THE UNIFORMITY OF DISTRACTOR RESPONSE DISTRIBUTIONS IN MULTIPLE-CHOICE QUESTIONS

John R. Dickinson
University of Windsor
MExperiences@bell.net

ABSTRACT

There exists considerable theory regarding desirable properties for multiple-choice questions; i.e., properties multiple-choice questions should exhibit. Despite the ubiquity of banks of multiple-choice questions accompanying virtually every introductory textbook in business–and having done so over numerous editions–little research has been published empirically evaluating the extent to which the questions do exhibit desirable properties. The questions' distractors, i.e., incorrect options, have been subject to even less investigation compared with the staple criteria of item difficulty and item discrimination. The present study investigates the extent to which the distributions of responses to distractors exhibit the desirable property of uniformity, i.e., are the distractors equally attractive.

INTRODUCTION

Banks of multiple-choice questions accompany virtually every introductory-level textbook in business, continuing the application of that question form as it approaches its centennial. It has a few roots, but Frederick J. Kelly is credited with culminating those in his 1915 "Kansas Silent Reading Test." (Samuelson 1987, p. 120) (Interestingly, the so-called "point system" whereby tests are scored by assigning one point for a correct answer preceded the multiple-choice question. It is the structured response feature of multiple-choice questions that is its contribution. [Rogers 1995, p. 217]) As surely, the evaluation of multiple-choice questions, the field known as *item analysis*, is just as old.

Despite the ubiquity of textbook-accompanying multiple-choice question banks today, very little published research has applied the well-established methods of item analysis to these questions. Exceptions include Dickinson (2013, 2011a, 2005) and Dickinson, Faria, & Whiteley (1991) who examined consumer behavior and principles of marketing banks of questions. Those studies mainly assessed the taxonomies into which questions are often classified; e.g., as to difficulty and skill type. The studies also report summary evaluation statistics such as percent correct and point-biserial correlations (i.e., discriminating ability) and, thus, do provide some norms for multiple-choice questions. However, the store of such norms remains limited.

The present study examines the distractors (or foils or misleads), i.e., the incorrect answer options, of two editions of a consumer behavior text, one edition of a second consumer behavior text, and two editions of a retailing text. More specifically, it is the distribution of examinees' responses to distractors that is of interest. Results for current editions of the texts might be considered by potential adopters. Those results, too, along with those for earlier editions serve to provide norms against which analyses of additional question banks might be compared.

Distractors are a defining property of multiple-choice questions and a challenge for test developers: "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) "The use of five alternatives is probably the upper limit...due to the difficulty in developing plausible distractors..." (Reynolds & Livingston 2012, p. 198)

IDEAL DISTRIBUTION OF DISTRACTOR RESPONSES

"*Item analysis* is a general term for a set of methods used to evaluate test items" (Kaplan & Saccuzzo 1982, p. 144) and in the field of item analysis the ideal distribution of responses to distractors is agreed:

- "Because all distracters [distractors] should be equally plausible to examinees who do not know the correct answer, every distracter on a specific item should also be selected by approximately the same number of examinees." (Aiken 1991, p. 79)
- "A perfect test item [would mean that] people who did not know the answer would choose randomly among the possible responses...each of the possible incorrect responses should be equally popular." (Murphy & Davidshofer 1988, p. 129)
- "...the incorrect alternatives should be equally attractive to subjects who do not know the correct

answer." (Gregory 2011, p. 145)

- Item 1 demonstrates the desired pattern of answers, with incorrect answers about equally dispersed [for both high scorers and low scorers]." (Gregory 2011, p. 145)

- "If people in this latter group [i.e., those who lack knowledge or skills tested] approach the item randomly, an equal proportion should select each alternative." (Friedenberg 1995, p. 286)

  The remaining people, those who are incorrect on the item, should be equally distributed across the different distractors." (Friedenberg 1995, p. 286)

- "...the proportion [of students] marking the best answer will contain the least possible proportion due to chance when all alternatives but the best answer are of equal difficulty..." (Horst 1933, p. 231)

The present research describes how closely responses to distractors of questions in the banks analyzed approach this ideal rectangular or uniform distribution.

