Combination of Formative and Summative Assessment Instruments in Elementary Algebra Classes: A Prescription for Success

Euguenia Peterson & M.Vali Siadat
Richard J. Daley College

The purpose of this study is to examine the effects of the implementation of formative assessment on student achievement in elementary algebra classes at Richard J. Daley College in Chicago, IL. The formative assessment is defined in this case as frequent, cumulative, time-restricted, multiple-choice quizzes with immediate constructive feedback. The impact of this endeavor is measured by summative external (the national standardized COMPASS Test) and internal (departmentally designed) assessment instruments. The research is guided by three primary hypotheses. First, the use of formative assessment could reveal the levels of conceptual understanding in a timely manner allowing for continuous readjustment of teaching and learning strategies, thus improving student academic achievement. Second, an increase in frequency of formative assessment will provide better results on summative assessment. Third, the results achieved on valid and reliable internal summative assessments are an accurate predictor of the external examination scores. Using a four-group experimental and correlational design and t-test for significance analysis, it was found that after approximately four months the students (N=222) who were regularly assessed with weekly quizzes achieved significantly higher scores on the final examinations than the students (N=1352) whose instruction did not include formative assessment at all. The higher outcomes attained were in conjunction with higher retention rates. An increased frequency of formative assessment did not produce a significant improvement in students' learning outcomes. The results obtained on internal summative assessments were also highly correlated to the external examination scores.

Introduction

It is widely recognized that many students entering post-secondary institutions, such as community colleges and universities, are deficient in mathematical background and learning skills that are presumed necessary for the successful completion of college level mathematics courses. According to the National Center for Education Statistics (2000) survey, 35% of first-time freshmen entering public two-year community colleges in fall of 2000 were enrolled in remedial mathematics. The term "remedial mathematics" means any subject from basic arithmetic to intermediate algebra. At Richard J. Daley College, 95% of enrolled students are not adequately prepared for a college-level algebra course (ACT, 2008). In an attempt to prepare students to perform at the college level in a more efficient manner, it is important to re-examine the assessment process (Stingler & Hiebert, 1999).

The present study examines the synergistic effect of formative and summative assessment instruments on academic success in the remedial Elements of Algebra classes (Math 110) at Richard J. Daley College. "Academic success" is defined as the acquisition and retention of the most fundamental mathematical knowledge. It is quantified as the cumulative score of 70% or above attained on the internal assessment instrument, specifically designed during this project, and a score of 29 or higher on the external assessment instrument, in the form of a standardized COMPASS test. The conclusions about the effectiveness of using formative assessment in remedial mathematics courses are made based on statistics such as the difference between two means (for test and control groups) on the internal summative tests, through a correlation coefficient between the results of the internal and external tests, and, when appropriate, through tests of the statistical significance of such measures. The normalized gain is also calculated based on the difference in performance on the pretest (COMPASS test taken prior to the course) and the posttest (COMPASS test taken upon the completion of the course).

The research is guided by three primary hypotheses. First, formative assessment in the form of frequent, cumulative, time-restricted, multiple-choice quizzes with immediate constructive feedback measures the level of conceptual understanding and improves student academic per-

formance on the summative assessment instruments. If this hypothesis holds true, then the second supposition is that an increase in frequency of formative assessment will provide better results on summative assessment. While the first two hypotheses evaluate the influence of formative assessment on the results of summative assessment, it would be prudent to expand this idea and assess the impact of these techniques on the results of a national standardized test. Therefore, the third hypothesis states that the results achieved on valid and reliable internal (departmentally designed) summative assessments are indeed an accurate predictor of the scores on external examination (the national standardized COMPASS Test) and vice versa.

