

## **The Relationship Between Teacher Assessments And NCLB Mathematics Testing**

Christopher Mark Herte, High School Supervisor of Mathematics, West Windsor-Plainsboro Regional Schools

### **Abstract**

This study examined the relationship between teacher assessments and grade eight New Jersey NCLB mathematics testing. It also examined how curricular changes affected student achievement for one New Jersey junior high school on the NCLB mathematics test.

The research considered teacher assessments and the New Jersey grade eight NCLB mathematics test for the 2003 and 2005 administrations. End of marking period grades and the midterm exam grades, expressed as percents, for one of the two lowest tracked math courses were collected and analyzed with the 2003 and 2005 NCLB test scores ( $n > 200$  each year). There is a need to determine how curricular changes affect in the relationship between teacher assessments and NCLB test scores.

There was little relationship between teacher assessments and the components of the 2003 NCLB math test. Two years later, after the curricular changes, the relationship between teacher assessment and NCLB testing increased and the percent of students who demonstrated proficiency on the NCLB test increased. The increases in NCLB testing were statistically significant.

This paper reviewed the methodology used, the findings of the study, and how the results may impact similar school districts. Suggestions for further research and action are presented.

### **Introduction**

Educational leaders have always been faced with the challenging task of increasing student achievement. However, now educational leaders must demonstrate increasing student achievement as measured on state assessments under the No Child Left Behind act of 2002 (NCLB).

The No Child Left Behind (NCLB) legislation was signed into law on January 8, 2002 by President George W. Bush initiating sweeping changes in the accountability of school districts to ensure that all children meet high academic standards. The accountability system requires that all public school districts administer tests in English language arts and mathematics in grades three through eight and one year in high school. State agencies were required to set three-year benchmark scores culminating with 100 percent of all children in the aggregate as well as the subgroups to pass state testing in grades three through eight and high school by the year 2014.

## Forum on Public Policy

Meeting these benchmarks in the aggregate as well as in each subgroup is the challenge of every public school.

In order to hold schools and districts accountable for the educational progress of their children, NCLB established serious sanctions for districts that continually fail to meet state established benchmarks for all children. Schools are required to meet each benchmark referred to as adequate yearly progress (AYP) each year for each subgroup and the entire school. Subgroups that are considered are ethnicity, special education, and English language learners. Sanctions that result from schools not meeting AYP range in severity from offering inter-district school choice to district restructuring for schools that continually fail to meet AYP.

The New Jersey Department of Education annually develops and administers the Grade Eight Proficiency Assessment (GEPA). This assessment consists of multiple choice, short response, and extended response items. The purpose of this assessment is, in part, for early identification of students that need remediation as well as how well the students, and school are meeting the state standards in mathematics for grade eight (NJDOE, 2006a). The GEPA is a secure test that may not be reproduced, distributed, or discussed by educators and students.

The difficulty for educational leaders is that these scores may be insufficient to determine specific areas for program improvement necessary to increase student achievement. It is therefore important to determine the relationship between Teacher Assessment and the GEPA. Following the 2003 administration of the GEPA, a study was conducted to determine this relationship for one school district and each of the five ability level courses. This study found that the relationship between teacher assessment and the GEPA was very weak (Herte, 2005). As a finding of the study of the 2003 administration of the GEPA a recommendation for further research was to conduct the study again after curriculum alignment with the state standards as

well as new textbook adoptions.

### *Discussion of Previous Study*

Herte (2005) reported on relationship of teacher assessments to the GEPA for the GEPA administered in 2003. Specifically, Herte reported that the relationship of teacher assessments in Algebra Part 1 could only explain 6.6 percent of the variance in number sense, 8.4 percent of the variance in spatial sense, 6.3 percent of the variance in data analysis, 5 percent of the variance in patterns and functions, and 10.9 percent of the variance in the GEPA scale score. This very weak relationship indicated that the content of Teacher Assessments in Algebra Part 1 needed to be aligned to the GEPA.

The Algebra Part 1 students earned grades that were above passing on teacher assessments when they could not achieve proficiency on the GEPA. Parents viewing only these higher report card grades could believe their children are achieving much higher than they actually are against the state standards. Aligning the teacher assessments both in content as well as expectation is necessary to obtain a clear picture of student achievement.

Herte (2005) recommended that similar districts should take several steps including curriculum alignment with the state standards, use of formative assessments to guide instruction, and to infuse non-algebraic topics in algebraic courses.

## **Theoretical Perspectives**

### *Teacher Assessment*

The NCTM (1997) suggested changes in the concept of program evaluation based upon assessment data. They included moving, “Toward detailed analyses of group data (e.g., examining variations in responses, and the disaggregation of data) and away from reporting only group means” (p. 67).