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.

For the five question banks nearly all questions have five response options. There are just a few exceptions (on the order of two or three questions per bank) and those exceptions are excluded here. Too, a few questions were deemed invalid in that the correct response was not clear in the text and those questions are excluded.

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.

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 8^{th} or 10^{th} question, with varying starting points) in an attempt to ensure that:

- a cross section of each chapter content was included among the examination questions,
- all three 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

Text	Total Multiple-Choice Questions
Levy, M. & Weitz, B. A. (2012), *Retailing Management*, Eighth Edition (LW 2012)	1211
Solomon, M. R., Zaichkowsky, J. L., & Polegato, R. (2011), *Consumer Behaviour*, Fifth Canadian Edition (SZP 2011)	1148
Levy, M. & Weitz, B. A. (2009), *Retailing Management*, Seventh Edition (LW 2009)	1332
Solomon, M. R., Zaichkowsky, J. L., & Polegato, R. (2008), *Consumer Behaviour*, Fourth Canadian Edition (SZP 2008)	1019
Hawkins, D. I., Mothersbaugh, D. L., & Best, R. J. (2007), *Consumer Behavior*, Tenth Edition (HMB 2007)	1624

ANALYSIS

Distractors may be viewed as categories, with so many observations (i.e., responses) in each category, and thus comprise a categorical or *qualitative* variable. As described above, the ideal distribution of distractor responses is a rectangular or uniform one. Specifically, each distractor should receive the same proportion of responses. Statistically, this means that the distribution of responses should be maximally *dispersed*.

There exist numerous measures of qualitative dispersion, such as the Index of Qualitative Variation (IQV, Mueller & Schuessler 1961). IQV equals $[J/(J-1)][1-\Sigma p_j^2]$ where J is the number of categories, i.e., the number of distractors, and p_j is the proportion of responses to distractor j. IQV is a normed measure in that it theoretically ranges from zero to one, inclusive.

IQV and other normed measures are not able to attain their theoretical maximum value of one when the total number of observations (n) is not an integer multiple of the number of categories (J), i.e., when (n Mod J)<>0. For example, 20 examinee responses across four distractors can be maximally distributed as 5-5-5-5 and IQV equals 1. However, 19 examinee responses can only be maximally distributed as, say, 5-5-5-4. Even though the 19 responses are distributed as maximally *as possible*, IQV will be less than 1 (IQV=0.9972). IQV as a measure of the idealness of distractor distributions, then, may be (and will be when [n Mod J]<>0) affected by this artifact. Angsta (Dickinson 2006) is a second normed measure of qualitative dispersion expressly designed to have the desirable property of equaling 1 when a distribution is dispersed as maximally as possible.

Both Angsta and IQV and several other measures of qualitative dispersion, however, are highly negatively skewed (Dickinson 2007). For example, the distribution 50 -22-7-1 has an Angsta dispersion of 0.70125. Within the 0 to 1, inclusive, range of Angsta it *appears* that the distribution is toward the maximum dispersion. In fact, however, of the 4,263 possible distributions of 80 observations among four categories, the 50-22-7-1 distribution is more dispersed than only 969 of the 4,263 distributions. Compared with those 4,263 all-possible-distributions, the 50-22-7-1 distribution has a dispersion greater than only 22.73 percent (=100*[969/4263]) of them. In this context it might be said that the distribution is only 22.73% of the "way" to maximum dispersion. This more intuitively interpretable measure of qualitative dispersion is the thusly named Intuit statistic (Dickinson 2011b, 2012).

Excluded Questions

As noted earlier, the few questions not having exactly five options, i.e., four distractors, were excluded from the sample of questions. Additional questions were excluded on two bases. First, where a question is answered correctly by all students there are no distractor responses and the notion of analyzing no responses is moot. Second, where a question is answered correctly by all but one student there is but a single distractor response. That single response might, anomalously, be viewed as being at once of minimum possible dispersion and of maximum possible dispersion. (Traditional measures of qualitative dispersion treat it as the former.) These questions are excluded here. Excluded questions are summarized in Table 3.

