Testing the Test: Validity and Reliability of Senior Exit Exam

Lizabeth Schlemer and Daniel Waldorf
California Polytechnic State University, San Luis Obispo

Abstract

A senior exit exam is considered an excellent direct measure of student learning for ABET assessment, but the usefulness of the information gathered is related to the validity and reliability of the test itself. The Industrial and Manufacturing Engineering Department at California Polytechnic State University, San Luis Obispo has used a content exam for several years. This paper will discuss test development, administration, and the role it plays in the assessment process. In addition, the test is evaluated using the standard psychometric techniques of reliability and validity. The results of the evaluation are used to refine the test. The importance of the evaluation of these types of instruments cannot be overstated as they often are used to guide curricular or other program improvements efforts.

INTRODUCTION

The Accrediting Board for Engineering and Technology (ABET) [1] encourages programs to use direct measures of performance when evaluating the achievement of learning outcomes prior to student graduation. Direct measures are those that assess achievement by observation of performance rather than by soliciting opinion about the achievement of a particular outcome. A standardized exam is a good direct measure. Others might include a third party evaluation of student projects or a manager's assessment of work done on co-op/internship. A standardized exam may be the most tempting for busy faculty trying to assess their program because it is fairly easy to administer, the results are naturally quantifiable, and the program can more or less guarantee a consistent rate of response. Such an exam, however, should be evaluated using a psychometric evaluation to study reliability, validity, and item correlation before the results are used to invest significant time and effort into improving a program.

Psychometric Evaluation

The aim of a psychometric evaluation of a test is to determine how well the instrument, in this case the test, is measuring the construct of interest, in this case the individual's ability. In every measurement whether it is a physical measurement like a micrometer or a psychological measurement like a survey, the resultant value has some amount of error. In order to evaluate the quality of the measurement device, psychometricians use two general characteristics: Reliability and Validity. Reliability describes the consistency of a measurement, while validity addresses the appropriateness of the instrument for measuring the desired construct[2]. A test can be completely consistent, yet measure the wrong thing. If students take a calculus test and every time they take the test they score at a consistently high or low level, the test would be deemed reliable. Yet the test probably would not be "valid" to assess a student's teamwork skill. This would be an example of a reliable, but invalid (for teamwork) test.

There is much written about psychometric evaluation of tests and instruments[3,4,5,6] . The typical evaluation includes addressing various types of validity and various measures of reliability. In addition, a complete analysis will include an item by item analysis.

Much work has been done evaluating surveys used in assessment of various individual characteristics. Many involved in engineering education will be familiar with the Felder's Index of Learning Styles[7] and particularly the accompanying psychometric evaluations of the instrument[8,9]. These studies give much information about the usefulness of the survey and the techniques used should also be used for evaluation of test instruments. Schimmel, King and Shamsuddin[10] discussed the development and use of a standardized exam for program improvement, but they stop short of evaluating this exam for psychometric soundness. Companies that administer standardize tests such as the SAT or the EIT/FE spend many hours and support much research to evaluate the psychometric properties, test bias, and other issues in order to produce the best possible test[11].

Reliability

Reliability coefficients measure the consistency of the instrument. Specifically it measures the consistency with which the test ranks test-takers in the same order on two test administrations. According to Chase[3] there are four methods of calculating reliability: test-retest, alternate forms, split-half, and internal consistency. In "test-retest" reliability, a test is administered twice with a period of time between test administrations. The correlation in scores is calculated. In the next method, "alternative forms," two forms of the same test are given to the same group of people. The correlation between scores is then calculated. In the "split-half" method, a test is divided into two tests each with half the items of the whole test. The score is calculated for each half test and then the correlation between the two scores is calculated. The last method, "internal consistency," is a method of determining how consistent each item in the test is with every other item in the test. This method is developed from the split-half method, but is extrapolated to include all possible subtests and the correlations between them. Internal consistency is usually calculated with Chronbach's Alpha[5]. There are obviously advantages and disadvantages to each method of measuring reliability, and often several methods are used in order to gather evidence of the test's reliability. The test-retest value may be influenced by learning that occurs between test administrations. The alternate form method could be influenced by fatigue when taking two tests at one sitting. The split-half and internal consistency methods require a homogeneous test, one that is measuring only one dimension of knowledge.

