Therefore, the 16 items with the best average similarity (approximating 5) and dissimilarity (approximating 1) ratings were selected.

Based on the ratings and comments received from the graduate student item review, 14 of the 16 items were further altered by the first researcher and a graduate student to strengthen the similarity or dissimilarity of the distractors. Items indicated by the graduate students as having more than one correct answer were carefully checked against the Department of Motor Vehicles handbook and corrected as needed.

---

\(^2\)Because the format of the item stem was also altered across versions, it was sometimes necessary to slightly alter the wording of the options in order to maintain grammatical flow between the item stem and the answer.
Stage 4

In the fourth stage, the four forms of the exam were created. Each version of the 16 items was randomized in a cyclical fashion (Keppel, 1991) until all versions of the items were equally distributed across the four forms, resulting in 16 items for each form. (For identification purposes, these forms were labeled Forms A, B, C, and D.) Each of the four versions of the 16 items is presented only once in each of the four forms of the exam. Great care was taken to ensure that each of the forms would be parallel to the others. Each item was presented in the same item order in each exam. However, item characteristics did vary in their presentation (i.e., placement of item in test) across the forms according to the partial-randomization method explained previously (see Keppel, 1991). Items containing the same combination of characteristics did not appear next to one another.

Stage 5

In the final stage of test development, a second item review was conducted to obtain similarity ratings for the final version of items using a 5-point scale ranging from 1 (not very similar) to 5 (very similar). Ten consultants working in the test development industry, none of whom had participated in the prior item assessments, completed the item ratings for the 16 items. Six questionnaires were returned. Interrater reliability, computed using an intraclass correlation, was .62. Pairwise correlations among the six raters are presented in Table 1.

Mean similarity ratings were computed for each of the distractors to ensure that each fell within the degree of similarity desired and are presented in Table 2.

### TABLE 1
Intercorrelation of Raters From the Final Item Review

<table>
<thead>
<tr>
<th>Rater</th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
<th>Rater 4</th>
<th>Rater 5</th>
<th>Rater 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 1</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater 2</td>
<td>.65*</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater 3</td>
<td>.76*</td>
<td>.62*</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater 4</td>
<td>.45*</td>
<td>.54*</td>
<td>.45*</td>
<td>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater 5</td>
<td>.74*</td>
<td>.59*</td>
<td>.74*</td>
<td>.49*</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>Rater 6</td>
<td>.73*</td>
<td>.61*</td>
<td>.80*</td>
<td>.43*</td>
<td>.71*</td>
<td>—</td>
</tr>
</tbody>
</table>

Note. $N = 6$.

* $p < .01$.