Distributions of Distractor Responses

Each multiple-choice question in this study has four distractors, each distractor, of course, attracting some number of examinee responses. Those frequencies comprise the distribution of distractor responses and the Intuit measure of qualitative dispersion was calculated for each of the sample questions.

RESULTS

Table 4 presents distributions of Intuit values

TABLE 2
SAMPLE QUESTIONS

Text	Bank Count	Sample Count	Sample as Percent of Bank	Questions per Exam [a]	Students per Exam [a]
LW (2012)	1211	624	51.5	52.3	38.0
SZP (2011)	1148	671	58.4	55.9	41.9
LW (2009)	1332	736	55.3	62.2	36.2
SZP (2008)	1019	674	66.1	56.2	39.9
HMB (2007)	1624	958	59.0	53.2	32.7

a Mean

across sample questions from the respective text banks plus summary statistics. Intuit values are readily interpreted: the higher the value the more closely a distribution of distractor responses comes to the ideal rectangular distribution.

Mean Intuit values for the respective question banks range from 0.525 (SZP 2008) to 0.491 (LW 2009) with mean values for the other three texts being in between. Overall, distractors for the five question banks are just somewhat more than half way to the ideal rectangular distribution. For each text, over 46 percent of the questions have less than half the ideal dispersion (Table 5).

Striking is the consistency of results in Table 4 across the five texts, with no text having markedly better or markedly worse distractor dispersions than any other text (though median values are slightly more varied than are the mean values). Relative to the ideal distribution, the distributions of distractor responses in these question banks are middling. And there is little basis for favoring one text bank over another.

Tables 5 and 6 provide more detailed insight into the performance of distractors for the various banks of questions. The two editions of SZP (2011, 2008) have slightly higher percentages of Intuit values greater than 0.8 than the other texts (Table 5). SZP (2011) has a materially lower percentage of questions (2.28%) where all of the distractor responses were concentrated in a single distractor option; that is, three of the four distractors attracted no responses. Across the banks, substantial percentages of

TABLE 3
EXCLUDED QUESTIONS

Text	Original Questions	100% Correct	One Distractor Response	Remaining Questions
LW (2012)	624	17 (2.72%)	36 (5.77%)	571
SZP (2011)	671	5 (0.75%)	8 (1.19%)	658
LW (2009)	736	21 (2.85%)	34 (4.62%)	681
SZP (2008)	674	4 (0.59%)	14 (2.08%)	656
HMB (2007)	958	19 (1.98%)	30 (3.13%)	909

TABLE 4
DISTRIBUTIONS OF DISTRACTOR RESPONSES

Intuit Range	LW (2012)	SZP (2011)	LW (2009)	SZP (2008)	HMB
0.9<Intuit<=1.0	10.51 [a]	10.48	11.60	11.28	10.45
0.8<Intuit<=0.9	9.28	11.85	8.22	11.59	8.69
0.7<Intuit<=0.8	8.76	8.81	8.52	8.69	10.56
0.6<Intuit<=0.7	13.49	10.94	12.04	12.80	12.76
0.5<Intuit<=0.6	6.30	9.42	9.99	9.45	8.58
0.4<Intuit<=0.5	10.33	8.97	8.96	10.98	11.77
0.3<Intuit<=0.4	9.98	10.94	8.08	9.91	9.79
0.2<Intuit<=0.3	8.41	8.97	9.99	6.71	6.60
0.1<Intuit<=0.2	9.63	9.57	8.22	6.40	9.68
0.0<=Intuit<=0.1	13.31	10.03	14.39	12.20	11.11
Mean	0.493	0.513	0.491	0.525	0.510
Median	0.500	0.531	0.513	0.556	0.524
Standard Deviation	0.308	0.294	0.310	0.296	0.297
Number of Questions	571	658	681	656	909
Distractor Responses	7140	11839	8753	10482	11636

a 10.51 percent of the 571 LW (2012) questions have an Intuit value greater than 0.9 and less than or equal to 1.0.

