An Evaluation of the Influence of Interim Assessments on Grade 8 Student Achievement in Mathematics and Language Arts*

Maria Pereira
Christopher Tienken

Abstract

A review of the literature pertaining to the effect and influence that interim assessments have on student achievement lacks quantitative data to determine the efficiency of their use in the classroom as a school reform tool. This study examined the strength and the direction of the relationships between interim pre and posttest assessments in language arts and mathematics in Grade 8 and student achievement on the New Jersey Grade 8 state standardized tests in those subjects. Analyses were conducted using simultaneous multiple regression models. All student data explored in this study pertained to 670 students in Grade 8 enrolled in four middle schools located in a suburban/urban central New Jersey community during the 2009-2010 academic school year. The results of the study revealed each school produced a combination of site specific results and the interim pretests accounted for the same or almost the same amount of variance in state test scores as the interim posttests.

1 NCPEA Publications

NOTE: This manuscript has been peer-reviewed, accepted, and is endorsed by the National Council of Professors of Educational Administration (NCPEA) as a significant contribution to the scholarship and practice of education administration. In addition to publication in the International

*Version 1.3: Oct 2, 2012 1:20 pm -0500
†http://creativecommons.org/licenses/by/3.0/

Journal of Educational Leadership Preparation,[1] Volume 7, Number 3 (Winter 2012), ISSN 2155-9635, this manuscript exists in the Connexions Content Commons as an Open Education Resource (OER). Formatted and edited by Theodore Creighton, Virginia Tech; Brad Bizzell, Radford University; and Janet Tareilo, Stephen F. Austin State University. The assignment of topic editor and double-blind reviews are managed by Editor, Linda Lemasters, George Washington University. The IJELP is indexed in the Education Resources Information Center (ERIC), sponsored by the United States Department of Education under Contract No. ED-04-CO-0005.

2 Sumario en español

Una revisión de la literatura que pertenece al efecto y la influencia que evaluaciones provisionales tienen en faltas de logro de estudiante los datos cuantitativos para determinar la eficiencia de su uso en el aula como una herramienta de reforma de escuela. Este estudio revisó la fuerza y la dirección de las relaciones entre interín pre y evaluaciones de examen posterior en artes de idioma y matemáticas en el Grado 8 y logro de estudiante en el Nueva Jersey Gradúan 8 estado las pruebas estandarizadas en esos sujetos. Analiza fueron realizados utilizando múltiples modelos simultáneos de retroceso. Todos los datos del estudiante exploraron en este estudio pertenecido a 670 estudiantes en el Grado 8 matriculado en cuatro colegios ubicados en una comunidad central suburbano/urbano de Nueva Jersey durante el 2008-2009 año escolar académico. Los resultados del estudio revelaron cada escuela produjo una combinación de sitio resultados específicos y las pruebas preliminares provisionales justificaron el mismo o casi la misma cantidad de variación en puntuaciones de estado como los exámenes posteriores provisionales.

NOTE: Esta es una traducción por computadora de la página web original. Se suministra como información general y no debe considerarse completa ni exacta.

3 Introduction

School and district administrators use the results from state assessments and other standardized assessment tools, such as interim assessments, to make decisions about things like revisions to curriculum and instruction programs, teacher quality, overall student achievement, and student placements in special programs and academic ability groups (Tienken, 2008a). Increasingly, school principals are turning to the use of student results from interim assessment products to make predictions about which students might pass their state mandated tests of skills and knowledge and which might need additional interventions to pass the test. The practice of school administrators targeting interventions to specific groups of students based on test results, especially those closest to their state's proficiency cut-score, is a national issue. Some have called the practice "educational triage" for the "bubble kids" (Booher-Jennings, 2005, p. 1).

School administrators around the nation are not immune to the pressures to raise student test scores in order to retain their jobs. There are multiple proposals for school administrator evaluation in place or under consideration in several states and all of them include the use of student test results as a major indicator for school administrator pay and job security. Accordingly school administrators in many states, including New Jersey, search for interventions to raise student achievement on state mandated tests. In some cases state education bureaucrats make recommendations to school administrators on the interventions they should use.