Validity

Validity is the extent to which the test measures the specific trait of interest. In this case we want the test to be a valid tool for assessing knowledge acquisition for our graduating Industrial Engineering (IE) and Manufacturing Engineering (MfgE) seniors. There are generally three types of validity used in educational psychometrics: Content, Criterion and Construct Validity[3]. The

first type, "content validity," is qualitative in nature. It attempts to judge the extent to which the test is a sample of the total content in the subject matter. For instance, in assessing math ability one would want to make sure that all subject areas of math are included in the test. In some situations a classification matrix will help to insure that that content on the test matches the content of the area of interest. The second type, "criterion validity," is the correlation of the test score to some other independent measure of the same trait. In the case of this exam a good criterion for comparison is the GPA of students. If the test score is highly correlated with GPA, this is evidence of criterion validity. The last type, "construct validity," is also qualitative; it looks at the extent to which the scores correspond with predictions based on psychological theory. For an ability test, evidence for construct validity might be that an individual would get a higher score on the test if they are closer to graduation than if they have just entered the program. This would be consistent with the theory of knowledge acquisition.

Item Correlation

There are many methods used to evaluate the correlation of test items with important quantities. Initially the difficulty for each item should be calculated. Difficulty is simply the relationship between each test question and the percentage of individuals who got the item correct[5]. A difficulty score is between zero and one, corresponding to no subjects answering the item correctly and all answering correctly, respectively. When items have difficulty of 0.0 or 1.0 they are not contributing to the test at all. The time consumed in answering them is not giving any additional information. It is better to have items with difficulties closer to 0.5. In addition to item difficulty, one must also ensure that every answer choice is appropriate and that distracters, incorrect choices, are not deceiving or inappropriate. If no one is picking a particular distracter or if everyone is picking a distracter, it may need to be revised. Item discrimination is also a common characteristic evaluated. This is calculated as the correlation of each item score with the total score on the test. Discrimination indicates the extent to which the item is behaving similarly to the test as a whole[5]. If an item has a high (close to 1.0) correlation, the item discriminates well. If a test item is negatively correlated with the total test score this would mean that when an individual gets that item correct he generally scores lower on the test – typically an undesirable situation. Other methods of item analysis include a factor analysis in order to discover correlation between underlying dimensions such as the ABET Criterion 3 a-k outcomes that the test is based on. In order to perform this type of analysis a large sample of individuals is needed.

Because ABET doesn't have any requirements for verification of a instrument, a senior exam is too often assumed to be a perfect instrument and improvement efforts launched as soon as a poor score is seen in a content area. It is true, however, that any conclusion drawn from the results of a test are only as good as the test itself. If a test measures performance inconsistently or worse yet measures the wrong kind of performance, the conclusions drawn may be incorrect. If the test results are to be used for initiating time-consuming and possibly disruptive program changes, the test should be thoroughly analyzed for psychometric soundness.

The objective of this study is to evaluate the senior exam used at Cal Poly to assess outcome achievement and revise the test based on the evaluation results. This paper will outline the methods used for evaluation, present the results, and describe the steps taken to interpret the results and use them to make changes to the exam. The evaluation methods include calculations of reliability, validity, and item analysis. An explanation will be given of the development and administration of the test at Cal Poly, and results will be presented based on tests given during the 2007-2009 period. Finally the paper will show how this evaluation has led to continuous improvement of the exam, to be offered to a new group of seniors in 2010.