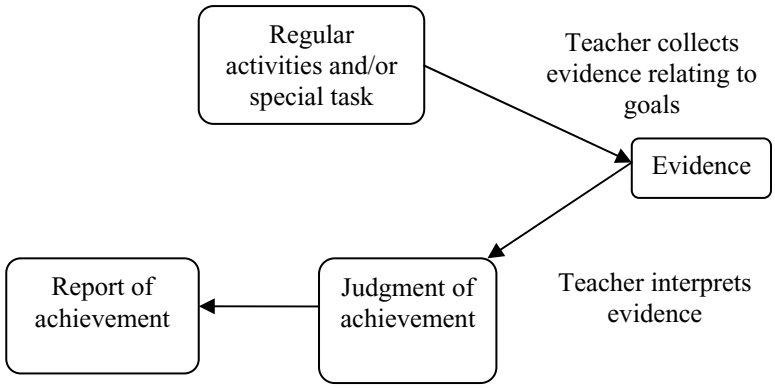

Figure 1. Summative Assessment

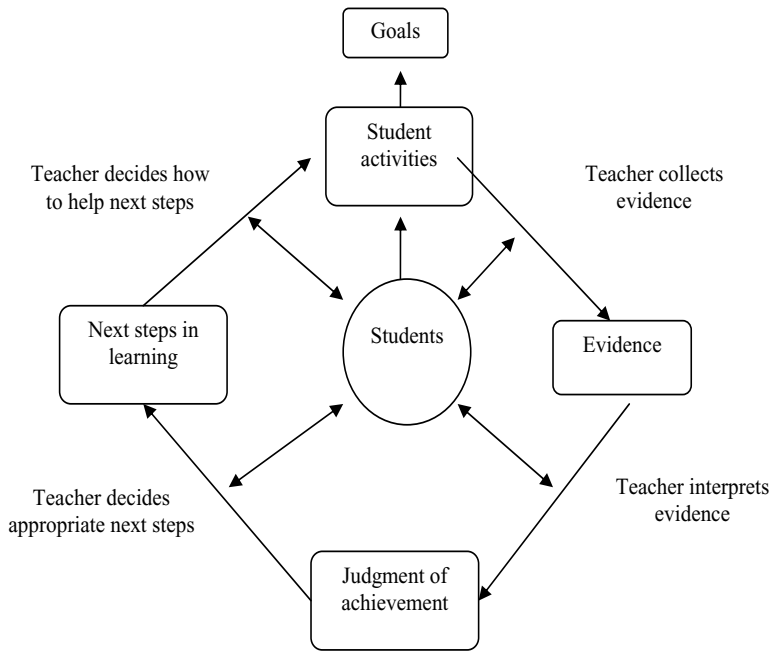

Figure 2. Formative Assessment Cycle

Literature

Definition of Formative and Summative Assessments

Black and William (1998) define assessment broadly as a set of activities that teachers and students undertake to get information that can be used diagnostically to alter teaching and learning to meet student needs. Under this definition, assessment encompasses teacher observation, classroom discussion, and analysis of student work, including homework and tests. In general, assessment can be divided into two types: formative and summative.

Scriven (1967) was the first to distinguish between formative and summative testing in his monograph on evaluation, but Bloom, Hastings, and Madaus (1971) extended the usage of these terms to their generally accepted current meaning. They defined those tests that were given at the end of each instructional period, such as midterm and final exams, as *summative assessment*. These examinations were designed for the purpose of grading and verifying the effectiveness of the curriculum (See Figure 1 and Bloom, Hastings, and Madaus, 1971). Hence, the focus of summative evaluation is on factual knowledge and the final outcomes only.

On the other hand, *formative assessment* involves systematic measurement of students' progress in the classroom and provides timely feedback to both the students and the instructor in order to guide their learning and teaching strategies toward successful completion of the course (See Figure 2). According to Hawthorne (1987), formative assessment is intended to modify and improve the educational program in progress by providing structured feedback. Although summative evaluations are more common in practice (Hawthorne, 1987; Mohr, 1995), comparison of the two figures suggests that formative and summative approaches are not mutually exclusive and that formative assessment should, in theory, prepare students to excel on summative tests.

Benefits of Formative Assessment

Currently, formative assessment is a highly recommended practice in education, but only several countries, such as Australia, Canada, and some European countries, promote formative assessment as a fundamental approach to education reform (*Formative Assessment Improving Learning in Secondary Classrooms,* 2005). There have been only three comprehensive and theo-

retically strong reviews of formative assessment in literature. Black and William's worldwide review (1998) is the most recent, and it references the findings from the two preceding works by Crooks (1988) and Natriello (1987). From Black and William's (1998) point of view, formative assessment enhances students' ability to learn, by allowing them to modify studying techniques. Classroom attendance and retention of learned material are also improved. The authors stress that formative assessment is more beneficial to low-achieving students and students with learning disabilities. Using this type of assessment, instructors have the opportunity to adjust teaching style and to experiment with new teaching techniques.

Vygotsky (1978), in his perspective of psychological development, provides the idea of formative assessment in the notion of scaffolding or acquisition of skills that occur by participating in activities with the support of a competent instructor. From Vygotsky's perspective, only competent instructors are able to assess their students' Zone of Proximal Development, which he defines as the true range of knowledge, skills, and capabilities that a student possesses. Social-Cognitive Learning Theory (SCLT; Bandura, 1986) complements the theory of scaffolding, stating that it is not sufficient to simply provide the right answer to the student. Instead, looking into the explanation of how to attain the right answer or reflecting on the process leading to an error, and explaining why that error may occur, would result in better performance in the future.

Components of the Formative Assessment Cycle

There are different variants of formative assessment that can be used to measure students' understanding of the material. Among them are different models of tests: group, paired, and individual, open- and closed-book, in-class and take-home. Different formats of tests also exist: fill-in-the-blank, true/false, short-answer, open-ended, multiple-choice, matching, arranging, identifying, ranking, and grouping. Other assessment tools are also available to instructors such as essays, concept maps, projects, interviews, surveys, portfolios, checklists, journals, labs, discussions, and homework (Whitaker, 1989). Some teaching techniques, such as teacher observation and classroom discussion, have an important place alongside analysis of tests and homework. Black and William (1998) encourage teachers to use questioning, but questions have to be reflective rather than simple and factual.

The formative assessment cycle is not complete without feedback. To demonstrate the effect of constructive feedback, Bangert-Drowns, Kulick and Morgan (1991) conducted a meta-analysis using 58 reports from 1960-1990. Most of them came from college-level samples of humanities and science classes. The results of this analysis indicated that feedback had the potential to raise student achievement from the 50th to the 59th percentile on different standardized assessments.

Since all the assessment instruments developed during this research utilized multiple-choice questions, this format of testing deserves a closer look. According to Angelo and Cross (2001), multiple-choice questions are appropriate for assessing students' mastery of details and specific knowledge. Carefully written multiple-choice items with well thought out distracters, indicating common misconceptions, can fairly accurately distinguish students with a good understanding of basic concepts from those who need additional practice (Sadler, 1989). This type of testing format could easily be implemented in mathematics courses. The advantages offered by this method are the increased ability to test a wide range of content and higher levels of cognition reliably in a relatively short time interval, and to score the items with higher speed. The statistical results, including item analysis of the tests, also provide a very accurate diagnostic tool for teachers to plan the next step.

