January 2018

Citation:

Peregrine Academic Services. (2018). *Assessment Service Validity and Reliability* (Report No. 2018-1). Gillette, WY: Author.

Peregrine Academic Services, LLC

Assessment Service Validity and Reliability

The need for program-level evaluation in higher education includes more than just accreditation, as other stakeholders also expect greater accountability through learning assessment (Murray, 2009). Although quantifying the inputs to higher education is important, perhaps even more important is measuring the change that occurs as a result of the educational experience. Continuous improvement can then be achieved when the results from the assessment are incorporated into instructional activities.

The purpose of this report is to describe the developmental history of the programmatic assessment services provided by Peregrine Academic Services to assess the retained knowledge of students enrolled in higher education for the purposes of program-level evaluation and to discuss exam validity and reliability. Conceptually and throughout the development, evaluation, and administration of the test bank, the developmental principles instituted by the entities listed were followed: American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (1985) and Cozby (2001).

EXAM SERVICES

Currently, Peregrine Academic Services provides eight distinct assessment services used for programmatic assessment, learning outcomes evaluation, assurance of learning, and satisfying accreditation requirements. The eight distinct services are:

1. Business (BUS), with undergraduate, masters, and doctoral level test banks
2. Global Business Education (GBE), with undergraduate and graduate test banks
3. Accounting and Finance (ACPC), with undergraduate and graduate test banks
4. Public Administration (PUB), with undergraduate and graduate test banks
5. Early Childhood Education (ECE), with undergraduate and graduate test banks
6. Healthcare Administration (HCA), one test bank used for both graduate and undergraduate assessments based on topic selection

7. Criminal Justice (CJ), with both graduate and undergraduate test banks

8. General Education (GEN ED), one test bank used for assessing the undergraduate GEN ED curriculum based on topic selections

Each of the exam services was developed and is maintained based upon the following procedures.

EXAM DEVELOPMENT

The exam services were each developed based upon the discipline-specific knowledge areas as defined by the accreditation/certification organization associated with the academic degree program, including the ACBSP, IACBE, AACSB, CAEP, ACJS, AACTE, AUPHA, CAHME, and NASPAA. The accreditation standards and principles from each organization were used to provide direction and focus related to the topic and subject identification. Program managers of schools typically structure their curriculum based upon the topics and most course-level and program-level learning outcomes are associated with the discipline topics (Cripps et al., 2011).

The overall construct for a summative program-level assessment for academic programs in higher education was developed in consultation with accreditation officials using identified scientific standards for measurement as described by Allen & Yen (1979). As a summative program-level instrument, the construct was to use foundational concepts associated with each discipline area for the exam questions in order to assess student retained knowledge and thus provide the institution with valuable information related to program-level learning outcomes.

The following procedures were taken, in order, to identify the specific exam concepts (a.k.a., subject areas) to include with the foundational aspects of each topic:

1. Academic officials were consulted for each of the represented disciplines.

2. Course curriculum at the undergraduate and graduate levels was reviewed.

3. Accreditation officials were engaged in order to determine accreditation-related expectations.

Based upon the concept identification activities, content for the exams was developed using a variety of techniques including subject-matter experts and commonly used course materials. Approximately 300-500 exam questions (multiple-choice with four or five responses) were developed for each exam topic in order to create the exam test bank. Unique test banks were developed based on the academic degree level as appropriate for the academic program. Exam questions focus on concept application at the foundational level, with only a limited number of questions that are definitional-based.

The test banks were then prepared for exam administration to a beta-test group of students at different universities, both within the US and when appropriate, outside of the US (the GBE

Exam Service). The exam is administered online with 10 questions per topic for a total of 100-120 questions per exam. Each exam is unique based upon a random selection of questions from the test bank. Exam questions are displayed one-at-a-time and ordered by topic. Topic order is also randomized for each exam.