questions–52.43% to 69.88%–had at least one distractor that did not attract any responses. That is, substantial percentages of questions have distractors that simply 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)

Table 6 results parallel those in Table 5, the difference being that in Table 6 it is distractors as individual options that are analyzed while in Table 5 it is the collection of distractors for each question that is analyzed. Again, the two editions of SZP (2011, 2008) distractors perform somewhat better than those for other texts. SZP (2011) has the lowest percentage (18.28%) of distractors that did not attract any responses and the lowest percentage (0.57%) that attracted all of the distractor responses.

DISCUSSION

Item analyses of published question banks serves several purposes. Most directly, of course, are the results for the specific banks analyzed, allowing for evaluation and comparison of those banks; information that might be considered by potential adopters. Three of the five banks in this study have been supplanted by more recent editions, obviously making the potential adopter view technically moot. However, where many of the questions in a preceding edition are repeated in the current edition, analyses of the preceding edition may still apply, if not perfectly accurately. Too, a corresponding series of analyses such as the present one would establish a track record, with improvement being a factor to be considered in adoption. Neither LW (2012, 2009) nor SZP (2011, 2008) displayed material improvement from the earlier to later edition.

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. Conditions for refining multiple-choice questions are supportive of doing so. This analysis of distractor dispersions and other types of item analyses might provide a *pro forma* for, and encouragement of, future analyses.

TABLE 5
ADDITIONAL RESULTS ACROSS QUESTIONS

Questions...	LW (2012)	SZP (2011)	LW (2009)	SZP (2008)	HMB (2007)
Intuit > 0.8	19.79%	22.34%	19.82%	22.87%	19.14%
Intuit <= 0.5	51.66%	48.48%	49.63%	46.19%	48.95%
Intuit = 1 (ideal)	5.43%	2.89%	4.99%	3.81%	5.61%
Intuit = 0 (all responses in a single distractor)	7.71%	2.28%	7.64%	4.27%	4.18%
At least one distractor of 0%	69.88%	52.43%	67.84%	54.88%	65.24%
Total questions	571	658	681	656	909

TABLE 6
ADDITIONAL RESULTS ACROSS DISTRACTORS

Distractors...	LW (2012)	SZP (2011)	LW (2009)	SZP (2008)	HMB (2007)
With 0% responses	27.28%	18.28%	27.13%	19.74%	24.37%
With 100% responses	1.93%	0.57%	1.91%	1.07%	1.05%
Total distractors	2284	2632	2724	2624	3636

Finally, analyses such as the present one provide necessary benchmarks or norms against which to compare similar analyses of other question banks.

REFERENCES

Aiken, Lewis R. (1991). *Psychological testing and assessment (7th ed)*. Needham Heights, MA: Allyn and Bacon. ISBN: 0-205-12864-5

Dickinson, John R. (2013), "How many options do multiple -choice questions *really* have? In Marian Boscia (Proceedings Editor), *Developments in business simulation and experiential learning*, Association for Business Simulation and Experiential Learning, Vol. 40, March 2013 (Bernie Keys Library), 171-175. ISBN: 0278-2375

Dickinson, John R. (2012). The standard error of the *Intuit* measure of qualitative dispersion. In Hale Kaynak (Coordinator), *2012 Proceedings*, Decision Sciences Institute 43rd Annual Meeting, 88201-88206.

Dickinson, John R. (2011a). An intuitive measure of qualitative dispersion. In Kaushik Sengupta (Proceedings Coordinator), *2011 Proceedings*, Decision Sciences Institute 42nd Annual Meeting, 541-546.

Dickinson, John R. (2011b). The difficulty and discriminating ability of a consumer behavior multiple-choice question bank. In Raji Srinivasan & Leigh McAlister (Eds.), *2011 Proceedings*, Vol. 22, American Marketing Association Winter Educators' Conference, 25-26.