In a 2008 memo released by the former New Jersey Department of Education (NJDOE) Commissioner of Education, Lucille Davy recommended the use of the *Learnia* computer-based formative assessment product, vended by Pearson, and claimed that research demonstrated that the formative assessment product was able to predict students who would have trouble on the state test and that the product had a positive impact (cause and effect) on student achievement. The memo did not state whether contextual factors such as student demographics, grade levels, or socio-economics influenced the effectiveness of the product. The memo and marketing for the product made it seem as if context did not matter.

[1] http://www.ncpeapublications.org/latest-issue-ijelp.html

We note that although Pearson and the former NJ Commissioner of Education used the term "formative assessment" to describe the *Learnia* product, Learnia or other "formative assessment" products on the market do not conform with the empirical definition of formative assessment. In fact, most products on the market are interim assessments. As we define later in the manuscript, there are important differences between the two types of assessments.

As a result of the former NJ Commissioner's recommendation, school administrators across the state were led to believe that computer-based pretest/posttest interim assessments were an effective intervention to predict student proficiency and raise student achievement regardless of school context and student demographic variables. Many school administrators mandated the use of interim assessments, the *Learnia* product in particular, in classrooms throughout their districts. During the 2007 through 2011 school years over 250 school districts in New Jersey, representing hundreds of schools, used the *Learnia* product. *Learnia* is the most popular formative assessment tool used in New Jersey and one of the most popular tools used nationwide. Millions of dollars have been spent on the product in New Jersey alone.

4 Purpose and Questions

Nationally, middle school student proficiency percentages on state mandated tests are generally lower than proficiency percentages for elementary school student results. Therefore, there is a heightened sense of urgency on the part of middle school administrators to seek interventions to raise achievement. Our purpose for this study was to evaluate the influence of the *Learnia* pretest and posttest interim assessments in language arts and mathematics on Grade 8 student scale-scores on state mandated New Jersey Assessment of Skills and Knowledge (NJASK) in Grade 8 language arts and mathematics.

We guided the study with the overarching question: When controlling for other factors found in the extant literature to influence middle school student achievement, how much variance in middle school student achievement on the NJ ASK8 language arts and mathematics sections do the results from pretest and posttest interim assessments in language arts and mathematics explain?

4.1 Significance of the Study to the Field of Education Administration

The results from the existing literature suggest that principals and other school administrators use the student results from interim assessments to make decisions about which students receive targeted intervention and which do not. However, there is little empirical literature on the efficacy of interim assessments for such use. The results of this study will provide school administrators empirical data they can use to make decisions about which resources to purchase, if in fact they choose to purchase interim assessment resources, and the efficacy of those resources for decision making purposes. Several of the existing studies on the topic are methodologically flawed, use small samples, and/or use simple pre-experimental designs that are not appropriate for drawing conclusions about efficacy (e.g. Takacs, 2010). Our design and methodology represent improvements over much of the existing literature.

5 Theoretical Framework

We draw upon production / function theory to guide our understanding of the topic and our analysis. The theory rests on the idea that the quantity of output (Q) is a function (f) of various inputs (X_1, X_2, X_3...). The symbolic representation for the theory is Q=f(X_1, X_2, X_3...). In the case of our study, Q represents the student results on the state mandated LA and math tests as a function of the various student and school inputs that contribute to student achievement. Interim assessments are one such input, among others.

A review of the empirical literature found support, albeit mixed in some cases, for four additional input variables: students' socioeconomic status (e.g., Coleman, et al., 1966; Koretz, 2009; Sirin, 2005), specific formative assessment strategies (Butler, 1988; Schunk, 1996) student gender, (Else-Quest, et al., 2010; Willingham and Cole, 1997) and teacher advanced degree status (Goldhaber & Brewer, 2002; Michel, 2008). Figure 1 presents the final theoretical framework used to guide this study that includes some variables

deemed by the NJDOE bureaucrats and the host district, and some of the variables found in the literature that influence student achievement.

