Transnational Journal of Business

Sponsored by:
LIBERTY UNIVERSITY
SCHOOL *of* BUSINESS

VOLUME 4 • SUMMER 2019 • A PUBLICATION OF ACBSP

VOLUME 4 (SUMMER 2019)

A PUBLICATION OF ACBSP

 ISSN 2572-2557

Accreditation Council for Business Schools and Programs (ACBSP)
11520 West 119th Street
Overland Park, KS, USA 66213

Sponsored by

DETERMINING TEST BANK RELIABILITY

Olin O. Oedekoven, Michael Napolitano, Joseph Lemmon and Charles Zaiontz

ABSTRACT

With online testing becoming increasingly popular in higher education, consideration needs to be made for how such testing will produce reliable results and how test bank reliability will be calculated, particularly if questions are served to students randomly from a test bank of questions. Traditional methods for determining exam reliability are not applicable when a test bank is used to randomly generate unique tests for exam participants using an online exam delivery platform. We developed a methodology that relies upon multiple measures that collectively determines the reliability of the test bank and identifies specific questions for remediation. These measures include Item Difficulty, Item Discrimination, and Item Interchangeability. Mean scores for each of these measures, individually and collectively, determine overall test bank reliability. If a test bank question fails any one of the tests, the question is marked for remediation by replacement or modification. The purpose of this paper is to describe a process by which test bank reliability can be calculated and the results used to ensure appropriate assessment reliability and comparative results over time based on using a test bank of questions rather than a fixed assessment instrument.

KEYWORDS: test bank reliability; online exams; standardized testing; reliability analysis; item analysis

AUTHORS: Olin O. Oedekoven, Michael Napolitano and Joseph Lemmon *Peregrine Academic Services*; Charles Zaiontz, *Real Statistical Analysis* Contact: oedekoven@peregrineacademics.com

Determining Test Bank Reliability

A critical requirement for student testing within higher education is that the assessment instruments used are both valid and reliable. With online testing becoming increasingly popular in higher education, consideration needs to be made for how such testing will produce reliable results and how test bank reliability will be calculated, particularly if questions are served to students randomly from a test bank of questions.

Reliability is the degree to which an assessment instrument produces stable and consistent results over time with different participants (Phelan & Wren, 2005). Assuming the instrumentation is valid and fit for purpose, forms of reliability seek to capture a person's consistent score on the area of interest,

with the results of repeated measurements not subject to significant change (de Vet, Mokkink, Mosmuller, & Terwee, 2017). A common approach is test-retest reliability in which measures of reliability are obtained by administering the same test twice over a period of time to a group of individuals (Cozby, 2001). The scores from Time 1 and Time 2 can then be correlated to evaluate the test for stability over time. Other standardized measures of instrument reliability include split-half reliability (Arkin, Gabrenya, Appelman, & Cochran, 1979), parallel forms reliability (Gabrenya & Arkin, 1980; Sharma, Dunn, Wei, Montie, & Gilbert, 2015), inter-rater reliability using measurements by different persons on the same occasion (Aronson & Carlsmith, 1968; de Vet et al., 2017), and internal consistency reliability using different sets of items from the same questionnaire (Berm, 1974; de Vet et al., 2017).

We provide online assessment services used by tertiary education institutions for programmatic evaluation of the academic degree programs of business, accounting, early childhood education, criminal justice, healthcare administration, and public administration, as well as general education. The assessment is administered by providing students with 60-120 questions randomly selected from a test bank that includes 3,000 – 5,000 questions. Test banks are organized by topics (academic disciplines) and subjects (learning outcomes for the academic discipline). Each student receives 10 questions per topic.

Given that the nature of the assessment process depends upon a test bank of questions randomly served to students rather than an exam with a fixed number of questions, traditional measures of exam reliability are not always practical. Instead, a three-measure approach was developed that includes Item Analysis, Item Discrimination, and Question Interchangeability to characterize test bank reliability and identify defective questions for replacement or modification. This three-measure approach takes into consideration the random selection of questions from the test bank so that each student receives a unique exam and customization of the exam through topic selection by adopting institutions.

