Validity of Student Self-Assessments

Sanjiv Sarin, Donald Headley
North Carolina A&T State University

Abstract

This paper examines the validity of self-assessment as a tool for measuring student abilities, in particular, whether self-assessments can be used as valid substitutes for instructor evaluations. Data is obtained from in-class student self-assessments and subsequent course tests that measure the same abilities. Correlation between self-assessment ratings and test scores are used to comment on the validity of self-assessments. Some observations are also made regarding the extent to which students over or under report their abilities on self-reports.

1. Introduction

Interest in valid methods for assessing student outcomes has grown in higher education. This is especially true in engineering education as a consequence of the new Engineering Criteria 2000 (EC 2000) requirements that became fully effective starting from the Fall 2001 visitation cycle. The new EC 2000 criteria represent a major shift in the philosophy behind accreditation of engineering programs. Instead of focusing on resources and inputs provided to an engineering program, the new accreditation criteria emphasize student learning, i.e., outcomes. The use of outcomes assessment data to guide the improvement of educational processes is a cornerstone of EC 2000[1].

Given the limitations of achievement tests in fulfilling outcomes assessment requirements, the assessment community has recommended several alternative approaches for assessing student outcomes. These include portfolios, capstone design project evaluations, student, alumni and employer surveys, and placement data of graduates. Yet, written surveys administered to current students are the most frequently used assessment instruments, due in part to two reasons – one, they are relatively inexpensive to conduct, and, two, a high response rate is almost guaranteed. A natural question is whether these student self-assessments are valid substitutes for test questions created and scored by an instructor.

This paper reports the results of a study undertaken to examine the validity of student self-assessment surveys as a tool for measuring the mastery of concepts. The term validity is defined as the correlation between self-assessment and test scores on the same measures. The paper reports results of data collected from three courses (two sophomore and one junior) taught over one academic year. The data represent over one hundred students in these courses. Students

were asked to subjectively rate themselves on very specific abilities. The same students were then given a test that measured the same abilities. Correlations between students' self-assessment and their performance on test questions are used to comment on the general validity of engineering student self-assessments.

2. Background

Interest in student self-reports has been motivated by the recognition that achievement tests alone are not sufficiently viable for outcomes assessment. This is due to several reasons. First, it is questionable whether a test can adequately measure all outcomes of interest, especially those involving soft skills such as team work and life long learning. Second, there is concern whether a single nationally administered test (or a small number of tests) can be used by a variety of institutions given that each institution has a different set of priorities and educational missions. However, before students' self-assessments can be used as proxies for tests and other direct assessments, it is important to establish their validity.

Several large-scale studies in this area have concluded with mixed results. Ewell et al[7], Pike[14] and Tsang[17] provided some degree of empirical support for student self reports as valid proxies for tests. However, the work of Astin[2] and Dumont and Troelstrup[6] has reported low correlations between self-reports and test scores. LeBold et al[11] utilized Purdue's Mathematics and Science Inventory to demonstrate the validity of student self-perceptions regarding their mathematics and chemistry development as they progress through several courses. LeBold et al also noticed strong relationships between self-reports and final course grades, in that, students who received D's or F's did not report appreciable pre-to-post test gains on their self-reports whereas students who received A's reported the maximum gains.

Comparisons between self-assessments and peer-assessments of the same entities have been conducted in the context of team based student work evaluation. Falchikov[8] studied self and peer assessments among psychology students and found high levels of agreement between peer scores with no under-scoring or over-scoring by self-scorers. A similar result was reported by Stefani[16] who compared self (and peer) assessment scores and tutor's scores in the biological sciences and found that student assessment (both self and peer) was as reliable as tutor assessment. Further, there was no evidence that higher achievers were under-scored or that lower achievers were over-scored. In contrast, in their study of business management students, Saavedra and Kwan[15] found that high performers were more discriminating than low performing students and were better able to assess performance. As part of the same study, the authors also concluded that below average students rated themselves significantly higher than the peer ratings they received. As part of a larger study aimed at identifying individual effort in teamwork, Kaufman, Felder and Fuller[9] concluded that individual self-ratings were statistically indifferent from ratings received from peers. The authors also reported that inflated self-ratings were less common than deflated self-ratings. Kruger and Dunning[10] studied the relationship between student confidence and their ability and concluded that those who are incompetent often appear to be more confident about

their abilities than students who are competent. On the other hand, students who were competent often underestimated their abilities. For a review of literature relating to student-based assessment, the reader is referred to the paper by Maskell[12].