---

3The cyclic counterbalancing procedure maintains the basic sequence of the items but shifts the position to the left for each subsequent sequence.
<table>
<thead>
<tr>
<th>Item No.</th>
<th>Option</th>
<th>Similar Option Set</th>
<th>Dissimilar Option Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>1</td>
<td>a</td>
<td>4.00</td>
<td>1.10</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>4.00</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>4.17</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>d</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>2</td>
<td>a</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>3.83</td>
<td>1.60</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>4.00</td>
<td>1.26</td>
</tr>
<tr>
<td></td>
<td>d</td>
<td>3.17</td>
<td>1.47</td>
</tr>
<tr>
<td>3</td>
<td>a</td>
<td>2.83</td>
<td>1.33</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>3.50</td>
<td>1.05</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>4.50</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>d</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>4</td>
<td>a</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>4.33</td>
<td>1.03</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>3.17</td>
<td>1.47</td>
</tr>
<tr>
<td></td>
<td>d</td>
<td>2.83</td>
<td>1.72</td>
</tr>
<tr>
<td>5</td>
<td>a</td>
<td>2.17</td>
<td>1.47</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>3.83</td>
<td>1.17</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>3.17</td>
<td>1.47</td>
</tr>
<tr>
<td></td>
<td>d</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>6</td>
<td>a</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>4.33</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>2.33</td>
<td>1.03</td>
</tr>
<tr>
<td></td>
<td>d</td>
<td>3.67</td>
<td>1.37</td>
</tr>
<tr>
<td>7</td>
<td>a</td>
<td>4.50</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>4.50</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>d</td>
<td>4.17</td>
<td>1.33</td>
</tr>
<tr>
<td>8</td>
<td>a</td>
<td>4.50</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>3.50</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>3.50</td>
<td>1.22</td>
</tr>
<tr>
<td></td>
<td>d</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>9</td>
<td>a</td>
<td>3.67</td>
<td>1.21</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>2.83</td>
<td>1.33</td>
</tr>
<tr>
<td></td>
<td>d</td>
<td>3.50</td>
<td>1.05</td>
</tr>
<tr>
<td>10</td>
<td>a</td>
<td>4.00</td>
<td>1.26</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>3.33</td>
<td>1.37</td>
</tr>
<tr>
<td></td>
<td>d</td>
<td>1.50</td>
<td>0.55</td>
</tr>
</tbody>
</table>
In addition, for ease of analysis, an overall similarity rating was computed for each of the items by averaging the mean similarity ratings for each of the distractors of an item. The means and standard deviations of the option sets for Items 1 to 16 from the item review are presented in Table 3. The ratings supported the characteristics of the items as we had intended, with mean similarity ratings ranging from 4.39 (Items 7 and 11) to 2.94 (Item 10), with the majority of the ratings being above 3.00. Likewise, the ratings support the intended item characteristics with the mean dissimilarity ratings ranging from 1.06 (Item 5) to 2.39 (Item 4), with the majority of the ratings falling below 1.99.
In addition to the mean ratings of the reviewers, the similarity–dissimilarity characteristics of the items were further assessed by correlating the reviewer ratings with the characteristic (similar or dissimilar) that we had intended. The intended characteristic of similarity of distractors to the correct answer—which
was dummy-coded to possess values of 1 (*dissimilar*) and 2 (*similar*)—correlated (.92, *p* < .01) to the mean similarity rating assigned by the item reviewers from the second review. Therefore, options that we had intended to be similar (2) were rated as similar by the reviewers (higher value ratings on a 5-point scale). Conversely, options that we had intended as dissimilar (1) were rated as dissimilar by the reviewers (lower value ratings on a 5-point scale).

The results from the item review suggest that the items do contain the characteristics that they were intended to possess. This supports the use of these items in the study for drawing inferences about item characteristics and examinee performance.

**Procedure**

The exams were given to the three school administrators for distribution to the teachers. The teachers were provided with an instruction sheet, consent forms, and all four versions of the randomized tests so that examinees within each test administration were assigned approximately the same number of each test form. Teachers were asked to administer the exams during class time, preferably to students who had completed a driver education course. The administrators read the script to the students. Students were told that participation was voluntary, anonymous, and confidential.

**RESULTS**

Five hundred twenty students participated in the third phase of the study; however, due to incomplete tests and aberrant responses (e.g., improbable responses to questions asking about the number of years of driving experience or number of traffic accidents), 27 students were omitted from the analysis, resulting in a final *N* of 493 students (234 males, 244 females, 15 did not report their gender), ranging in age from 13 to 18 years (*M*=16.03 years, *SD*=1.06). This sample size is quite similar if not larger than other published reports on this topic, for example, Hudson and Campion (1994) had approximately 196 participants. Each student completed one of the four versions of our mock driver’s license exam (Form A had 127 examinees, Form B had 125 examinees, Form C had 123 examinees, and Form D had 118 examinees). In terms of class breakdown, there were 33 freshmen, 93 sophomores, 134 juniors, and 216 seniors. Of the 493 students, 10.5% of the students had participated in traffic school in the past 24 months, 50.9% had participated in a driver instruction course in the past 24 months, 55.8% had taken the written driver’s license exam administered by the Department of Motor Vehicles, 50.3% had passed the exam, and 31.5% had passed the road test. In addition, students averaged .58 years of driving experience, acquired .11
(SD=.45) traffic tickets in the past 9 years, and had been involved in .25 (SD=.81) traffic accidents in the past 9 years.

Test score reliability for the four forms, estimated using coefficient alpha, were .59, .65, .64, and .66 for Forms A through D, respectively. The alphas are moderate (c.f. Nunnally & Bernstein, 1994), which is likely a function of item heterogeneity in a brief 16-item test.

To examine validity, we split the sample into two groups. The first group consisted of students who had completed a driver instruction course. The second group had not completed a driver instruction course (M=9.69, SD=2.57), and who were therefore expected to perform better, did indeed answer 2.46 more items correctly than those who had not taken a course (M=7.24, SD=2.57), F(1, 467)=90.83, p=.00, partial η²=.16. This significant result provides some evidence of the validity.