An exam proctor is not required to administer the exam and there are several exam integrity measures that are built into the process, including:

- Randomized questions
- Randomized topic order
- Timed response periods for questions
- Full restriction to copy/paste from the exam window

The exams were then beta-tested with students at different universities. The beta-test included approximately 1,000 exams. The psychometric analysis (Kuder & Richardson, 1937; Nunnally & Bernstein, 1994) of beta-test data:

- Facilitated the creation of the normed scoring/grading scale
- Identified exam questions with substandard quality, which were then eliminated
- Established how subject-level scores can be combined to generate topic-level scores;
- Established the average completion time requirements, both per question and by per exam

The exams were then administered to additional students from other universities. Once an additional sample of 5,000 completed exams was obtained, further psychometric analyses were conducted on the test bank and additional refinements and improvements of exam questions were made. Periodically, a similar review of the test bank is conducted to ensure the quality of the exam and the test bank. More information regarding reliability processes is presented later in this paper.

VALIDITY

Validity is defined as the extent to which the exam results are relevant and meaningful for the purpose of the exam (Cronbach, 1971), and in this case, to assess the student retained knowledge of the selected program topics in order to assist university program managers with evaluating learning outcomes. To ensure that the exam service is valid and fit for its purpose, evidence was

collected from the first stages of test development and the first beta-testing along with ongoing validation efforts. Some of the exam validity measures include:

- Content validity was performed with exam questions written and reviewed by academic professionals within each academic discipline. All exam questions were linked to the established topics.
- All exam questions have a subject-level designation with 4-8 subjects per topic. Subject-level designation and subsequent reporting allows for direct measurement of learning outcomes based upon institution's own defined criteria.
- Regarding criterion-related validity, exam questions are based upon the accreditation requirements for program-level assessments as defined and described by the associated accreditation organization.
- Exam responses are either correct or incorrect with only one possible correct choice.
- Exam scores are determined by summarizing the percent correct: per subject, per topic, and by total score.
- Test bank quality reviews eliminated substandard questions following the initial beta-testing.
- Regarding face validity, there have been more than 5,000 reviews of the services in their online delivery format by higher education officials representing over 400 academic institutions (both within the US and outside of the US) as of August 2015.
- For construct validity, the exam service was designed in consultation with accreditation officials.

RELIABILITY

Reliability is defined as the extent to which the exam results can be relied upon and that the results will be similar on repeated occurrences. Reliability processes that are employed to ensure a reliable service are extensive.

Item Analysis

Item analysis is a technique that evaluates the effectiveness of items (i.e. questions) in a test. The two principal measures used in item analysis are item difficulty and item discrimination.

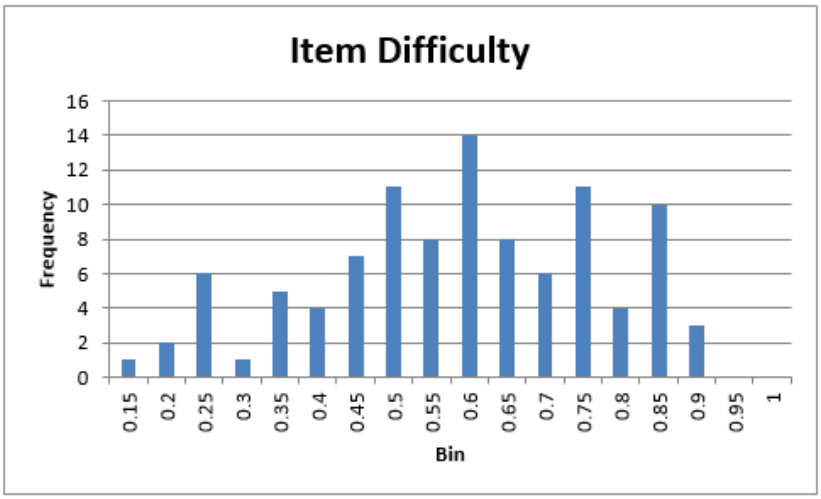
Item Difficulty

The difficulty of an item (i.e. a question) in a test is the percentage of the sample taking the test that answers that question correctly. This metric takes a value between 0 and 1 (or 0-100%). High values indicate that the question is easy, while low values indicate that the question is difficult.

> **Example 1**: 56 of the 100 students who were given a test containing question Q1 answered the question correctly. The item difficulty for this question is therefore 56/100 = 56%.

Targets and Evaluation Criteria: A target item difficulty of 60% has been set with an acceptable range of 40 – 80%. We periodically examine the item difficulty of all the questions in the test bank. Any question whose item difficulty is outside this range is eliminated or modified and retested.

