DISTRACTOR DISCRIMINATION IN MULTIPLE-CHOICE QUESTION BANKS

John R. Dickinson, Odette School of Business, University of Windsor, Windsor, Ontario, Canada N9B 3P4, MExperiences@bell.net, 519.253.4232, Ext. 3104

ABSTRACT

Despite the ubiquity of multiple-choice question banks accompanying textbooks and an established tradition of methods for analyzing the effectiveness of such questions, the latter has only rarely been applied to the former. Distractors are the incorrect answer options of a multiple-choice question with their effectiveness warranting analysis. This study examines the ability of distractors in five question banks to discriminate between better and poorer students. Implications pertain to the question banks examined. As well, necessary norms and a pro forma for similar analyses of other question banks are provided.

Keywords: Multiple-choice questions, item analysis, distractors, item discrimination

INTRODUCTION

Virtually every introductory-level marketing (and business generally) textbook is accompanied by a bank of test questions. The questions may be of different types, such as true-false, fill-in-the-blank, and short answer essays. Invariably, though, most of the questions are of the multiple-choice type. Too, though certainly not entirely the case, textbooks in many areas–such as those in this research–are in the maturity stage, having gone through numerous editions over many years with question banks evolving in step with successive editions. Longstanding as these texts and question banks are, in education and psychology the evaluation of questions generally has a decades longer history, a watershed being the 1930s (Rogers, 1995, Chapter 3). A cornerstone of educational/psychological testing is item analysis: “An analysis of the responses given by a group of examinees to the individual items on a test...” (Aiken, 1991, p. 73) Yet published item analyses of textbook-accompanying question banks are rare. Question bank descriptions generally do not describe methods by which questions were developed and (empirically) refined. Only a few independent assessments of question banks have been published: Dickinson (2011, 2005), Dickinson, Faria, & Whiteley (1991).

For several published question banks, the present research analyzes a critical facet of multiple-choice questions. Namely, the question distractors or incorrect answer options. More specifically, the objective of the research is to ascertain the effectiveness of distractors in discriminating between better and worse students.
BACKGROUND

The desired role of distractors in discriminating between better and worse students is widely recognized:

- “Poorer and better students often choose different alternatives when incorrect, which in theory may provide diagnostically useful information.” (Nunnally & Bernstein, 1994, p. 304)

- “A test developer can supplement the item-discrimination approach by inspecting the number of examinees in the upper- and lower-scoring groups who choose each of the incorrect alternatives...a good item should show proportional dispersion of incorrect choices for both high- and low-scoring subjects.” (Gregory, 2011, p. 145)

- “The second question [key to distractor analysis] involves discrimination. Effective distractors should attract more examinees in the bottom group than in the top group...That is, distractors should demonstrate negative discrimination!” (Reynolds & Livingston, 2012, p. 233)

- “By charting the number of testtakers in the U and L groups who chose each alternative, the test developer can get an idea of the effectiveness of a distractor by means of a simple eyeball test...Item 2 signals a situation in which a relatively large number of members of the U group choose a particular distractor choice...” (Cohen & Swerdlik, 2010, pp. 259-260)

The present research, then, examines five published multiple-choice question banks with respect to the extent to which the question distractors serve the above noted discrimination purpose.

QUESTION BANKS

Data are analyzed for five multiple-choice question banks, three for consumer behavior texts (including two editions of one text) and two for successive editions of a retailing management text. See Table 1. A few questions were deemed invalid in that the correct response was not clear in the text and those questions are excluded. Too, a small number of questions have fewer than the prevalent five options and those questions were excluded.

SAMPLE QUESTIONS

Multiple-choice questions are arranged in the test question banks according to the order in which the question content appears in the textbook. For each examination specific multiple-choice questions were selected on a systematic sampling basis (every 8th or 10th question, with varying starting points)
in an attempt to ensure that:

- a cross section of each chapter content was included among the examination questions,
- respective exams were of comparable composition, and
- a representative sample of master bank questions was obtained.

The data base of sample questions is summarized in Table 2.

### TABLE 1: MULTIPLE-CHOICE QUESTION BANKS ANALYZED

<table>
<thead>
<tr>
<th>Text</th>
<th>Total Multiple-Choice Questions</th>
</tr>
</thead>
</table>

### TABLE 2: SAMPLE QUESTIONS

<table>
<thead>
<tr>
<th>Text</th>
<th>Bank Count</th>
<th>Sample Count</th>
<th>Sample as Percent of Bank</th>
</tr>
</thead>
<tbody>
<tr>
<td>LW (2012)</td>
<td>1196</td>
<td>149 *</td>
<td>12.5</td>
</tr>
<tr>
<td>SZP (2011)</td>
<td>1158</td>
<td>505</td>
<td>43.6</td>
</tr>
<tr>
<td>LW (2009)</td>
<td>1332</td>
<td>736</td>
<td>55.3</td>
</tr>
<tr>
<td>SZP (2008)</td>
<td>1019</td>
<td>674</td>
<td>66.1</td>
</tr>
<tr>
<td>HMB (2007)</td>
<td>1624</td>
<td>958</td>
<td>59.0</td>
</tr>
</tbody>
</table>

* This relatively small sample of questions is due to the newness of this text edition.
MULTIPLE-CHOICE EXAMS

For all of the courses for which data are available, two midterm exams and one final exam were administered. The exams were not cumulative. The first midterm exam covered about the first third of the chapters (6 or 7 chapters depending on the specific text), the second midterm covered the middle third of the chapters, and the final exam covered the remaining chapters (5, 6, or 7 chapters). Exams comprised only multiple-choice questions from the relevant master bank. All exams were worth 20 percent of students’ final weighted averages for the course. Parameters of exams for each text are presented in Table 3.