## Forum on Public Policy

In a study of elementary and secondary mathematics and English teachers, McMillan and Nash (2000) found that there was variability among grading and assessment practices. Through interviews conducted and responses coded, McMillan and Nash (2000) found six themes regarding how teachers decide to use specific assessment and grading practices. These themes were 1) teacher beliefs and values, 2) classroom realities, 3) external factors, 4) teacher decision making rationale, 5) assessment practices, and 6) grading practices. The model they identified described teachers' need for flexible assessment and grading practices so that they could individually accommodate each student. McMillan and Nash reported that external pressure of recent mandated statewide testing caused teachers to increase the role of objective assessments and grading practices into their repertoire. Teachers reported that they made their assessments more aligned to the statewide assessment format. This enabled students to be more comfortable with the format on these statewide assessments.

The ability level class could also have an impact on student grades. McMillan and Nash (2000) noted, "The reality of poor student attitudes and inappropriate behavior, especially in remedial and standard classes, seemed to have both a direct and indirect impact on students' grades" (p. 13). One teacher's comments about the possibility of the students passing the statewide assessment during their interview were reported by McMillan and Nash (2000), "Given these remedial classes, there's no way they will pass it (SOL test), unless they make a total 900% change in their attitude and in their behavior, they're not" (p. 14).

McMillan and Nash (2000) observed that teachers reported their belief that their assessment of student achievement and assigning grades gave a better understanding of the depth of student knowledge. This was predicated on the basis that they used multiple assessments and of different variety so that students could demonstrate their achievement in various ways. The

teachers, to guide instruction, used formative assessments extensively. The nature of formative assessments gave teachers necessary feedback as to when students had mastered concepts or when more or different instruction was necessary.

Boston (2002) examined the role of formative assessments and their diagnostic use to teachers and students in providing feedback. Boston (2002) stated, “Teachers can build in many opportunities to assess how students are learning and then use this information to make beneficial changes in instruction” (p. 1). Noting the importance of formative assessment Boston (2002) stated, “While state tests provide a snapshot of a student’s performance on a given day under test conditions, formative assessment allows teachers to monitor and guide students’ performance over time in multiple problem-solving situations” (p. 1). This sentiment was mirrored by Guskey (2003, February), who stated, “Teachers who develop useful assessments, provide corrective instruction, and give students second chances to demonstrate success can improve their instruction and help students learn” (p. 7).

Guskey (2003, February) pointed out that the best assessments are the tests, quizzes, and assignments that teachers give on a regular basis. Teachers trust these assessments since they were developed by the teacher in order to address the curriculum. Additionally, the results are available to the teacher in order to alter the classroom instruction. Guskey (2003, February) pointed out that many teachers have not received instruction in creating assessments and as such may, “construct their own in a haphazard fashion” (p. 7). Teachers should test what they teach, rather than teaching to the test. The need for corrective action following student assessment is critical. Guskey (2003, February) stated, “assessments must be followed by high-quality, corrective instruction designed to remedy whatever learning errors the assessment identified,” (p. 9).

## Forum on Public Policy

Diagnostic assessment of student learning will help ensure success for all students. Gandal and McGiffert (2003, February) stated, “Just as medical tests help diagnose and treat patients, rigorous and meaningful education assessments can help ensure the academic health of all students” (p. 39). The limitation of using only teacher assessment of students’ achievement doesn’t allow a healthy check of how students are meeting state standards. The teacher assigned grade in one school can mean something drastically different in another. Yet with this limitation, instruction focusing on students’ weaknesses can improve student achievement. Gandal and McGiffert acknowledged the importance of statewide testing but also recommend the use of teacher assessments to improve student achievement and improve instruction. Districts must begin to use ongoing assessments that give immediate feedback to the teachers and students. This will enable teachers to address students’ weaknesses.

### *NCLB Security Issues*

Policies that require exams to be secure or closed and not have items disclosed to the public can have negative results if insufficient data are reported. While these policies may be necessary for testing reliability where test items are used in subsequent test administrations, the ability to improve student achievement is reduced to the reporting results that are available to school districts and parents. There is a need to identify specific individual student weaknesses in order to improve individual student achievement and to target topics or skills that need additional or alternate instruction. When educational professionals only have several numbers to describe how a student or students achieve it is very difficult to target instruction. Additionally, there have been difficulties with state testing that have yielded dramatic results.

Bowman (2003, November 19) reported that Florida’s 1<sup>st</sup> District Court of Appeals ruled in 2003 that a father did not have the right to view the graduation test his son had repeatedly

## Forum on Public Policy

failed. The father was only allowed to view his son's score on the Florida Comprehensive Assessment Test (FCAT). The father, Steven O. Cooper, sued the State of Florida in 2001 after being denied access to his son's test booklet and answer sheets. Florida education officials argued that creating a new test each year would be prohibitively expensive. Bowman (2003, November 19) reported that Governor Jeb Bush praised the Court of Appeals decision saying, "The Florida Comprehensive Assessment Test has been a catalyst for student achievement, and today's decision allows us to maintain meaningful standards, while giving parents and educators the ability to monitor student gains" (p. 5).