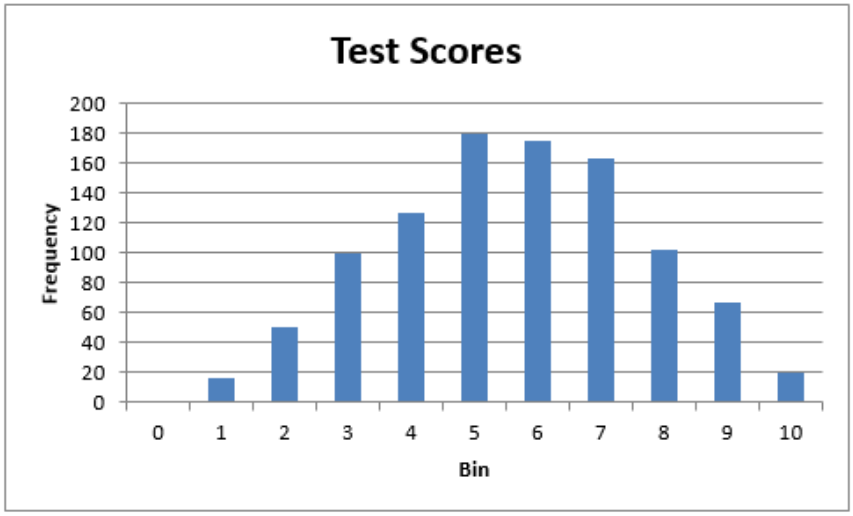
> **Example 2**: The questions in the test bank shown in Figure 1 have a range of item difficulty from 11.5% to 89.6%. The mean item difficulty is 5.62, which is just below our target. 15 questions have an item difficulty below 40% and 13 have an item difficulty above 80% .



Mean	0.562037
Standard Error	0.018564
Median	0.56701
Mode	0.5
Standard Deviation	0.186563
Sample Variance	0.034806
Kurtosis	-0.52525
Skewness	-0.32424
Range	0.780449
Maximum	0.895833
Minimum	0.115385
Sum	56.76575
Count	101
Geometric Mean	0.523631
Harmonic Mean	0.474279
AAD	0.150581
MAD	0.133048
IQR	0.261679

Figure 1 – Item Difficulty Distribution

Item difficulty is related to the difficulty of the test as a whole, which in this case is either the test score or the percentage correct, i.e. the number of correct answers divided by the number of questions (10 in our case). The distribution of test scores in a 1,000-student sample is as follows:



Mean	5.629
Standard Error	0.06468
Median	6
Mode	5
Standard Deviation	2.045371
Sample Variance	4.183543
Kurtosis	-0.56501
Skewness	-0.06333
Range	10
Maximum	10
Minimum	0
Sum	5629
Count	1000
Geometric Mean	#NUM!
Harmonic Mean	#NUM!
AAD	1.687034
MAD	1
IQR	3

Figure 2 – Distribution of Test Scores

Item Discrimination

Item discrimination is a measure of how well an item (i.e. a question) distinguishes between those with more skill (based on the subject that the test measures) from those with less skill. We track two measures of item discrimination: the index of discrimination and the point-biserial correlation.

Index of Discrimination

A principal measure of item discrimination is the index of discrimination (a.k.a., the discrimination index). This index is measured by selecting two groups: high skill and low skill based on the total test score. We assign the high skilled group to be those students whose total score is in the top 27% and the low skilled group to those students in the bottom 27%.

The discrimination index for a specific question is the percentage of students in the high skilled group who answer that question correctly minus the percentage of students in the low skilled group who answer the question correctly.

The discrimination index takes values between -1 and +1. Values close to +1 indicate that the question does a good job of discriminating between high performers and low performers. Values near zero indicate that the question does a poor job of discriminating between high performers and low performers. Values near -1 indicate that the question tends to be 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

Example 3: 100 students are asked to respond to question Q1. The scores on this question (1 for correct and 0 for incorrect) along with the total scores for the 100 students is shown in Figure 3.

