A *TaxI* Analysis of an International Marketing Multiple-Choice Question Bank

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

ABSTRACT

Published banks of multiple-choice examination questions are ubiquitous. Many of the banks classify the questions into three levels of difficulty. The accuracy of those difficulty classifications, though, has been only sparsely investigated. This study assesses the accuracy of those classifications for a question bank accompanying a widely adopted international marketing textbook. The research employs a recently introduced statistic, TaxI, and complementary analyses associated with TaxI.

INTRODUCTION AND PURPOSE

Banks of multiple-choice type questions are a staple accompanying virtually all marketing textbooks and those of other business disciplines. Very often, in these banks multiple-choice type questions are classified by the authors as to level of difficulty. (Often, too, questions are classified on a second dimension such as skill, etc.) However, in no instances known to this author are the bases, much less empirical support, for these classifications published. (This observation is not to suggest that the classifications are baseless, just that those bases are not made public.)

Question difficulty may be a major consideration when instructors select specific questions to comprise an exam. As a basis for selecting questions, question difficulty may serve a variety of purposes. These include setting the overall difficulty of the exam (Aiken 1991, p. 194), motivating examinees (Cronbach 1984, p. 39; Henrysson 1971, p. 139; Reynolds & Livingston 2012, pp. 223, 225), and ordering questions (Cohen & Swerdlik (2010, p. 255; Rogers 1995, p. 391) . Presumably, then, the published difficulty taxonomy should serve as a useful guide to question selection.

That usefulness, though, is contingent on the published difficulty classifications being accurate. Recently, Dickinson (2013) has introduced a statistic, the *Taxonomy Index* (*TaxI*) that measures that accuracy. The *TaxI* statistic is accompanied by several related complementary analyses. The present study applies that suite of analyses to the question bank accompanying a widely adopted–now in its seventeenth edition–international marketing text, Cateora *et al.* (2016, hereafter Cateora) toward describing the accuracy of the difficulty classifications. That accuracy, of course, should be of interest to adopters of the text. Also, though, the analyses applied here may provide encouragement of and a *pro forma* for similar analyses of other question banks and also provide a contribution toward establishing norms for such banks.

THE CATEORA TEST QUESTION BANK

The Cateora test question bank contains (1) true-false, (2) multiple-choice, and (3) essay questions. The present study includes only the multiple-choice type of question. As with most such published taxonomies, question difficulty is categorized by Cateora into three levels: Easy, Medium and Hard. (Their taxonomy has two additional dimensions: Bloom's types and AACSB categories.) The nominal population of Cateora multiple-choice questions numbers 1180 questions. Creditably, in the course of administering exams, only two sample questions were deemed invalid on the basis of there being no clear correct answer in the text. Thus, the research population here comprises 1178 questions.

QUESTION DIFFICULTY

In the present study is question difficulty is operationalized as the percent of correct responses for a given question. Percent correct is a near-universally prescribed measure of item difficulty (Anastasi & Urbina 1997, p. 173; Gregory 2011, p. 141; Guilford 1954, p. 418; Gulliksen 1950, p. 366; Henrysson 1971, p. 139; Nunnally & Bernstein 1994, p. 301). The mean percent correct over all 425 questions in this study was 81.28, with no questions being answered incorrectly by all students and 25 questions being answered correctly by all students.

EXAMINATIONS

Providing data for the present analyses were a total of six examinations administered across three sections of an international marketing course taught by the same instructor using a common format and evaluation scheme. The undergraduate course is typically taken in the fourth year of a student's university program and has as its prerequisites two semester-long principles of marketing courses.

The Cateora textbook comprises 19 chapters. For each class the first midterm examination covered chapters one through six, the second midterm exam covered chapters seven through twelve, and the noncumulative final exam covered chapters thirteen through nineteen. All exam questions were selected from the Cateora test bank as described below. Each of the three exams counted for one-third of the students' final course grades.

Exams were scored as the percent of questions answered correctly; no penalty was deducted for incorrect answers. In the very few instances where a question was not answered (16) or multiple answers were given, these were considered to be incorrect answers, both for exam scoring purposes and for the present research.

Class sizes were 33 to 47 students, there being an average of 39.26 answers to each of the Cateora questions.

SAMPLING METHOD

The present research concerns only multiple-choice type questions. Cateora multiple-choice questions are arranged in the test question bank 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, ranging from 10 to 13 questions per chapter. 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 Cateora questions was obtained.

The total sample of questions was 425 or 36.08 percent of the total population of 1178 Cateora multiple-choice type questions. The essential analyses for the present study are founded on questions in each of the three difficulty levels. The sample question counts *vis-a-vis* the bank question counts are reported in Table 1.