Theoretical Foundation

Internal consistency reliability is a measure of reliability used to evaluate the degree to which different test items that probe the same construct, skill, knowledge base, etc. produce similar results. The most commonly used tests of internal consistency reliability are Split-half reliability (Wagner & Flamos, 1988) and Cronbach's alpha (Cronbach, 1971; de Vet et al., 2017; Leppink & Pérez-Fuster, 2017). Kuder-Richardson 20 is also used, but it is essentially a restricted version of Cronbach's alpha (Sengathir & Manoharan, 2013).

The process of obtaining split-half reliability is begun by *splitting in half* all items of a test that are intended to probe the same area of knowledge to form two *sets* of items (Wagner & Flamos, 1988). The entire

test is administered to a group of individuals, the total score for each set is computed, and finally the split-half reliability is obtained by determining the correlation between the two total set scores.

One problem with the split-half method is that the reliability estimate obtained using any random split of the items is likely to differ from that obtained using another (Wagner & Flamos, 1988). A solution to this problem is to compute the split-half reliability coefficient for every one of the possible split-halves and then find the mean of those coefficients. This is the motivation for Cronbach's alpha (de Vet et al., 2017).

In a test bank assessment, each student is assessed based on a fixed number of questions selected from the test bank at random. For example, if the test bank contains 100 questions and each student is assessed based on 10 of these questions selected at random, then there are over 17 trillion possible tests, and so it is unlikely that any two students will receive the exact same set of questions. Since the sets of items are different, the split-half and Cronbach's alpha measurements of reliability cannot be calculated (Leppink & Pérez-Fuster, 2017; Wagner & Flamos, 1988). This means that a different approach to measuring internal consistency reliability for test bank assessments is required.

For this purpose, the Question Interchangeability test was defined. Inherently by question interchangeability, it is meant the ability to substitute one question in the test bank for another without significantly affecting the total score that an individual would receive on the test. The objective is to weed out any questions that fail the question interchangeability test.

To arrive at a specific Question Interchangeability test measurement for any particular question Q in the test bank, a two-tailed *t*-test can be performed between the total score of all the students who had question Q in their test versus the total score of the students who did not have question Q in their test.

Given the large number of students being assessed (sample sizes range from 10,000 to more than 100,000), it was found that effect size was a better metric than test significance. Thus, the Question Interchangeability index is defined to be Cohen's effect size for this test and considers a Question Interchangeability index of .20 (a small effect size) or less to be acceptable and a larger value to be unacceptable.

The Online Programmatic Assessment Service

The online programmatic assessment services are used to assess retained knowledge of students at the academic program level. Adopting schools employ these services to evaluate the effectiveness of their academic programs, identify areas for improvement, and demonstrate program outcomes to external stakeholders such as accreditation agencies.

School officials map their programmatic learning outcomes to the test bank using topic selection,

typically 6-12 topics depending on the curriculum included within the academic program. The exam is administered to students toward the end of their academic program, usually just before graduation. Each student receives a unique exam administered through a secure online exam platform that has embedded exam integrity measures. Each exam topic includes 10 questions randomly selected from the test bank. Questions are administered in groups based on the topics. Topic order is presented randomly.

Assessment results are used primarily in aggregate format to understand the academic program's strengths and opportunities for improvement based on the assessment criteria (targets) set by the adopting institution. The specific results from an adopting school can be compared to all other schools that have employed the same instrument for external benchmarking. Test bank reliability, as determined through regular psychometric analyses of the test bank, is essential so that school officials can perform appropriate comparisons between students and between student groups over time for longitudinal analysis and for external benchmarking.

Calculating Test Bank Reliability

Item Analysis is used to evaluate the effectiveness of items in a test. For the reliability analysis report, *items* are the test bank questions. Two measures are used for Item Analysis: Item Difficulty and Item Discrimination. In an exam situation, Item Difficulty (Question Difficulty) is the percentage of the sample (of students) answering a question correctly. This measure takes a value between 0 and 1 (or 0-100%). High values indicate the question is easy, while low values indicate the question is difficult. A target Item Difficulty of 60% was established with an acceptable range of 35 – 80%. Item Difficulty is examined periodically for all questions in the test bank and those whose item difficulty is outside this range are replaced or modified. While the Item Difficulty measures the difficulty of each question in a test bank, the Test Scores Difficulty measures the total exam score, or percentage correct for each student in order to understand the distribution of these measurements for each test bank topic.