Q1	Total	Q1	Total	Q1	Total	Q1	Total	Q1	Total
1	8	0	4	0	7	0	8	1	7
0	5	1	9	1	8	1	6	0	6
0	5	0	3	0	2	1	4	1	2
1	4	0	7	1	6	0	5	0	6
0	1	1	9	1	7	0	6	1	6
1	8	1	8	0	5	0	7	0	5
0	9	1	10	1	6	1	8	1	10
1	5	0	4	1	4	1	9	1	7
0	4	0	5	1	7	0	3	0	7
1	3	1	5	1	6	1	7	1	6
1	8	0	3	0	5	0	6	1	9
0	6	1	7	1	8	1	7	0	6
1	8	0	4	0	3	0	5	1	10
0	0	1	4	1	9	1	10	1	6
1	5	1	6	0	5	1	4	1	7
1	8	1	6	1	9	0	4	1	6
1	7	1	9	0	2	1	3	1	8
0	8	0	5	0	4	0	5	0	2
1	8	0	7	0	1	1	3	0	4
0	2	0	9	1	7	1	9	1	6

Figure 3 – Scores for Q1 vs. total scores

A histogram of the total scores for these 100 students is shown in Figure 4.

Frequency Table

item	freq	cum
0	1	1
1	2	3
2	5	8
3	7	15
4	12	27
5	14	41
6	17	58
7	15	73
8	13	86
9	10	96
10	4	100
	100	

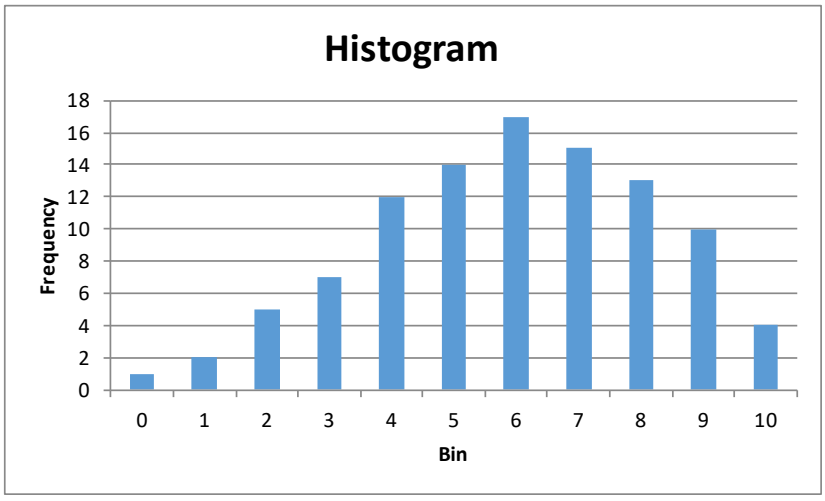


Figure 4 – Histogram of total scores

We now show how to calculate the index of discrimination. The best 27 total scores are 8, 9 and 10 (high skill group) and the worst 27 scores are 0, 1, 2, 3 and 4 (low skill group). Of the 27 students who got a score of 8 or higher, 23 students answered question Q1 correctly. Of the 27 students who got a score of 4 or lower, 9 students answered Q1 correctly.

Thus the index of discrimination is

$$\frac{23}{27} - \frac{9}{27} = \frac{14}{27} = .5185$$

Targets and Evaluation Criteria: We use the following guidelines for the index of discrimination.

- Less than 0%: Defective item
- 0 – 19.9%: Poor discrimination
- 20 – 29.9%: Acceptable discrimination
- 30 – 39.9%: Good discrimination
- 40% or more: Excellent discrimination

Defective questions and questions with poor discrimination are eliminated or modified and retested. We evaluate questions in the acceptable discrimination range to determine which should be eliminated or modified and retested.

> **Example 4**: A 10 question multiple choice test is given to 40 students. Each question has four choices (plus blank if the student didn't answer the question). We now interpret questions Q1 through Q6 based on the data in Figure 5 where the 20 students with the highest exam scores (High skill) are compared with the 20 students with the lowest exam scores (Low skill). The correct answer for each question is highlighted. We use D for the discrimination index and Df for the item difficulty.

