A Quality Outcome Assessment of Learning Discrete Mathematics Course: The Case of University Students

Marn-Ling Shing[1], Chen-Chi Shing[2], Lee-Pin Shing[3], Lee-Hur Shing[4]

[1]Early Child Education Department and Institute of Child Development Taipei Municipal University of Education 1 Ai-Kuo West Road, Taipei, Taiwan, R.O.C

[2]Information Technology Department, Radford University, Radford Box 6933, VA 24142

[3]Biology Department, Virginia Tech Blacksburg, VA 24061

[4]Business Information Technology Department, Virginia Tech Blacksburg, VA 24061

ABSTRACT

Teaching a mathematics foundation course such as Discrete Mathematics for an information technology curriculum is always a challenge. The challenge may be identifying students' mathematical backgrounds early and then using different teaching techniques in the classroom. An even bigger challenge is that many topics have to be covered effectively in a short semester course. This paper provides a standard quantitative methodology for conducting an outcome assessment using Discrete Mathematics as a case study. It starts with creating an ABET accredited course outcome based on different learning levels. And then it shows how to design assessment instruments, how to determine the sample size, how to collect data and how to analyze and validate the data.

General Terms Outcome assessment.

Keywords Assessment standards; quality assurance; outcome assessment; discrete mathematics assessment.

1. INTRODUCTION

Mathematics is used in everyday life including all fields of information technology. It forms the foundation of information technology and is the basis for software developers to create efficient tools to help people solve complicated problems. Thus, all undergraduate students in all seven concentrations other than information systems and web development in the Information Technology Department at Radford University are required to take the Discrete Mathematics course. The course has three credit hours and is offered by the department in one section every semester. The course is cross-listed with the Mathematics/Statistics department. Students who have finished the first course in a Principles of Programming class with a minimum grade of "C" and a Calculus or a pre-Calculus course are allowed to register for the course. More than 90% of the students in the class are Caucasian male. Their ages range from eighteen to twenty four years old. Their average SAT scores were around 1000 with average SAT MATH scores around 500. The course is mainly oriented towards computer applications; therefore, only few mathematics majors take the class. There are usually around 40 students in the class every semester.

In addition to some basic material such as database relations, data representations and Boolean algebra, the Discrete Mathematics course[11] covers topics such as logic, proofs, sets, functions, algorithm complexity, mathematical induction, counting, recurrence relation, finite state machine, graph theory, trees and matching.

The instructor prepares thirty-two lectures in fourteen weeks plus twenty-seven homework sets. Homework assignments are graded daily to assess the students' learning. Because the computer science concentration in the information technology department (other than the distance education program) has been ABET (Accreditation Board for Engineering and Technology) accredited since 1990, the instructor must use, in addition to daily assignments, appropriate instruments such as tests to assess the course outcome. These outcomes, required by ABET [14], are identified at the beginning of the course. This paper attempts to describe the assessment process for the Discrete Mathematics course. In contrast to existing papers that are either on assessment design [13], content-based assessment [7] or on service quality analysis [8], this paper addresses the complete process of a quality assurance outcome-based assessment for a course.

In the next few sections we will first identify the ABET course outcomes and then describe the design and implementation of an assessment plan using three tests and a final exam. Finally, we will analyze the teaching effectiveness using the statistical software package SAS [1].

2. ABET COURSE OUTCOMES

The definition of outcome –based assessment is given by Rigby (2006) in the following: “Outcome … reflects the performance(s) students are expected to demonstrate to indicate achievement of outcomes, e.g., identify, solve, list and select. Outcome-based assessment can go beyond providing feedback on student achievement… and can provide effectiveness of instruction.” [10]

We will proceed a course assessment based on the definition. ABET first requires accredited colleges to identify outcomes for their programs, e.g., the computer science program at Radford University. Then colleges need to design course outcomes so that they are in line with the program outcomes. To identify the learning levels of the students with respect to these course outcomes, usually Bloom’s taxonomy [3] is used. There are six different levels in the taxonomy. They are, from the lowest to the highest, knowledge, comprehension, application, analysis, synthesis, and evaluation. At the knowledge learning level students should be able to remember facts. At the comprehension level they are expected to understand the meaning. At the application level students should be able to apply facts to other situations. At the analysis level they can break down facts into pieces. At the synthesis level they can assemble those individual pieces back to a whole piece. Finally, at the evaluation level students are able to assess situations and make judgment based on certain criteria. Concerning the outcomes, the Discrete Mathematics course has been designed such that they can be assessed for each major topic covered in the course. There are seven outcomes identified and listed in the ABET course syllabus in Table 1.

Table 1. ABET Course Outcome

Number	Course Outcome
1	Demonstrate an ability to design mathematical argument
2	Demonstrate an ability to write mathematical proofs
3	Apply mathematical induction and design a recursive algorithm
4	Apply combinatorial analysis to solve counting problems
5	Analyze complexity of algorithms
6	Apply discrete structures to solve problems
7	Choose grammars and finite state machines to model computations

Outcome #1 requires students to be able to translate an English sentence into symbolic logic and perform predicate calculus, draw inference and then translate the results back into English sentences. Students can prove a mathematical theorem using either direct or indirect proof or proof by contradiction to meet Outcome #2. Outcome #3 requires students to prove a theorem by mathematical induction and write an algorithm using recursion. Outcome #4 requires students to use permutation and combination to do counting. Students need to analyze an algorithm using “big O” notation [11] to satisfy Outcome #5. Outcome #6 requires students to apply a structure such as a tree to solve a problem. Outcome #7 requires students to use finite state machines to solve problems. In terms of Bloom’s taxonomy, Outcomes #3, 4 and 6 fall into the application learning level, Outcome #5 falls into the analysis level, and Outcomes #1, 2, and 7 fall into the higher synthesis level. According to Bloom's taxonomy, these outcomes are classified in Table 2.