Item Discrimination is a measure of how well an item (a question) distinguishes between those with more knowledge from those with less knowledge. Two measures are used for item discrimination: the Discrimination Index and the Point-Biserial Correlation. The Discrimination Index is the principal measure of item discrimination and is determined for each question. This is done by first selecting two groups of students based on their overall test scores: those with high knowledge and those with low knowledge levels. The high knowledge group consist of students whose exam score is in the top 27% and the low knowledge group consist of those in the bottom 27%.

The second step, performed for each question, is to calculate the percentage of students in the high

knowledge group who answer the question correctly minus the percentage of students in the low knowledge group who answer the question correctly. The Discrimination Index takes values between -1 and +1. The closer the value is to +1, the better the question discriminates between high and low performing students. Conversely, values near 0 indicate that the question does a poor job of discriminating between high and low performers. Negative values indicate that the question is often answered correctly by those who perform the worst on the overall test and incorrectly by those who perform the best on the overall test, which is clearly not desirable. The following guidelines are used when analyzing Discrimination Index results: Less than 0: Defective item; 0 – .199: Poor discrimination; .20 – .299: Acceptable discrimination; .30 – .399: Good discrimination; and .40 or more: Excellent discrimination.

The second measure of Item Discrimination is the Point-Biserial Correlation (also called the Item-Total Correlation) which is equal to the Pearson's Correlation Coefficient between the scores on the entire exam and the scores on the single item, i.e., a question (1 = correct answer; 0 = incorrect answer). The following guidelines are used when analyzing the Point-Biserial Correlation Coefficients: Less than 0: Defective item; 0 - .099: Poor discrimination; .10 – .199: Fair discrimination; .20 – .299: Good discrimination; and .30 or more: Excellent discrimination. When reviewing the quality of questions, both the Discrimination Index and the Point-Biserial Correlation Coefficient are taken into account.

Because the questions for each student are chosen at random from the questions in the test bank, the usual measures of reliability (split-half, KR20, and Cronbach's alpha) cannot be used in the traditional sense (Leppink & Pérez-Fuster, 2017). Instead, Question Interchangeability is used as the principal measure of reliability. Question Interchangeability refers to the ability to substitute a question in the test bank for another without significantly affecting the total score that an individual would receive on the exam. The objective is to eliminate or modify questions that fail the question interchangeability test. Question Interchangeability is determined for each question based on Cohen's *d* effect size measurement for a two-tailed *t*-test.

Cohen's Effect Size *d* (Algina et al., 2006) is calculated based on a two-tailed *t*-test comparing the total score for all the students who had that particular question in their exam versus the total score of the students who did not have that question in their exam. Cohen's Effect Size *d* measures the size of this difference, using the following criteria: small ($d \approx .20$), medium ($d \approx .50$), or large ($d \approx .80$). Test bank questions with a Question Interchangeability measurement of $d > .20$ are replaced or modified. Since the sample sizes for the questions are very large, the *t*-test result by itself is likely to show a significant difference even when the actual difference is very small. For this reason, the effect size measure-

ment is used as the criterion instead of the *p*-value from the *t*-test.

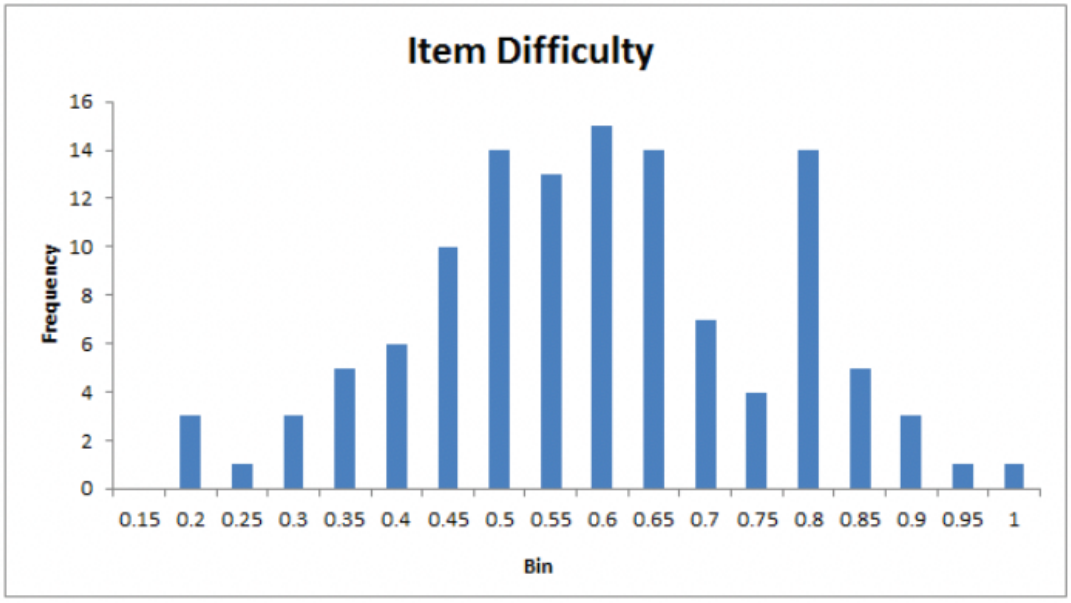
The minimum sample size used for statistically valid analyses (based on reducing Type II errors) is at least 30 uses of the question and at least 100 completed exams. If a test bank topic includes questions that fail to meet this minimum sample size, this is noted in the topic summary.