Frequency of Formative Testing

There are many reasons to believe that frequent testing is beneficial to learning. For example, from the Grover, Becker, and Davis (1989) perspective, frequent testing improves regular study habits, encourages processing of information at a deeper level, and reduces anxiety. Daniel Willingham (2004) insists that any math concept studied for one semester and never reviewed again could be forgotten in 3-4 years. Yet, if this concept was reviewed and applied for 3-4 consecutive years, it might be retained for more than 50 years. Willingham believes that many basic mathematical operations must be practiced until they become automatized. Acquisition of this automaticity helps promote higher levels of thinking and learning.

In 1989, Dineen, Taylor, and Stephens investigated the effect of frequent testing on high school mathematics students. The group tested on a daily basis consistently outscored the group tested weekly. A greater difference was observed for students enrolled in the less difficult courses, indicating that frequent testing may be more effective with weaker students. The authors conclude that newly learned material should be tested within a week of first exposure.

Recent Progress in the Use of Formative Assessment

In an attempt to improve student performance in remedial mathematics courses, Siadat and Sagher (1997) developed the Keystone model that is based on rigorous assessment. Although authors did not use such specific terms as formative and summative assessment, their model

was founded on these basic concepts. The Keystone model used the novel technique of cumulative assessment in all its quizzes and tests. Cumulative tests require students to review material on which they have already been tested, thus reinforcing what they have previously learned (Siadat, Sagher, & Hagedorn, 2000 and Siadat, Sagher, Hagedorn, & Musial, 2001). Cumulative tests also give students a chance to integrate and synthesize course content (Crooks, 1988; Jacobs and Chase, 1992). The Keystone Project was piloted at Daley College in the 1998-1999 academic year. The research involved 11 tests (N = 332) and 9 control classes (N = 311). The study reported dramatic improvement in three areas: student performance on the final exam, student study skills, and student retention rate (Siadat, Musial, and Sagher, summer 2000, and Siadat, Sagher, Hagedorn, & Musial, 2001). While the original Keystone research provided highly impressive results, the study had to be expanded to encompass a larger student population.

Research Procedure

Research Design

In order to examine the effects of formative assessment on student achievement in elementary algebra classes and to test the hypotheses formulated in the introduction, a longitudinal experimental study was conducted. The following study was based on the quantitative research method with elements of experimental and correlational design models (See Table 1). According to Gay, Mills, and Airasian (2006), experimental and correlational studies are the most valid approaches to solve educational problems. Using this design, subjects were pre- and post-tested, in order to measure change over a period of intervention. The present experimental study differed from the initial Keystone research in scale and use of two summative instruments (internal and external) instead of one. It has to be noted that the effect of frequency of formative assessment was not tested in the original Keystone study.

Participants and Instructional Groups

The present study took place during the 2004 - 2006 academic years. The study was conducted in Elements of Algebra classes taught at Richard J. Daley College. Conditions that remained constant throughout all sections of the course were: frequency and duration of interaction with the instructor (110-minute classes, two times per week, for 16 weeks), academic objectives, textbook, content covered, and homework assignments. All students were assessed with common departmental midterm and final exams, and a common COMPASS test. The presence of formative assessment was the only difference between the test and control groups. Frequent formative assessment of learning with immediate feedback, as an independent variable, was used in the test groups.

The study included 1,574 students taught by 25 instructors. The students were of mixed gender, various ages, race, ethnicity, ability, and economic backgrounds. The sample size selected for this experiment was adequate for the type of research conducted (Krejcie & Morgan, 1970). The students were self-assigned to instructional groups upon enrollment. Individuals in control groups were expected to be comparable to those in the test groups.

Table 1. Research Design

Group	Assignment	Number of Students	Experimental Design Model	Correlational Design Model	
			Treatment (independent variable)	Summative Assessment (dependent variable)	Standardized National Exam
I. Test/Keystone-Weekly	Self-assigned	154	Keystone method with weekly cumulative quizzes. As formative assessment	common departmental midterm and final exams as internal summative assessment	COMPASS Test as external summative assessment
II. Test/Keystone-Biweekly	Self-assigned	68	Keystone method with cumulative quizzes twice a week. As formative assessment	common departmental midterm and final exams as internal summative assessment	COMPASS Test as external summative assessment
III. Control/Trained Instructor	Self-assigned	80	Traditional method without formative assessment	common departmental midterm and final exams as internal summative assessment	COMPASS Test as external summative assessment
IV. Control/Non-Trained Instructor	Self-assigned	1272	Traditional method without formative assessment	common departmental midterm and final exams as internal summative assessment	COMPASS Test as external summative assessment

The test group included 10 sections taught by 2 instructors, who received specific training prior to the experiment and utilized the Keystone methodology throughout the course. The control group included 50 sections taught by 23 instructors, who used a traditional method of teaching. The test group was further broken down into two categories according to the frequency of assessment. The "Keystone-weekly" group received quizzes every week while the "Keystone-biweekly" group was quizzed each class period, which amounted to two quizzes per week. In order to ensure that the only independent variable was the Keystone method, in other words, to eliminate personal teaching effect as a variable, it was proposed to further subdivide the control group into two categories: classes taught by instructors trained in the Keystone method but who would not be applying their training to the control sections, and classes taught by instructors without training.

Assessment Instruments

Several instruments, such as frequent cumulative quizzes, common departmental midterm and final exams, and the standardized COMPASS test, were used throughout the current study. Their purpose was to measure the performance of the students in the classes under study. Several standards were also applied to all quizzes, midterms, and final exams. These standards were in the form of written, closed-book tests that disallowed calculator use. The standardized assessment tool in the form of the COMPASS test was in electronic form utilizing the online computer testing system where students were allowed to use any type of calculators.