Further analysis of district-specific characteristics of the initial variables included in this framework resulted in the elimination of teacher advanced degree status. We found based on an analysis of personnel records that over 90% of the teachers had a Masters degree in education. Thus, we later eliminated the variable of teacher degree status because at the school level there was almost no variance in the status of the language arts and mathematics teachers for the district.

An additional possible influential variable was identified: Academic Support Instruction (ASI), also known as Basic Skills Instruction. In an effort to specialize instruction based on the student's individual needs district personnel provide ASI in language arts and mathematics in Grades 6-8. ASI Mathematics and ASI Language Arts courses are conducted during the regular school day and replace student's mandatory mathematics and language arts classes. We added ASI as a variable because the district administration obviously believed it influenced achievement, otherwise they would not have expended tax dollars on the program. Although the empirical literature suggests mixed results for ASI-type programs nationally, the district administration at this site never conducted an evaluation of the local ASI program. Therefore, we could not know if it did or did not influence achievement and we added it into our models. Because there is ambiguity between the terms formative assessment and interim assessment we reviewed the empirical literature for each term and provide defining characteristics to resolve the ambiguity of terms and to assist the reader in understanding the results of our study.

5.1 Formative Assessment

According to Perie, Marion, and Gong (2007) "formative assessment is a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes" (p.1). There are specific characteristics of formative assessments that separate them from interim assessments. For example, the frequency of administration and the manner in which the results permit teachers to modify instruction help to determine if the assessment is formative or interim.

State standardized test are considered summative assessments. Summative assessments are generally given one time at the end of a unit or at the end of the school year to evaluate students' performance against some type of content standards. The results from these assessments do not provide teachers with information that could be integrated into instruction daily.

Asking questions in class is a type of formative assessment that can help teachers monitor and adjust their lessons in "real time" to improve student achievement of the intended instructional goals. The most effective formative assessment occurs frequently and it is structured to provide opportunities for the students to practice the technique of self-evaluation, and reflection in order to monitor and adjust their individual learning (Sadler, 1989; Schunk, 1996). When looking at studies involving web-based formative assessment programs, small (e.g. 0.03) or statistically non-significant effect sizes were reported (Buchanan, 2000; Sly, 1999; Velan, Rakesh, Mark, & Wakefield, 2002; Wang, 2007). One reason is that web-based formative assessment products are often stand alone programs. They are the end, so to speak. They are not integrated into the teaching and learning process. They are simply another activity to accomplish.

When looking at self-reflection and self-assessment as a formative assessment strategy integrated into the teaching and learning processes, statistically significant effect sizes of .73-1.97 have been reported across the Grades 4-12 spectrum (Sadler, 1989; Schunk, 1996; White & Frederiksen, 1998). The type of formative assessment strategy seems to matter greatly in terms of the influence on student achievement.

Perie, Marion, and Gong (2007) contend that formative assessment activities possess the following characteristics:

1. They are embedded within the learning activity and linked directly to the current unit of instruction.
2. They are small-scale (a few seconds, a few minutes, less than a class period) and short-cycle (they are often called "minute-by-minute" assessments or formative instruction).
3. Tasks presented may vary from one student to another depending on the teacher's judgment.

4. They are ongoing and fluid in nature. (p.1)

5.2 Interim Assessments

There has been an increase in use of interim assessments by schools and districts in the United States in an effort to improve student achievement (Goertz, Olah, & Riggan, 2009). Dunn and Mulvenon (2009) delineated the differences between formative and interim assessments:

5.2.1

Formative assessment is defined as assessment used by teachers and students to adjust teaching and learning, as compared to interim assessment that informs policymakers or educators at the classroom, school, or district level about student achievement levels and curriculum effectiveness. (p.5)

Perie, Marion, and Gong (2007) consider interim assessments to be:

5.2.2

Interim assessments (1) evaluate students' knowledge and skills relative to a specific set of academic goals, typically within a limited time frame, and (2) are designed to inform decisions at both the classroom and beyond the classroom level, such as the school or district level... As such, the timing of the administration is likely to be controlled by the school or district rather than by the teacher, which therefore makes these assessments less instructionally relevant than formative assessments. (pp.1-2)