Performing Test Bank Reliability

Sixteen unique test banks are provided and maintained for programmatic analysis and standardized testing. The datasets of each test bank are aggregated into their applicable academic programs, academic degree levels, and demographic characteristics of the schools using the assessment service. Each dataset must have a minimum of 100 completed exams. Outliers, including incomplete exams, are excluded from the dataset. The aggregation of data occurs yearly, on a sliding scale, spanning four years.

Performing Item Difficulty

Item Difficulty and descriptive statistics are first calculated for each academic program and topic, an example of which is shown in Figure 1 where the topic is comprised of 119 questions and offered 11,770 times. Seven questions were below, and 10 questions were above the acceptable range, resulting in 86% of the questions within the acceptable range for Item Difficulty (35-80%). Mean Item Difficulty for this example was 57.12%.

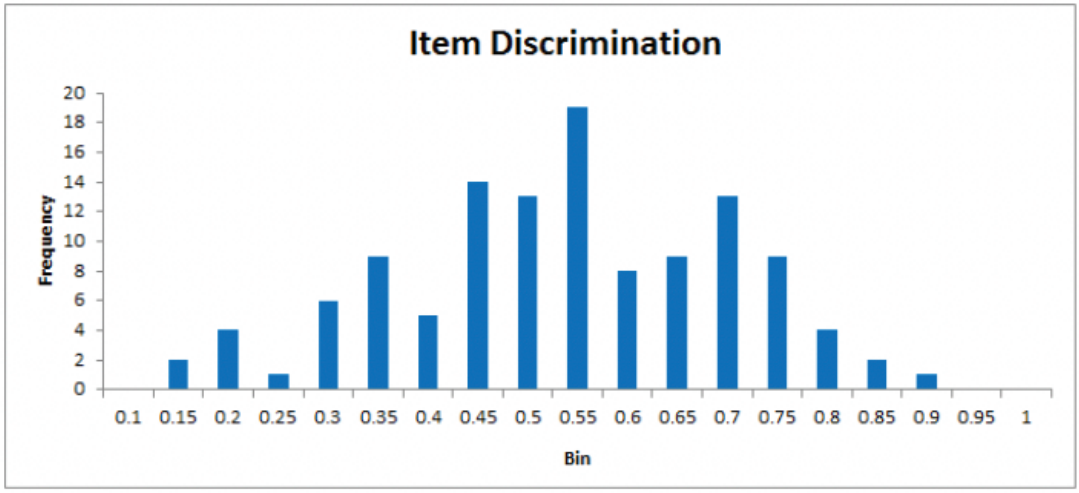


Mean	**0.5712**
Standard Error	0.01561
Median	0.57426
Mode	0.77528
Standard Deviation	0.17025
Sample Variance	0.02898
Kurtosis	-0.4144
Skewness	-0.0749
Range	0.79498
Maximum	0.96739
Minimum	0.17241
Count	119
Geometric Mean	0.54256
Harmonic Mean	0.5089
AAD	0.13705
MAD	0.11681
IQR	0.23137

Figure 1. Item Difficulty distribution for an exam topic.