	F	G	H	I	J	K	L	M	N	O	P
2		Q1				Q2				Q3	
3		High	Low			High	Low			High	Low
4	A	5	5		A	2	4		A	4	2
5	B	4	5		B	12	11		B	10	14
6	C	6	4		C	3	1		C	3	1
7	D	5	6		D	3	4		D	3	3
8	--	0	0		--	0	0		--	0	0
9		20	20			20	20			20	20
10											
11	Df	0.25			Df	0.575			Df	0.6	
12	D	0			D	0.05			D	-0.2	
13											
14		Q4				Q5				Q6	
15		High	Low			High	Low			High	Low
16	A	2	2		A	0	1		A	11	9
17	B	1	3		B	1	1		B	0	1
18	C	2	3		C	19	16		C	9	9
19	D	12	8		D	0	2		D	0	1
20	--	3	4		--	0	0		--	0	0
21		20	20			20	20			20	20
22											
23	Df	0.5			Df	0.875			Df	0.5	
24	D	0.2			D	0.15			D	0.4	

Figure 5 – Item Analysis for multiple choice test

For Q1, Df = .25 and D = 0. The four choices were selected by approximately the same number of students. This indicates that the answers were selected at random, probably because the students were guessing. Possible reasons for this are that the question was too difficult or poorly worded.

For Q2, D = 0.05, indicating there is no differentiation for this question between the students who did well on the whole test and those that did more poorly. The question may be valid, but not reliable, i.e. not consistent with the other questions on the test.

For Q3, D is negative, indicating that the high skilled students are doing worse on this question than the low skilled students. One cause for this may be that the question is ambiguous, but only the top students are getting tricked. It is also possible that the question, although perfectly valid, is testing something different from the rest of the test. In fact, if many of the questions on the test have a negative index of discrimination, this may indicate that you are actually testing more than one skill. In this case, you should segregate the questions by skill and calculate D for each skill.

Too many students did not even answer Q4. Possible causes are that the question was too difficult or the wording was too confusing. If the question occurs at the end of the test, it might be that these student ran out of time or got too tired to answer the question or simply didn't see the question.

Too many students got the correct answer to Q5. This likely means that the question was too easy (Df = .875).

For Q6, approximately half the student chose the incorrect response C and almost no one chose B or D. This indicates that choice C is too appealing and B and D are not appealing enough. In general, one of the incorrect choices shouldn't garner half the responses and no choice should get less than 5% of responses.

Point-biserial Correlation

Another measure of item discrimination is the point-biserial correlation coefficient (aka the item-total correlation) which is the Pearson's correlation coefficient between the scores on the entire test and the scores on the single item (where 1 = correct answer and 0 = incorrect answer). Two versions of this measurement are calculated: one where the total score is used and the other, the corrected point-biserial correlation coefficient (a.k.a., the corrected item-total correlation), where the total score without the item under consideration is used.

Example 5: Calculate the two versions of the point-serial correlation coefficient for question Q1 in Example 3.

In general, suppose that n students take a test containing the question under study, where n_0 answer the question incorrectly and n_1 answer the question correctly (thus $n = n_0 +$

n_1). Suppose also that m_0 is the mean of the total scores for the n_0 students who answered the question under study incorrectly, m_1 is the mean of the total scores for the n_1 students who answered the question correctly and s is the standard deviation of all the total scores. Then the point-biserial correlation coefficient can be calculated by the formula

$$r = \frac{m_1 - m_0}{s}\sqrt{\frac{n_0 n_1}{n(n-1)}}$$

For Example 5 the (uncorrected) point-biserial correlation coefficient is

$$r = \frac{m_1 - m_0}{s}\sqrt{\frac{n_0 n_1}{n(n-1)}} = \frac{6.82 - 4.77}{2.27}\sqrt{\frac{44(56)}{100(99)}} = .45$$

The corrected point-biserial correlation coefficient is

$$r = \frac{m_1 - m_0}{s}\sqrt{\frac{n_0 n_1}{n(n-1)}} = \frac{5.82 - 4.77}{2.10}\sqrt{\frac{44(56)}{100(99)}} = .25$$

Targets and Evaluation Criteria: In addition to the discrimination index evaluation criteria, we use the following guidelines for the uncorrected point-biserial correlation:

- Less than 0: Defective item
- 0 - .10: Poor discrimination
- .10 – .20: Fair discrimination
- .20 – .40: Good discrimination
- .40 or more: Excellent discrimination

We see that the uncorrected point-biserial correlation for Example 4 is excellent. We generally check that the corrected point-biserial correlation is at least .2 as well, which it is for Example 5.

Reliability Coefficients

Because the questions for each student are chosen at random from the questions in the test bank for that subject, the usual measures of reliability (split-half, KR20, and Cronbach's alpha) cannot be used, although we will have more to say about this later. Instead, we use question interchangeability as our principal measure of reliability.