A three-way mixed-model analysis of variance was conducted to evaluate the effect of option similarity (Option) and item-stem format (Stem) on exam performance, which was operationalized as the number of correctly answered items on the exam (p value). (See Table 4 for the means and standard deviations of the four test conditions.) For each examinee, four total scores of four items were compared to one another. Each score represented one of the item-stem and distractor similarity combinations. To determine if the four forms of the test were parallel, test form was entered as a between-subjects factor. The Option × Stem × Form interaction effect was significant, F(1, 489)=8.38, p=.00, partial η²=.05. However, because the effect size was very small (c.f. Cohen, 1992), it was thought that the four forms could be treated as essentially parallel for testing this study’s hypotheses.

A significant main effect for Option was found, F(1, 489)=150.99, p=.00, partial η²=.24. More items were answered correctly when the distractors were written dissimilarly to the correct answer (M=.61) than when the distractors were written similarly (M=.49). Therefore, Hypothesis 1 was supported.

The Item Stem effect was not significant, F(1, 489)=1.85, p=.17. Hypothesis 2 was not supported. Here, p values did not differ significantly from when the

<table>
<thead>
<tr>
<th>Item Stem</th>
<th>Similar</th>
<th>Dissimilar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statement</td>
<td>0.50</td>
<td>0.62</td>
</tr>
<tr>
<td>Question</td>
<td>0.49</td>
<td>0.61</td>
</tr>
</tbody>
</table>

(See Table 4 for the means and standard deviations of the four test conditions.)
stem was written as an open-ended statement ($M=.56$) or as a question ($M=.55$). Furthermore, Option and Stem were not found to have a significant interaction, $F(1, 489) = .41, p = .52$.

To further assess the impact of our constructs on item performance, we decided to examine Differential Item Functioning (DIF). DIF occurs when a significant number of examinees with equal proficiency, but with different group membership, have a different probability of answering an item correctly (Hidalgo & Lopez-Pina, 2004). Although DIF is commonly used for detecting item bias between existing groups of examinees (e.g., groups who differ according to gender or ethnicity), we used the procedure to determine if items showed systematical differences in item responses between our artificially developed groups of examinees (i.e., groups that were formed based on their exposure to our manipulated variables). We attempted to determine if groups of examinees that had been assigned similar option items (the focal group) would have a lower probability of answering the item correctly than would those who had been assigned the dissimilar option items (the reference group). In addition, we attempted to determine if groups of examinees who had been assigned open-ended statements would have a lower probability of answering the item correctly than those who had been assigned the questions. Ability level was not expected to vary across the constructs (our grouping variables) because examinees were exposed to all manipulations.

To assess DIF, the Mantel-Haenszel (MH) procedure proposed by Holland and Thayer (1988) was used. The MH procedure has been found to be a powerful procedure in identifying uniform DIF (Hidalgo & Lopez-Pina, 2004; Mazor, Clauser, & Hambleton, 1994). The MH procedure tests the hypothesis that the odds of correctly answering an item are equal for two groups along the proficiency scale; it is distributed as a chi-square distribution with 1 degree of freedom. (For an extensive description of the MH procedure, see, for example, Hidalgo & Lopez-Pina, 2004.) To obtain a measure of effect size of the DIF, delta-alpha-MH (Holland & Thayer, 1988) may be used. Using this effect size, a zero value indicates no DIF, a negative value indicates that the item benefits the reference group above the focal group, and a positive value indicates that the item favors the focal group over the reference group. Therefore, in our analyses, a zero value would indicate no DIF, a negative value would indicate that examinees who had dissimilar options had an advantage over those who had similar options, and a positive value would indicate an advantage for the examinees who had the similar options. According to the guidelines of Zwick and Ercikan, (1989) an item with a delta-alpha-MH with an absolute value smaller than 1 is

---

4The Logistic Regression procedure (c.f. Hidalgo & Lopez-Pina, 2004; Swaminathan & Rogers, 1990) and the adjusted MH procedure (c.f. Mazor et al., 1994) were also applied. The procedures gave similar results and therefore will not be discussed.
classified as a Type A item (or negligible DIF); a value between 1 and 1.5 is classified as a Type B item (moderate DIF); and a value larger than 1.5 is a Type C item (large DIF). Type B and C items should also result in a significant MH test.

The findings from the MH procedure are reported in Table 5. For the Option manipulation, we found 3 Type B items, 9 Type C items, and 4 nonsignificant items. One of the Type C and one of the Type B items showed a reverse effect in favor of the similar options. In general, the MH procedure showed that our manipulation of item characteristics for option similarity produced 12 instances of DIF in favor of the dissimilarly written options over the similarly written ones. For the stem manipulation, only one significant Type B item was found. No interaction effect was tested, because no interaction had been hypothesized.