Performing Item Discrimination

The next measure performed is the Discrimination Index for each question, as shown in Figure 2 based on the same data set used for Figure 1. The resulting distribution shows a range between 0.15 and 0.9. Overall, 89% of the questions offered showed good or excellent discrimination.



Mean	**0.510571**
Standard Error	0.0150905
Median	0.5174867
Mode	N/A
Standard Deviation	0.1646177
Sample Variance	0.027099
Kurtosis	-0.392312
Skewness	-0.198701
Range	0.7783369
Maximum	0.8916494
Minimum	0.1133125
Count	119
AAD	0.1324
MAD	0.11729
IQR	0.229274

Figure 2. Item Discrimination distribution for an exam topic.

One issue in calculating the Discrimination Index is to determine the 27% cutoff in a very large dataset (n > 10,000) when there are ties, in which case the cutoff may exist in the middle of a value. To avoid this shortcoming, interpolation against the upper and lower bounds of each category is administered. Consider the situation shown in Figure 3 for seventy-seven students who had question 12 in their test, where the second row (Q12) represents the score for that question for each student 0 (incorrect) or 1 (correct) and the third row (Total) shows the corresponding total score on the exam for that student.

Student	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
Q12	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	1	0	0	0	0	1	1	1	1
Total	1	1	2	2	2	3	3	3	3	3	3	4	4	4	4	4	4	4	5	5	5	5	5	5	5	5
Student	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ
Q12	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Total	5	5	5	6	6	6	6	6	6	6	6	6	6	6	6	7	7	7	7	7	7	7	7	7	8	8
Student	BA	BB	BC	BD	BE	BF	BG	BH	BI	BJ	BK	BL	BM	BN	BO	BP	BQ	BR	BS	BT	BU	BV	BW	BX	BY	
Q12	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
Total	8	8	8	8	8	8	9	9	9	9	9	9	9	9	9	9	9	10	10	10	10	10	10	10	10	

Figure 3. Seventy-seven students that received Question 12 for the exam topic.

Calculating the Discrimination Index for the data in Figure 3 is as follows.

Let $z = 0.27 * 77 = 20.79$

Thus, 20.79 is the cutoff for both the low skilled and high skilled groups. 20.79 is rounded up to 21 to find the total score cutoffs. The score of 5 is the cutoff value for the low skilled group and 8 for the high skilled group.

Students A through R are clearly in the low skilled group; only 7 out of these 18 students answered Q12 correctly. 2.79 more students are needed to satisfy this group (for a total of 20.79), but 11 students (S through AB) have a score of 5. Since 7 of these students answered Q12 correctly, it is calculated 2.79 * 7/11 = 1.775 as the Q12 contribution. Adding 7 (the contribution from students A through G), yields 8.775 correct answers in the low skilled group.

Students AY through BY are in the high skilled group; all 19 of these students answered Q12 correctly. 1.79 more students are needed for this group (for a total of 20.79), but 8 students (AY through BF) have a score of 8. Since all these students answered Q12 correctly, it is calculated 1.79 * 8/8 = 1.79. Adding 19 (the contribution from Students BG through BY) yields 20.79 correct answers in the high s killed group.

Thus, the Discrimination Index for question 12 is:

$$\frac{20.79}{20.79} - \frac{8.775}{20.79} = \frac{12.015}{20.79} = 0.5779$$

The Point-Biserial Correlation is the next measure calculated, using:

$$r = \frac{m_1 - m_0}{s}\sqrt{\frac{n_0 n_1}{n(n-1)}} = \frac{7.096 - 3.133}{2.44}\sqrt{\frac{62(15)}{77(76)}} = 0.6475$$

where r = Point-Biserial Correlation coefficient, m_1 = mean of exam scores of students who answered question 12 correctly, $m0$ = mean of exam scores of students who answered question 12 incorrectly, s = standard deviation of the exam scores for question 12, $n0$ = count of incorrect answers, $n1$ = count of correct answers, and n = count of answers (i.e., $n0 + n1$).

The Discrimination Index (0.5779) and Point-Biserial Correlation (0.6475) yield excellent discrimination for this question. The same process is repeated for each question included within the test bank.

Performing Question Intergangeability

The effect size is the final measure calculated by taking the difference between the two means divided by the pooled standard deviation. The two means are comprised of the percent score of students who had the question offered, and the percent score of those who did not, as shown in Figure 4.