It bears mentioning here that self-assessments are also conducted in employment settings to comply with a variety of company regulations. For instance, many organizations require employees to perform self-appraisals prior to a formal annual evaluation by a supervisor, and to report employee conformance to substance use policies. Reported work in this area also sheds light on the degree to which self-reports can be used as substitutes for direct measures. For example, Cook et al[5] studied 1,200 employees of a steel manufacturing company to assess illicit drug use through three methods – self-reports, urinalysis and hair analysis. Results of this study indicated that actual prevalence of drug use (as confirmed by urine and hair samples) was approximately 50% higher than estimates given in self-reports.

In addition to the validity of self-reports, there is also a larger and more useful context in which self-assessments are being studied. This involves the impact of self-assessment on learning itself. Recent research summarized by Black and William[3] demonstrates that student self-assessment skills enhance student achievement and learning. Realizing that self-assessment skills are necessary for enhancing learning, some instructors (see for example, Brackin and Gibson[4]) require students to demonstrate these skills as part of their course requirements.

The purpose of this paper is to extend past work in studying the correlation between self-reports and test scores. The paper reports the results of a yearlong experiment to study this correlation in an engineering program.

3. Approach

The subjects for this experiment were students enrolled in three courses offered by the Industrial Engineering department at North Carolina A&T State University during the Fall 1999 and Spring 2000 semesters. All three courses were taught by one of the authors. Catalog descriptions of these courses and prerequisite information are available from the authors upon request. Enrollment values and other pertinent details about these courses are presented in Table 1.

Table 1: Course Information

Course Number	Course Title	Course Level	Semester	Number of students
INEN 270	Engineering Statistics	Sophomore	Fall 1999	32
INEN 330	Operations Research	Junior	Fall 1999	35
INEN 260	Engineering Economy	Sophomore	Spring 2000	39

The data collection consisted of a written survey followed by a test in each of the courses. The following procedure was used. In the last week of the semester, a written survey was

administered to allow students to self-assess their proficiency in material learned in each course. This was immediately followed by a test that carefully measures the same proficiencies that were addressed in the written surveys. Each survey and test consisted of five items. The survey responses carried no points towards course grade, but the tests were worth 15% of the course grade. Copies of the written survey and corresponding test for each course are not presented here in the interest of space, but are available from the primary author. In addition to student responses on the survey and test questions, their overall grade on the course was also recorded. Both the surveys and the tests were encoded to promote anonymity of responses as well as to allow the computation of correlation between responses on corresponding responses by individual students.

Students were asked to rate their ability in a very specific area on a scale of 1 (strongly disagree that I can do this) to 5 (strongly agree that I can do this). Each question on the survey corresponds to a specific question on the test. The test questions were graded on a scale from 0 (incorrect or no answer) to 5 (complete and accurate answer). This was appropriate since the test questions required students to show assumptions, steps and the final answer. Therefore, an answer to a question can be "more correct" or "more wrong." The correlation between survey ratings and test scores can then be used to comment on the validity of the self-report surveys as surrogates for the tests.

As an example, in INEN 270, one of the questions on the survey was as follows:

Given a probability distribution function for a discrete random variable, I can compute its expected value.

Students were asked to select one of the following five choices in response to this question:

Strongly Disagree	*Disagree*	*Neutral*	*Agree*	*Strongly Agree*
1	*2*	*3*	*4*	*5*

The corresponding question on the test is as follows:

The number of orders per day for a component used in an assembly plant is a random variable with the following probability distribution:

No. of Orders	*100*	*110*	*120*	*130*	*140*	*150*
Probability	*0.156*	*0.200*	*0.250*	*0.196*	*0.143*	*0.055*

What is the expected (average) number of orders in a day?