### DISCUSSION

We have known for some time on an informal basis, and more recently on a more formal basis (Smith & Smith, 1988) that, all else being equal, increased similarity

<table>
<thead>
<tr>
<th>Item</th>
<th>$\chi^2_{MH}$</th>
<th>$p$</th>
<th>$\Delta_{MH}$</th>
<th>Type</th>
<th>$\chi^2_{MH}$</th>
<th>$p$</th>
<th>$\Delta_{MH}$</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10.81</td>
<td>0.00</td>
<td>1.68</td>
<td>C</td>
<td>0.57</td>
<td>0.45</td>
<td>-0.42</td>
<td>ns</td>
</tr>
<tr>
<td>2</td>
<td>0.00</td>
<td>0.97</td>
<td>-0.08</td>
<td>ns</td>
<td>0.11</td>
<td>0.73</td>
<td>-0.25</td>
<td>ns</td>
</tr>
<tr>
<td>3</td>
<td>35.78</td>
<td>0.00</td>
<td>-3.13</td>
<td>C</td>
<td>0.00</td>
<td>0.99</td>
<td>0.00</td>
<td>ns</td>
</tr>
<tr>
<td>4</td>
<td>16.96</td>
<td>0.00</td>
<td>-2.34</td>
<td>C</td>
<td>0.68</td>
<td>0.41</td>
<td>-0.51</td>
<td>ns</td>
</tr>
<tr>
<td>5</td>
<td>0.00</td>
<td>0.98</td>
<td>-0.01</td>
<td>ns</td>
<td>0.09</td>
<td>0.76</td>
<td>-0.27</td>
<td>ns</td>
</tr>
<tr>
<td>6</td>
<td>46.83</td>
<td>0.00</td>
<td>-3.41</td>
<td>C</td>
<td>0.00</td>
<td>0.95</td>
<td>0.03</td>
<td>ns</td>
</tr>
<tr>
<td>7</td>
<td>9.29</td>
<td>0.00</td>
<td>-1.48</td>
<td>B</td>
<td>0.76</td>
<td>0.38</td>
<td>-0.46</td>
<td>ns</td>
</tr>
<tr>
<td>8</td>
<td>52.05</td>
<td>0.00</td>
<td>-3.56</td>
<td>C</td>
<td>0.00</td>
<td>0.96</td>
<td>0.02</td>
<td>ns</td>
</tr>
<tr>
<td>9</td>
<td>133.46</td>
<td>0.00</td>
<td>-5.92</td>
<td>C</td>
<td>0.91</td>
<td>0.34</td>
<td>0.47</td>
<td>ns</td>
</tr>
<tr>
<td>10</td>
<td>15.91</td>
<td>0.00</td>
<td>-2.31</td>
<td>C</td>
<td>4.30</td>
<td>0.04</td>
<td>1.21</td>
<td>B</td>
</tr>
<tr>
<td>11</td>
<td>0.08</td>
<td>0.77</td>
<td>-0.18</td>
<td>ns</td>
<td>2.09</td>
<td>0.15</td>
<td>0.71</td>
<td>ns</td>
</tr>
<tr>
<td>12</td>
<td>5.16</td>
<td>0.02</td>
<td>-1.34</td>
<td>B</td>
<td>0.01</td>
<td>0.90</td>
<td>-0.13</td>
<td>ns</td>
</tr>
<tr>
<td>13</td>
<td>20.64</td>
<td>0.00</td>
<td>-2.66</td>
<td>C</td>
<td>3.62</td>
<td>0.06</td>
<td>1.12</td>
<td>ns</td>
</tr>
<tr>
<td>14</td>
<td>9.62</td>
<td>0.00</td>
<td>-1.50</td>
<td>C</td>
<td>0.00</td>
<td>0.95</td>
<td>-0.07</td>
<td>ns</td>
</tr>
<tr>
<td>15</td>
<td>3.32</td>
<td>0.07</td>
<td>-1.02</td>
<td>ns</td>
<td>0.59</td>
<td>0.44</td>
<td>0.45</td>
<td>ns</td>
</tr>
<tr>
<td>16</td>
<td>7.32</td>
<td>0.01</td>
<td>1.29</td>
<td>B</td>
<td>0.37</td>
<td>0.54</td>
<td>0.33</td>
<td>ns</td>
</tr>
</tbody>
</table>

Note: MH=Mantel-Haenszel, ns=nonsignificant, Type B=moderate effect size, Type C=large effect size. $\chi^2_{MH}$ has $df=1$.