Goertz, Olah, and Riggan (2009) conducted an exploratory study in which they investigated the use of interim assessments, as well as the policies that support their use in the classroom. Included in their study were 45 elementary school teachers (Grade 3 and 5) selected by a purposive sample of nine schools located in two Pennsylvania school districts. The researchers purposely selected an urban and a suburban district to study in an effort to gather information on how "policy supports for assessment and instructional improvement function in these different environments" (p.2). The results showed "that interim assessments are useful but not sufficient to inform instructional improvements" (Goertz, Olah, & Riggan, 2009, p.8). The authors uncovered minimal evidence indicating that the interim assessments they "studied help teachers develop a deeper understanding of student's mathematical learning—a precursor to instructional improvements. Most items in the assessments did not provide actionable information on students' misunderstandings" (Goertz, Olah, & Riggan, 2009, p.8).

According to Perie, Marion, and Gong (2007) the best commercially prepared interim assessment programs can:

1. Provide an item bank reportedly linked to state content standards.
2. Assess students on a flexible time schedule wherever a computer and perhaps internet connections are available.
3. Provide immediate or very rapid results.
4. Highlight content standards in which more items were answered incorrectly.
5. Link scores on these assessments to the scores on end-of-year assessments to predict results on end-of-year assessment. (p.14)

The purpose and intended uses of an assessment determine whether it is classified as a formative assessment or an interim assessment. Far from being definitive, the literature on the effectiveness of formative and interim assessments to raise student achievement is mixed. The results in the literature suggest that specific strategies of formative assessment might be helpful, but the literature provides less certain guidance for the use of interim assessment and calls into question any large-scale use of the strategy without further evaluation of its efficacy with children.

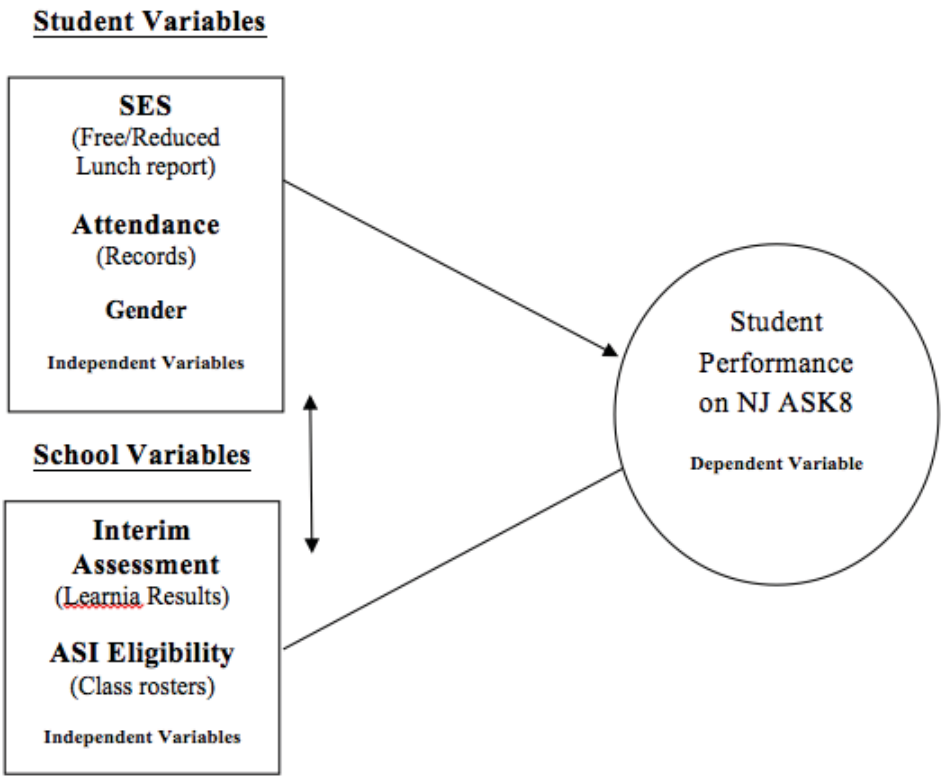

Figure 1. Production / Function Theoretical Framework

6 Design and Methodology

"Non-experimental research is frequently an important and appropriate mode of research in education" (Johnson, 2001, p.3) due largely in part to the inability of education researchers to perform randomized experiments and quasi-experiments on children. We used a non-experimental, cross-sectional, explanatory design. An explanatory study must meet the following criteria: (a) The researchers attempted to develop or test a theory about a phenomenon to explain "how" and "why" it operates; (b) The researchers attempted to explain how the phenomenon operates by identifying the possible factors that influence change in it (Johnson, 2001).