77 Scores with Question 12 offered

3	5	1	5	5	...
3	2	5	3	2	...
2	1	4	3	3	...
6	6	5	7	7	...
8	4	9	9	10	...
7	8	4	7	10	...

1,100 Scores without Question 12 offered

9	9	7	9	7	3	5	7	7	8	4	...
4	5	9	1	10	6	9	8	6	5	10	...
7	9	9	8	5	8	8	7	10	3	6	...
1	2	4	4	9	7	7	5	4	4	6	...
6	6	7	5	7	10	7	8	3	8	5	...
5	10	8	5	3	4	6	7	1	1	10	...

Figure 4. Scores of students having and not having Question 12 for the test bank topic.

Cohen's *d* is calculated as follows:

$$S_{pooled} = \sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(76)2.4357 + (1099)2.1963}{77 + 1100 - 2}}$$

$$= \frac{185.1132 + 2413.7337}{1175} = 2.212$$

$$d = \frac{M_2 - M_1}{S_{pooled}} = \frac{|\,6.325 - 5.664\,|}{2.212} = \mathbf{0.2988}$$

Due to the effect size being above the target of 0.2, this question would be flagged for review despite the first two measures showing excellent discrimination.

Reporting Test bank Reliability

Using the calculations performed during the annual aggregation of the data, a report was created to review test banks segregated by topic, subject, and academic degree level. The report reviews each question included within the test bank. Highlighted questions that require attention (modification or replacement) are based on four criteria: Item Difficulty, the Discrimination Index, Point-Biserial Correlation, and Question Interchangeability.

An example of the results is shown in Table 1. Highlighted questions allow test bank reviewers to quickly identify questions for remediation. The question can then be deactivated or modified as needed. Four criteria are used for question highlighting: Item Difficulty (outside target range 35-80%), the Discrimination Index (below acceptable threshold of .20), Point-Biserial Correlation (below the acceptable threshold of .10), and Cohen's *d* (greater than the acceptable threshold of .20).

Table 1 *An Example Report of Test Bank Reliability.*

Subject: Profit, Loss, Cash Flow, and Margins			Item Discrimination				Question Interchangeability	
Question ID	Item Difficulty	Number of Times Question Offered	Discrimination Index	Evaluation	Point-biserial Correlation	Evaluation	Cohen's *d*	Evaluation
15599	0.423	6,032	0.614	Excellent	0.502	Excellent	0.026	Pass
15622	0.619	6,160	0.526	Excellent	0.427	Excellent	0.100	Pass
15644	0.399	6,200	0.451	Excellent	0.388	Good	0.029	Pass
15648	0.275	5,882	0.379	Good	0.369	Good	0.100	Pass
15650	0.390	6,006	0.537	Excellent	0.456	Excellent	0.036	Pass
15686	0.382	5,855	0.412	Excellent	0.360	Good	0.032	Pass
15689	0.678	6,257	0.503	Excellent	0.422	Excellent	0.125	Pass
15732	0.269	6,067	0.444	Excellent	0.437	Excellent	0.108	Pass
15734	0.518	5,902	0.453	Excellent	0.369	Good	0.042	Pass
58169	0.293	4,837	0.546	Excellent	0.495	Excellent	0.058	Pass
87441	0.163	3,191	0.175	Poor	0.218	Good	0.120	Pass
	Subject Difficulty Mean	Number of Times Question Offered	Discrimination Index	Evaluation	Point-biserial Correlation	Evaluation	Cohen's *d*	Evaluation
Averages:	0.401	5,671.727	0.458	Excellent	0.404	Excellent	0.070	Pass
Number of questions in table: 11			Number of questions in testbank: 11					

A summary report of the reliability analyses includes two tables that provide a summary of the data at the topic level for each test bank. The first table, as exemplified in Table 2, is a summation of data in the full report at the topic level. The second table, as exemplified in Table 3, shows statistics on the summary data at the topic level.