\(a\)Indicates a significant effect opposite to the hypothesized direction.
among the alternatives of a multiple-choice item will result in increased item difficulty. In much of the research however, all else has really not been equal. That is, distractor similarity/dissimilarity always covaried with other factors that could easily affect the difficulty of test items (e.g., content relevance, plausibility).

This study attempted to control for extraneous variables by keeping item content constant and by keeping the wording of the distractors relatively close across item sets so that they could be more directly compared. Under these relatively controlled conditions, we were able to produce an average difference in difficulty ($p$ values) of about .12. Such a difference, while not dramatic, is relatively substantial both from a statistical standpoint, in that the effect accounted for about a quarter of the total variance, as well as from a practical standpoint, in that a change of .12 difficulty could be quite noticeable in a testing environment (particularly in high-stakes testing). We performed further analyses of the $p$ values at the item level and we found that even distractor sets with lower similarity ratings produced a significant effect. For example, although Item 10 has an average similarity rating of only 2.94 for the similar set of options and an average similarity rating of 2.11 for the dissimilar set of options, the $p$ values still differed significantly by .15. Despite the same item stem and content, 29% of examinees were able to answer Item 10 correctly when the options were dissimilar, whereas only 14% were able to answer it correctly using the similar set of options, $F(1, 489)=17.81, p=.00$. In addition, the item still managed to achieve a relatively large effect size as identified by the MH procedure. This indicates the strong effect of option similarity on item difficulty even when the options differ only moderately from one another in similarity.

When the item sets differ greatly from one another, such as for Item 9 (in which the similar set of options had an average similarity rating of 3.33 and the dissimilar options had an average similarity rating of 1.78), the difference in difficulty (.48 in this case) is much greater. Seventy percent of the examinees were able to answer Item 9 correctly when the options were dissimilar to one another and only 22% answered correctly when the options were similar to one another. In addition, the MH procedure indicates a relatively large effect size. Such a large difference in performance between the two conditions suggests that performance may have been due more to the characteristics of the items than to the knowledge level of the examinee. Items containing dissimilar options may be easier to guess correctly; therefore, performance may have resulted from test-wiseness rather than knowledge of the material.

One possible limitation of our study was that our sample was drawn during summer school session, so the students may not have been representative of the general student population. However, we also acknowledge that many high school students also attend summer school for advanced education classes, so it is not clear whether this sample is representative of the population of students who seek a driver’s license. We hoped to spark their interest and hold their attention
during the short time we asked them to spend taking the test, but we would recom-
mend caution in generalizing our findings until replications with other popula-
tions in other content domains come forward.

Another possible limitation could have been a relationship between similarity
and plausibility, that is, our distractors may have been less plausible when they
were less similar. To assess this limitation, we examined the relationship between
plausibility and similarity and found a high correlation ($r = .81$), which could
indicate that we were assessing both similarity and plausibility. We do caution
against placing too much emphasis on this correlation, because the raters were
provided the correct answer and the same raters provided the ratings of similarity
and plausibility, which could have produced common method bias. One possible
explanation is that the raters could have assessed the plausibility of the distrac-
tors in relation to how similar they were to the correct answer. Meaning that, in
hindsight, less-similar distractors may have seemed less plausible. Future
research may examine this in a different way by gathering the plausibility and
similarity ratings separately and by not indicating the correct answer. More stud-
ies should be conducted on both option similarity and plausibility. Plausibility is
another interesting item characteristic, which Haladyna et al. (2002) agree is an
area “that is long overdue for study” (p. 327).

We recommend continuing this line of research by which improvements in
item-writing methods are guided by empirical research, that is, showing a link
between item formats and examinee performance. The relationship of item char-
acteristics to examinee performance is particularly interesting, because it allows
researchers to test what item characteristics influence item difficulty. Instructing
item developers to write distractors that are similar in content and structure to the
correct answer is more specific and clear than simply asking them to write plausi-
ble response options. In our study we found support for the differential effect of
only one item characteristic, option similarity. Contrary to Eisley’s (as cited in
Haladyna et al., 2002) findings, we did not find a difference in item difficulty for
the item stem (open-ended statement vs. question). The effect size was negligi-
ble. However, this finding may still be useful for item writers, because it provides
empirical evidence that both formats can be used interchangeably (as recom-
mended by Haladyna et al., 2002). We must also mention that this finding could
also be applicable only to the type of open-ended items that we used. Our open-
ended items used clear and detailed statements and are perhaps better thought of
as implied questions. In contrast, some open-ended items may serve as disguised
multiple true–false items (e.g., Which of the following statements is true?) that
are then followed by a heterogeneous set of options (Downing, 2005).