Table 2. Course outcomes learning levels

Outcome #	1	2	3	4	5	6	7
Knowledge							
Comprehension							
Application			x	x		x	
Analysis					x		
Synthesis	x	x					x
Evaluation							

The course topics are very broad and extensive. It is, therefore, very difficult to assess all the outcomes in a final exam alone. We decided instead to assess them in three tests and a final exam. In order to help students achieve these outcomes, the instructor needs to know the students' algebraic background at the beginning of the semester. For example, if the students do not know how to use algebra to prove a theorem abstractly, the instructor may have to give more step-by-step proof examples in order to prepare the students for writing mathematical proofs. On the other hand, some students may know some topics before they take the class. Thus, the exam results cannot really measure the students' outcomes. Therefore, to measure teaching effectiveness, we must use pre- and post-tests and compare the differences in student competencies. To both assess course outcomes and help with the instructor's teaching, measuring students' algebraic knowledge must be included in the pre-test. In addition, the pre-test must be able to measure the students' knowledge of the material to be taught in the course.

In the next section, we show how to design the outcome assessment instruments which can measure the required course outcomes.

3. THE DESIGN OF ASSESSMENT INSTRUMENTS

Any course outcome assessment must be practical, valid and reliable [10]. A practical assessment is able to collect information within limited time and cost. Due to limited resources, we decided to use multiple choice questions. A test bank is suggested to store all possible questions and the instructor can randomly select appropriate questions. A valid assessment refers to an assessment able to measure what is intended to be measured. To account for the students' pre-knowledge about the course content, pre- and post-test assessment instruments are used to collect the necessary data for analysis [5]. Another advantage of using pre- and post-tests is that the pre-knowledge can be estimated by students' performances. And a reliable assessment must have consistent measurements. For an entire course assessment, it may be preferable that the percentage of each assessed course outcome depend on the percentage of the instruction devoted to the topics that are related to the outcome [3]. In this paper, however, we will not use different weights for the separate course outcomes.

The pre-test instrument questions are given in the Appendix. The pre-test usually also collects some background information needed for drawing inference about the types of students in the study (Pre-test Question #26-28). Because many course outcomes need to be measured and because there are many students taking the course, we use multiple choice questions in all the tests and the final exam. In addition, for measuring whether students can do mathematical proofs, the multiple choice questions must be able to test the steps of a mathematical proof.

3.1 Pre-Test and Post-Test

A pre-test is given to the students at the first day of class as a closed book exam in order to measure the students' pre-knowledge of the course topics. It is graded as an extra credit quiz to increase the students' incentive of taking the pre-test without putting too much weight on it.

3.1.1 Measuring Algebraic background

In order to achieve the validity of the assessment we create five instrument questions from five areas in algebra. Those five areas are: solving a two-variable linear equation (Question #1), solving a word problem (Question #2), simplifying a rational expression (Question #3), manipulating exponents (Question #4), and finding the sum of an arithmetic series (Question #5).

3.1.2 Measuring Course Outcomes

We will only measure the first six course outcomes in this paper. To measure each outcome, we have two ways to design the instruments. We can either use the same questions in the post-test as in the pre-test or use a different question in the post-test but from a set of similar questions. For the latter, we can create a question bank from which we can randomly choose the necessary number of questions when needed. A sample question bank for Question #6 is given in Table 3.

Table 3. A Sample Question Bank corresponding to Pre-test Question #6

Q1	Suppose *h* and *c* are these propositions: *h*: *I go swimming* *c*: *it is a cold day.* Express in symbols the compound proposition *I don't go swimming when it is a cold day.* $h \rightarrow c$, B. $c \rightarrow \neg h$, C. $\neg c \rightarrow h$, D. $\neg h \rightarrow c$
Q2	Which implication is logically equivalent to the implication? $\neg \boldsymbol{r} \rightarrow \boldsymbol{s}$? A. $\neg \boldsymbol{s} \rightarrow \boldsymbol{r}$, B. $\neg \boldsymbol{s} \rightarrow \neg \boldsymbol{r}$, C. $\boldsymbol{r} \rightarrow \neg \boldsymbol{s}$, D. $\boldsymbol{s} \rightarrow \neg \boldsymbol{r}$
Q3	The implication . $\boldsymbol{q} \rightarrow \neg \boldsymbol{p}$ is true for all possible assignments of truth values to ***p*** and ***q*** except for which assignment? A. ***p*** true, ***q*** true, B. ***p*** true, ***q*** false, C. ***p*** false, ***q*** true, D. ***p*** false, ***q*** false

To achieve content validity the pre-test questions in the assessment instrument must be matched with course outcomes. These relationships are shown in Table 4.

Table 4. Pre-Test Questions and Measured Course Outcomes

Pre-Test Question #	Topic	Outcome #
6	Logic (conditional)	1
7	Logic (conditional)	1
8	Logic (conditional)	1
9	quantifier	1
10	quantifier	1
11	quantifier	1
12	proof	2
13	proof	2
14	function	6
15	function	6
16	function	6
17	big oh	5
18	big oh	5
19	big oh	5
20	induction	3
21	induction	3
22	induction	3
23	counting	4
24	counting	4
25	counting	4

We see from Table 4 that there are six questions to assess Outcome #1. On the other hand, there are only two questions to assess Outcome #2. The number of questions used will affect the variances of the outcome assessments and hence the reliability of the assessment. By measuring the relative amount of variance contribution by each question to the total variance, the correlation of all questions can be estimated. Cronbach alpha is one of such measure.