Table 2 *An Example of the Reliability Data Summary for the Topics Included within the Test Bank*

			Item Discrimination				Question Interchangeability	
Topic	Number of Times Questions Offered	Difficulty Mean	Discrimination Index	Evaluation	Point-biserial Correlation	Evaluation	Cohen's *d*	Evaluation
Accounting	591,615	0.511	0.428	Excellent	0.344	Good	0.059	Pass
Business Ethics	607,826	0.510	0.472	Excellent	0.405	Excellent	0.098	Pass
Business Finance	541,846	0.446	0.472	Excellent	0.404	Excellent	0.060	Pass
Business Integration and Strategic Management	531,309	0.555	0.488	Excellent	0.410	Excellent	0.051	Pass
Business Leadership	471,696	0.505	0.437	Excellent	0.368	Good	0.058	Pass
Economics: Macroeconomics	277,779	0.470	0.545	Excellent	0.470	Excellent	0.099	Pass
Economics: Microeconomics	264,101	0.492	0.529	Excellent	0.458	Excellent	0.122	Pass
Global Dimensions of Business	426,625	0.460	0.445	Excellent	0.379	Good	0.058	Pass
Information Management Systems	466,534	0.567	0.476	Excellent	0.414	Excellent	0.073	Pass
Legal Environment of Business	497,830	0.522	0.433	Excellent	0.371	Good	0.086	Pass
Management: Human Resource Management	141,253	0.569	0.588	Excellent	0.493	Excellent	0.162	Pass
Management: Operations/Production Management	170,946	0.505	0.632	Excellent	0.537	Excellent	0.133	Pass
Management: Organizational Behavior	188,174	0.586	0.604	Excellent	0.473	Excellent	0.131	Pass
Marketing	621,912	0.453	0.464	Excellent	0.381	Good	0.057	Pass
Quantitative Research Techniques and Statistics	457,952	0.481	0.500	Excellent	0.426	Excellent	0.061	Pass
	Averages:	0.509	0.488	Excellent	0.413	Excellent	0.083	Pass

Table 3 *An Example of the Reliability Data Summary for the Questions Included in Each Topic within the Test Bank*

Topic	Number of Questions*	Number of Times Questions Offered*	Difficulty Mean	Difficulty Index	Discrimination Index*	Point-biserial Correlation*	Question Interchangeability*
Accounting	113	591,615	0.511	92.92 %	100 %	100 %	100 %
Business Ethics	122	607,826	0.510	85.25 %	98.36 %	99.18 %	93.44 %
Business Finance	100	541,846	0.446	75.00 %	98.00 %	100 %	99.00 %
Business Integration and Strategic Management	134	531,309	0.555	95.52 %	99.25 %	100 %	100 %
Business Leadership	104	471,696	0.505	83.65 %	97.12 %	100 %	100 %
Economics: Macroeconomics	143	277,779	0.470	78.32 %	100 %	100 %	91.61 %
Economics: Microeconomics	117	264,101	0.492	73.50 %	100 %	100 %	83.76 %
Global Dimensions of Business	96	426,625	0.460	75.00 %	95.83 %	100 %	100 %
Information Management Systems	107	466,534	0.567	85.98 %	99.07 %	99.07 %	95.33 %
Legal Environment of Business	184	497,830	0.522	81.52 %	95.65 %	100 %	96.74 %
Management: Human Resource Management	50	141,253	0.569	84.00 %	100 %	100 %	64.00 %
Management: Operations/Production Management	42	170,946	0.505	80.95 %	97.62 %	100 %	80.95 %
Management: Organizational Behavior	72	188,174	0.586	84.72 %	100 %	100 %	73.61 %
Marketing	66	621,912	0.453	77.27 %	98.48 %	100 %	100 %
Quantitative Research Techniques and Statistics	97	457,952	0.481	82.47 %	96.91 %	100 %	96.91 %
	Averages:	417,159.87	0.509	82.41 %	98.42 %	99.88 %	91.69 %

Using Test Bank Reliability

The decision to replace or modify a test question is based on Item Discrimination, Question Interchangeability, and Item Difficulty. If an item falls below any of the desired thresholds, the item is either replaced or modified.

The decision to modify a specific test bank topic, which includes 100-400 questions, is based on the descriptive statistics for the topic and the summary of the item results for the questions included within the topic. If the topic-level results fall below desired thresholds, the entire topic is reviewed and modified with all questions included within the topic evaluated.