This study has important implications for test developers of all high-stakes
tests. In particular, it has implications for those who do not perform additional
test-control techniques, such as test-centered standard-setting methods or equat-
ing to adjust for differences in difficulty between/among forms. (Standard setting
is discussed more fully below.) Our findings should serve as a reminder that constructing alternate forms based on content alone without considering distractor similarity may result in forms of different difficulty such that some examinees will be disadvantaged. If equating is not performed, for example, then alternate forms might be made as parallel as possible with regard to inclusion of numbers of items with similar/dissimilar distractors.

Our recommended approach is to develop tests so that examinees are measured on the knowledge they possess. As we have discovered, the properties of an item can influence test takers’ scores. Instead of responding correctly based on their knowledge, test takers may receive credit for an item simply because of test-item construction. Van Heerden and Hoogstraten (1979) had participants respond to item scales that contained no item stem. They found that persons who are motivated to obtain a high score on the test will study the characteristics of the response options in order to make a best guess as to the correct answer. This implies that motivated examinees who are not knowledgeable of the item content may be more sensitive to item characteristics than those examinees who know the answer and do not pay attention to item characteristics. Ideally, tests should be designed to allow those who know the test content to perform well and those who are not knowledgeable to be unable to answer correctly by eliminating implausible response choices. Therefore, we could argue that item writers should be instructed to write response options that are similar in content, theme, meaning, words, and sentence length to ensure the fairness of the item. The objective would not be to increase the difficulty of the item, but to enhance the validity of the score. However, more research should be conducted before making any definite conclusions.

In addition, this research is not only helpful for item writers, but also extends to other areas of high-stakes testing (such as credentialing exams that often use test-based methods for standard setting). Test-based standard setting is the process by which passing scores on a high-stakes exam are designated by a group of subject matter experts (SMEs). One example of a standard-setting technique is the Angoff method (Angoff, 1971), which is a means of setting cut scores that can be employed for almost any type of exam (e.g., multiple-choice and short-answer). It defines the cut score as the lowest score the minimally acceptable candidate (MAC) is likely to achieve on an exam. Candidates scoring below this level are believed to lack sufficient knowledge, skills, or abilities to perform in the area being tested. The Angoff method employs a group of SMEs to set cut scores by estimating the probability values of the MAC for each item.

Standard-setting methods such as the Angoff (Angoff, 1971) have been criticized for various reasons; for example, researchers question the ability of SMEs to make multiple quantitative judgments from the qualitative information (e.g., the characteristics of the item) they are provided (Cizek, 1993; Impara & Plake, 1998). Standard-setting practice can be guided by research that focuses on the
effects of varying item characteristics on examinee performance and standard-setting. SMEs can be provided with relevant item characteristics during their estimation of probability levels of examinee performance. In relation to our findings, SMEs could be informed that option sets that contain distractors that are more similar to the correct answer in content, wording, theme, and length have been shown to be more difficult. This could help them to estimate more realistic cut scores for the items. Future studies could assess the impact of providing such information on the validity of passing scores.

CONCLUSION

Our investigation of the effect of response option similarity and item-stem format on item difficulty resulted in finding a positive relationship between similar option items and item difficulty, and no relationship between item-stem format and item difficulty. Response options that are dissimilar from the correct answer may be easier to eliminate as plausible options suggesting that test performance could result from test-wiseness rather than knowledge of the test material. This provides support that particular item characteristics can strongly impact the validity of the test score. We recommend that special attention be given to item characteristics when developing tests, particularly for high-stakes testing situations.

REFERENCES


AUTHOR NOTE


M. Evelina Ascalon is now in the department of Human Capital Processes and Evaluation, Credit Suisse, Zurich, Switzerland, as well as an adjunct researcher at the Erasmus University Rotterdam. We express appreciation to all those who volunteered to participate in and helped us with this study, and to those who provided us valuable feedback on earlier versions of this article.

Correspondence concerning this article should be sent to Evelina Ascalon evelina.ascalon@creditsuisse.com.