3.1.2.1 Cronbach's Alpha

Cronbach's alpha coefficient is commonly used as an index to measure internal consistency of a psychometric test score [2]. It describes how a group of questions assessing the same outcome are correlated with each other. However, it does not measure the homogeneity of the assessment [9]. Suppose that there are n items measuring an outcome, the standardized Cronbach alpha (or Spearman-Brown correction formula), which normalizes the variance of each item to be one, is defined as

$$\alpha = n*\bar{r} /(1+(n-1)*\bar{r}), \qquad \text{(EQ 3.1)}$$

where $\bar{r}$ is the average of the n(n-1)/2 entries of the upper or lower triangular Pearson correlation matrix. Alpha can take on values from -1 to 1. However, any negative alpha value is not meaningful [10]. According to Schmitt [12], the value increases as a function of the number of questions, n, for a fixed $\bar{r}$ value. If the questions are independent of each other, then all the entries of the upper triangular Pearson correlation matrix are equal to zero. Therefore, alpha = 0. On the other hand, if each question measures the same outcome, i.e. they are completely correlated, then all the entries of the upper triangular Pearson correlation are equal to one. In this case, alpha = 1. This means the more correlated those questions are, the higher is the internal consistency. Thus the more reliable is the assessment. In practice, the alpha value should be at least 0.7 [12]. We believe the alpha value is more meaningful for post-test since in the pre-test students more likely have guessed the answers. Based on the data in Table 12 in Section 4, Table 5 lists the Pearson correlation matrix for the six items measuring Outcome 1 of Table 4.

Table 5. Pearson's correlation matrix among Q6 – Q11 for Table 4

Q #	**6**	**7**	**8**	**9**	**10**	**11**
6	1.0	0.88	0.62	0.74	0.10	0.80
7	0.88	1.0	0.88	0.94	0.33	0.87
8	0.62	0.88	1.0	0.81	0.71	0.90
9	0.74	0.94	0.81	1.0	0.37	0.68
10	0.10	0.33	0.71	0.37	1.0	0.42
11	0.80	0.87	0.90	0.68	0.42	1.0

We see from Table 5 that Question 10 has a low correlation with all other questions except with Question 8 and $\bar{r}$ =1/15(0.88451+0.62292+0.74800+0.10363+0.80700+0.88065+0.94777+0.33151+0.87115+0.81429+0.71335+0.90373+0.37540+0.68763+0.42012)= 0.6741. Thus, the Cronbach *alpha* = 6*0.6741/(1+5*0.6741)=0.9254. The values for n, $\bar{r}$, and alpha for Outcomes 1 to 6 are given in Table 6.

Table 6. Average correlation and Cronbach's alpha for measuring Outcome 1-6 according to Table 4

Outcome #	**n**	$\bar{r}$	**Cronbach's alpha**
1	6	0.6741	0.9254
2	2	0.46	0.6301
3	3	0.7732	0.9109
4	3	0.5249	0.7682
5	3	0.4183	0.6833
6	3	0.3016	0.5644

For Outcome #2 in Table 6 the alpha value is lower than 0.7. This can create a reliability problem of the test in measuring the validity of Outcome #2. The reason for the low value may be a low correlation between questions 12 and 13 (see Table 4) or the fact that we used only two questions. In order to get some idea about the relationship between alpha and n, and to obtain, for a given value of r bar, an expression for the minimum value of n, i.e. the number of questions used for an outcome, we obtain form (EQ 3.1)

$$\alpha /(1-\alpha) = n*\bar{r} /(1+(n-1)*\bar{r} - n*\bar{r}),$$

or

$$n = \alpha (1-\bar{r})/ (\bar{r} (1-\alpha)) \qquad \text{(EQ 3.2)}$$

Thus, in order to attain the alpha value 0.7, given that r bar=.46 (see Table 6) we get from (EQ 3.2), n=0.7*(1-0.46)/(0.46(1-0.7))=2.74. Therefore, the number of questions must be at least 3 when the average correlation among questions is at least 0.46. For Outcome #5 we see from Table 6 that the alpha value is also lower than 0.7 but very close to 0.7. The Pearson correlation matrix for the questions measuring Outcome #5 is given in Table 7.

Table 7. Pearson's correlation matrix for measuring Outcome #5

Q #	**17**	**18**	**19**
17	1.0	0.24217	0.03468
18	0.24217	1.0	0.97805
19	0.03468	0.97805	1.0

Table 7 shows that Question 18 and 19 are highly correlated, but both are not highly correlated with Question 17. It seems to indicate that there are two factors involved. One of the factor involves with Question 18 and 19 only. Cronbach (1951) stated that alpha is an underestimate of reliability unless the correlation matrix is unidimensional (or for a single factor) [2]. Schmitt (1996) [12] suggested to use the upper limit of validity for the correction, which is equal to the square root of alpha or $\sqrt{0.6833}$ =0.83. After the correction, the internal consistency is larger than 0.7.

The last row in Table 6 shows that the alpha value for Outcome #6 is 0.5644. The corresponding Pearson correlation matrix can be found in Table 8.

Table 8. Pearson's correlation matrix for measuring Outcome #6

Q #	14	15	16
14	1.0	0.04458	0.76377
15	0.04458	1.0	0.09640
16	0.76377	0.09640	1.0

Table 8 shows that Questions 14 and 16 are correlated, but both are not correlated with Question 15. It seems to indicate again that there are two factors involved with the corrected alpha $\sqrt{0.5644}$ =0.75. After the correction, the internal consistency is larger than 0.7. However, if the questions involved are not really one-dimensional, then the corrected alpha can be an over-estimate [10]. Another way to handle the multi-dimensionality among correlations of those questions is to delete some unrelated questions using an additional SAS output of PROC CORR with Cronbach's alpha as given in Table 9.