TABLE 3: EXAM PARAMETERS

<table>
<thead>
<tr>
<th>Text</th>
<th>Number of Exams</th>
<th>Mean Questions per Exam</th>
<th>Mean Students per Exam</th>
<th>Mean Score (% Correct)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LW (2012)</td>
<td>3</td>
<td>49.7</td>
<td>36.7</td>
<td>69.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(52, 46)</td>
<td>(39, 35)</td>
<td>(70.6, 68.9)</td>
</tr>
<tr>
<td>SZP (2011)</td>
<td>9</td>
<td>56.1</td>
<td>41.2</td>
<td>58.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(60, 50)</td>
<td>(54, 28)</td>
<td>(61.9, 54.5)</td>
</tr>
<tr>
<td>LW (2009)</td>
<td>12</td>
<td>61.3</td>
<td>36.2</td>
<td>67.44</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(70, 55)</td>
<td>(49, 27)</td>
<td>(73.0, 58.5)</td>
</tr>
<tr>
<td>SZP (2008)</td>
<td>12</td>
<td>56.2</td>
<td>39.9</td>
<td>61.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(60, 48)</td>
<td>(49, 32)</td>
<td>(67.1, 57.5)</td>
</tr>
<tr>
<td>HMB (2007)</td>
<td>18</td>
<td>53.2</td>
<td>32.7</td>
<td>62.74</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(56, 47)</td>
<td>(42, 25)</td>
<td>(69.2, 56.6)</td>
</tr>
</tbody>
</table>

Numbers in parentheses are maximum and minimum, respectively.

ANALYSIS

Item discrimination is made more clear when only extreme students are included in the analysis. Excluding mid-performing students, of course, reduces the effective sample size. Where percent correct values are normally distributed, the optimal balance between discrimination and reliability due to the reduction in the number of observations occurs when the upper and lower 27 percent of students are analyzed (Anastasi, 1968, p. 169; Cureton, 1957; Guilford, 1954, p. 428).

This paradigm was adopted here. Students were divided into high performing and low performing groups based on their overall exam scores. Analyses, then, were carried out exam by exam and, specifically, using the highest scoring 27 percent of students and the lowest scoring 27 percent. (The
same analyses were carried out using a median split, i.e., all students. Results are alluded to in the “Discussion” below.)

Analyses for this study are of three types. First, incidences where distractors attracted no responses are identified. These distractors do not satisfy their essential purpose and, of course, cannot discriminate among students.

“The key [to distractor analysis] is to examine each distractor and ask two questions. First, did the distractor distract some examinees? If no examinees selected the distractor it is not doing its job. An effective distractor must be selected by some examinees. If a distractor is so obviously incorrect that no examinees select it, it is ineffective and needs to be revised or replaced.” (Reynolds Livingston, 2012, p. 233)

Second, for each distractor the percents of the high performing and low performing student subgroups, respectively, selecting the distractor were calculated. Effective distractors would have the latter percent being greater than the former percent, i.e., Low Group % > High Group % (Table 4, row 1). Dysfunctional distractors would have the former percent being greater than or equal to the latter percent, i.e., High Group % >= Low Group % (Table 4, row 2). Where the two proportions are equal the distractor does not discriminate, explaining the assignment of “equal” to the second comparison described.

Third, the basic comparisons of high and low performing student subgroups just described do not reflect the degree or extent to which low performing students select a distractor more frequently than do high performing students and vice versa. As a stronger assessment of distractor discriminating ability, then, similar comparisons were made between high and low performing students only on a more extreme basis. Specifically, results are reported where the percent of one group is not just greater than the second group, but is more than twice that of the second group.

RESULTS

No-Responses Distractors

Substantial percentages of distractors do not, in fact, distract at all. Across the five question banks, about 30 to 40 percent of distractors attracted no responses (Table 4, row 5). Not only do these distractors not serve their basic purpose, but student time is wasted reading what are apparently implausible options. And central to the purpose of this study, these distractors cannot discriminate between high and low performing students.

Basic Discrimination

Regarding effectively discriminating distractors, across the five question banks, just under half of the distractors discriminate as intended, i.e., students having low overall exam scores selecting the
distractor more frequently than students having high overall exam scores. (Table 4, row 1) This is fairly consistent across the five banks, though the difference between the two extremes of 49.7 percent (SZP, 2011) and 42.1 percent (LW, 2009) is material.

Dysfunctionally, for between 14.4 percent (LW, 2012) and 22.3 percent (SZP, 2008) of the distractors actually attracted more high performing students than low performing students (Table 4, row 2).

