Perspectives In Learning

Volume 16 | Number 1 Article 2

8-23-2017

The Trouble with Test Banks

Harvey Richman
Columbus State University, richman_harvey@columbusstate.edu

Molly Hrezo
Columbus State University, hrezo_molly@columbusstate.edu

The Trouble with Test Banks

Harvey Richman
Columbus State University

Molly J. Hrezo
Columbus State University

Abstract

We compared the psychometrics of quiz questions randomly selected from a test bank with the psychometrics of quiz questions the instructor had selected from the bank for quality and modified (if necessary). On multiple psychometric indices, the instructor selected/modified questions were superior to questions randomly selected from the test bank. Most notably, when compared with instructor written/modified questions, randomly selected bank questions were nearly 6.5 times more likely to contain a distractor that drew more responses than the correct answer. Details and implications are discussed.

Most instructors assign a textbook to their students. And, the great majority of textbooks are accompanied by companion test banks which are widely used. Tarrant, Knierim, Hayes, and Ware (2006) noted that, in a large sample of over 2,700 questions being evaluated for quality, only about 14% were instructor generated. The fact that quality multiple choice questions are difficult to construct (Hansen & Dexter, 1997) likely increases instructor reliance on publisher supplied test banks. Several authors have called into question the quality of test bank items (e.g., Bailey, Karcher, & Clevenger, 1998; Hansen & Dexter, 1997; Moncada & Moncada, 2010). Moncada and Harmon (2004) suggest that care be taken when choosing items from a test bank because poor test items can result in unreliable assessment of outcomes and students' feeling that the test questions were "ambiguous and unfair."

It is our belief that the proliferation of online education with automated testing makes test item quality an increasingly important topic. It is well known that student cheating is a potential problem when computerized online testing is used. One strategy a professor may elect to use to curb it is to have a unique exam generated for each student with items randomly selected from an electronic test back provided by the text publisher. Online LMS (learning management system) providers, such as Desire2Learn, tout this feature when promoting their products. The problem is that if there are a significant number of items with poor psychometrics in the bank: (1) overall reliability and validity of the exam may be lowered and more importantly, and (2) a given student, through no fault of his or her own, could have the misfortune of being dealt a particularly bad version of the test and suffer the consequences that go with it.

Numerous studies have addressed the quality, or lack thereof, in multiple choice items drawn from test banks (e.g., Bailey et al., 1998; Hansen & Dexter, 1997; Moncada & Harmon, 2004). What the majority of the studies in this literature have in common is

that the conclusions drawn have relied on either post hoc analyses of test item data sets for their psychometric properties or ratings of test items by trained judges. The present research utilized a true experimental design in a real-world classroom setting to observe differences in psychometric properties between randomly selected bank items and items selected and modified for quality by the instructor.

We hypothesized that the psychometrics of instructor selected and modified items would be superior to the psychometrics of items randomly selected from a test bank, as would be done in automated randomized testing.

Methods

Participants

Participants were students enrolled in one of three successive sections of the same upper level psychology course at a state university. Sample 1 *n* was 35, Sample 2 *n* was 32, and Sample 3 *n* was 27 for a total sample *N* of 94. As these were samples of convenience, demographic data were not collected. However, based on demographic data from several large studies conducted within the same student population, we estimate the mean age of the sample to be 21 years with a gender distribution of 70% female and 30% male.

Measures and Procedures

Students, at their option, completed two 20-item multiple choice (ABCD) quizzes. Students earned course "extra-credit" commensurate with their performance on the two quizzes. The quizzes covered material from two text chapters unused for the course. There were no lectures and no study guides to prepare with, only reading of the two text chapters. This procedure assured no advantage for the instructor-prepared quiz. One quiz (hereafter Quiz I, for Instructor) was comprised of instructor selected and modified (for quality) bank questions. The second quiz (hereafter Quiz B, for Bank) was comprised of questions randomly drawn from the same test bank. Order of presentation of quizzes I and B was counterbalanced so that half of the students completed Quiz I first and the other half completed Quiz B first. Responses were recorded on Scantron sheets and graded using Parscore software.

Results

Item analyses were conducted using Parscore software. Additional analyses (quiz - course final average correlations) utilized Microsoft Excel 2010. Formulas for item analyses and interpretations were adapted from Friedenberg (1995). When the three samples were collapsed for an analysis, the derived statistic was appropriately weighted for the three sample sizes. Samples combined ($N = 94$) mean for Quiz I was 12.56 ($SD = 3.51$) or 63% correct. Mean for Quiz B was 10.81 ($SD = 3.30$) or 46% correct. Tables 1 and 2 show item difficulty "*p*" analysis results with total sample average *p* values of .63 and .54 for Quizzes I and B, respectively. The *p* statistic examines the proportion of test takers correct on a given item. Its value can range from 0 to 1.0 with moderate p values in the .4 to .6 range being desirable.