EXPERIMENTATION

Test Development

The exam is actually two exams, one developed for IE and one for MfgE. They were developed to address all eleven of the ABET Criterion 3a-k program outcomes as well as the discipline-specific outcomes (ABET Criterion 9) specified for accreditation of each program. Since the Industrial and Manufacturing Engineering (IME) Department at Cal Poly has approximately eleven full-time faculty at any given time, it was natural to assign a program outcome to each faculty member, usually close to their area of interest. The faculty member was thus granted "ownership" of the outcome and was expected to identify the most appropriate assessment tools for measuring student success, including both the senior exit exam and other tools. The faculty were then asked to accumulate suitable questions specifically related to their program outcome for the exam. Although many questions were posed that apply to both the IE and MfgE exams, some of the questions were specific to major. In this way approximately 155 questions were gathered to cover the eleven outcomes for both exams. Ninety of the questions can be administered to either IE or MfgE students. Thirty-one are specific to IE students, and 33 are specifically for MfgE students. The questions came from a variety of sources, including individual course assignment and exam questions or modified examples from other engineering review materials. Most questions are multiple choice with 4 or 5 options. On each test there are 3 or 4 short answer questions. Updated questions are solicited periodically.

Administration

The one-hour exam at Cal Poly is given every quarter to students finishing their senior project. The exam for each major is typically offered as three different versions each drawing over 50 questions from the ever-expanding bank of test questions that cover virtually all of the program outcome areas and skills. Students are not advised to prepare for the exam, but they are notified that a portion (usually 10%) of their senior project grade will be affected by their performance on the exam. Thereby, skipping the exam or purposefully answering incorrectly (e.g., to get finished sooner) is discouraged. Since the content tested typically spans their educational career, a score of 50% or better is considered minimally acceptable from a program assessment point of view. Such a score is typically used as a passing score on the national EIT/FE exam, and it is the intent to

provide questions of comparable difficulty. Questions are continuously revised in an attempt to standardize the level of difficulty, but it is acknowledged that some variation exists in this regard.

To study test-retest reliability, individuals who took the test in the Spring of 2009 were recruited to retake the test in Fall of 2009. The students were given the exact same test version that they had taken in the Spring. Alternative test forms were not administered back to back because it seemed that too much fatigue and not enough effort would result from the students for the second administration of the test.

Statistical Analysis

A common, commercially-available statistical analysis software package was used to analyze the results. The analysis tools used included correlations, standard one-way and two-way ANOVA, and paired t-test with a presumed level of significance of .05. Cronbach's Alpha, a measure of internal consistency, was also calculated using the standard software.

RESULTS AND DISCUSSION

The IME department typically graduates approximately 65 IE students and 15 MfgE students each academic year. The exams in the current form have been administered for two years, resulting in 160 individuals' test results. The average score on the exams is 65.7%; the standard deviation is 8.7%. A distribution of the scores is illustrated in the histogram in Figure 1.

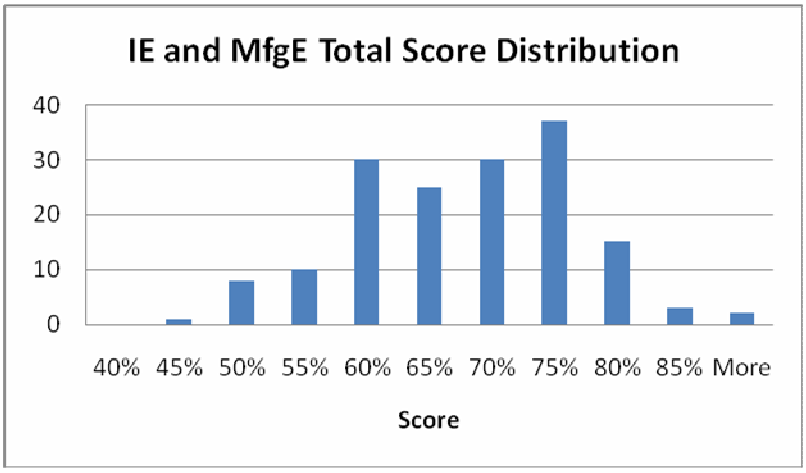

Figure 1: Distribution of scores 2007-2009, IE and MfgE majors

The scores by gender and ethnicity are included in Table 1 below. There are no obvious patters to suggest bias in the scores based on ethnicity or gender, although both males and whites score higher than the other groups. Cell values include the average score and the standard deviation in parenthesis followed by the number of individuals in each group.