Blair (2004, January 14) reported that in Chicago, the U.S. Court of Appeals confirmed the ruling of the lower court that a teacher and a publication editor did not have the right to publish several social studies and English tests that the Chicago school system had been piloting for three years. The publication editor argued that the public had the right to determine whether the exams were appropriate for the students. Judge Richard Posner wrote in the 3-0 decision that the teacher and editor had a right to be critical of the tests but not to publish the tests in their entirety (p. 5).

New Jersey has a secure test policy where teachers and parents may not view the three state administered tests: High School Proficiency Assessment (HSPA), the GEPA, and the NJ Assessment of Student Knowledge (NJ ASK). The NJDOE (2005b) warned, "Examiners, proctors, and other school personnel are NOT to look at, discuss, or disclose any test items before, during, or after the test administration" (p. 2). The NJDOE (2006b) further warned, "Security breaches may have financial consequences for the district, professional consequences for staff, and disciplinary consequences for students" (p. 8). The NJDOE (2006b) explained the

## Forum on Public Policy

reason for this as some of the items on the assessment will reappear in subsequent administrations and it is necessary to maintain the stability of the test.

### *GEPA Reporting*

The NJDOE supplied the following reports: *Individual Student*, *Summary of School Performance*, *School Performance by Demographic Groups*, *School Student Rosters*, and *Summary of District Performance*. The NJDOE (2005b) reported the number of points each student received on each of the four core content clusters, knowledge, problem solving, and the scale score used to determine the proficiency level. The NJDOE (2005b) reported, “Cluster Data: Cluster data are provided to help identify students’ strengths and weaknesses” (p. 17). The NJDOE (n.d.a) lists the skills and concepts that comprise each of the four content clusters.

The cluster, Number and Numerical Operations, consists of at least 15 separate skills or concepts. The cluster, Geometry and Measurement, consists of at least 19 separate skills or concepts, several of which have numerous components. The cluster, Patterns and Algebra, consists of at least 10 skills and concepts. The cluster, Data Analysis, Probability, and Discrete Mathematics, consists of at least 14 skills or concepts.

Reporting individual and group scores are essential for the improvement of curriculum and instruction to meet the challenging standards set by NCLB. The GEPA reporting consists of a scale score and six sub-scores. Four sub-scores measure the four mathematics content clusters. The remaining two sub-scores are knowledge, which is the sum of the total points earned on the four previously mentioned sub-scores, and problem solving skills.

### *Assessment Errors*

Errors in scoring can have great implications for states, districts, schools and students. The United States General Accounting Office (2002) noted that errors had been detected in

## Forum on Public Policy

contractor scoring by local district officials, parents, and individuals at state agencies. “Based on erroneous scores calculated by a contractor, one state sent thousands of children to summer school in the mistaken belief that their performance was poor enough to meet the criterion for summer intervention” (p. 16). The U.S. General Accounting Office (2002) also noted, “based on a contractor’s erroneous scoring, a state incorrectly identified several schools as ‘in need of improvement,’ a designation that carries with it both bad publicity and extra expense” (p. 16).

In June 2003, the New York State Department of Education had difficulties with the Math A test, which is a graduation requirement. Based upon a survey by the State Department of Education, only 37 percent of the students taking the exam passed it. Richard Mills, Commissioner of Education for New York State, was quoted in Dillon (2003, June 25), “I think we made some mistakes with this exam, and it’s up to us to identify and correct them” (p. B4). Due to the immediacy of graduations, Seniors were exempted from passing the test. Mills established an independent panel to review the exam and analyze its results with the charge to make specific recommendations. Some of the recommendations included: to revise the mathematics standards making them clearer and easier for teachers to apply; to produce a suggested scope and sequence K-12 curriculum; and to establish a new Math A exam. By the end of August a new scoring chart for the June 2003 exam was created ensuring an increase in most students’ scores.