Cumulative quizzes, as formative assessment instruments, were mandatory only for test groups. Each quiz contained 10 to 15 multiple-choice questions with five options each. Quizzes were administered at the beginning of class periods, were time-restricted, and reflected previously covered material and homework assignments. The quizzes were graded on an absolute scale – no partial credit was given for any item on the quiz. Upon conclusion of each quiz, answer scan-sheets were collected for processing, but the students were allowed to keep the question sheets as a guide for further review and study at home. A brief review of the quiz problems and their answers were provided immediately after each quiz. To help students learn from their mistakes, troublesome questions were repeated on follow-up quizzes until mastery was achieved on the topic. Students from the Keystone-weekly test group were tested once a week and had a total of 13 quizzes during the semester. The Keystone-biweekly test group received 26 quizzes. Control group students were assessed after one or two chapters or had no quizzes at all depending on the instructors' preferences. Two major differences between quizzes offered in test and control groups were cumulative vs. non-cumulative assessment and availability of immediate feedback after each quiz.

All students, regardless of group, were given the same internal summative tests (departmental midterm and final exams). The midterm and final exams were collaboratively developed by instructors from all sections and administered by the coordinator. Midterm exams had 25-multiple choice questions and the final examinations consisted of 50 questions, also multiple choice. To maintain the uniformity of the procedures and integrity of grading, the scoring of individual student performance on the midterm and final exams was done by the coordinator of the course using ParScore software (Scantron Corp., 2008). Feedback in the form of rosters with various types of scoring systems (raw, percent, norm-referenced, and criterion-referenced) and item analysis were provided to each instructor.

In conjunction with internal summative course examinations, all students were required to take the external summative computerized COMPASS test developed by ACT, Inc. (2008). This assessment tool served the additional purpose of finding the correlation between the internal departmental summative assessment instrument results and the external standardized test.

Data Analysis Procedures

The experimental study involved a variety of descriptive and inferential statistics for the analysis of collected data. The descriptive statistics included counts, frequencies, means, variances, and standard deviations of dependent variables. All variables were converted into percent values, which allowed a direct comparison of variables. Both contingency tables and bar graphs were constructed to reveal patterns in the data. Inferential statistics used a *t-test* at selected probability level (95% confidence interval) to reject the null hypothesis and determine whether the difference between group results was significant.

The normalized gain was used to compare the pre- and post-test results collected from the COMPASS examinations. The normalized gain (g) for a treatment is defined as g = actual average gain/maximum possible average gain (Gery, 1972; Hake, 1998; Hovland, Lumsdaine, & Sheffield, 1949). The strength and direction of the linear relationship between two variables are measured by the Pearson correlation coefficient. After the predetermined time, the students' skills were measured and compared to determine which treatment, if either, produced higher skill levels.

Results and Discussion

It is evident that initial preparation of the students enrolled in all sections of Elementary Algebra classes is nearly equal when the scores in the pre-test column of Table

2 are compared. The normalized gain value is calculated using the formula described earlier. Thus, if a class averaged 18 on the pre-test and 36 on the post-test, then the class-average normalized gain (g) = (36 - 18)/(29 - 18) = 18/11 = 1.64. The number "29" in the denominator of the fraction is derived from the table of the COMPASS test cut-off scores.

The data in Table 2 demonstrate that scores in the post-test column have increased from the initial exam for all groups participating in the experiment. This proves that all students gained knowledge. However, as is clearly seen, the results for both test groups are higher than those for the control group. The average score on the post-test for the control group shows that many students did not pass the COMPASS test and did not qualify to proceed to the next level of mathematics upon the completion of a full semester of traditional instructions and assessment. Based on the average scores of the test groups, the majority of the students passed the COMPASS test at the end of the semester and became eligible to enroll in the Intermediate Algebra course.

Table 3 illustrates student performance on internal summative examinations. Both test groups demonstrated higher achievement on internal summative exams than control groups – 70th vs. 60th percentile for the midterm exam and 66th vs. 50th percentile for the final exam. The performance of the test groups was approximately 16% higher at the end of the semester than that of the control groups.

As can be seen from Table 3, both test groups exhibited smaller standard deviations (SD) in the final exam as compared to those of the control groups (13% and 11% compared to 22% and 18%). Standard deviation shows how tightly all of the scores are clustered around the mean in a given set of data. If the standard deviation is small, the bell-shaped curve of normal distribution is steep and relatively few scores have a propensity toward one extreme or the other. In addition to these findings, it is apparent from Table 3 that all groups had lower results on the final internal exam when compared to those on the midterm. This could be explained by the increased amount of material tested at the end of the semester (seven vs. three chapters).

The significance of the difference between the results of summative assessments is investigated using a t-*test.* The degrees of freedom (d*f*) for a *t*-test are determined according to the following formula: $(N_1 + N_2 - 2)$, where N_1 is the number of subjects in group 1 and N_2 is the number of subjects in group 2. The calculated *t*-values were compared with the critical *t*-values found in Gay, Mills, and Airasian (2006) for the $p < 0.05$ level of significance. The critical *t*-value of 1.960 corresponds to the d*f* greater than 120.

The results displayed in Tables 4 and 5 demonstrate a significant difference between the means for external and internal summative examinations when either of the two Test groups is compared to the control/non-trained Instructor group. The difference is found to be not significant between the means of two control groups, which only varied in the training level of the instructors. Consequently, the only factor that had a substantial impact on the obtained results was the method of teaching using formative assessment and not the human teaching factor.