Table 9. Cronbach Coefficient Alpha with Deleted Variable among Q14 – Q16 for Table 8

Q #	**Correlation**	**Cronbach's alpha**
14	0.545885	0.175848
15	0.075061	0.866068
16	0.595115	0.085352

From Table 8 we see that Question 15 has very low correlation with Questions 14 and 16. In addition, from Table 9 we see that deleting Question 15 would slightly increase the alpha value. This indicates Question 15 did not contribute much to the measuring of the internal consistency. Therefore, in order to increase the alpha value, Question 15 can be deleted or re-designed in a future study. After deleting Question 15 from the study, the alpha value increases to over 0.7 as shown in Table 10.

Table 10. Average correlation and Cronbach's alpha for measuring Outcome 6 after deleting Question 15

Outcome #	**n**	$\bar{r}$	**Cronbach's alpha**
6	2	0.7638	0.8661

3.1.2.2 Difficulty and Discrimination Indices

There are two more factors that can affect the reliability of an assessment. They are the difficulty index and the discrimination index of each question. The difficulty index measures how difficult a question can be by comparing the performance of "good students" with that of "bad students". And the discrimination index measures whether "bad students" are guessing and getting higher scores than "good students". In order to calculate both indices, we need to first use the same number of students in the upper group (or upper 27% of the test scores) and the lower group (or bottom 27% of the test scores) in an assumed normal student population [6]. Both indices for a specific question are calculated as shown below:

Difficulty index = (number of correct answers by the upper group + number of correct answer by the lower group)/ (total number of students in both upper and lower groups *4* 0.27)

Discrimination index=(number of correct answers by the upper group – number of correct answer by the lower group) / (number of students in each group * 4 * 0.27)

For example, suppose 100 students took the post-test. We first rank them in terms of total score. If 20 students out of 27 students in the upper 27% group answered question 1 correctly and 15 students in the lower 27% group answered the question correctly then the difficulty index for Question #1 = (20+15) / (54*4*0.27) = 0.60. And the discrimination index for Question #1 = (20-15) / (27 * 4 * 0.27) = 0.17.

The value of the difficulty is between 0 and 1. In practice, if it is below 0.2, the question is considered to be too difficult to affect the consistency (i.e. reliability) of the assessment [9]. Some authors even suggest this value to be 0.4 [11].The values of discrimination index ranges from -1 to 1. If all lower group students are guessing a question right and all upper group students are guessing it wrong, then the discrimination index for the question = -1. This would not be the purpose of an assessment. In practice, any question with a negative discrimination index must be discarded. When the discrimination indices for each question are 0, this indicates the assessment is either too easy (difficulty index is close to 1) or too difficult (difficulty index is close to 0). This would also decrease the reliability of the assessment. Both indices will be calculated only for the post-test.

Examples of selecting "good" questions to measure course outcomes based on difficulty and discrimination indices can be found in Rigby and Dark [10].

In the post-test, the question numbers used to assess Outcome #1 are different from those in the pre-test and they are given in Table 11.

Table 11. Post-Test Questions and Measured Course Outcomes

Q #	Match Pre-test Q #	Topic	Outcome #
1,2,3	6	Logic (conditional)	1
4,5,7	7	Logic (conditional)	1
14	8	Logic (conditional)	1
8	9	quantifier	1
9,10	10	quantifier	1
11,12	11	quantifier	1

4. ANALYSIS

In this section we will first examine the graph of student performances on our instrument questions and then test whether the students attained the first five course outcomes.

4.1 Descriptive Statistics

The data collected for the pre-test results in this section are for students in Fall 2006. The summary is shown in Table 12.

Table 12. Pre-Test Data

Pre-Test Question #	#students correct	%students correct	Topic	Category
1	8	0.3636363	algebra	2 variable linear equation
2	12	0.5454545	algebra	word problem
3	9	0.4090909	algebra	rational expression
4	12	0.5454545	algebra	exponents
5	9	0.4090909	algebra	arithmetic series
6	7	0.3181818	logic(conditional)	
7	8	0.3636363	logic(conditional)	
8	9	0.4090909	logic(conditional)	
9	8	0.3636363	quantifier	

10	4	0.1818181	quantifier	
11	5	0.2272727	quantifier	
12	2	0.0909090	proof	
13	3	0.1363636	proof	
14	3	0.1363636	function	
15	5	0.2272727	function	
16	11	0.5	function	
17	1	0.0454545	big oh	
18	5	0.2272727	big oh	
19	6	0.2727272	big oh	
20	4	0.1818181	induction	
21	0	0	induction	
22	4	0.1818181	induction	
23	6	0.2727272	counting	
24	8	0.3636363	counting	
25	3	0.1363636	counting	

The algebraic background in Fall 2006 is given in Figure 1.