Conclusions

Typical measures of exam reliability, including test-retest reliability (Leppink & Pérez-Fuster, 2017), parallel forms reliability (Sharma et al., 2015), inter-rater reliability (de Vet et al., 2017), internal consistency reliability (Bonett & Wright, 2015; Mokkink et al., 2010), and split-half reliability (Arkin et al., 1979; Wagner & Flamos, 1988), are not feasible when the exam is administered by randomly selecting questions from a test bank that includes several thousand questions because the statistical assumptions for each of these methods cannot be met.

Therefore, a specific process was derived for determining test bank reliability based on a combination of Item Discrimination, Question Interchangeability and Item Difficulty. These measures of reliability provide an accurate representation of the overall test bank reliability. The process guides test bank maintenance activities that include modifying or replacing defective questions. The process described in this paper for determining test bank reliability when questions are served randomly to students can be used in a variety of academic situations including for both formative and summative assessment.

REFERENCES

Algina, J., Keselman, H. J., & Penfield, R. D. (2006). Confidence interval coverage for Cohen's effect size statistic. *Educational and psychological measurement*, *66*(6), 945-960.

Arkin, R. M., Gabrenya, W. K., Jr., Appelman, A. J., & Cochran, S. (1979). Self-presentation, self-monitoring, and the self-serving bias in causal attribution. *Personality and Social Psychology Bulletin*, *5*, 73-76.

Aronson, E., & Carlsmith, J. M. (1968). Experimentation in social psychology. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (Vol. 2). Reading, MA.: Addison-Wesley.

Bonett, D. G., & Wright, T. A. (2015). Cronbach's alpha reliability: Interval estimation, hypothesis testing, and sample size planning. *Journal of Organizational Behavior*, *36*(1), 3-15. https://doi.org/10.1002/job.1960

Cozby, P.C. (2001). *Measurement Concepts. Methods in Behavioral Research* (7th ed.). California: Mayfield Publishing Company.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.). *Educational Measurement* (2nd ed.). Washington, D. C.: American Council on Education.

de Vet, H. C., Mokkink, L. B., Mosmuller, D. G., & Terwee, C. B. (2017). Spearman–Brown prophecy formula and Cronbach's alpha: Different faces of reliability and opportunities for new applications. *Journal of Clinical Epidemiology, 85*, 45-49.
https://doi.org/10.1016/j.jclinepi.2017.01.013

Gabrenya, W. K., Jr., & Arkin, R. M. (1980). Self-monitoring scale: Factor structure and correlates. *Personality and Social Psychology Bulletin*, 13-22.

Leppink, J., & Pérez-Fuster, P. (2017). We need more replication research - A case for test-retest reliability. *Perspectives on Medical Education*, *6*(3), 158-164. http://doi.org/10.1007/s40037-017-0347-z

Mokkink, L. B., Terwee, C. B., Patrick, D., Alonso, J., Stratford, P., Knol, D. L., et al. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology, 63*, 737-745.

Phelan, C., & Wren, J. (2005). Exploring reliability in academic assessment. Resource document. University of Northern Iowa. Retrieved from https://chfasoa.uni.edu/reliabilityandvalidity.htm

Sengathir, J., & Manoharan, R. (2013). A split half reliability coefficient based mathematical model for mitigating selfish nodes in MANETs. *2013 3rd IEEE International Advance Computing Conference (IACC), Advance Computing Conference (IACC), 2013 IEEE 3rd International*, 267. http://doi.org/10.1109/IAdCC.2013.6514233

Sharma, P., Dunn, R. L., Wei, J. T., Montie, J. E., & Gilbert, S. M. (2015). Evaluation of point-of-care PRO assessment in clinic settings: Integration, parallel-forms reliability, and patient acceptability of electronic QOL measures during clinic visits. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation*, *25*(3), 575-583. http://doi.org/10.1007/s11136-015-1113-5

Wagner, E. E., & Flamos, O. (1988). Optimized split-half reliability for the Bender Visual Motor Gestalt Test: Further evidence for the use of the maximization procedure. *Journal of Personality Assessment*, *52*(3), 454.

ACBSP GLOBAL BUSINESS ACCREDITATION

ACCREDITATION COUNCIL FOR BUSINESS SCHOOLS & PROGRAMS

World Headquarters • 11520 West 119th Street • Overland Park, KS 66213 USA

USA • Belgium • Perú • China

acbsp.org • membership@acbsp.org