Another important set of data supporting the described findings is the passing rate on the internal summative assessments such as midterm and final departmental exams (see Table 6). Passing rate is defined as the number of students who attained 70% or above on the internal exams compared to the total number of students who participated in these assessments. The midterm results are consistent with the final exam

Table 2. Student Performance on External Summative Assessment Instrument

Group	N	COMPASS Test (passing raw score = 29)		
		Raw Score on Pre-Test	Raw Score on Post-Test	Normalized Gain
Test/Keystone Weekly	65	20.0	33.0	1.44
Test/Keystone Biweekly	58	19.9	34.1	1.56
Control	138	20.3	28.3	0.92

Table 3. Student Performance on Internal Summative Examinations

Group	N	Midterm Exam		Final Exam	
		Mean	SD	Mean	SD
Test/Keystone Weekly	154	70%	13%	66%	13%
Test/Keystone Biweekly	68	66%	12%	59%	11%
Control/Trained Instructor	80	61%	18%	50%	22%
Control/Non-Trained Instructor	1272	57%	18%	49%	18%

Table 4. The t-test Means on the External Summative Examination

Group Comparison	d*f*	*t*-value vs. t_{cr}-value and significance
Test/Keystone Weekly vs. Control/Non-Trained Instructor	306	1.98 > 1.96 significant
Test/Keystone Biweekly vs. Control/Non-Trained Instructor	335	3.67 > 1.96 significant
Test/Keystone Weekly vs. Test/Keystone Biweekly	67	0.422 < 2.00 non-significant

results. As is seen, the classes treated with cumulative frequent quizzes have a higher passing rate than the control groups.

It may be argued that higher performance of the test groups on the summative exams was the result of higher drop rates of students in these classes. To investigate this assumption, we performed retention rate analysis on the test and control groups. A retention rate represents the ratio of the number of students who completed the course to the total enrollment number. Table 7 presents retention rates for students in the test and control groups. As the data clearly shows, the withdrawal rates are 10% higher in the control groups, where the average performance was below that of test groups. Therefore, better performance in the test groups was achieved in conjunction with higher retention rates; so administering frequent formative assessment may improve attendance and enhance retention rates in elementary algebra classes. These findings support the work of Ramaprasad (1983) and Jacobs and Chase (1992).

When the retention rates, passing rates, and the scores acquired using the external and internal summative assessment instruments are examined collectively, it could be concluded that the students from the test groups achieved higher learning outcomes than the students from the control groups (See summarized findings in Table 8). Frequent testing may help the instructor to monitor and guide student performance. The students also tend to become actively involved in their own learning process.

The evidence accumulated during this study is in agreement with Black and William's (1998) and Siadat, et al., (1997, 2000, summer 2000, 2001, & 2008) findings. The first research hypothesis that states that the use of formative assessment in the form of frequent, cumulative, time-restricted, multiple-choice quizzes with the immediate constructive feedback reveals the levels of conceptual understanding in a timely manner and improves student academic performance on

Table 5. The t-test Means on the Internal Summative Examination

Group Comparison	*df*	Midterm Exam		Final Exam	
		t-value	Difference	*t*-value	Difference
Test/Keystone Weekly vs. Control/Non-Trained Instructor	1424	8.70 > 1.96	significant	11.40 > 1.96	significant
Test / Keystone Biweekly vs. Control/Non-Trained Instructor	1338	4.08 > 1.96	significant	4.54 > 1.96	significant
Control/Trained Instructor vs.Control/Non-Trained Instructor	1350	1.93 < 1.96	not significant	0.48 < 1.96	not significant
Test/Keystone Weekly vs. Test/Keystone Biweekly	220	2.17 > 1.96	significant	3.88 > 1.96	significant

Table 6. Student Performance as Measured by the Passing Rate on Internal Summative Exams

Group	N	Midterm Exam 70% or higher	Final Exam 70% or higher
Test/Keystone Weekly	154	55%	36%
Test/Keystone Biweekly	68	56%	33%
Control/Trained Instructor	80	42%	27%
Control/Non-Trained Instructor	1272	27%	17%

Table 7. Student Performance as Measured by Retention Rate

Group	N	Midterm Retention Rate	Final Retention Rate
Test/Keystone Weekly	154	90%	71%
Test/Keystone Biweekly	68	84%	81%
Control/Trained Instructor	80	81%	64%
Control/Non-Trained Instructor	1272	71%	61%

Table 8. Summary of Experimental Results

Group	External Summative Assessment	Internal Final Summative Assessment		
	Normalized Gain	Final Exam Mean	Passing Rate (70% or higher)	Retention Rate
Test/Keystone Weekly	1.44	66%	36%	71%
Test/Keystone Biweekly	1.56	59%	33%	81%
Control/Trained Instructor	-	50%	27%	64%
Control/Non-Trained Instructor	0.92	49%	17%	61%

the summative assessment instruments (See Section I) is strongly supported by provided results.