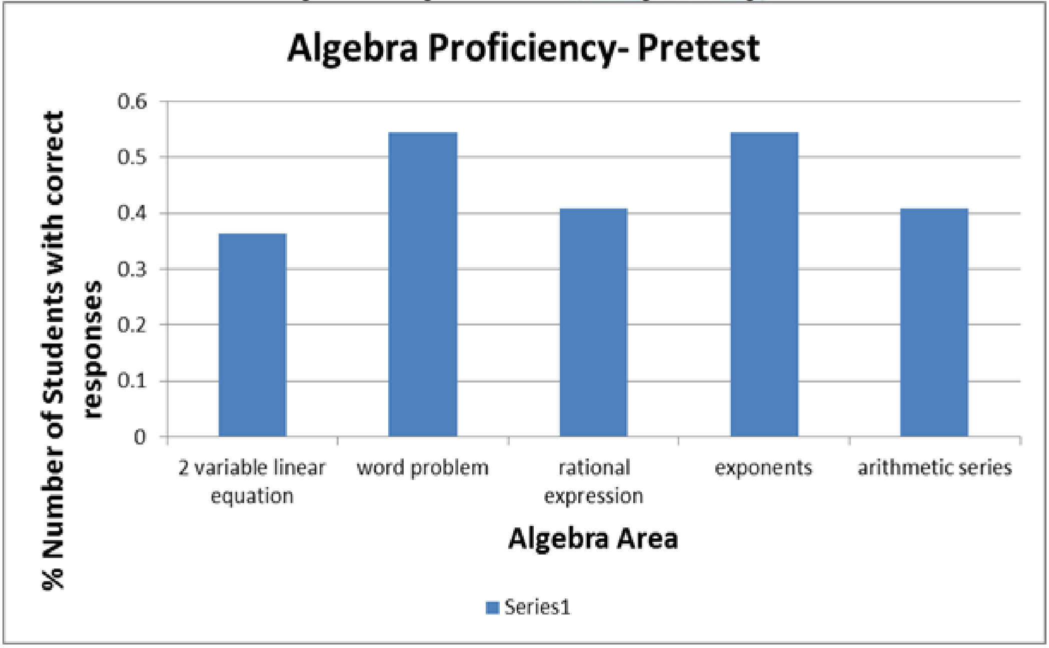


Fig 1: Students Algebraic Background

Figure 1 shows that the percentages of correct responses are between 35 to 55%. It seems that the students are not ready for mathematical proof using algebra. The students' pre-knowledge are summarized in Figure 2.

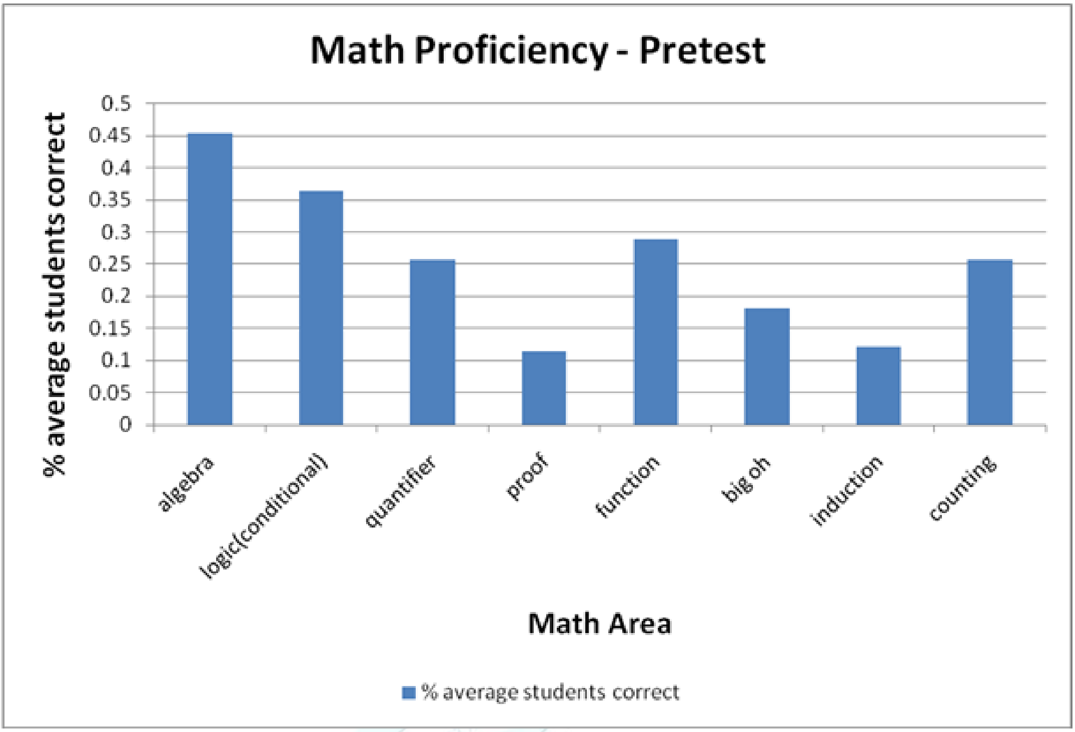
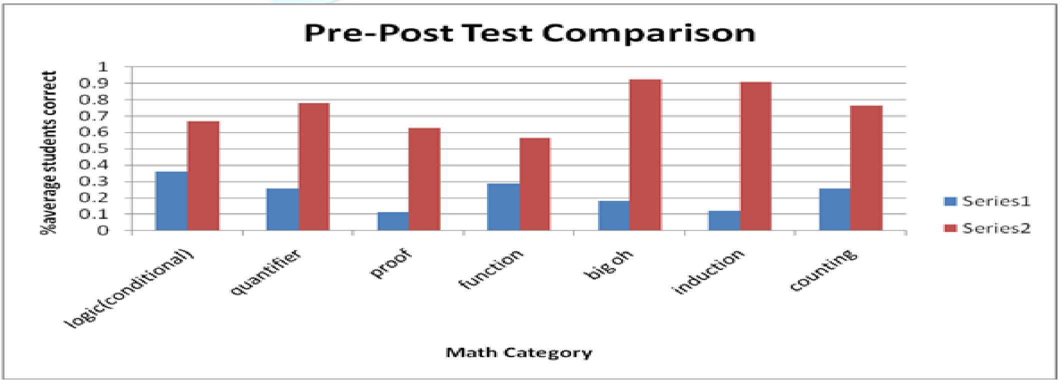


Fig 2: Students Pre-Knowledge Summary

Figure 2 shows that students are more prepared for algebra than for any of the topics for the course at the beginning of the semester. In Figure 3 we compare the post-test results with the pre-test results where series 1 represents pre-test and series 2 represents post-test



.

Fig 3: Students Pre-Test and Post-Test Comparison

From Figure 3 it can be seen that the percentage of the students' correct responses in each Mathematics category in the post-test is clearly better than that in the pre-test. The means and standard deviations for each Mathematics area in the pre- and post-tests are shown in Table 13.