Table 1

Item Difficulty Analyses for Quiz I

	M	*SD*	*Min*	*Max*
Sample 1	.64	.17	.28	.91
Sample 2	.61	.19	.25	.88
Sample 3	.63	.16	.26	.93

Note: Average *p* value was .63.

Table 2

Item Difficulty Analyses for Quiz B

	M	*SD*	*Min*	*Max*
Sample 1	.34	.21	.21	.85
Sample 2	.51	.21	.25	.81
Sample 3	.53	.18	.26	.85

Note: Average *p* value was .54.

Tables 3 and 4 show item discrimination analysis results with total sample average *D* values of .39 and .36 for Quizzes I and B, respectively. The *D* statistic compares proportion of test takers correct on an item in a high performing group with proportion correct in a low performing group. Its value can range from 0 to 1.0 with values closer to 1.0 being more desirable.

Table 3

Item Discrimination Analyses for Quiz I

	M	*SD*	*Min*	*Max*
Sample 1	.42	.24	.11	.78
Sample 2	.30	.28	-.11	.78
Sample 3	.54	.25	.00	.86

Note: Average *D* value was .39.

Table 4

Item Discrimination Analyses for Quiz B

	M	*SD*	*Min*	*Max*
Sample 1	.38	.18	.00	.67
Sample 2	.36	.17	.00	.67
Sample 3	.49	.21	.14	.72

Note: Average *D* value was .36.

Tables 5 and 6 show item-total correlation (point-biserial) data with total sample average r_{pb} values of .39 and .36 for Quizzes I and B. The r_{pb} values can range from 0 to 1.0 with values closer to 1.0 being more desirable.

Table 5

Item-Total Correlations for Quiz I

	M	*SD*	*Min*	*Max*
Sample 1	.38	.18	.11	.73
Sample 2	.31	.23	.15	.65
Sample 3	.46	.19	.01	.61

Note: Average r_{pb} value was .39.

Table 6

Item-Total Correlations for Quiz B

	M	*SD*	*Min*	*Max*
Sample 1	.35	.14	.10	.60
Sample 2	.32	.15	.12	.60
Sample 3	.40	.15	.07	.61

Note: Average r_{pb} value was .36.

Table 7 shows internal consistency reliability analysis results (Kuder-Richardson 20) with total sample *KR20* values of .67 and .63 for Quizzes I and B, respectively. Internal consistency reliability indicates the extent to which all test items are drawn from the same domain. This value for a given test can range from 0 to 1.0 with values closer to 1.0 being more desirable.

Table 7

Quiz Internal Consistency Reliabilities

	Quiz I	Quiz B
Sample 1	.69	.63
Sample 2	.53	.55
Sample 3	.82	.72
Average	.67	.63

Table 8 reports, for each quiz, the number of times a distractor (wrong answer) drew more hits than the correct answer. Averaged over the three samples, this event occurred on 0.67 of the 20 items on Quiz I (3.35% of questions) and on 4.33 of the 20 items on

Quiz B (21.65% of questions). In an effort to assess quiz validity, we correlated Quizzes I and B with student overall final course averages. All samples combined (N = 94), these correlations for Quizzes I and B were .31 and .28 respectively.

Table 8

Number of Times per Quiz that a Distractor Had Higher Endorsement than the Correct Answer

	Quiz I		Quiz B	
Sample 1	1	0.5%	4	20%
Sample 2	0	0.0%	5	25%
Sample 3	1	0.5%	4	20%
Average	0.67	3.35%	4.33	21.65%

Discussion

The pattern of results observed was, overall, consistent with the hypothesis that the psychometrics of instructor selected/modified questions would be superior to those for questions drawn randomly from the test bank. The instructor written/modified questions were superior to randomly drawn bank questions in terms of item discrimination "*D*," item total (point-biserial) correlations, internal consistency reliability (*KR20*), and correlation between quiz and course overall average (to address validity). However, we must note that all of these differences were small in magnitude.

Differences on the item difficulty index "*p*" require a bit more interpretation. Although the initial average *p* value for Quiz B questions (at .54) was closer to the theoretical ideal value of .5, which maximizes variability. After "correction for guessing", however, (Friedenberg, 1995) the ideal *p* value becomes .625, almost the exact average value of the Quiz I items (.63). This value is closer to the value we would look for in the real world classroom because, while an average *p* value of .5 maximizes variability, it also results in an average grade of 50. With this value, the instructor would have to shift or curve the exam scores considerably to achieve a reasonable distribution of grades. Thus, Quiz I was superior to Quiz B in terms of item difficulty "*p*" as well.