Question Interchangeability

By question interchangeability, we mean the ability to substitute one question in the test bank with another without significantly affecting the total score that an individual would receive on the test. Our objective is to weed out any questions that fail the question interchangeability test.

For each question Q in the test bank, the specific question interchangeability test we use is to perform a two-tailed t-test 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. We set the significance level at 5%. Thus, any question that shows a significant difference based on this statistical test will be viewed as failing the question interchangeability test.

We know that some perfectly good questions will fail the test, but we want to be on the safe side. In fact, we are willing to reject 5% of the questions (Type I error) even though they meet the interchangeability criterion (i.e., the null hypothesis is true).

Although the data are not normal, it is not highly skewed either, and so the t test should be adequate. Even though the variances for each of the two samples (total scores for tests containing question Q vs. those that do not contain question Q) are generally quite similar, we use the t test with unequal variances just to be on the safe side.

Furthermore, we perform a Mann-Whitney test as well on each question and flag for further investigation questions that do not fail the t test version of the question interchangeability test but do fail the MW version of the test, once again to be on the safe side.

For our purposes, we also need to be concerned about Type II errors, i.e., cases where we do not reject a question even though it does not actually meet the interchangeability criterion. If we have, say 100 questions, in a test bank and each test consists of 10 questions, then on average each question occurs in 10% of the students' tests. If we test 2,500 students then we should be able to limit our Type II error to 5% (i.e. statistical power of 95%) and still be able to detect an effect of size .24 or more (with a sample of 4,000 we should be able to detect an effect size of .19).

Given that the pooled variance for each t-test is about 4.18, an effect size of .19 is equal to a difference of sample means of about .39 and an effect size of .24 is equal to a difference of sample means of about .49.

> **Example 6**: In general, we conduct the question interchangeability test on each of the approximately 100 different questions in a test bank based on the responses to tests consisting of 10 randomly selected questions taken by at least 1,000 students.
>
> In order to demonstrate how this is done, we conduct the question interchangeability test for a much simpler situation, consisting of the questions in a 10 question test bank based

on the responses to tests consisting of 3 randomly selected questions taken by 24 students. The data are summarized in Figure 6.

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Score
1		1					1	1			3
2		0	1					0			1
3	1								0	1	2
4				0	1				1		2
5			0	1			1				2
6						1	0		1		2
7					1	1	1				3
8			1	0				0			1
9	0				1					1	2
10		0		1			0				1
11	1						0			1	2
12			0	1		1					2
13				1		0		0			1
14		0		0		0					0
15	1		1		0						2
16	0		0	0							0
17	0				0			0			0
18		1					0		0		1
19			1		1					1	3
20				0		0		0			0
21		0						0		1	1
22		1			1		1				3
23							1	0		0	1
24	1					1			1		3
Correct	4	3	4	4	5	4	5	1	3	5	1.583333
Total	7	7	7	9	7	7	9	8	5	6	24
t-test	0.973463	0.666403	0.970823	0.024386	0.118838	0.975508	0.107669	0.048105	0.21982	0.422047	

Figure 6 – Question interchangeability test

For example, we see that 7 students answered question Q1, and so 24 – 7 = 17 students did not answer Q1. From Figure 7 we see the total scores for the students that answered question Q1 and the total scores for those that did not.

The mean total scores for these two groups are 1.571429 and 1.588235, which as we can see from Figure 6 are not significantly different (p-value = .973463).

w/ Q1	w/o Q1
2	3
2	1
2	2
2	2
0	2
0	3
3	1
	1
	2
	1
	0
	1
	3
	0
	1
	3
	1
1.571429	1.588235

Figure 7 – Total scores for student with and w/o Q1

We see from Figure 7 that question Q4 fails the question interchangeability test since there is a significant difference in the total scores for those who answered question Q4 from those that didn't (p-value = .024386 < .05 = α). Similarly Q8 fails the test (p-value = .048105). All the other questions pass the test.

Other Reliability Measures

Other exam reliability measures include:

- Detailed peer review and sensitivity analysis included as part of exam question development procedures.
- Exam scoring is 100 objective based upon automated item marking (questions are either correct or incorrect).
- Secure electronic item banking. The exam services meet the security requirements for Management of Information Technology (MIS) Sarbanes-Oxley (SOX) compliant organizations.
- Strict client confidentiality of client-specific data and reports is maintained within a SOX-compliant framework of security measures.
- Quality assurance procedures are in place at every stage of the process, from exam question development through delivery, scoring, and reporting.