The U.S. General Accounting Office (2002) recommended to the Secretary of Education, Rod Paige, that,

Assessment results are a key part of the mechanism for holding both schools and states accountable for improving educational performance. Thus, ensuring the completeness and accuracy of assessment data is central to measuring students’ progress and ensuring accountability. Without adequate oversight of assessment scoring, efforts to identify and improve low-performing schools could be hindered

by lack of confidence in assessment results or uncertainty regarding whether particular schools have been appropriately identified for improvement. (p. 19)

*Assessment Irregularities*

Popham (2006, April 19) pointed out that with the increased stakes of state testing under NCLB some educators are more likely to resort to testing infractions to demonstrate improved test scores. Hurst (2004, October 6) reported that the number of testing irregularities in Nevada's public schools has increased by more than 50 percent from 2002-03 school year to the 2003-04. The majority of the 121 incidents occurred at the secondary level where students have access to technology such as cell phones that take pictures and can text message. Hurst (2004, October 6) further reported that in answering this finding Keith Rheault, Nevada state superintendent, maintained that with the increased demand for schools to meet adequate yearly progress puts more pressure on teachers and students. In Austin, Texas the district pled No-Contest and paid a fine of \$5,000, where district administrators were alleged to have manipulated state testing data as reported by Keller (2002, January 16). Hoff (2003, November 5) cited that 21 teachers were caught cheating from 1998 through mid-2002. Hoff (2003, November 5) also reported that Robert Schaeffer, director of the Center for Fair & Open Testing, maintained that one could predict that some teachers and students will resort to cheating when the pressure to perform is increased. Manzo (2005, January 19) quoted Walter M. Haney, professor of education at Boston College, "Even if there's not outright fraud, where people become so obsessed with raising test scores on one relatively narrow test, cheating and other improprieties are likely to occur" (p. 14).

Assessment of students occurs for various purposes. The main purpose of assessment is to improve student achievement. Additional purposes are to improve instruction, to alter instruction (as in formative assessment) and to determine the mastery of content and skills by

## Forum on Public Policy

students. It is important that the assessments give an accurate measure of student achievement and be reliable from one administration to another. It is necessary to have security measures in place and uniform testing conditions in order to have results that will be meaningful.

### *Curricular Changes*

The district in this study took several actions following the results of the 2003 GEPA administration. The district provided summer staff workshops for teachers to rewrite assessments and to examine the role of non-algebraic topics in algebraic courses during the summers of 2003 and 2004. During the fall of 2003, the district established a committee of teachers and administrators to examine new textbooks for all grade eight math courses. Following an analysis of the NJ state standards and textbooks, two textbooks were selected and piloted in two classes for approximately two months. The committee reconvened and selected one of these texts, *Algebra 1*, authored by Larson, Boswell, Kanold, and Stiff (2004) to be used in the three ability level math courses: Algebra Part 1, Algebra 1, and Algebra 1 Honors. These new textbooks replaced the ten-year old textbooks formerly used for these grade eight mathematics courses.

Teachers using these new textbooks received one day of initial training from the publishing company. Students in Algebra Part 1 began using these new textbooks in September, 2004. During the summer of 2004, teachers and administrators met and wrote the curriculum guide to be used by all teachers teaching Algebra Part 1. During two staff development days in November, 2004 teachers met to discuss the progress of the new materials and curriculum and developed a common midterm exam and other assessments. Teachers met monthly for departmental meetings as well as informally in planning sessions.

## **Method**

### *Setting*

The setting for this study was a suburban public school district with an enrollment of over 9000 students from grades Kindergarten through 12<sup>th</sup> grade located in central New Jersey. The district has eight elementary schools comprised of grades Kindergarten through fifth grade, a middle school comprised of grades six and seven, a junior high school consisting of eighth and ninth grade students and a high school with students in grades ten through twelve. The ethnicity of the district is 68 percent white, 24 percent Asian, 3 percent Hispanic, 3 percent African American, and 2 percent Other. The socioeconomic status of the community is primarily middle and upper middle class with many residents commuting to New York City for employment. The total cost per pupil, including transportation, was \$11,073 during the 2002-2003 school year and \$12,021 during the 2004-2005 school year. The New Jersey state average was \$11,646 and \$12,567 during the same school years.

### *Research Question*

Following curricular changes, what is the relationship of teacher assessments to the New Jersey NCLB Grade Eight Proficiency Assessment (GEPA)?

### *Independent Variables*

Teacher Assessments in mathematics consisted of four components: 1) First marking period grade; 2) Second marking period grade; 3) Third marking period grade; and 4) Midterm exam. The marking period grade is defined as the weighted average assigned to a student by the student's teacher during a consecutive ten-week period. These three variables are first marking period grade, second marking period grade, and third marking period grade. Students in the same course were administered common assessments as part of the departmental practice. Each

## Forum on Public Policy

marking period grade was based upon the math department grading policy of the school district consisting of a weighted average of 50 percent major assessments (tests), 25 percent minor assessments (quizzes), and 25 percent performance assessments (homework completion and class participation). Each marking period grade was calculated automatically using Intergrade software that resulted in a numerical percentage from 0-100.

The Midterm exam was the percentage correct that a student answered on a common criterion referenced test created by all teachers instructing students in Algebra Part 1 within the school district. The same midterm exam was administered to all students in Algebra Part 1, as is the departmental practice of the school district.