Table 23. Pre-Test and Post-Test Means and Standard Deviation in Each Math Category

	Mean/Std	Big Oh	Conditional	Counting	Function	Induction	Proof	Quantifier
PreTest	**mean**	0.1818182	0.3636364	0.2575758	0.2878788	0.1212121	0.1136364	0.2575758
	std	0.1202614	0.0454545	0.1143914	0.1892424	0.1049728	0.0321412	0.0946212
PostTest	**mean**	0.9242167	0.6666667	0.7619000	0.5681833	0.9047667	0.6287833	0.7803000
	std	0.0262261	0.0660526	0.0476000	0.1419616	0.0476500	0.0749769	0.0656014

Although it appears from Table 13 that we have attained our goal that the students have achieved the expected results for the first six outcomes, we must use statistics to test the student outcomes (see Section 4.2) The difficulty and discrimination indices for the post-tests to assess outcome #1 were given in Table 14.

Table 34. Difficulty and Discrimination Indices for Post-test

Post-test Question #	Discrimination Index	Difficulty Index
1	0.1266	0.4470
2	0.1916	0.4310
3	0.1916	0.4310
4	0.3932	0.3832
5	0.1266	0.4470
6	0.3932	0.3192
7	0.0640	0.4630
8	0.2556	0.4150
9	0	0.4795
10	0.3932	0.3192
11	0.1916	0.4310
12	0.5108	0.3512
13	0.1916	0.4310
14	0.1916	0.2714

Since the discrimination indices are close to 0 in Questions #7 and #9 (see Table 14), these two questions may not be good questions to assess Outcome #1. They should be discarded in any future assessment.

4.2 Test Statistics

In this section we show how to use the statistical package SAS 9.1 to draw correct inferences alluded to in the previous section. We want to test the null hypothesis: No difference between pre- and post-test results in all categories versus the alternative hypothesis: Post-test results in all categories are higher than pre-test results The data collected are from Fall 2006 to Spring 2010. The dependent variable is the average percentage of the students' correct responses in each Mathematics category. We assume that the observations are independent within each Mathematics category. There are three classification categories involved. The first one is type of test (pre-test and post-test). The second one accounts for the different semesters. The third contains all Mathematics categories. The outcome effectiveness can be measured by the mean difference between the post-test results and the pre-test results. We use the SAS procedure GLM (General Linear Model) to perform a three-way Analysis of Variance (ANOVA) [5]. Before using the GLM procedure, we must make sure that the data are normally distributed . The data for pre-test results throughout the years are shown in Figure 4, representing all Mathematics categories. Specifically, in Figure 4, series 1 represents 2 variable linear equation, series 2 word problems, series 3 rational expression, series 4 exponents, series 5 arithmetic series, series 6 logic (conditional), series7 quantifier, series 8 proof, series 9 function, series 10 big oh, series 11 induction, and series 12 counting.

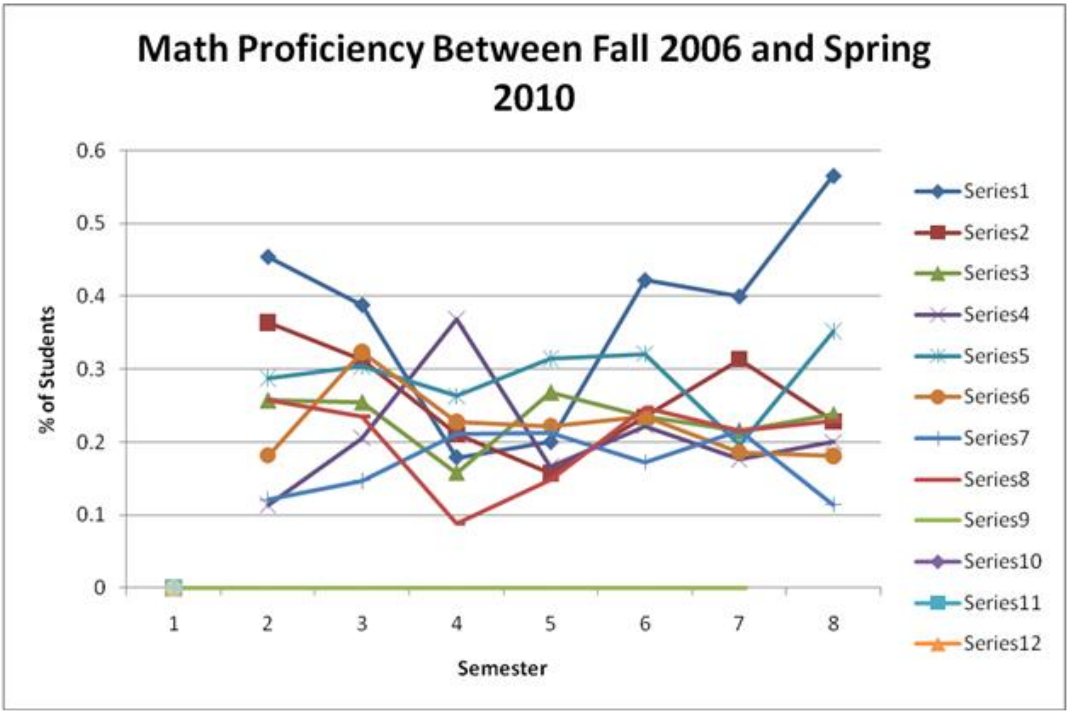


Fig 4: Pre-Test and Post-Test Means in Each Math Category

Figure 4 shows that the students' pre-test performance in each mathematical category throughout those years seems to fall mostly between 15% and 35%. For instructional purpose this will give the instructor some information about students' Math background before the class.