Furthermore, the data in Tables 2, 3, and 8 demonstrate the difference in student performance between the two test groups: Keystone Weekly and Keystone Biweekly. The normalized gain derived from the external summative assessment instrument for the Keystone Biweekly group is (g) = 1.56, while the gain of the Keystone Weekly group is (g) = 1.44. There is no statistically significant difference between these values (See Table 4). In contrast to the external summative instrument, the internal summative assessment results show that the Keystone Weekly group achieved 4% higher results on the midterm exam than the Keystone Biweekly group (Table 3). This difference increases up to 7% for the final exam (See Tables 3 and 8). The *t*-test (Table 5) illustrates a statistically significant difference for this set of data. In spite of the better gain attained by the Keystone Weekly group, the Keystone Biweekly group maintained a 10% higher retention rate compared to the Keystone Weekly group. This may indicate that better gains of the Keystone Weekly were due, at least in part, to higher attrition of weaker students. The lower final internal exam performance of Biweekly Keystone group could also be attributed to the type of feedback provided to the students. It is possible that the feedback that followed every quiz was not constructive enough to help build a better understanding of algebraic concepts, or that there was a large variability in proportion of previously covered material on each quiz. Nevertheless, the present findings are inconclusive due to a high number of variables that may have influenced the results. On these grounds, the second research hypothesis, which states that more frequent formative assessment provides better results on summative assessment, cannot be accepted and would need further investigation in a better controlled environment.

Table 9. Pearson's Correlation Coefficients for the Test/ Keystone Weekly Group

		Midterm Exam	COMPASS Test as Post-Test	Final Exam
Midterm Exam	Pearson Correlation Coefficient	1	.497(*)	.610(**)
	Sig. (2-tailed)		.022	.003
	N	21	21	21
COMPASS Test as Post-Test	Pearson Correlation Coefficient	.497(*)	1	.607(**)
	Sig. (2-tailed)	.022		.003
	N	21	21	21
Final Exam	Pearson Correlation Coefficient	.610(**)	.607(**)	1
	Sig. (2-tailed)	.003	.003	
	N	21	21	21

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

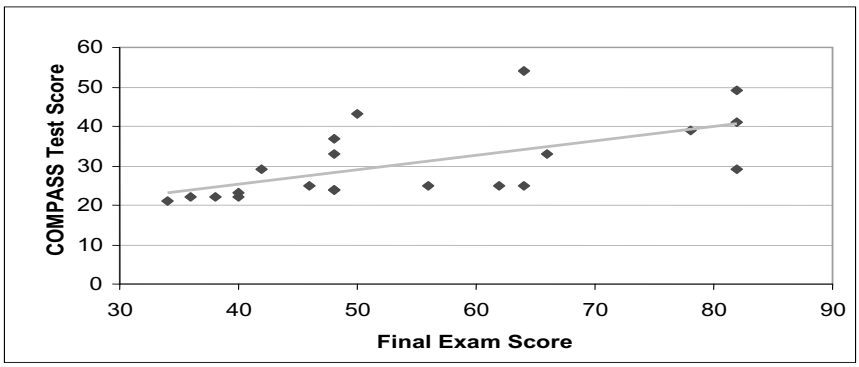

Figure 3. Scatter Plot of the Final Exam and COMPASS Test Scores for the Test/Keystone Weekly Group.

Tables 9, 10, and 11 contain the data gathered for the evaluation of the degree of correlation between the COMPASS post-test and midterm/final exam scores using the linear Pearson's correlation test. The scatter plots in Figures 3, 4 and 5 are included for the visual demonstration of the presence of a possible relationship between the two variables.

Analysis of our experimental data illustrates the existence of a positive linear relationship, as expressed by the correlation coefficient *r*, between the external and internal summative examination scores for all groups. The correlation coefficient may range from 0 to 1, where 0 signifies no linear relationship between the two variables, while 1 means that the two variables are strongly correlated. For this study, the correlation coefficient between COMPASS and final exam for the Test/Keystone Weekly group was r = 0.607, for the Test / Keystone Biweekly was r = 0.561, and for the control/non-trained instructor group was r = 0.578. Results for all three groups are statistically significant. The existence of a significant correlation between the outcomes achieved on external and internal assessments supports the third hypothesis. As such the scores received on departmentally designed summative as-

sessments are a high predictor of the scores on the national standardized COMPASS test, and vice versa. The establishment of a strong positive relationship between the two tests allows a fair judgment of students' skills in algebra, allowing an appropriate placement of students on the continuum of mathematics courses.

Conclusions

The present longitudinal study addresses the issue of the general lack of basic skills in mathematics of the incoming students in the community college setting. It provides a glimpse into the remedial Elements of Algebra (Math 110) course at Daley College in order to contribute to the understanding of how implementation of formative assessment can help students achieve academic success as measured by summative assessment instruments. The experimental research is based on the Keystone method developed by Siadat and Sagher in 1997, but greatly expanded this earlier study by recruiting a significantly larger sample of participants and utilizing a broader spectrum of assessment instruments in the form of the national standardized COMPASS test and departmentally designed exams.

The results of the current investigation indicate that the use of formative assessment, consisting of frequent, cumulative, time-restricted, multiple-choice quizzes with immediate constructive feedback, reveal the level of conceptual understanding in a timely manner. This, in turn, allows for continuous readjustment of teaching and learning strategies leading to an improvement of student academic achievement. In the performed study, the test groups had consistently higher final exam scores and passing and retention rates. Furthermore, the results of internal summative assessments proved to be highly correlated with the external examination scores due to the positive and significant correlation between the two sets of data.