Neither cause nor trends in performance can be determined. The data for this study were collected from five middle schools that serve students who lived in home and neighborhood environments that exhibited lower middle class to upper middle class wealth characteristics. Generalizations should not be made to schools that serve less wealthy or wealthier populations than those served by the schools in the study.

We analyzed student level data by school building and not aggregated to the district level due to the differences in student demographic and achievement among the schools. The dependent variables were the Grade 8 student scale scores from the language arts and mathematics sections. The independent variables were student attendance, gender, eligibility for free or reduced lunch, eligibility for the ASI basic skills remediation program, pretest interim assessment scores in language arts and mathematics, and posttest interim assessment scores in language arts and mathematics.

6.1 Sample and Data Source

All data explored in this study pertained to students in Grade 8 enrolled in four middle schools located in a ethnically and economically diverse suburban/urban central New Jersey community during the 2009-2010

academic school year. The final student sample population for this study was 670 Grade 8 students. Data from students who met the following criteria were included in the study (a) general education program; (b) received a valid score on both sections of the 2009 NJ ASK language arts and mathematics tests; (c) received a valid score on both sections of the *Learnia* pretest and posttest assessments in language arts and mathematics during the 2008-2009 school year; and (d) do not qualify for the district's English language learners (ELL) program.

Table 1

Demographic Characteristics of Grade 8 Students by School [1] *Reported in Number and Percentage of Population*

School	Total	White	Black	Asian	Hispanic	Economically Disadvantaged
A	227	97 (42%)	49 (22%)	33 (15%)	47 (21%)	80 (35%)
B	229	165 (72%)	19 (8%)	22 (10%)	22 (10%)	31 (14%)
C	200	70 (35%)	27 (14%)	32 (16%)	70 (35%)	89 (45%)
D	252	97 (38%)	29 (12%)	99 (39%)	27 (11%)	58 (23%)

[1] One student categorized as Pacific Islander attended School A. There were no other students categorized as Pacific Islander in the other three schools. There was one student categorized as American Indian in School B and School C.

Concerning sample size for regression, Field (2009) references Green (1991) for establishing a minimum acceptable sample size. Field wrote,

6.1.1

"...if you want to test the model overall, then he [Green] recommends a minimum sample size of $50 + 8k$, where k is the number of predictors. So, with five predictors, you'd need a sample size of $50 + 40 = 90$. If you want to test the individual predictors then he suggests a minimum sample size of $104 + k$, so again taking the example of 5 predictors you'd need a sample size of $104 + 5 = 109$" (Field, 2009, p.222).

We needed a minimum of 112 cases to meet Field's (2009) and Green's (1991) threshold for sample size to ensure proper power to test the full model. We exceeded the minimum number of cases required for each model. We needed a minimum of 214 cases at each site to ensure proper power to test individual predictors. We exceeded the minimum requirement in three of the four schools and failed to meet the minimum requirement by 14 cases in School C.

6.2 Instrumentation

Instrumentation for the study consisted of the NJ ASK in Grade 8 and the computer-based interim assessment product *Learnia*. "The New Jersey Department of Education is required by federal law to ensure that the instruments it uses to measure student achievement for school accountability provide reliable results" (NJDOE, 2009, p. 116). Because high-stakes decisions are made often using solely a student's test scores (Tienken, 2008a, b), it is important to discuss the standard error of measurement (SEM) associated with these assessments. When considering the conditional SEM, Tienken (2008b, citing Harville, 1991) summarized that it is "an estimate of the amount of error or lack of precision one must consider when interpreting a test score" (p.37). He stated that "the SEM describes how far the reported results may differ from a student's true score" (p.37). To clarify on the issues of reliability and SEM, the NJDOE (2009) published in the following in the NJ ASK Technical Report:

6.2.1

Although the conceptualization of reliability and SEM is relatively straightforward, issues underlying the estimation of reliability are not. Reliability can be estimated via the correlation of scores on parallel forms or from test-retest data, or it can be estimated from a single test administration using any one of a variety of techniques (e.g., Cronbach, 1951; Kuder & Richardson, 1937). A very popular technique for estimating reliability from a single test administration is Cronbach's coefficient alpha (p.117).