Clearly, the question posed on the test was designed to measure the same ability referred to in the self-assessment survey. Furthermore, the test is administered immediately after the students have completed the survey. This was done to avoid bias due to time lag and any associated changes in student abilities over time.

4. Data Analysis And Discussion

The data was analyzed with two primary objectives. One, to explore whether student self-assessments are significantly correlated with achievement test scores and if so, whether self-assessments can serve as useful proxies for test scores. Second, to verify whether high performing engineering students demonstrate an ability to perform more accurate self-assessments as compared to low performing engineering students.

To assess relationships between the ordinal level self-confidence and test scores, the nonparametric Spearman's correlation coefficient test was used. Correlations between student confidence reported on the survey and test scores for each course as well as for the aggregate over all courses are reported in Table 2. Each coefficient is based on data pairs of the five survey and corresponding test items. For example, for the INEN 270 sample, the number of pairs is 160 (5 times 32 students). The hypothesis tested is a two-tailed test (i.e., the alternative hypothesis is that there is a relationship between self-confidence and test scores) with an assumed Type I error (α) = 0.05. Therefore, the results are deemed significant if the p-value exceeds $\alpha/2$ or 0.025. The formal statement of the hypothesis is as follows:

H_0: Spearman's $\rho = 0$

H_1: Spearman's ρ ? 0

Table 2: Correlation Coefficients (Spearman's) of Self Confidence and Test Scores

Sample	Observed Value	Conclusion
INEN 270	0.258	Significant
INEN 330	0.221	Significant
INEN 260	0.125	Not significant
All courses	0.184	Significant

Although the correlations between self-assessments and test scores are generally significant, their absolute values are low and suggest a low correlation between the two variables. If the purpose of student self-assessments is formative in nature, their use can be justified. However, the use of self-assessments to satisfy summative evaluation requirements is questionable.

To distinguish between the self-assessment abilities of high performing versus low performing students, the data was sorted to separate responses of high performing (those receiving course grades of A or B) and those of low performing (those receiving course grades of C or lower). Each student's self confidence rating (CR) and corresponding test score (TS) was transformed

into a deviation D = CR – TS. With this transformation, a deviation of zero represents perfect ability to perform self-assessment (assuming that the instructor's grade has no error). A value of $D > 0$ implies over-confidence whereas $D < 0$ implies under-confidence. In order to test for sample differences in the deviation statistic D, the D values were analyzed using a two-tailed Mann-Whitney statistical test. The obtained standardized z statistic's observed value was 6.1, and indicates significantly different unequal populations. The average value of D for high performers is 0.1 and the same statistic for low performers is 1.2. Thus, on average, low performers appear to exhibit a higher degree of over-confidence, whereas high performing students are more discriminating. These results provide confirmatory support for the previously reported finding that competent students can more accurately provide self-assessments.

5. Summary

Students in three courses were administered a written self-assessment survey followed by an instructor-graded test, both intended to measure the same set of abilities. The study finds statistically significant correlations between student self-assessments and test scores. However, the strength of the correlation is weak. Additional investigation of the data reveals that high performing students possess higher ability to perform self-assessments as compared to low performing students.

Based on the results of this study and past work done in this area, it can be concluded that self-assessments can be used as surrogates for achievement tests in formative assessment. It can also be argued that self-assessments performed by high performing students be used instead of aggregated self-assessments of all students.

6. Acknowledgment

The research leading to this paper was supported in part by a grant from SUCCEED, an NSF coalition that aims to revitalize undergraduate engineering education. This support is gratefully acknowledged. The first author would also like to thank his students in INEN 260, 270 and 330 for voluntarily participating in this study.