Dickinson, John R. (2007). The asymmetry of measures of qualitative dispersion. In Pavur, Robert J. (Proceedings Coordinator), *2007 Proceedings*, Decision Sciences Institute 38th Annual Meeting.

Dickinson, John R. (2006). A new statistic for item analysis. In Harlan E. Spotts, Harlan (Ed.), *Proceedings*, Vol. XXIX, Annual Conference of the Academy of Marketing Science, 206.

Dickinson, John R. (2005). An assessment of a consumer behavior multiple-choice question taxonomy. In Kathleen Seiders, & Glenn B. Voss (Eds.), *Marketing theory and applications*, Vol. 16, American Marketing Association Winter Educators' Conference, February, 22-23.

Dickinson, John R., Faria, A. J., & Whiteley, T. Richard (1991). An empirical investigation of the validity of McCarthy's multiple-choice question classification matrix. *Journal of marketing education*, Vol. 13 (Summer), 54-66.

Friedenberg, Lisa (1995). *Psychological testing: design, analysis, and use*. Boston: Allyn and Bacon. ISBN: 0-205-14214-1

Gregory, Robert J. (2011). *Psychological testing: history, principles, and applications (6th ed.)*. Pearson. ISBN-10: 0205782140, ISBN-13: 9780205782147

Hawkins, Del I., Mothersbaugh, David L., & Best, Roger J. (2007). *Consumer behavior (10th ed.)*. Boston: McGraw-Hill Irwin. ISBN-13: 978-0-07-310137-8, ISBN-10: 0-07-310137-0

Horst, Paul (1933). The difficulty of a multiple choice test item. *Journal of educational psychology*, Vol. 24, Issue 3, 229-232.

Kaplan, Robert M. & Saccuzzo, Dennis P. (1982). *Psychological testing*. Monterey, CA: Brooks/Cole Publishing Company.

Levy, Michael & Weitz, Barton A. (2012). *Retailing management (8th ed.)*. New York: McGraw-Hill Irwin. ISBN-13: 978-0-07-353002, ISBN-10: 0-07-353002-6

Levy, Michael & Weitz, Barton A. (2009). *Retailing management (7th ed.)*. New York: McGraw-Hill Irwin. ISBN-13: 978-0-07-338104-6, ISBN-10: 0-07-338104-7

Mueller, John H. & Schuessler, Karl F. (1961). *Statistical reasoning in sociology*. Boston: Houghton Mifflin.

Murphy, Kevin R. & Davidshofer, Charles O. (1988). *Psychological testing*. Englewood Cliffs, NJ: Prentice Hall. ISBN: 0-13-732587-8

Reynolds, Cecil R. & Livingston, Ronald B. (2012). *Mastering modern psychological testing: theory and methods*. Pearson. ISBN-10: 020548350X, ISBN-13: 9780205483501

Rogers, Tim B. (1995). *The psychological testing enterprise: an introduction*. Pacific Grove, CA: Brooks/Cole Publishing Company. ISBN: 0-534-21648-X

Samelson, Franz (1987). Early mental testing. In Michael M. Sokal (Ed.), *Psychological testing and American society, 1890-1930*. New Brunswick: Rutgers University Press, pp. 113-127. ISBN: 0-8135-1193-3

Solomon, Michael R., Zaichkowsky, Judith L., & Polegato, Rosemary (2011). *Consumer behaviour (5th Canadian ed.)*. Toronto: Pearson Prentice Hall. ISBN: 978-0-137-01828-4

Solomon, Michael R., Zaichkowsky, Judith L., & Polegato, Rosemary (2008). *Consumer behaviour (4th Canadian ed.)*. Toronto: Pearson Prentice Hall. ISBN-13: 978-0-13-174040-2, ISBN-10: 0-13-174040-7

Wesman, A. G. (1971). Writing the test item. In Thorndike, Robert L. (Ed.), *Educational measurement (2nd ed.)*. Washington, D.C.: American Council on Education, 81-129. ISBN: 0-8268-1271-6