4.2.1 Test for Normality

We first draw a Q-Q plot (see Figure 5)and a histogram (see Figure 6) for the pre- and post-tests results based on the percentage of average number of students with correct answers in each Mathematics category to show that the observations are from a normal distribution.

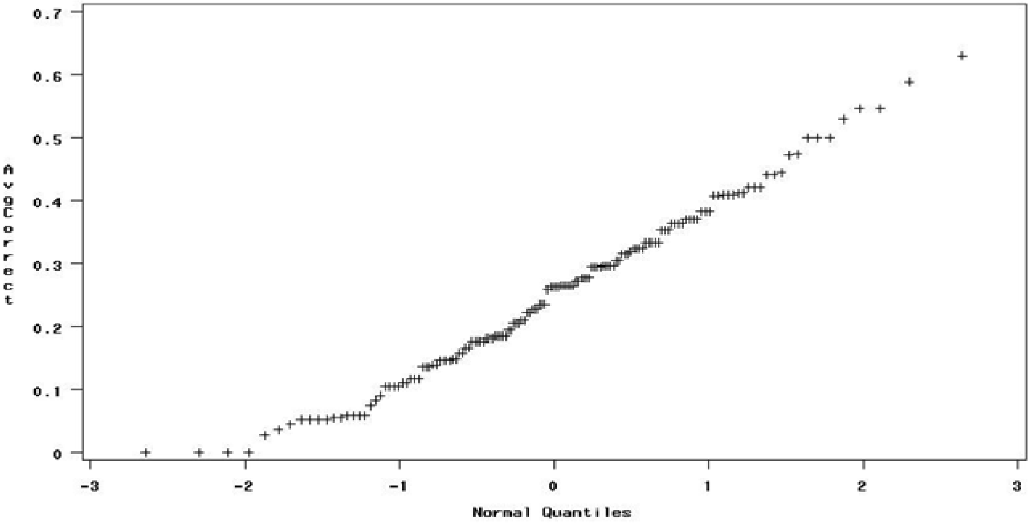


Fig 5: Q-Q Plot of All Observations

Figure 5 shows that the Q-Q plot is close to a line, which means that the observations are close to samples from a normal distribution.

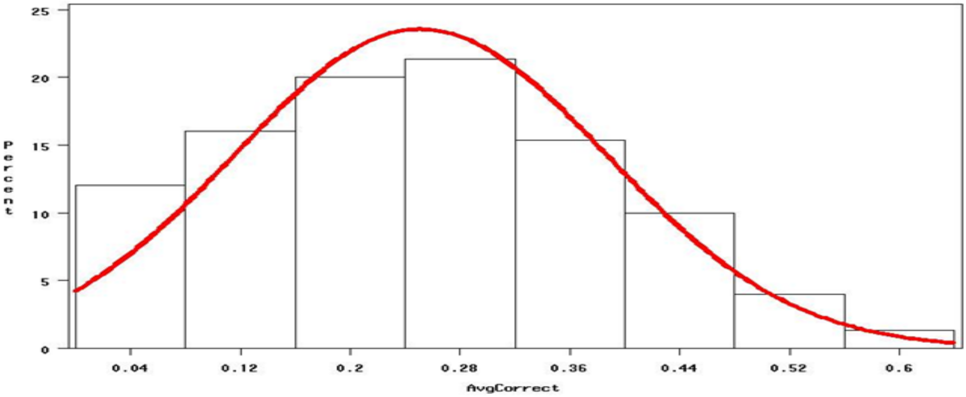


Figure 6 also indicates the possibility of the sample distribution being a normal distribution. In the next section we proceed to use an ANOVA test as given in the next section.

4.2.2 ANOVA Test

The SAS output of three way ANOVA test with three classifications (Test Type, different semesters, and Mathematics categories) and their interactions is displayed in Table 15.

Table 45. 3-way ANOVA Test

Source	Df	F Value	Pr > F
Semester	6	13.18	<.0001
Type	1	734.97	<.0001
Semester*Type	6	10.89	<.0001
category	6	1.60	0.1484
Semester*category	36	1.05	0.4032
Type*category	6	3.90	0.0011
Error	36	1.41	0.0759

In Table 15 the model accounts for 84.6% (R-Square) of the total sum of squares If Question #15 would be dropped, the R-square could increase to 85.5%. Type I sum of squares, which adds each factor into the model sequentially, are used over in the analysis because there are no missing data and the data are balanced. The p-values are less than 0.001 for Semester, Type and the interaction of Semester and Type. This means that different semesters produce different results, and they are affected by different type of tests. The types of tests are affected also by different Mathematics categories. In addition, using Figure 3 we conclude that the post-test result is higher than the pre-test result. However, there are no difference among the Mathematics categories. In order to find out which Semester means are significantly different Duncan multiple comparisons along with related means are given Table 16.

Table 56. Duncan's Multiple comparison

Duncan Grouping	Mean	N	Semester
A	0.53023	40	Spring07
A	0.51234	40	Spring06
A	0.49280	40	Fall06
B	0.41397	40	Spring09
B	0.40933	40	Fall09
C B	0.37192	40	Spring08
C	0.33443	40	Fall07

Table 16 shows that there are 40 observations are counted for evaluating the means for each semester. And there are three non-significant groups: group1 includes Spring07, Spring06, Fall06 and group2 includes Spring09, Fall09 and group3 includes Spring08, Fall07. Between those three groups creates the significant effect. Similarly, there are 140 observations used in calculating the means for pre-test and post-test. The average score for the post-test is 64.88% and that for the pre-test is 22.69%. In the next section we find the relationship, contribution and implication to the field of the services and standards.