Table 10. Pearson's Correlation Coefficients for the Test/ Keystone Biweekly Group

		Midterm Exam	COMPASS Test as Post-Test	Final Exam
Midterm Exam	Pearson Correlation Coefficient	1	.475(*)	.767(**)
	Sig. (2-tailed)		.000	.000
	N	50	50	50
COMPASS Test as Post-Test	Pearson Correlation Coefficient	.475(*)	1	.561(**)
	Sig. (2-tailed)	.000		.000
	N	50	50	50
Final Exam	Pearson Correlation Coefficient	.767(**)	.561(**)	1
	Sig. (2-tailed)	.000	.000	
	N	50	50	50

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

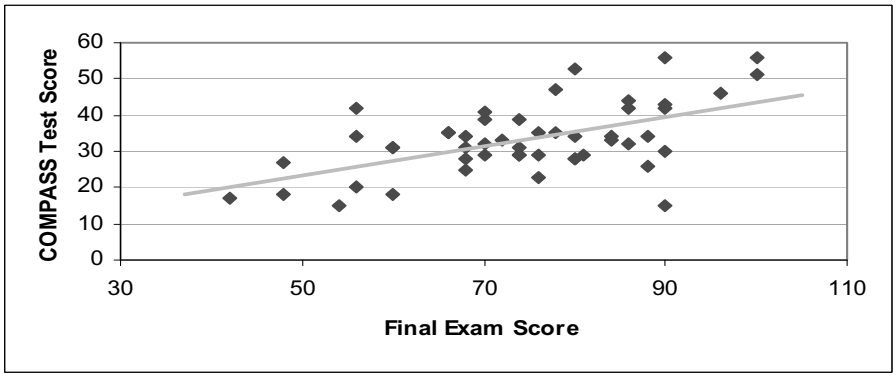

Figure 4. Scatter Plot of the Final Exam and COMPASS Test Scores for the Test/Keystone Biweekly Group

Table 11. Linear Correlation Test for the Control/Non-Trained Instructor Group

		Midterm Exam	COMPASS Test as Post-Test	Final Exam
Midterm Exam	Pearson Correlation Coefficient	1	.567(**)	.576(**)
	Sig. (2-tailed)		.000	.000
	N	287	287	287
COMPASS Test as Post-Test	Pearson Correlation Coefficient	.567(**)	1	.578(**)
	Sig. (2-tailed)	.000		.000
	N	287	287	287
Final Exam	Pearson Correlation Coefficient	.576(**)	.578(**)	1
	Sig. (2-tailed)	.000	.000	
	N	287	287	287

Correlation is significant at the 0.01 level (2-tailed).

However, an increased frequency of formative assessment did not produce a significant improvement in students' learning outcomes. This result was unexpected, and may be contrary to the findings in literature. Lack of constructive feedback and strong variability in proportion to previously covered material in new quizzes may have been the main causes of the findings. Additional research should be undertaken to accumulate more data so as to provide a definitive conclusion about the impact of more frequent quizzing on the final exam results. It is suggested a follow-up cohort study be conducted on those students who participated in the current study, in order to examine a long-term vs. short-term effect of frequent quizzes on student retention of basic algebra knowledge, and to track their success in the higher level math classes.

The findings of this study are consistent with previous educational research and Theories of Learning, Vygotsky's "Zone of Proximal Development" and Bandura's "Social-Cognitive Learning Theory," (Vygotsky, 1978; Bandura, 1986). Formative assessment seems to be the most promising classroom assessment technique with both cognitive and psychological benefits for students. This study has confirmed that interactive formative assessment is effective in raising the achievement level of participating students, and that the combination of formative and summative assessment instruments in elementary algebra enhances student success. Due to the relatively large sample size of the current study, the results could be easily extended to the general student population. It is reasonable to expect that similar improvements could be achieved in any remedial course and in any setting; therefore, the findings may have important implications for faculty at any college involved in teaching remedial courses.

Recommendations

Traditional teaching of remedial/developmental mathematics, especially in view of an ever increasing influx of unprepared students into college mathematics classes across the nation, has failed to meet the present day educational challenges. The present study has shown that a unique combination of formative assessment techniques, along with readjustment of teaching practices to develop students' mastery, can substantially improve performance and retention of students in basic algebra classes. Implementation of this approach does not require a major allocation of financial and material resources by the college. Computer application support to develop a test bank and perform statistical analysis, and a commitment of the teacher to engage in frequent assessment of learning and instruction, is all that is needed.

The best reward for the teachers is experiencing their students achieving success. Success of students is a motivating factor for both students and their instructors. It encourages students to work harder to achieve more and it inculcates enthusiasm in teachers to continue with their innovative ways. Investment in technological support of instruction and providing training in the use of dynamical assessment of teaching and learning, as presented in our study, will bear great dividends when student performance and retention is improved. Colleges and universities struggling with remedial education need to revamp these programs in mathematics if they are to stem the rampant failure and attrition rates that currently exist on their campuses. This research study presented an alternative with an established track record. It is hoped that our educator colleagues and peer institutions seriously consider its implementation.

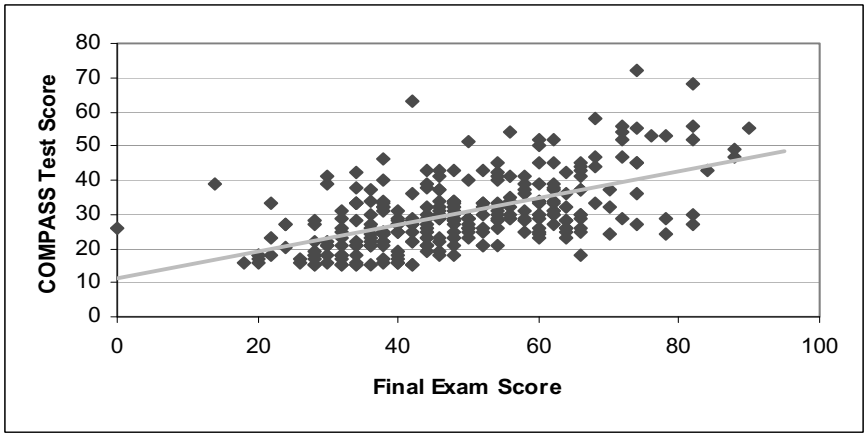

Figure 5. Scatter Plot of the Final Exam and COMPASS Test Scores for the Control/Non-Trained

References

ACT, Inc. (2008). College Placement Test retrieved December 26, 2008, from COMPASS http://www.act.org/compass/

Angelo, T., & Cross, P. (2001). *Classroom assessment techniques*. San Francisco: Jossey-Bass.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice-Hall.