Table 1: Scores by gender and Ethnicity

Gender	Score (stdev) n
Male	66.3% (8.6%) 125
Female	63.3% (9.2%) 34
Total	65.7% (8.7%) 160

Ethnicity	Score (stdev) n
White	67.3% (8.2%) 112
Asian	61.3% (9.2%) 18
Hispanic	62.1% (9.2%) 25
Other	63.7% (8.7%) 7
Total	65.7% (8.7%) 160

The histogram in Figure 1 aggregates the scores for both IE and MfgE students. The scores for the three separate versions of each test are summarized below. Cell values include the average score and the standard deviation in parenthesis.

Table 2: Scores by Major and Test Version

	Score (stdev)			Score (stdev)
IE version A	68.0 % (8.8%)		MfgE Version A	70.0% (8.8%)
IE Version B	64.4% (8.9%)		MfgE Version B	64.4% (9.0%)
IE Version C	63.2% (7.9%)		Mfge Version C	66.7% 7.7%
All IE versions	65.3% (8.8%)		All MfgE Versions	67.1% (8.7%)
IE and MFGE			65.7% (8.7%)	

The average score on the six versions of the test vary from 63.2% to 70.0%. A standard one-way analysis of variance (ANOVA) on score by the six versions was conducted resulting in a p-value of 0.056. This indicates the effect of the versions is not significant overall at the chosen level (.05) but may be close enough to be considered a cause for concern, especially if the results are used to evaluate individuals. The ANOVA results indicate that 3.6% of the variance (adjusted r^2) in scores

can be explained by the version given. Practically speaking, students who get the MfgE version A test will on average score 7% more than the students given IE version C. Because this test is used primarily for program evaluation and not individual evaluation, this result is not a cause for concern. If at any future time the results are used for judging individuals, the test versions should be changed.

Since the test is meant to assess achievement of the ABET Criterion 3 a-k outcomes, sub-scores are also calculated in Table 3 below. When used at Cal Poly for program assessment, the absolute level of achievement of each of the outcomes is considered. For instance the average score for outcome b - ability to design and conduct experiments, as well as to analyze and interpret data – is below the threshold 50% level. This was a flag to evaluate further the achievement of this outcome. In this case, the faculty agreed that the items provided for outcome b were significantly more difficult that the other items and did not properly reflect expected level of achievement. The outcome b questions are currently under revision.

Table 3: GSE score and sub-scores IE and MfgE

Outcome	IE (n= 130)	Mfge (n = 32)	Total (n=162)
a	67.0% (21.5%)	73.7% (23.4%)	69.2% (20.8%)
b	48.3% (22.2%)	43.8% (31.4%)	48.2% (23.8%)
c	63.9% (21.9%)	69.8% (24.6%)	66.6% (19.9%)
d	76.5% (19.2%)	54.9% (29.8%)	73.2% (21.9%)
e	62.6% (26.2%)	70.3% (30.7%)	64.9% (26.4%)
f	52.3% (25.8%)	52.7% (26.8%)	53.0% (25.4%)
g	76.0% (26.9%)	74.0% (26.4%)	76.5% (21.2%)
h	76.8% (26.9%)	72.1% (22.0%)	76.9% (24.8%)
i	69.2% (35.1%)	68.8% (50.3%)	74.4% (31.6%)
j	83.7% (19.5%)	70.4% (25.5%)	82.1% (19.5%)
k	60.6% (17.3%)	70.6% (24.3%)	62.9% (18.7%)
Total	65.3% (8.8%)	67.1% (8.7%)	65.7% (8.7%)