### *Dependent Variables*

GEPA consists of five subscales: 1) Number sense; 2) Spatial sense; 3) Data analysis; 4) Patterns and functions and 5) GEPA knowledge.

Number sense, based upon the New Jersey Core Curriculum Content Standard 4.1 (Number and Numerical Operations), was the GEPA subscale that measured the numerical skills of grade eight students. The range of scores for this scale was 0-12 and was converted to a percentage based on a total of 12 points (NJDOE, 2005b).

Spatial sense, based upon New Jersey Core Curriculum Content Standard 4.2 (Geometry and Measurement), was the GEPA subscale that measured the spatial and measurement skills of grade eight students. The range of scores for this scale was 0-12 and was converted to a percentage based on a total of 12 points (NJDOE, 2005b).

Patterns and functions, based upon New Jersey Core Curriculum Content Standard 4.3 (Patterns, Functions, and Algebra), was the GEPA subscale that measured the algebraic skills of grade eight students. The range of scores for this scale was 0-12 and was converted to a

percentage based on a total of 12 points (NJDOE, 2005b).

Data analysis, based upon New Jersey Core Curriculum Content Standard 4.4 (Data Analysis, Probability, Statistics, and Discrete Mathematics), was the GEPA subscale that measured the data analysis, probability, statistics, and discrete mathematics skills of grade eight students. The range of scores for this scale was 0-12 and was converted to a percentage based on a total of 12 points (NJDOE, 2005b).

GEPA knowledge was the sum of the four sub-scores 1) Number sense 2) Spatial sense 3) Patterns & Functions and 4) Data analysis. The range of scores for this scale was 0-48 points and is directly converted to the GEPA Scale Score (NJDOE, 2005b).

GEPA scale score measured how prepared the student was toward the New Jersey Core Curriculum Content Standards (NJDOE, 2003b). The New Jersey Department of Education has identified levels of proficiency. The range of scores for this scale was 150 to 300. Students with scores within the range of 150-199 are considered “partially proficient.” Students with scores within the range of 200-249 are considered “proficient”, while students with scores within the range of 250-300 are considered “advanced proficient” (NJDOE, 2003c).

### *Selection of Subjects*

This study is limited to one school district and the students who are assigned to the mathematics course Algebra Part 1 during the 2002-2003 school year (2003 cohort) and the 2004-2005 school year (2005 cohort). The ability level mathematics course, Algebra Part 1, was selected due the number of students enrolled as well as having the highest number of students failing to demonstrate proficiency on the GEPA. The total number of students at this school taking the GEPA was 778 in 2003, and 723 in 2005. The number of students that were enrolled in Algebra Part 1 for the first three marking periods and who took the GEPA was 254 in 2003,

and 218 in 2005.

*Procedure*

Approval for the access and use of student data was obtained in writing from the superintendent of the school district prior to collecting any data. The request for use of the student data outlined the purpose of the study and how the analysis would be reported. The results and analysis were made available to the school district for its curricular purposes.

Teacher assessment data were collected electronically from the district's database. These data included the first marking period grade, second marking period grade, third marking period grade, and midterm exam. These data were pared to the student database, which included demographic data using Microsoft Excel. The GEPA scale score and sub-scores were manually entered into the Excel spreadsheet with the student name and teacher assessment for each student. These data were exported to the Statistical Package for the Social Sciences (SPSS) for analysis.

The present study was conducted to determine what the relationship was between the teacher assessment and the GEPA for two cohorts of eighth grade students during the 2002-2003 and 2004-2005 school years. The relationships between teacher assessments to each component of the GEPA were examined separately by calculating the variance for each contributing variable using stepwise multiple regressions.

**Results**

The relationship between teacher assessments and the GEPA was stronger for the 2005 administration than the 2003. For the GEPA component, number sense, the relationship went from  $R^2 = .066$  in 2003 to  $R^2 = .164$  in 2005, as illustrated in Table 2. Using step-wise

regression, the midterm exam was the only component for teacher assessment used in both the 2003 and 2005 regression models.

For the GEPA component, spatial sense, the relationship went from  $R^2 = .084$  in 2003 to  $R^2 = .221$  in 2005. Using step-wise regression, the midterm exam was the only component for teacher assessment used in the 2003 regression model. The 2005 regression model used the third marking period grade and then the midterm exam. This indicated that the third marking period grade in 2005 had a stronger relationship to spatial sense than in 2003.

For the GEPA component, patterns and functions, the relationship went from  $R^2 = .050$  in 2003 to  $R^2 = .200$  in 2005. Using step-wise regression, the midterm exam was the only component for teacher assessment used in the 2003 regression model. The 2005 regression model used the third marking period grade and then the midterm exam. This indicated that the third marking period grade in 2005 had a stronger relationship to patterns and functions than in 2003.