Bangert-Drowns, R. L., Kulick, J. A., & Morgan, M. T. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research*, *61*(2), 213-238.

Black, P., & William, D. (1998). Assessment and classroom learning. *Assessment in education*, *5*(1), 7-74.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on the formative and summative evaluation of student learning*. New York, NY: McGraw-Hill.

Crooks, T. J. (1988). The impact of classroom evaluation practices in students. *Review of Educational Research, 58(4)*, 438-481.

Dineen, P., Taylor, J., & Stephens, L. (1989). The effect of testing frequency upon the achievement of students in high school mathematics courses. *School Science and Mathematics*, *89*(3), 197-200.

Formative assessment improving learning in secondary classrooms (2005), Paris: OECD Publishing.

Gay, L. R., Mills, G. E., & Airasian, P. (2006). *Educational research: Competencies for analysis and application*, Columbus, Ohio: Pearson/Merrill Prentice Hall.

Gery, F. W. (1972) "Does mathematics matter?" In Research Paper in Economic Education, edited by A. Welch. New York: *Joint Council on Economic Education*, pp. 142-157.

Grover, C. A., Becker, A. H., & Davis, S. F. (1989). Chapters and units: Frequent versus infrequent testing revisited. *Teaching of Psychology*, *16*(4), 192-194.

Hake, R. R. (1998). Interactive-engagement vs. traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *Am. J. Phys*., *66(1)*, 64-74.

Hawthorne, E. M. (1987). *Evaluating employee training programs: A research-based guide for human resources managers*, New York, Westport, and London: Quorum Books.

Hovland, C. I., Lumsdaine, A and Sheffield, F. (1949). *Experiments on mass communication*. Princeton, NJ: Princeton University Press.

Jacobs, L., & Chase, C. (1992). Developing and using tests effectively. San Francisco: Jossey-Bass.

Krejcie, R.V., & Morgan, D. W. (1970). Determining sample size for research activities. *Educational and Psychological Measurement*, *30(3)*, 608.

Mohr L. B. (1995). *Impact analysis for program evaluation (2nd Ed.)*. Thousand Oaks, London, New Delhi: SAGE Publication, International Educational and Professional Publisher.

National Center for Education Statistics. Remedial education at degree-granting postsecondary institutions in Fall 2000. Washington, DC: U.S. Department of Education, Institute of Education Sciences (NCES 2004-010).

Natriello, G. (1987). The impact of evaluation processes on students. *Educational Psychologist*, *22(2)*, 155- 175.

Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science*, *28*(1), 4-13.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*(2), 119-144.

Scantron Corp.(2008). Products:Parscore. Retrieved December 26, 2008, from http://www.scantron.com/downloads/ParSCORE.pdf

Scriven, M. (1967). *The Methodology of evaluation*. Washington, DC: American Educational Research Association.

Siadat, M. V., & Sagher, Y. (1997). Building study and work skills in a college mathematics classroom. (*ERIC Document Reproduction Service.* ED 449834).

Siadat, M. V., Sagher, Y., & Hagedorn, L. (2000). Building study skills in a college mathematics classroom. *The Journal of General Education, 49*(2), 132-155.

Siadat, M. V., Musial, P. M., & Sagher, Y. (Summer 2000). Mathematics: Keystone to student learning. *Academic Exchange Quarterly, 4*(2), 15-22.

Siadat, M. V., Sagher, Y., Hagedorn, L., & Musial, P. M. (2001). *Sound instruction: Ready to use classroom practice*. California: Rapid Intellect Group.

Siadat, M. V., Musial, P. M., & Sagher, Y. (2008). Keystone method: A learning paradigm in mathematics. *Problems, Resources, and Issues in Mathematics Undergraduate Studies (PRIMUS), 18(4*), 337-348.

Stingler, J. W., & Hiebert, J. (1999). *The teaching gap: Best ideas from the world's teachers for improving education in the classroom.* New York, NY: Free Press.

Vygotsky, L. S. (1978). *Mind and society: The development of higher mental processes.* Cambridge, MA: Harvard University Press.

Whitaker, U. (1989). *Assessing learning: Standards, principles, and procedures*. Philadelphia, PA: Council for Adult and Experiential Learning.

Willingham, D. T. (2004). Practice makes perfect – But only if you practice beyond the point of perfection, *American Educator*, 28 (1) 31-39.

Euguenia Peterson, *Ph.D. is a professor in the Department of Mathematics at Richard J. Daley College. She can be contacted at 7500 S. Pulaski Rd., Chicago, IL 60652, (773) 838-7718, or epeterson@ccc.edu.*

M. Vali Siadat, *Ph.D., D.A., is a professor in the Department of Mathematics at Richard J. Daley College. He can be contacted at 7500 S. Pulaski Rd., Chicago, IL 60652, (773) 838- 7658, or vsiadat@ccc.edu.*