Test Reliability

Several reliability measures were calculated. The first was based on nine students who retested for reliability. All but one student scored higher in the retest. The average increase in score was 3.4%, although the increase was not statistically significant based on a paired t-test. A computation of correlation between the scores on the first test and those on the retest resulted in 0.810 (p = .008). Table 4 below illustrates the reliability (correlation coefficient) of the individual sub scores by outcome. The individual outcome scores are not as reliable in general as the entire test. It could be that the sub-scores are less reliable or it could be because in general the longer the test the more reliable it is[3]. An example of the misuse of score information without concern for the psychometric soundness of the measure can be seen when closely examined the outcome b values. As mentioned before, the average percent correct for outcome b was a low 48.2%. But referring to Table 4, the reliability of this outcome is also very low, implying inconsistent or low reliability on this sub-score.

Table 4 Test-retest Reliability

Outcome	Correlation coefficient
A	.632
B	.078
C	.228
D	.573
E	.654
F	.867*
G	.887*
H	.641
I	.791*
J	.788*
K	.511
Total	.810*

*Significant at a .05 level

Although slight changes are made each time the tests are administered, the results from a small sample of 24 test takers, who were given the same IE version A in the Spring of 2009, were used to calculate the split half reliability and internal consistency, expressed by Cronbach's alpha. The split-half reliability was 0.680 and Cronbach's alpha is 0.666.

Table 5 below summarizes the reliability results. Because these values are correlations, the higher the value the better, but usually values between 0.7 and 0.8 are considered acceptable.

Table 5 Summary of Reliability Measures

Test – retest (6 moths)	0.810
Split-half Reliability	0.680
Cronbach's Alpha	0.666

Test Validity

Content Validity. The objective here is to evaluate the extent to which the test samples the whole content of knowledge for IE and MfgE students. One way to do that is to enumerate the desired outcomes and make sure each is evaluated. The test was developed using the ABET a through k outcomes so a natural list would include these. Table 5 below includes the number of questions out of the 50 to 56 that evaluate each to the outcomes.

Table 6: Number of questions in the test that pertain to ABET a through k

Outcome	IE	MFGE
a – Application of Math, Science and Engineering	6	8 or 9
b – Design of Experiments	3, 6 or 7	5
c – Design	7 or 8	9 or 10
d – Multi-disciplinary Teams	4 or 5	3 or 4
e – Solve Engineering problems	3	2
f – Ethics	4	4 or 5
g – Communications	4	3
h – Broad education	2 or 3	3 or 4
i – Lifelong Learning	1, 2 or 3	1, 2 or 3
j – Contemporary Issues	3, 5 or 6	4, 5 or 6
k – Engineering Tools	6 or 7	3, 4 or 5

Other categorization of questions may also be helpful to evaluate content validity. For instance the curriculum could be examined to make sure that each skill or subject area is tested.

Criterion Validity. In order to evaluate the test on the basis of criterion validity the correlation between test score and GPA was calculated. The correlation of 0.244 ($p = 0.002$) was obtained for all IE and MFGE students. Although this seems like a low correlation for engineers who are used to dealing with objects, a correlation in this range when human subjects are being measures is at an acceptable level. To put this in perspective the correlation between SAT scores and first year college GPA is often calculated between .10 and .30[12].

When the test administrations generate more data the correlation of the sub-scores with course grades could generate more criterion validity evidence. At this point there is insufficient data to check these correlations.

Construct Validity. The extent to which the measure matches theory on ability tests can be seen in the increase in scores on the test-retest administration. The increase of 3.4% is not statistically significant, but given that 8 of the 9 individuals showed an increase in scores, is an indication on construct validity.

Test Item Correlation

There are 154 items in the test pool; each test has between 50 and 56 items. There is one item in the pool that has a 0.0 difficulty and 16 that have a 1.0 difficult; these items should be revised or eliminated. The range of difficulty values can be seen in Figure 2. Only 39 items out of 154 have a difficulty below 50%.