For the GEPA component, data analysis, the relationship went from  $R^2 = .063$  in 2003 to  $R^2 = .278$  in 2005. Using step-wise regression, the midterm exam was the only component for teacher assessment used in the 2003 regression model. The 2005 regression model used the midterm exam and then the third marking period grade. This indicated that the third marking period grade in 2005 had a stronger relationship to data analysis than in 2003.

GEPA knowledge also had a stronger relationship where the relationship went from  $R^2 = .109$  in 2003 to  $R^2 = .336$  in 2005. Using step-wise regression, the midterm exam was the only component for teacher assessment used in the 2003 regression model. The 2005 regression model used the third marking period grade and then the midterm exam. As with the previous three variables, this indicated that the third marking period grade in 2005 had a stronger

relationship to GEPA knowledge than in 2003.

Since there were stronger relationships between teacher assessments and GEPA it was important to determine whether there were significant changes in the components of each. Table 1 illustrates the means for teacher assessments and GEPA for 2003 and the 2005 cohort. There was a decrease for the first marking period grade,  $M = 81.278$  in 2003 to  $M = 77.807$  with  $SD = 10.011$  and  $SD = 10.099$  for 2003 and 2005, respectively. Each component of teacher assessment had a lower mean in 2005 than in 2003. Conversely there were increases in the means for each component of GEPA from 2003 to 2005. Number sense had the greatest increase of almost 17 percentage points with  $M = 44.59$ ,  $SD = 21.45$  in 2003 to  $M = 61.58$ ,  $SD = 19.16$  in 2005. Based on the changes from 2003 to 2005 it was necessary to determine whether these changes were significant.

A random sample consisting of 35 students from the 2003 cohort and 35 students from the 2005 cohort were selected. The null hypothesis was that there were no statistically significant differences between the scores from the 2003 and 2005 cohorts. Independent samples  $t$  tests were conducted with the null hypothesis tested at the  $p < .05$  level. The differences between the 2003 and 2005 components of teacher assessments were not statistically significant with  $p$  values of  $p = 3.18$  for the first marking period grade,  $p = .655$  for the second marking period grade,  $p = .691$  for the midterm exam, and  $p = .120$  for the third marking period grade.

Most of the differences between the 2003 and 2005 components of the GEPA were statistically significant at the  $p < .05$  level. The differences in number sense, spatial sense, data analysis, and GEPA knowledge from 2003 to 2005 were all significant. With  $p$  values less than .05 the null hypothesis was rejected. These  $p$  values were  $p < .0005$  for number sense,  $p = .009$  for spatial sense,  $p < .0005$  for data analysis, and  $p = .001$  for GEPA knowledge. The  $p$  value for

patterns and functions was  $p = 3.26$  and therefore, the null hypothesis could not be rejected. Based on these findings, there were significant increases from 2003 to 2005 for most components of GEPA.

Were significant increases in GEPA components reflected in the proficiency of the Algebra Part 1 students on the GEPA? In 2003, of the 254 students in Algebra Part 1, 52.8 percent of the students were identified as proficient or advanced proficient indicating that these students should not need remedial instruction. In 2005 the percentage increased. Of the 218 students in Algebra Part 1, in 2005, 68.3 percent were identified as being proficient or advanced proficient as illustrated in Figure 2.

Further analysis was conducted to determine whether this increase in the percent of students scoring proficient or advanced proficient was significant. Random samples of 35 students were selected from both the 2003 cohort and the 2005 cohort. The value of 0 was entered in the variable, proficiency level, for students who scored partially proficient and the value of 1 was entered for students who scored either proficient or advanced proficient. The Mann-Whitney nonparametric test was conducted on these data with the  $p$  value set to  $p < .05$  necessary to determine significance. The mean rank was 30 and the sum of the ranks was 1050 for the 2003 GEPA. The mean rank was 41 and the sum of the ranks was 1435 for the 2005 GEPA. Based on this, the percent of students proficient or advanced proficient in 2005 was statistically greater than in 2003 with a  $p$  value of  $p = .009$ .

### **Discussion and Conclusions**

In a climate of accountability and sanctions for schools that do not demonstrate Adequate Yearly Progress (AYP) under No Child Left Behind (NCLB) it is incumbent on educational leaders to passionately pursue methods and programs that improve student achievement. In this

## Forum on Public Policy

study, the relationship between teacher assessments and NCLB testing was determined for two testing administrations of the New Jersey GEPA. Following the 2003 GEPA administration, educational professionals aligned the curriculum to NJ state standards, researched and adopted new textbooks, and participated in staff development on assessment. The results for the 2005 GEPA were reported as having a stronger relationship between teacher assessments and the GEPA. Additionally, there was a statistically significant increase in most of the GEPA components. There was also a statistically significant increase in the percent of students that scored proficient or advanced proficient on the 2005 GEPA.