5. IMPLICATIONS

This paper provides a detailed standard quality assurance quantitative methodology for practitioners for conducting an outcome assessment, using a Discrete Mathematics as a case study. It starts with creating an ABET accredited course outcome based on different learning levels. And then it shows how to conduct a pilot study to decide on the the sample size for reliability using (EQ 3.2).. It explains how to use pre- and post-test design assessment instruments, how to collect data and how to analyze and validate the data. The methodology can be applied not only in the regular classroom setting, but also in the distance education setting if the pre-test data can be collected in the same way as that in the post-test setting. The process can also identify the students' mathematical background at the beginning of the semester and can help teachers with the teaching. For this reason, this paper will provide a critical tool for services and standards experts to conduct an outcome assessment for a Discrete Mathematics course. The data used in the paper were collected directly from computer scanned forms. And they were graded and tabled along with difficulty and discrimination indices by the authors' own software. The grades tables were fed into SAS programs to check reliability and validity and to obtain ANOVA tables.

6. CONCLUSIONS, LIMITATIONS AND DIRECTIONS FOR FUTURE RESEARCH

In this paper we use criterion-based assessment theory of assessing students' understanding of material after instruction to find both on how well the students understand the outcomes and identify which students need remediation [7]. From the data analysis in Section 4, we find that at the beginning of each semester, students' don't have enough background in algebra. The students' performance on proof area is the worst in the pre-test results. The instructor has to spend more time on teaching students to write mathematical proof. At the end of the semester we find that the students have attained the first six course outcomes. In addition, we can check the students' background in algebra throughout the years. Using a one-way ANOVA test (Table 17), we find that the algebraic backgrounds are significantly different (p=0.0027<0.05) throughout those years.

Table 67. 1-way ANOVA Background Test

Source	Df	F Value	Pr > F
Semester	5	5.03	0.0027

We can also check whether any of the five fields in the students' algebraic background used in the assessment are different. A one-way ANOVA test (Table 18) shows that there are no differences

Table 78. 1-way ANOVA Algebra Field Test

Source	Df	F Value	Pr > F
Question	4	0.69	0.6079

The outcome assessment done is this paper is reliable and valid to assess five outcomes, although it would be better to have at least three questions used to assess Outcome #2. Even though we have assessed mainly the first five course outcomes using three midterm tests and the final exam, we can use as an auxiliary instruments the students' portfolio that include graded homework assignments. For example, in order to see whether students can write a mathematical proof of a theorem, we may need to see their actual proof writing in their turned-in assignments that are kept in the portfolio. To assess all seven course outcomes, we only need to add more questions to the assessment instrument. Those questions must be randomly selected beforehand and be tested for their effectiveness to assess those intended outcomes. Although the data analysis shows a significant result between pre-test and post-test statistically, the data should be collected so that we have each student's score in each mathematics category for both pre-test and post-test in each semester in order to test the student subject effect within each semester using semester and test type as classification in a two-way ANOVA. This can be done in a future study.

The assessment process is an on-going process. Some bad questions such as question 15 in the post-test should be eliminated. This improves both the reliability and the validity of the assessment, The difficulty and discrimination indices are based on normal distribution population assumption. For small sample, they may not be useful. We believe that practitioners can use the procedures proposed in this paper to conduct a quality outcome assessment for any course. However, the entire assessment should be handled by a dedicated office such as Institutional Research Office in the college because the instrument design, data collection and analysis are long and costly processes. The entire process involves the establishment of a question database so that enough questions are available to be picked randomly. The same questions must appear in both pre- and post-tests. The post-tests can be conducted in the form of quizzes and tests. And they must be carefully mapped so data can be collected and analyzed to avoid missing data. Missing data will complicate the inference from the data analysis. Besides, the samples collected for validity may be taken in ten semesters if only one class is offered each semester for a 30-40 students' classes. The methods of this paper have not been used in the distance education setting because those sections are not taught by the authors and the validity analysis is not robust for a small number of students.. For future studies, the authors will try to map the methodology to other types of standards such as IEEE, Six Sigma, 5-S, SACS Assessment Standards and NCATE Standard.

7. ACKNOWLEDGMENTS

This paper is in memory of my mother Sharn-Yun Chen for her life time encouragement, support and insatiable hunger for knowledge. Without her this paper would have never been completed.

8. REFERENCES

[1] Cody, R., Smith, J. 1997. Applied Statistics and the SAS programming Language, Prentice Hall Inc, New Jersey.

[2] Cronbach, L.J. 1951. Using Coefficient alpha and the internal structure of tests. Psychometrika, 16, (3), 297-334.

[3] Dark, M.J. 2004. Evaluation. In Education and Technology: An Encyclopedia (Kovalchik and Dawson, Eds). ABC CLIO: Santa Barbara, CA.

[4] Gamst, G., Meyers, L., Guarino, A. 2008. Analysis of Variance Designs, Cambridge University Press, New York.

[5] Hinkelmann, K., Kempthorne, O. 1994. Design and Analysis of Experiments, Vol. 1, John Wiley & Sons Inc , New York.

[6] Kelley, T.L. 1939. The selection of upper and lower groups for the validation of test items. Journal of Educational Psychology, 30, (1), 17-24.

[7] Maples, G., Heady, R , Zhu, Z. 2008 “Exemplars and the need for content-based publication standards”. International Journal of Services and Standards, 4(3), 269-283.

[8] Miller, R., Hardgrave, B, Jones, T.. 2008 “Quality assessment and accreditation of engineering programmes”. International Journal of Services and Standards, 4(1), 1-15.

[9] Revelle, W. and Zinbarg, R. 2009. Coefficients Alpha, Beta, Omega, and the GLB: Comments on Sutsma. Psychometrika, 74, (1), 145-154.

[10] Rigby, S., Dark, M.J. 2006. Using Outcomes-Based Assessment Data to Improve Assessment and Instruction: A Case Study. ACM SIGITE Newsletter, 1, (3), 10-15.