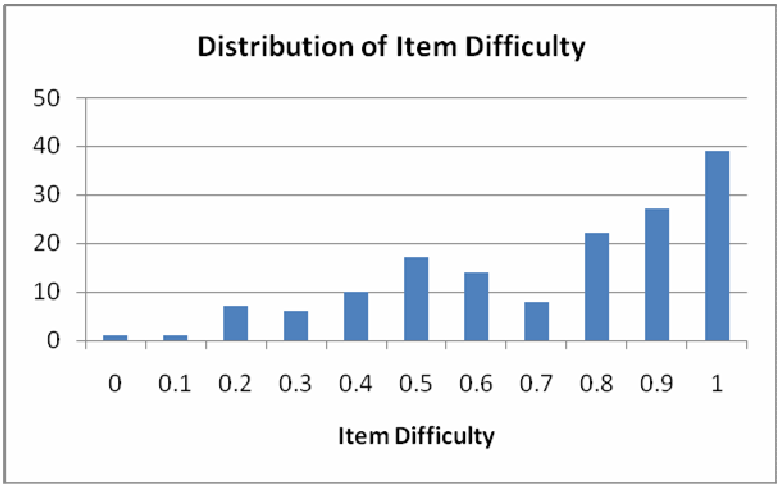

Figure 2: Item Difficulty

Items were also analyzed for discrimination. There are 15 items that have a negative correlation with the total score, but none of these correlations are statistically significant at the 0.05 level. The average correlation coefficient is 0.222. The distribution of discrimination values are shown in Figure 3. Of the 109 items for which the data is sufficient to calculate a correlation, 37 showed values significant at the 0.05 level, all positive.

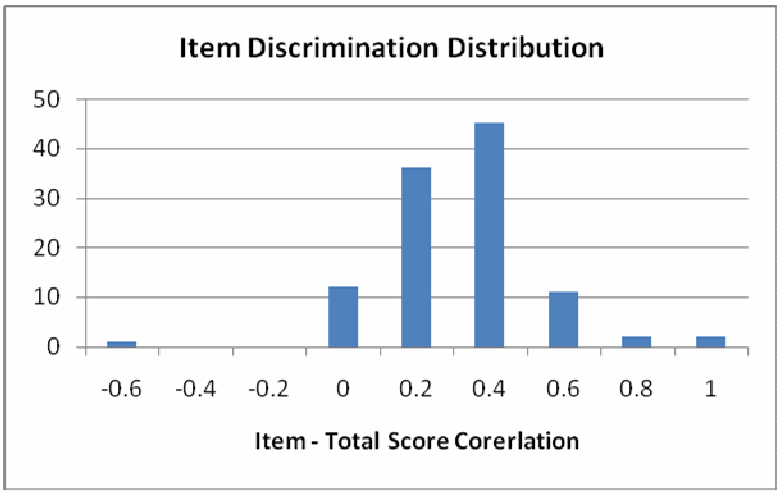

Figure 3: Item Discrimination

Test Revision

The results of applying psychometric analysis tools to study the performance of the exam can be used to revise and improve the test itself. Several revisions are in progress. Items that have difficulty of 0.0 or 1.0 and those items with negative discrimination will be dropped. The difficulty of each version will be balanced by carefully choosing items with appropriate difficulty. Faculty are being asked to review their outcome questions for correctness, timely use of topics, and distracters based on the results shown above. A careful analysis of response to item choices could lead to rewording or enhancing items. The test will be re-administered in March 2010 and again in June 2010. It is expected that the test's psychometric properties will increase with this thorough evaluation and subsequent update.

Although this study leads to important conclusions, statistical significance could be enhanced with more test administrations. In addition, there are limits due mainly to the interpretation of qualitative information. These interpretations depend on individuals familiar with psychometrics and test evaluation.

SUMMARY AND CONCLUSIONS

A psychometric analysis was conducted for the regular administration of a senior exit given by the IME Department at Cal Poly for the purpose of program assessment. The analysis evaluated the exam for reliability, validity, and other measures of correlation that indicate the overall effectiveness of the exam. Results were obtained through standard statistical methods used on data from two years of quarterly exams and a small set of retests.