There are several possible causes for these increases, including that one or more of the actions taken by the educational professionals in the district were effective. Another possibility could be that the mathematical achievement of the students in the 2005 cohort was higher than the 2003 cohort prior to the 2003 administration of the GEPA. However, the district administration used the same criteria, including standardized test scores to place students into the Algebra Part 1 course. There is a possibility that the 2005 GEPA test items were not as difficult as those administered in 2003. However, the New Jersey Department of Education takes steps to maintain the statistical stability of each testing administration. These findings were limited to the two cohorts of students, taking Algebra Part 1 and the GEPA.

It is necessary to align curriculum to the state standards for two reasons. The first is that these are the topics and skills that the state department of education has outlined as essential for all students. Secondly, in order to increase student achievement and making AYP, precious instructional time should be devoted to those topics and skills that are identified in the state standards. In states where there are insufficient data reported on student achievement and where NCLB testing is secure and not released, this method of determining the relationship may yield

## Forum on Public Policy

important data for program improvement. However, the need to determine and address individual student weaknesses is not met with this method alone. Similar districts may find that common assessments, such as a midterm exam or district developed instrument, can yield valuable data for individual students through the use of item analyses.

The need for security of assessments balanced with useable data supplied by departments of education is imperative in meeting AYP goals.

### *Further Research*

In this era of data driven instruction it is important to find how the accessibility of data for educational professionals informs instruction, program improvements, and increases student achievement. A qualitative study could be conducted to determine how educational professionals utilize achievement data. A study could be conducted to determine what the relationship is between teacher assessments and NCLB testing for schools that have meet AYP goals and have high student achievement. A study could be conducted to determine the relationship between the Scholastic Aptitude Test (SAT) and NCLB high school testing. A study of a school or district that continually does not meet AYP goals to determine the relationship between teacher assessment and NCLB testing could be conducted.

### **References**

- Blair, J. (2004, January 14). Court rules against editor for publishing Chicago tests. *Education Week*, 23, p. 5.
- Boston, C. (2002). The Concept of Formative Assessment. College Park, MD: ERIC Clearinghouse on Assessment and Evaluation. (ERIC Document Reproduction Service No. ED470206)
- Bowman, D. H. (2003, November 19). Florida court rejects father's bid to view state test. *Education Week*, 23, p. 5.
- Dillon, S. (June 25, 2003). Citing flaw, state voids math scores. *The New York Times*.
- Gandal, M. & McGiggert, L. (February, 2003). The power of testing. *Educational Leadership*, 60, 39-42.
- Guskey, T. R. (February, 2003). How classroom assessments improve learning. *Educational Leadership*, 60, 7-11.
- Herte, C. M. (2004). The Relationship of teacher assessment in grade eight mathematics to the New Jersey grade eight proficiency assessment in mathematics and the no child left behind accountability. ProQuest

## Forum on Public Policy

- Information and Learning Company. Ann Arbor, MI, (UMI No. 3139638).
- Hurst, M. D. (2004, October 6). Nevada report reveals spike in testing irregularities. *Education Week*, 24, p. 19, 22.
- Keller, B. (2002, January 16). Austin cheating scandal ends in no-contest plea, fine. *Education Week*, 21, p. 3.
- Larson, R., Boswell, L., Kanold, T.D., & Stiff, L. (2004). *Algebra 1*. Boston, MA: McDougal Littell.
- Manzo, K. K. (2005, January 19). Texas takes aim at tainted testing program. *Education Week*, 24, p. 1,14.
- McMillan, J. H., & Nash, S. (2000). *Teacher classroom assessment and grading practices decision making*. Richmond, VA: Metropolitan Educational Research Consortium. (ERIC Document Reproduction Service No. ED447195)
- National Council of Teachers of Mathematics. (1997). *Assessment Standards for School Mathematics*. Reston, VA: Author.
- National Council of Teachers of Mathematics. (2000). *Principles and Standards for School Mathematics*. Reston, VA: Author.
- New Jersey Department of Education. (2003a). *GEPA Student Preparation Booklet*. Trenton, NJ: Author.
- New Jersey Department of Education. (2003b). *GEPA Test Manual*. Trenton, NJ: Author.
- New Jersey Department of Education. (2003c). *School and district guidelines: Interpretation and use of GEPA results*. Trenton, NJ: Author.
- New Jersey Department of Education (2005a). 2004-05 New Jersey School Report Card. Retrieved July 5, 2006 from [www.state.nj.us/rc/rc05](http://www.state.nj.us/rc/rc05)
- New Jersey Department of Education. (2005b). *GEPA Score Interpretation Manual: March 2005 Grade Eight Proficiency Assessment (GEPA)*. Trenton, NJ: Author.
- New Jersey Department of Education. (2006a). *GEPA Student Preparation Booklet*. Trenton, NJ: Author.
- New Jersey Department of Education. (2006b). *GEPA Test Manual: Grade Eight Proficiency Assessment March 2006*. Trenton, NJ: Author.
- New Jersey Department of Education. (n.d.a.). Mathematics Standards. Retrieved September 9, 2003 from [www.njpep.org/standards/revised\\_standards/Math\\_newstandards](http://www.njpep.org/standards/revised_standards/Math_newstandards)
- Popham, W. J. (February, 2003). The seductive allure of data. *Educational Leadership*, 60, 48-51.
- Popham, W. J. (2006, April 19). Educator cheating on No Child Left Behind tests: can we stop it? *Education Week*, 25, pp. 32-33.
- United States General Accounting Office. (2002). *Title I: Education needs to monitor states' scoring of assessments. Report to the Secretary of Education*. Washington, DC: General Accounting Office.