**Stronger Discrimination**

Table 4 reports distractors where low performing students were more than twice as likely to select a distractor than were high performing students (row 3), i.e., effective discrimination, and where high performing students were at least twice as likely to select a distractor than were low performing students (row 4), i.e., dysfunctional discrimination.

The percentages of effective distractors satisfying this more demanding criterion (Table 4, row 3) were lower, of course, than those satisfying the basic criterion (Table 4, row 1), but not dramatically so. In this respect, these distractors are strongly discriminating. The percentages of more extremely dysfunctional distractors (Table 4, row 4) are about half those of basic inversely discriminating distractors (Table 4, row 2). Though dysfunctional distractors have a presence in the question banks, they are not extremely dysfunctional.

**TABLE 4: DISTRACTOR DISCRIMINATING ABILITY**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Low Group &gt; High Group</td>
<td>45.6</td>
<td>49.7</td>
<td>42.1</td>
<td>47.7</td>
</tr>
<tr>
<td>2</td>
<td>High Group &gt;= Low Group</td>
<td>14.4</td>
<td>21.6</td>
<td>17.1</td>
<td>22.3</td>
</tr>
<tr>
<td>3</td>
<td>Low Group &gt; 2*High Group</td>
<td>39.3</td>
<td>40.6</td>
<td>35.2</td>
<td>39.3</td>
</tr>
<tr>
<td>4</td>
<td>High Group &gt;= 2*Low Group</td>
<td>6.7</td>
<td>10.1</td>
<td>8.2</td>
<td>10.9</td>
</tr>
<tr>
<td>5</td>
<td>Distractors With No Responses</td>
<td>39.9</td>
<td>28.8</td>
<td>40.9</td>
<td>30.0</td>
</tr>
<tr>
<td>6</td>
<td>Total Distractors</td>
<td>596</td>
<td>2020</td>
<td>2944</td>
<td>2696</td>
</tr>
</tbody>
</table>

**DISCUSSION**

Item analyses of published multiple-choice questions are rare in the literature. Rarer still are analyses of distractor options specifically. The present study serves the purpose of evaluating two
current question banks, providing information for consideration by potential adopters. Comparisons of these current edition results with those of their earlier editions might signal a refinement of the questions. Both LW (2012 compared with 2009) and SZP (2011 compared with 2008) show modest improvement with respect to distractors performing as they are expected to (Table 4, row 1) as well as modest improvement in there being fewer dysfunctional distractors (Table 4, row 2). Across the five question banks, the two editions of SZP (2011 and 2008) anomalously have both the highest percentages of properly and improperly discriminating distractors (Table 4, rows 1 and 2, respectively). (This is possible in light of those two question banks having markedly fewer distractors receiving no responses (Table 4, row 5) than the other three banks.)

Results here generally indicate the potential for further improvement of the question distractors. Between distractors that do not discriminate (Table 4, row 5) and distractors that dysfunctionally discriminate (Table 4, row 2) over half of the distractors analyzed here do not serve to beneficially discriminate between high and low performing students. This is the case for each of the five question banks. As explained earlier, results in Table 4 are for extreme student subgroups, the highest performing 27 percent and the lowest performing 27 percent. A median split of students, based on their overall exam scores, of course includes all students in the analysis. On this basis, the percentages of distractors that do not discriminate or discriminate dysfunctionally range from 47.1 (SZP, 2011) to 55.0 (LW, 2009) across the five texts.

As evidenced by the texts included in this research, many texts have evolved over multiple editions, each edition presenting the opportunity to revise the accompanying bank of multiple-choice questions. Too, requisite data for several types of item analysis would seem to be plentiful and readily available.

That writing effective distractors is challenging is widely recognized: “The major short-comings of multiple-choice questions are, first, the difficulty of writing good distractor options...” (Gregory, 2011, p. 140); “When an individual item is being written, the number of potentially meaningful, relevant distractors is far more limited [than the universe of items]; the law of diminishing returns very quickly takes over...the search for good distractors after three or four good ones have already been found is likely to be frustrating and fruitless.” (Wesman, 1971, p. 99-98); “The use of five alternatives is probably the upper limit...due to the difficulty in developing plausible distractors...” (Reynolds & Livingston, 2012, p. 198) Still, conditions for refining multiple-choice questions are supportive of doing so.

Analysis of published multiple-choice question banks provides information that might be considered by potential adopters. Depending on the frequency with which new editions are published, presumably with an at least somewhat revised bank of questions, it may be infeasible to gather sufficient data and to publish analyses of those data before a new edition of the question bank is published. (The HMB, 2007 text has been supplanted by two editions.) However, where many of the questions in a preceding edition are repeated in the current edition, analysis of the preceding edition questions may still apply, if not perfectly accurately. Too, a corresponding series of analyses

668773-7
such as the present one would establish a track record, with improvement being a factor to be considered in adoption.

Finally, analyses of both current and past editions of question banks provide necessary benchmarks or norms against which to compare similar analyses of other question banks.

REFERENCES


Dickinson Distractor Discrimination