References

1. ABET, *Criteria For Accrediting Programs In Engineering In The United States*, Accreditation Board for Engineering and Technology, Inc., 111 Market Place, Suite 1050, Baltimore, Maryland 21202-4012. *http://www.abet.ba.md.us/EAC/eac2000.html*
2. Astin, A.W., *What Matters in College: Four Critical Years Revisited.* Jossey-Bass Publishers, San Francisco, CA, 1993.
3. Black, P. and William D., "Inside the Black Box: Raising Standards through Classroom Assessment," Phi Delta Kappan, 80(2), 139-148, 1998.
4. Brackin, P.M. and Gibson, J.D., "Techniques for Assessing Industrial Projects in Engineering Design Courses", Proceedings of the Annual Conference of the ASEE, Albuquerque, NM, June 24-27, 2001.

5. Cook, R.F., Bernstein, A. Arrington, T.A., and Marshall, G., "Methods for Assessing Drug Use Prevalence in the Workplace: A Comparison of Self-Reports, Urinalysis and Hair Analysis," The International Journal of Addictions, 30(4), 403-426, 1995.
6. Dumont, R.G. and Troelstrup, R.L., "Exploring Relationships between objective and subjective measures of instructional outcomes," Research in Higher Education, 12, 37-51, 1980.
7. Ewell, P.T., Lovell, C.D., Dressler, P. and Jones, D.P., *A Preliminary Study of the Feasibility and Utility for National Policy on Instructional Good Practice Indicators in Undergraduate Education.* US Government Printing Office, Washington, D.C., NCES 94-437, 1994.
8. Falchikov, N., "Group Process Analysis: Self and Peer Assessment of Working Together as a Group," Educational and Training Technology International, Vol. ETTI 30.3, 275-284, 1991.
9. Kaufman, D.F., Felder, R.M. and Fuller, H., "Accounting for Individual Effort in Cooperative Learning Teams," Journal of Engineering Education, 89(2), 133-140, 2000.
10. Kruger, J., & Dunning, D., "Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments," Journal of Personality and Social Psychology, 77, 1121-1134, 1999.
11. LeBold, W.K., Budny, D.D. and Ward, S.K., "How Do Students Grade Their Learning?" Proceedings of the Frontiers in Education Conference, Atlanta, GA, 1995.
12. Maskell, D., "Student-based Assessment in a Multi-disciplinary Problem-based Learning Environment," Journal of Engineering Education, 88(2), 237-241, 1999.
13. Mehta, S. and Danielson, S., "Self-Assessment by Students: An Effective, Valid, and Simple Tool?" Proceedings of the Annual Conference of the ASEE, St. Louis, MO, June 18-21, 2000.
14. Pike, G.R., "The Relationship between Self-Reports of College Experiences and Achievement Test Scores," Research in Higher Education 36, 1-22, 1995.
15. Saavedra, R. and Kwan, S.K., "Peer Evaluation in Self-Managing Work Groups," Journal of Applied Psychology. 78(3), 450-462, 1993.
16. Stefani, L.A.J., "Peer, Self and Tutor Assessment: Relative Reliabilities," Studies in Higher Education. 19(1), 69-75, 1994.
17. Tsang, E., "Assessing Student Learning For a Materials, Manufacturing and Design Lab," Proceedings of the Annual Conference of the ASEE, Seattle, WA, June 28-July 1, 1998.

Author Biographies

SANJIV SARIN

Sanjiv Sarin is a Professor of Industrial Engineering and Associate Dean for the College of Engineering at North Carolina A&T State University. He received a bachelor's degree in Chemical Engineering from the Indian Institute of Technology, New Delhi and a Ph.D. in Industrial Engineering from the State University of New York, Buffalo. He is a member of ASEE and IIE, and a registered Professional Engineer in North Carolina.

DONALD B. HEADLEY

Donald Headley is a Visiting Professor at North Carolina A&T State University. He is a Human Factors Research Psychologist with the Army Research Laboratory – Human Research and Engineering Directorate, Aberdeen Proving Ground, Maryland. He received a B.S. in Research Psychology from the University of Massachusetts, Amherst, and a M.S. and Ph.D. in Research Psychology from Oklahoma State University. He has completed postdoctoral work in statistics, program evaluation, and research administration.