The most dramatic and compelling of our findings involved an analysis of question distractors (incorrect answers). As can be seen in Table 8, for Quiz I, there was, averaged over samples, a distractor that drew more responses than the correct answer on 0.67 of the 20 quiz items (3.35% of the questions). For Quiz B, there was, averaged over samples, a distractor that drew more responses than the correct answer on 4.33 of the 20 quiz items (21.65% of the questions). Said another way, more than 20%, nearly a quarter, of the Quiz B items were invalidated by this problem. We believe, our findings are consistent with, and reinforce, the findings reported earlier in this paper. One limitation of the present study is that it involved only one Instructor and one Text/test bank. Future replications might include multiple instructors and multiple texts across a broader range of subject areas. It should be noted that the test bank used in this study was, in fact, written by the text author. Not all are. Thus, another direction for future research might involve an examination of the quality of banks written by third parties.

So, how can instructors better insure that their exams will contain reliable and valid test items? There is clearly a need for a better understanding of psychometrics and test construction among college and university instructors. DiBattista and Kurzawa (2011), based on survey data, concluded that "Unfortunately, most postsecondary instructors are not trained in the principles of testing, and only about one-third of them even understand terms such as item discrimination and reliability."

Regardless of their initial skill level, instructors can improve their ability to select higher quality test bank items and to modify them if necessary in addition to becoming better test item writers themselves. These skills can be enhanced through a variety of strategies, such as faculty development (Naeem, Vleuten, & Alfaris 2012) and peer review (Malau-Aduli & Zimitat 2012). Guidelines are available to aid in developing these skills (e.g., Hansen & Dexter, 1997).

In summary, our findings were consistent with our hypothesis, though, overall, the magnitude of differences between the instructor selected/modified and randomly selected bank items were small. However, putting all other results aside, we believe that the findings from our distractor analysis alone are cause for concern. Recall that nearly one quarter of the 20 items on the randomized bank quiz (Quiz B) had a distractor that drew more responses than the correct answer, invalidating the questions. Additionally, we believe this study to be important because these differences were observed in a true experimental context in a real world classroom setting. We assume that similar studies must be rare if they exist because we were unable to locate any. As noted previously, we believe that the use of computerized testing with automated randomization of bank questions will increase dramatically as time goes by, making clear the need for additional attention and research in this area.

References

Bailey, C. D., Karcher, J. N., & Clevenger, B. (1998). A comparison of the quality of multiple-choice questions from CPA exams and textbook test banks. *The Accounting Educators' Journal*, *10*(2), 12-28. Retrieved from http://www.aejournal.com/ojs/index.php/aej/article/view/4

DiBattista, D., & Kurzawa, L. (2011). Examination of the quality of multiple-choice items on classroom tests. *Canadian Journal for the Scholarship of Teaching and Learning*, *2*(2). doi: 10.5206/cjsotl-rcacea.2011.2.4

Friedenberg, L. (1995). *Psychological testing: Design, analysis, and use*. Boston, MA: Allyn and Bacon.

Hansen, J. D., & Dexter, L. (1997). Quality multiple-choice test questions: Item-writing guidelines and an analysis of auditing test banks. *Journal of Education for Business*, *73*(2), 94-97.

Malau-Aduli, B. S., & Zimitat, C. (2012). Peer review improves the quality of MCQ examinations. *Assessment & Evaluation in Higher Education*, *37*(8), 919-931. doi: 10.1080/02602938.2011.586991

Moncada, S. M., & Harmon, M. (2004). Test item quality: An assessment of accounting test banks, *Journal of Accounting and Finance Research*, *12*(4), 28-39.

Moncada, S. M., & Moncada, T. P. (2010). Assessing student learning with conventional multiple-choice exams: Design and implementation considerations for business faculty. *International Journal of Education Research*, *5*(2), 15-19.

Naeem, N., Vleuten, C., & Alfaris, E. A. (2012). Faculty development on item writing substantially improves item quality. *Advances in Health Science Education*, *17*(3), 369-376. doi: 10.1007/s10459-011-9315-2\

Tarrant, M., Knierim, A., Hayes, S. K., & Ware, J. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing

assessments, *Nurse Education Today, 26*(8), 662-671. doi: 10.1016/ j.nedt.2006.07.006

HARVEY RICHMAN is Professor of Psychology in the Department of Psychology at Columbus State University in Columbus Georgia. Dr. Richman has over 20 years of teaching experience with interests in personality and abnormal behavior, quantitative assessment, psychometrics, and pedagogy.

MOLLY J. HREZO recently received her BS degree in psychology from Columbus State University.