- Reliability stability was confirmed during the initial beta-testing of the CPC-based COMP exam when selected student groups were administered the same exam twice with no statistically significant difference in scores ($p<.05$).
- Abandoned exams are excluded from summary reports. Only completed exams are used when reporting summarized results used for internal and external benchmarking.
- Institutional use of the exam services ranges from as few as 5 per institution to as high as 8,000+ per year per institution with multiple years of use and multiple cohorts of students from the same institution. A comparison of exam results with individual student GPA scores based on a sample of 200 exam scores from three different institutions showed a direct and positive correlation of GPA with the Exam Score.
- Reliability equivalence was established based upon the Alabama State University study (McNeal et al., 2012) that included students completing both the CPC-based COMP exam and the ETS MFT.

Summary of Validity and Reliability

From conception of the service, through development and beta-testing, and with ongoing quality assurance practices in place, the strategic goal of the programmatic assessment service is to provide colleges and universities with valid and reliable assessment instruments that can be incorporated into the program and appropriately used to measure learning outcomes in order to fulfill several accreditation and accountability requirements. The customizable exam service is comprehensive for the academic program as defined by the program's accreditation organizations.

Validity is maintained through regular and systematic psychometric analysis. Reliability is ensured through the security and maintenance of the online delivery platform, with automated reporting of scores and results, and with ongoing and regular psychometric processes.

References

Accreditation Council for Business Schools and Programs (2014). *ACBSP standards and criteria for demonstrating excellence in baccalaureate/graduate degree schools and programs*. Overland Park, KS: Author.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: Authors.

Association to Advance Collegiate Schools of Business. (2013). *Eligibility procedures and accreditation standards for business accreditation*. Tampa, FL: Author.

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

Cripps, J., Clark, C., & Oedekoven, O (2011). The undergraduate common professional component (CPC): Origins and process. *International Journal of Business & Management Tomorrow*, 1(1), 1-24.

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

Gould, R. (2002) *Bootstrap hypothesis test*, Stats 110A, UCLA http://www.stat.ucla.edu/~rgould/110as02/bshypothesis.pdf

Hanlon, B. and Larget, B. (2011) *Power and sample size determination*, University of Wisconsin, Madison (www.stat.wisc.edu/~st571-1/10-power-2.pdf)

Hesterberg, T., Monaghan, S. et al (2003) *Bootstrap methods and permutation tests*, Chapter 18 of The practice of business statistics, W.H. Freeman and Company.http://bcs.whfreeman.com/pbs/cat_160/PBS18.pdf

Howell, D. (2007) *Resampling statistics: randomization and the bootstrap.* http://www.uvm.edu/~dhowell/StatPages/Resampling/Resampling.html

Howell, D. C. (2010). *Statistical methods for psychology* (7th ed.). Wadsworth, Cengage Learning.

International Assembly for Collegiate Business Education. (2013). *Self study manual*. Lenexa, KS: Author.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, 2, 151-160.

León, R. (2004) *Nonparametric tests and bootstrapping*, Statistics 571: Statistical Methods, University of Tennessee. http://web.utk.edu/~leon/stat571/2004SummerPDFs/571Unit14.pdf

McNeal, R. C., Oedekoven, O. O., & Prater, T. (2012). *Evaluating comprehensive exams using common professional components.* Proceeding from the ACBSP Region 8 annual meeting. Ashville, NC: Accreditation Council for Business Schools and Programs.

Murray, F. (2009). An accreditation dilemma: The tension between program accountability and program improvement in programmatic accreditation. *New Directions for Higher Education*, *145*, 59-68.

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

University of Arizona, College of Medicine Phoenix (2015) *Assessment and Evaluations – Item Analysis.* http://phoenixmed.arizona.edu/students/assessment/assessment-and-evaluations-item-analysis

University of Illinois at Champagne-Urbana (2015) *Test Item Performance: The Item Analysis.* http://cte.illinois.edu/testing/exam/testperf.html

Zaiontz, C. (2015) *Real statistical analysis using Excel.* http://www.real-statistics.com