## Forum on Public Policy

Table 1:

**Teacher Assessments and GEPA by Year**

	Year of Test	Mean	Std. Deviation	Std. Error Mean
Marking Period 1	2003	81.278	10.011	.628
	2005	77.807	10.099	.684
Marking Period 2	2003	79.656	10.755	.675
	2005	76.982	12.654	.857
Midterm Exam	2003	73.917	11.207	.703
	2005	70.982	14.889	1.008
Marking Period 3	2003	81.652	12.018	.754
	2005	78.060	11.035	.747
Number Sense	2003	44.59	21.45	1.35
	2005	61.58	19.16	1.30
Spatial Sense	2003	40.58	20.84	1.31
	2005	44.50	21.73	1.47
Patterns & Functions	2003	58.69	19.67	1.23
	2005	62.23	17.65	1.20
Data Analysis	2003	51.97	19.73	1.24
	2005	63.76	18.28	1.24
GEPA Knowledge	2003	48.85	16.76	1.05
	2005	58.02	15.32	1.04

2003: N = 254, 2005: N = 218

Table 2:

**Regression Comparisons 2003<sup>a</sup> and 2005<sup>b</sup>**

Component	GEPA			
	Year	R	R <sup>2</sup>	Standard Error
Number Sense	2003	.256	.066	2.494
	2005	.405	.164	17.559
Spatial Sense	2003	.289	.084	2.398
	2005	.471	.221	19.265
Patterns & Functions	2003	.224	.050	2.305
	2005	.447	.200	15.863
Data Analysis	2003	.251	.063	2.296
	2005	.527	.278	15.605
GEPA Total Points	2003	.330	.109	22.93
	2005	.579	.336	12.547

<sup>a</sup>n = 254    <sup>b</sup>n = 218

Figure 1:

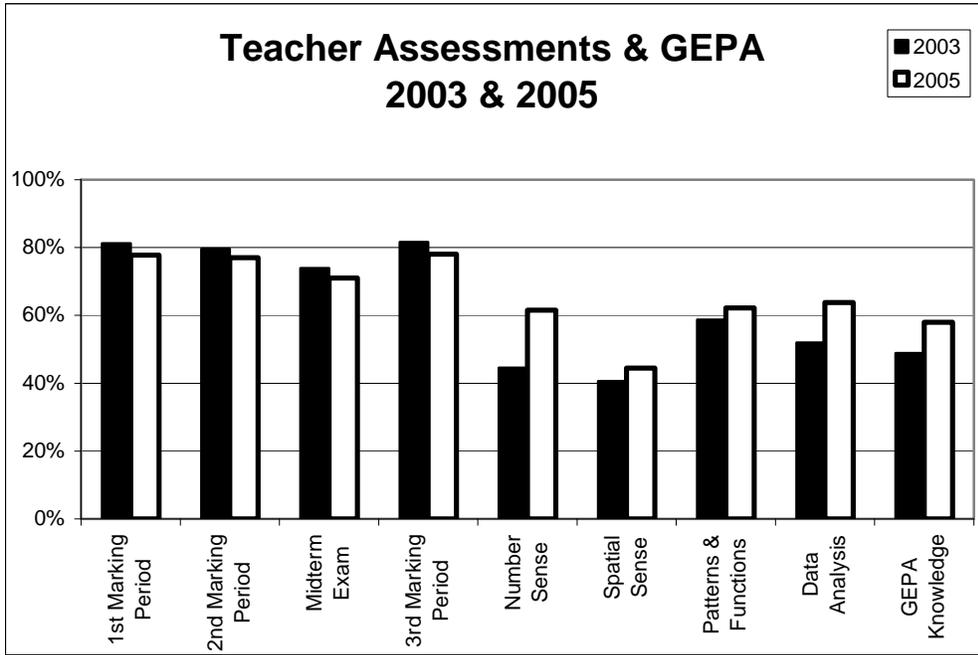


Figure 2:

