ISSN 2164-2877 (print)
ISSN 2164-2885 (online)

The Relationship between Classified Difficulty and Implausible Distractors in Multiple-Choice Questions

J. Alexander Smith
Oklahoma City University
Meinders School of Business
Oklahoma City, OK, USA

John R. Dickinson
University of Windsor
Windsor, Ontario N9B 3P4
Canada

Corresponding Author: J. Alexander Smith, asmith@okcu.edu

ABSTRACT

Published banks of multiple-choice questions are ubiquitous, the questions in those banks often being classified into levels of difficulty. The specific level of difficulty into which a question is classified might or should be a function of the question's substance. Possibly, though, insubstantive aspects of the question, such as the incidence of incorrect answers that are readily dismissed, also affect the difficulty level into which a question is classified. The present research investigates the relationship between classified question difficulty and the incidence of implausible incorrect answer options.

Introduction

Assessment of teaching and learning is not a new idea, but it has gained increased importance in the last two decades. This is primarily due to increased demands of accountability by funding organizations (e.g. governments) and requirements by accreditation associations (Bollag 2006; AACSB 2013a). The Association to Advance Collegiate Schools of Business (AACSB), for example, has integrated assessment, what they call assurance of learning, as a critical component of accreditation standards. Per AACSB, assurance of learning is defined as, "… processes for demonstrating that students achieve learning expectations for the programs in which they participate" (AACSB 2013b, p. 29). This makes intellectual sense as the students should be receiving what they pay for in terms of their education.

Multiple choice exams are a common method for evaluating student learning and can be used as an assessment tool (Santos, Hu, and Jordan 2014). These exams have the benefit of being easy to distribute and grade. Questions in published multiple-choice question banks are commonly classified into three levels of difficulty. "One of the most important responsibilities of the [instructor] is to define the level and the distribution of the difficulties of the items that are to compose the final test" (Tinkelman 1971, p. 62). With this, instructors might well rely on the published difficulty classifications when constructing an exam. Thus, instructors have a fundamental interest in the validity of the published classifications. In turn, serving that interest is the responsibility of those who write and classify questions. (It is possible too, of

ISSN 2164-2877 (print)
ISSN 2164-2885 (online)

course, that instructors themselves might analyze published questions and refine or discard them accordingly.) Since publishers rarely describe how questions are classified, the validity of those classifications is difficult to assess. One available means, though, is through the analysis of certain elements of the questions; specifically, the incorrect answer options, i.e., distractors, and the (im)plausibility of those distractors. A question may to be classified as, *easy*, not because the object of the question is easy because the question stem and correct answer are purposely framed, to provide cues. Instead, the classification is selected because of the insubstantive aspect that some of the distractors are so obviously incorrect that they do not serve their purpose of attracting responses from students who do not know the correct answer.

The present research used samples from selected question banks, and shows there does exist a relationship between classified difficulty and the presence of implausible distractors. Implications are that those who write and classify published questions might be mindful of this relationship when composing distractors in the first place and then strive to ensure that the question difficulty classification is not simply due to the presence of implausible distractors. For particular question banks, question writers/publishers might conduct analyses similar to those in this research prior to publication. The analysis may focus on not only the percent of correct responses, but also more broadly on distractor responses. Those analyses, too, can be ongoing after publication with the goal of refining subsequent editions of question banks. Likewise, for a particular bank of interest, instructors might also conduct such analyses. This could be for the purpose of refining/discarding questions, but also more broadly as a measure of the validity of the published question bank with an eye to continuing or discontinuing its use.

Background

Distractors (or foils or misleads), i.e., the incorrect answer options, are an integral component of multiple-choice questions. As such, distractors play a major role in determining the properties of the questions. "The content of an item can be altered radically by changing the distractors, while keeping the correct response the same" (Cronbach 1971, p. 454). The essential purpose of distractors is to attract responses from examinees who do not know the correct answer. Distractors that fail to distract, then, do not serve their basic purpose.

> 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)

Recognition of the importance of effective distractors is widespread. "Make all distractors plausible and attractive to examinees who lack the information or ability tested by the item" (Wesman 1971, p. 116). (Later paraphrased by Millman & Greene [1989, p. 353]: "Make all

ISSN 2164-2877 (print)
ISSN 2164-2885 (online)

options plausible and attractive to examinees who lack the information or ability referenced by the item.") "Distractors that are hardly ever chosen are too transparently incorrect and can be omitted or, preferably, replaced" (Nunnally & Bernstein 1994, p. 301) and, "...adding distractors that fail to distract cannot improve the utility of the item" (Wesman 1971, p. 100). "In multiple-choice tests he [the test writer] learns which distractors (wrong answers) or misleads are not functioning, as shown by their relative unpopularity" (Guilford 1954, p. 417)

The common implication of these observations is that implausible distractors compromise the essential effectiveness of multiple-choice questions. The more specific compromise in the present research is the material degree to which obviously incorrect answers and published question difficulty classifications are related. That is, to some material extent, the published classifications appear to be a function of some distractors not serving their essential purpose. This, in turn, implies a need for greater care in writing effective distractors originally and refining ineffective distractors from one edition of the question bank to the next edition.

This recognized importance notwithstanding, Dickinson (2013) has shown that for samples of questions from several question banks there is a substantial presence of ineffective distractors. Across five question banks, the percent of sample questions having at least one distractor attracting no responses ranged from 53.53% to 70.89%. The percent of questions with at least one distractor attracting ten percent or less of total responses ranged from 97.02% to 99.16%.

The primary implication of the use of implausible distractors is that the measurement of examinees is compromised. "This [measuring students' levels of comprehension] does not result if the test questions are such that little real knowledge is needed by the testee because of the ease of eliminating ridiculous or remote possibilities in the incorrect choices" (Weitzman & McNamara 1946, p. 517)

Implausible distractors, then, are one basis of the present research. Banks of multiple-choice questions accompany most introductory level textbooks in business. The questions, typically, are classified into three levels of difficulty–*easy, medium, hard*–and often on other dimensions such as skill type. Those classified levels of difficulty are the second basis of the present research. The specific focus is to investigate, in selected banks of questions, whether classified level of difficulty is related to the presence of implausible distractors.

Various scenarios might lead to such a relationship. In classifying a given question, the question writer may consider what he/she deems to be the effectiveness of its distractors. Possibly, too, the writer may actively construct the effectiveness of distractors toward some desired difficulty level. Perhaps a question is classified without conscious regard for its distractors, yet the distractors exert a subtle influence nevertheless.

The essential effect of implausible distractors is to render a question easier to answer. Fundamentally, then, it was hypothesized that the number of implausible distractors and the

ISSN 2164-2877 (print)
ISSN 2164-2885 (online)

published classified difficulty of a question would be inversely related.

Data

Multiple-choice question banks accompanying six texts were examined. Among the six were two editions of a consumer behavior text plus a second consumer behavior text and three editions of a retailing text. The texts, the total number of multiple-choice questions in the respective banks, and the number of questions sampled from each question bank are reported in Table 1.

Table 1
Bank and Sample Question Counts

Text	Total Questions	Sample Questions	Percent of Total
Levy, Weitz, & Grewal (2014, LWG), *Retailing Management*, Ninth Edition	1229	307	25
Levy & Weitz (2012, LW), *Retailing Management*, Eighth Edition	1211	624	51.5
Solomon, Zaichkowsky, & Polegato (2011, SZP), *Consumer Behaviour*, Fifth Canadian Edition	1148	671	58.4
Levy & Weitz (2009, LW), *Retailing Management*, Seventh Edition	1332	736	55.3
Solomon, Zaichkowsky, & Polegato (2008, SZP), *Consumer Behaviour*, Fourth Canadian Edition	1019	674	66.1
Hawkins, Mothersbaugh, & Best (2007, HMB), *Consumer Behavior*, Tenth Edition	1624	958	59

Courses.

Providing data for the present analyses were undergraduate courses in retailing and consumer behavior typically taken in the third year of a student's university program. Both courses have prerequisites of two semester-long principles of marketing courses. The same instructor taught all classes on a project basis. During the time under study, no changes to the course format, content or methodology occurred.