Overall, the scores for individuals taking the test are at an appropriate level. Although not significant at the 5% level, there is bias due to test version at the 0.056 level. This bias is not a problem if the test results are used for program improvement, but if the results are used for

individual evaluation this bias may be a concern. This can be corrected by revising the test versions so that they have equal average difficulty. But there does not seem to be any bias in the scores based on gender or ethnicity, indicating a fair test.

Reliability of the test as measured by test-retest correlation, split half, and Cronbach's alpha show acceptable levels of correlations. The test-retest value of 0.81 is an indication of an especially reliable test. Even with a sample size of nine, the value is statistically significant at a 0.05 level. The split half and Cronbach's alpha values are somewhat lower, although still in an acceptable range. The slightly lower values suggest a somewhat non-homogeneous test. Because the test is intended to measure eleven ABET defined outcomes, it may indeed be heterogeneous.

The validity of the test is indicated by several pieces of evidence. The content is aligned with the outcomes that it is trying to measure. The correlation between the test score and GPA is in an acceptable range and statistically significant at a 0.05 level. In addition, the evidence of construct validity from the increase in the test-retest scores, although not statistically significant, shows that the test is probably measuring an ability that increases over time.

The item analysis points to a fairly good pool of items. There are some instances where the items are not contributing any information (difficult of 0.0 or 1.0), or the items are not discriminating well (item-total score correlation below 0.0). If these items are omitted or revised, an increase may be seen in the test reliability and validly.

As more individual data is collected, there will be opportunities to perform more extensive statistical evaluations. This may include factor analysis, or further investigation into the sub-scores. In addition, more work will be done to verify the reliability measures such as more individuals taking the retest or more data points for internal consistency calculations.

The analysis done in this project could be the basis for evaluating all instruments used for program improvement. It may be appropriate for ABET to develop seminars or workshops addressing these techniques.

Bibliography

1. http://www.abet.org/ URL accuracy confirmed on January 4, 2010
2. Thorndike, R. M. (2005), *Measurement and Evaluation in psychology and education, 7th edition.* Person , New Jersey.
3. Chase, C., (1999) *Contemporary Assessment for Educator,* Longham, New York.
4. Brown, F. G. (1970). *Principles of Educational and Psychological Testing.* Dryden Press, Hinsdale, IL.
5. Hopkins, K. D., (1998) *Educational and Psychological Measurement and Evaluation, 8th edition*. Allyn and Bacon, Boston.
6. Mehrens, W. A., and Lehman, I. J., (1984) *Measurement and Evaluation in education and Psychology, 3rd edition*, Holt, Rinehart and Winston, New York.
7. Flder, R. M. and Silverman, L.K. (1988) "Learning and Teaching Styles in Engineering Education," *Engr. Education, 78*(7), 674-681.

8. Felder, R.M. and Spurlin, J.E. (2005) "Applications, Reliability, and Validity of the Index of Learning Styles," *Intl. Journal of Engineering Education, 21*(1), 103-112.
9. Litzinger, T.A., Lee, S.H. Wise, J.C. and Felder, R.M. (2007) "A Psychometric Study of the Index of Learning Styles," *J. Engr. Education, 96*(4), 309-319.
10. Schimmel, K.A., King, F. G., Ilias, S., (2003) Using Standardized examinations to assess engineering programs. Proceedings of the 2003 American Society of Engineering Education Annual Conference and Exposition.
11. Buckendahl, C. W, and Plake, B. S. (2006) "Evaluating Tests", in *Handbook of Test Development* edited by Downing S. M. and Haladyna, T. M., Lawerence Erlbaum, Mahwah, New Jersey.
12. Geiser, S. and Studley, R., (2002) "UC and the SAT: Predictive Validity and Differential Impact of the SAT I and SAT II at the University of California." *Educational Assessment,* 8(1) p1-26.