TABLE 1
SAMPLE AND POPULATION QUESTIONS

Classified Difficulty	Bank Count	Sample Count	Sample as Percent of Bank
Easy	683	246	36.02
Medium	434	161	37.10
Hard	61	18	29.51

TAXI ANALYSIS AND RESULTS

At the core of the analyses here is the *Taxonomy Index* (*TaxI*) statistic. The information on which the *TaxI* statistic is based is that which would normally be available with the administration of exams comprising samples of questions from a text-accompanying bank of multiple-choice questions. As noted above, those questions are virtually always classified into three levels of difficulty.

The aim of the procedure resulting in the *TaxI* statistic is to determine to what extent the *a priori* classifications are borne out empirically. The key feature of the *TaxI* statistic is the operational defining of what, empirically, constitutes easy, medium, and hard difficulty levels. Those definitions are drawn from the sample of questions itself, i.e., questions that have been included in exams that have been administered.

Suppose, for example, results are available for, say, 30 questions classified as Easy, 50 questions classified as Medium, and 20 questions classified as Hard. It is a simple enough matter to rank those 100 questions on the percent of students answering the respective questions correctly. (There do exist other measures of item difficulty, but percent correct is the most common.) Then ascertained is the number of questions classified as Easy that are among the questions having the highest 30 percents correct. That is the number of classified Easy questions that are classified correctly. Likewise, ascertain the number of classified Medium questions that are among the middle 50 rank ordered percents correct and the number of classified Hard questions that are among the lowest 20 ranked questions. The three counts of questions correctly classified–across the respective difficulty levels–are then summed and that sum is divided by the total number of sample questions. The resulting proportion is the definition of *TaxI*. *TaxI* is the proportion of questions correctly classified.

Tied Percents Correct

The key to *TaxI* is the number of questions taxonomy-classified into each of three difficulty levels, say, for example, 30 Easy, 50 Medium, and 20 Hard. *TaxI*, then, is the result of an algorithm that first rank orders questions according to percent correct and then determines within the highest 30 ranked questions the number that are classified Easy. And so on. Possibly, though, the

30th and 31st ranked questions have equal percents correct. If the two questions are of different taxonomy classifications then the resulting *TaxI* value will be affected depending on which questions are ranked 30th and 31st. In such cases, the *TaxI* algorithm reorders the tied questions so that the value of *TaxI* is maximized. This protocol serves to most favorably present the published taxonomy.

Proportional Chance Criterion

Interpretation of the *TaxI* statistic stands on its own as what it is–the proportion of multiple-choice questions that are correctly classified on the dimension of difficulty in the published taxonomy. Too, different published question banks and taxonomies may be compared on the basis of *TaxI*. A relevant benchmark for *TaxI*, though, is the proportion of questions that could be correctly classified as easy, medium, or hard on the basis of chance alone; that is, without invocation of the taxonomy classification for any given question.

$$C_{pro} = p(Easy)^{2} + p(Medium)^{2} + p(Hard)^{2}$$

Let p(Easy), p(Medium), and p(Hard) be the proportions of all questions in the sample classified in the published taxonomy in the respective three difficulty levels. Selecting a question from the bank at random and randomly–according to the bank proportions–anticipating the question will be empirically easy, medium, or hard yields the proportional chance criterion:

Comparing *TaxI* with C_{pro} is a check against the former value being at least partly an artifact of the distribution of questions across the three levels of difficulty in the question bank. As the distribution deviates more from uniform, the random chance that a question from the bank will have an empirical difficulty the same as the taxonomy-classified difficulty increases.

***TaxI* and C_{pro} for Cateora**

For the sample of questions from the Cateora bank, *TaxI* equals 49.65 percent. That is, just under half the questions are classified correctly. C_{pro} equals 48.03 percent. Using the published taxonomy classifications as a guide to question difficulty is barely more accurate than anticipating question difficulty randomly.

Classification Matrix

From the data processed to arrive at *TaxI* may also be formed a classification matrix, published taxonomy difficulty level comprising the rows and measured or empirical difficulty level comprising the columns (Table 2).

TABLE 2
CLASSIFICATION MATRIX (QUESTION COUNTS)

		EMPIRICAL		
		Easy	**Medium**	**Hard**
TAXONOMY CLASSIFIED	**Easy**	0.5813[a] (143)	0.3537 (87)	0.0650 (16)
	Medium	0.5652 (91)	0..4224 (68)	0.0124 (2)
	Hard	0.6667 (12)	0.3333 (6)	0.0 (0)

[a] Of the 246 sample questions classified in the published taxonomy as Easy, 143 or 58.13 percent are classified correctly.