ISSN 2164-2877 (print)
ISSN 2164-2885 (online)

Examinations

For each class the first exam covered approximately the first third of the chapters, the second exam covered about the middle third of the chapters, and the noncumulative final exam covered the last third of the chapters. Reflecting the project basis for both courses, exams were based solely on the assigned textbook. That is, exams entirely comprised multiple-choice questions drawn from the published question bank.

Each of the exams counted for 20 percent of the students' final course grades (the project counting for the remaining 40 percent). Exams were scored as the percent of questions answered correctly; no penalty was deducted for incorrect answers. In the very few instances where multiple answers were given, those questions were excluded from the present research. Numbers of exams and students are reported in Table 2.

Table 2
Exams and Students

Text (de-identified)	Student-Exams [a]	Questions per Exam [b]	Students per Exam [b]	Score [b]
LWG (2014), 9th	260	51.2	43.3	73.3
LW (2012), 8th	456	52.3	38.0	69.5
SZP (2011), 5th	503	55.9	41.9	58.2
LW (2009), 7th	434	61.3	36.2	67.4
SZP (2008), 4th	479	56.2	39.9	61.1
HMB (2007), 10th	588	53.2	32.7	62.7

a A student-exam is one student taking one exam.
b mean

Sampling Method

Multiple-choice questions are arranged in the test question bank per 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. Questions were sampled on a chapter-by-chapter basis. Parameters guiding the sampling were a total of 50-60 questions per exam (in light of the 90-minute class period), the total number of published questions for the chapter, and the anticipation that the text would be adopted for a certain number of classes within and across successive semesters.

For example, for the first exam of a given text, the first question in the chapter was selected followed by every n-th question. Consider a chapter having 80 published multiple-choice questions and the anticipation that questions from the text eventually would be drawn for six classes. For the first exam, the first question in the chapter was selected followed by every nth

ISSN 2164-2877 (print)
ISSN 2164-2885 (online)

question. For the example parameters just given, n equaled 9 and questions 1, 10, 19, 28, 37, 46, 55, 64, 73 were for the first exam. For the sixth exam, questions 6, 15, 24, 33, 42, 51, 60, 69, 78 were selected. Within a given question bank, no questions were repeated across exams.

This systematic sampling approach was an attempt to ensure that:

- a cross section of each chapter content was included among the examination questions,
- all respective midterm and final examinations were of comparable composition, and
- a representative sample of the text bank questions was obtained.

Counts of test bank and sample questions are reported in Table 1. All questions analyzed had five answer options: the correct answer plus four distractors.

Analysis

The purpose of this research is to investigate whether classified question difficulty level and the incidence of implausible distractors are related. Classified question difficulty level is drawn from the published question bank. Distractors attracting no or few responses were measured in three ways for each question:

- The number of distractors attracting zero responses. (Where all four distractors attracted zero responses, all students answered the question correctly.)
- The number of distractors attracting less than or equal to 5 percent of total responses (the total including correct responses).
- The number of distractors attracting less than or equal to 10 percent of total responses.

Percentages of questions with distractors attracting each of the three levels of implausibility are reported in Table 3.

ISSN 2164-2877 (print)
ISSN 2164-2885 (online)

Table 3
Percent of Questions by Level of Implausible Distractors (count)

Text	0%	≤5%	≤ 10%
LWG (2014), 9th (n=307)	68.7 * (n=211)	90.6 (n=278)	99.3 (n=305)
LW (2012), 8th (n=624)	72.4 (n=452)	91.5 (n=571)	98.9 (n=617)
SZP (2011), 5th (n=671)	53.4 (n=358)	86.3 (n=579)	97.0 (n=651)
LW (2009), 7th (n=736)	70.2 (n=517)	91.4 (n=673)	99.0 (n=729)
SZP (2008), 4th (n=674)	56.1 (n=378)	85.0 (n=573)	97.6 (n=658)
HMB (2007), 10th (n=958	67.0 (n=642)	91.5 (n=877)	99.2 (n=950)

* 68.7 percent or 211 of the 307 sample questions had at least one distractor that attracted zero percent of student answers.

The type of analysis was rank correlation, using each of the above three operational definitions of implausible distractors separately. The data were first organized into cross-tabulation tables, the rows comprising published question difficulty level–*easy, medium, hard*–and the columns comprising the number of qualifying distractors–ranging from 0 to 4. These tables may be seen to be ordered contingency tables and analyzed accordingly (Gibbons 1993, pp. 60-80).

Numerous measures of rank correlation for ordered tables are available. Perhaps the most commonly used is Spearman's rho and that is reported here. (As with rho, all the other available types of rank correlations were negative – as was hypothesized – and were statistically significant, $p<.001$.)

Both the rows and columns of the ordered table are arranged in ascending order. The rows are arranged in order of increasing difficulty. The columns are arranged in order of increasing ineffectiveness. With this arrangement, the correlation between classified question difficulty and the effectiveness of its distractors was expected to be negative.

ISSN 2164-2877 (print)
ISSN 2164-2885 (online)

Results

Table 4 presents the rank order correlations between published difficulty classification level (1 to 3) and the number of distractors (0 to 4) attracting the noted percent of responses. (The percent of responses is based on all responses, including responses to the correct answer option.)

Table 4
Spearman Rho Rank Correlations between Classified Question Difficulty and the Number of Implausible Distractors

Text	0% [a]	≤ 5% [b]	≤ 10% [c]
LWG (2014), 9th	-0.0692 **	-0.0961 *	-0.0717 **
LW (2012), 8th	-0.1593	-0.2187	-0.1655
SZP (2011), 5th	-0.2426	-0.3342	-0.3144
LW (2009), 7th	-0.2317	-0.2739	-0.2335
SZP (2008), 4th	-0.2456	-0.2981	-0.3554
HMB (2007), 10th	-0.2159	-0.2502	-0.2690

a Implausible distractors defined as those attracting 0% of responses.

b Implausible distractors defined as those attracting ≤5% of responses.

c Implausible distractors defined as those attracting ≤10% of responses.

All one-tail p-values < .001, except * p<.05, and ** p<.12

Except for LWG (2014) all the correlations are highly statistically significant (one-tail p-value < .001). Though approaching statistical significance, the larger p-values for LWG (2014) may be a function of the relatively smaller sample of questions from that bank. Possibly, too, they may reflect a refinement of the questions in that edition of the question bank. The correlations are material, with all but five having absolute values greater than 0.20. For these samples of multiple-choice questions, published classified difficulty and the number of implausible distractors are related.

Discussion

The results of this study show that, for the samples of question banks analyzed, published classified question difficulty and the number of implausible distractors are (inversely) related.

Normatively, plausible distractors should affect question observed or measured difficulty (Dickinson 2015). Here, though, it is implausible or ineffective distractors that are related to

ISSN 2164-2877 (print)
ISSN 2164-2885 (online)

classified question difficulty. Ineffective distractors have several undesirable implications as noted earlier. Instructors relying on classified question difficulty to compile exams might be cautioned that the classification is partly due to an undesirable property of the questions.

Question writers might be similarly cautioned, though writing distractors is difficult, as is recognized by many. "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, pp. 99-100). ". . . preparation of an additional distractor may well require disproportionate additional effort on the part of the item writers" (Tinkelman 1971, p. 74). "The use of five alternatives is probably the upper limit . . . due to the difficulty in developing plausible distractors..." (Reynolds & Livingston 2012, p. 198).

The results of this research, of course, do not necessarily hold for all published banks of multiple-choice questions. There exist any number of guides for writing multiple-choice questions (Gregory 2011, p. 140; Haladyna 2004; Reynolds & Livingston 2012, pp. 197-202; Wesman 1971). The many different question writers, though, are not necessarily in lock-step with those guides. Nor do those guides encompass relevant human characteristics of item writers such as subject expertise, ingenuity, empathy with target students, and straightforward expression.

The consistency of the results across the several test banks (those of multiple editions no doubt having several duplicated questions), though, suggests some reliability of the findings.

Data for replicating this research are plentiful and easily obtained. Such replication might further support (or not) the essential result of this study. In addition, publishers might carry out similar investigations of their question banks. Many texts publish periodic editions (LWG being in its ninth edition, SZP now being in its sixth edition, and HMB now being in its thirteenth edition). Refining the distractors (or other properties) of multiple-choice questions from edition to edition would soon see improved question banks, of benefit to publishers specifically and pedagogy generally.

References

AACSB. (2013a). Assurance of learning standards: An interpretation. Tampa, FL. AACSB International.

AACSB. (2013b). Eligibility procedures and accreditation standards for business accreditation. Tampa, FL. AACSB International.

ISSN 2164-2877 (print)
ISSN 2164-2885 (online)

Bollag, B. (2006, December 06). "Fears of possible federal learning standards grow as liberal-arts accreditor is penalized," *Chronicle of Higher Education*. Retrieved from www.chronicle.com/article/fears-of-possible-federal/119555

Cronbach, L. J. (1971). Test validation. In Thorndike, Robert L. (Ed.), *Educational Measurement*, Second Edition. Washington, D.C.: American Council on Education, 443-507.

Dickinson, J. R. (2013). How many options do multiple-choice questions *really* have? *Developments in Business Simulation and Experiential Learning*, 40, 171-175.

Dickinson, J. R. (2015). The effect of the *real* number of options on the difficulty of multiple-choice questions. *Developments in Business Simulation and Experiential Learning*, 42, 23-26.

Gibbons, J. D. (1993). *Nonparametric measures of association*. Newbury Park, CA: Sage Publications, Inc.

Gregory, R. J. (2011). *Psychological testing: History, principles, and applications*, Sixth Edition. Boston: Pearson.

Guilford, J. P. (1954). *Psychometric methods*, Second Edition. New York: McGraw-Hill Book Company.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*, Third Edition. New York: Routledge

Hawkins, D. I., Mothersbaugh, D. L., & Best, R. J. (2007). *Consumer behavior*, Tenth Edition. Boston: McGraw-Hill Irwin.

Levy, M., Weitz, B. A., & Grewal, D. (2014). *Retailing management*, Ninth Edition. New York: McGraw-Hill Irwin.

Levy, M. & Weitz, B. A. (2012). *Retailing management*, Eighth Edition. New York: McGraw-Hill Irwin.

Levy, M. & Weitz, B. A. (2009). *Retailing management*, Seventh Edition. New York: McGraw-Hill Irwin.

Millman, J. & Greene, J. (1989). The specification and development of tests of achievement and ability. In Linn, Robert L. (Ed.), *Educational measurement*, Third Edition. New York: American Council on Education and Macmillan Publishing Company, 335-366.

Nunnally, J. C. & Bernstein, I. H. (1994). *Psychometric theory*, Third Edition. New York: McGraw-Hill.

Reynolds, C. R. & Livingston, R. B. (2012). *Mastering modern psychological testing: Theory and methods*. Boston: Pearson.

ISSN 2164-2877 (print)
ISSN 2164-2885 (online)

Santos, M. R., Hu, A., & Jordan, D. (2014). Incorporating multiple-choice questions into an AACSB assurance of learning process: A course-embedded assessment application to an introductory finance course, *Journal of Education for Business*, 89, 71-76.

Solomon, M. R., Zaichkowsky, J. L., & Polegato, R. (2011). *Consumer behavior*, Fifth Canadian Edition. Toronto: Pearson Prentice Hall.

Solomon, M. R., Zaichkowsky, J. L., & Polegato, R. (2008). *Consumer behavior*, Fourth Canadian Edition. Toronto: Pearson Prentice Hall.

Tinkelman, S. N. (1971). Planning the objective test. In Thorndike, Robert L. (Ed.), *Educational measurement*, Second Edition. Washington, D.C.: American Council on Education, 46-80.

Weitzman, Ellis & McNamara, Walter J. (1946), Apt use of inept choice in multiple choice testings, *Journal of Educational Research*, *39*(7), 517-522.

Wesman, A. G. (1971). Writing the test item. In Thorndike, Robert L. (Ed.), *Educational measurement*, Second Edition. Washington, D.C.: American Council on Education, 81-129.