The classification matrix serves several purposes. First, while *TaxI* succinctly describes the overall accuracy with which multiple-choice questions are classified, the corresponding proportions for each classified difficulty level separately may be found on the diagonal of the classification matrix.

Second, diagnostic information as to the nature of misclassifications is provided. For example, questions classified as Medium are more likely to actually, i.e., empirically, be of easy difficulty (0.5652) than medium (0.4224) and questions classified as Hard are far (infinitely!)more likely to actually be of easy difficulty (0.6667) than hard (0.0). Guiding the refinement of the existing bank of questions for subsequent editions, for Cateora *et al.* (2016) some questions classified as Medium are insufficiently difficult and all questions classified as Hard are insufficiently difficult.

Third, the classification matrix provides an indication of the degree of misclassifications. Per *TaxI*, 49.65 percent of the 425 sample questions are classified correctly. From Table 2, 43.76 percent of the 425 questions are misclassified into a difficulty level adjacent to the published classification and 6.59 percent are misclassified into nonadjacent levels. 43.76 percent and 6.59 percent are the "off-by-one" (OB1) and "off-by-two" (OB2) percentages.

Rank Correlation

Both the rows and the columns of the classification matrix are of the ordinal scale type. That is, the classification matrix is an "ordered contingency table" (Gibbons 1993, p. 60). With that, an ordered table rank correlation, such as Spearman's rho, may be applied to measure the strength of association between taxonomy-classified difficulty and empirical difficulty. A positive correlation would be expected. For Table 2, though, rho equals -0.0232. Despite that correlation being negative, the related one-tail p-value is 0.6832. The p-value prevails and the correct interpretation is that empirical and classified difficulty are not related.

DISCUSSION

The *Taxonomy Index* (*TaxI*) is a recently introduced protocol for assessing the accuracy of published multiple-choice question difficulty taxonomies. Given the ubiquity of such taxonomies, such assessment seems warranted.

For the example presented here (Cateora *et al.* 2016) the proportion of correctly classified questions is *TaxI* = 0.4965. That accuracy of the published difficulty classifications is barely greater than what would be achieved by randomly classifying questions; $C_{\text{pro}} = 0.4803$. The taxonomy is not a useful guide when selecting questions to comprise an exam based on difficulty. This limited usefulness is corroborated by the ordered table rank correlation (one-tail p-value = 0.6832).

Additional complementary findings–pattern of misclassifications, "off-by" statistics–serve a diagnostic purpose. They should prove useful in guiding the refinement of the question bank and classifying the difficulty of the questions.

Results here are specifically relevant to those interested in the Cateora multiple-choice question bank. More generally, the results are the beginning of what, with similar analyses of other question banks, might become norms for evaluating the taxonomies that accompany the banks.

REFERENCES

Aiken, Lewis R. (1991). *Psychological testing and assessment* (7th ed.). Boston: Allyn and Bacon.

Anastasi, Anne & Urbina, Susana (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Cateora, Philip R., Gilly, Mary C., Graham, John L., & Money, R. Bruce (2016). *International marketing* (17th ed.). New York: McGraw-Hill Education. ISBN: 978-0-07-7842166-1, MHID: 0-07-784216-2

Cohen, Ronald Jay & Swerdlik, Mark E. (2010). *Psychological testing and assessment* (7th ed.). McGraw Hill. ISBN-13: 9780073129099

Cronbach, Lee J. (1984). *Essentials of psychological testing* (4th ed). New York: Harper & Row, Publishers, Inc. ISBN: 0-06350249-6

Dickinson, John R. (2013). *TaxI*: A statistic describing the accuracy of multiple-choice question difficulty Classifications. In Kubacki, Krzystof (Ed.), *Proceedings of the 2013 Academy of Marketing Science Annual Conference*, Vol. XXXVI, p. 337.

Gibbons, Jean Dickinson (1993). *Nonparametric measures of association*. Newbury Park, CA: SAGE Publications, Inc. ISBN: 0-8039-4664-3

Gregory, Robert J. (2011). *Psychological testing* (6th ed.). Boston: Allyn and Bacon.

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill Book Company.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Henrysson, Sten (1971). Gathering, analyzing, and using data on test items. In Thorndike, Robert L. (Ed.), *Educational Measurement* (2nd ed.). Washington, D.C.: American Council on Education.

Nunnally, Jum C. & Bernstein, Ira H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill, Inc.

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

Rogers, Tim B. (1995). *The psychological testing enterprise: An introduction*. Pacific Grove, CA: Brooks/Cole Publishing Company.