The Cronbach's alpha for full-test reliability of the language arts and mathematics sections exceeds .80 and the conditional SEM is approximately 10 scale score points at the proficiency cut-score for individual scores.

As of 2010, *Learnia* was available to school districts in California, New Jersey, and Texas. *Learnia* is marketed as a web-based "formative" assessment tool produced by Pearson. *Learnia* is designed to help diagnose the academic strengths and weaknesses of individual students in language arts and mathematics.. Assessments delivered by *Learnia* align with the New Jersey Core Curriculum Content Standards (NJCCCS) as well as with the criteria set forth by Measurement Incorporated (MI), the corporate vendor that develops NJ ASK state mandated standardized test. Reliability data for *Learnia* were not available.

6.3 Data Collection

Student-level NJ ASK test scores were collected from the district database. The free and reduced lunch report, attendance records, and ASI class rosters were provided by the District's Data Analyst. *Learnia* pre and post assessment results were collected via the computerized *Learnia* summary report feature. All reports were coded to guarantee confidentially. Each coded student identifier represented a record. Each complete report contained the following data unique to each record: NJ ASK test scores (language arts and math), *Learnia* pre assessment scores (language arts and math), *Learnia* post assessment scores (language arts and math), free and reduced lunch identification, attendance record, and ASI eligibility. Incomplete cases with records void of a least one component of data were excluded from the study.

6.4 Data Analysis

Because of differences in the demographics and academic makeup of the students in the four middle schools, and to meet the simultaneous regression assumption of data independence, we analyzed the data from the four schools separately instead of aggregating the scores across the schools. Separate analyses were also necessary because the district leadership interpreted the test results on the individual school level and made decisions for each school based on that school's results, not based on district aggregated test scores. Resources are allocated in the district based on the individual schools' output, not the aggregate of the district's middle schools. Although the district is categorized as lower middle class, the four schools used in the study have characteristics representative of working class to upper class socio-economics.

Because not all the independent variable were continuous (i.e., free/reduced lunch eligibility, attendance policy violations, and ASI eligibility), we used dichotomous dummy coding for categorical variables. The following recoding was used for the student variable of SES: 1 = eligible for free/reduced lunch, 0 = not eligible; for the student variable of attendance: 1= student exceed district policy of 16 absences, 0= student did not exceed 16 absences; for the school variable of ASI eligibility: 1= eligible for ASI services, and 0= not eligible. Gender was coded as 0=male, 1= non-male.

We ran scatter plots and correlation matrices in order to check the assumption of linear relationships of each predictor with the dependent variable. We conducted Pearson Product Moment correlations between the dependent variables and the predictor variables for each subject (M and LA) for each school to determine the relationships among the variables prior to conducting simultaneous multiple regression as a preliminary check for multicollinearity. We do report Pearson correlations for the sake of brevity but they can be requested from the authors. As part of the simultaneous multiple regression, all appropriate student and school variables were entered at the same time. From this procedure we were able to ascertain which predictors contributed statistically significantly to the multiple regressions. Multicollinearity issues were tested by computing the variance inflation factor (VIF) from the data loaded into the regression models. Multicollinearity was not an issue in the models as the tolerance values for the predictors were not exceedingly low (<1-R^2) for any variable except NJ ASK8 Math. From the t value and the p value, we determined if one specific variable statistically significantly contributed to the prediction equation for NJ ASK scores from all the independent variables.

7 Results

We present results by school and subject within each school.

7.1 School A Language Arts

A statistically significant model emerged (F= 31.209, p=$\leq$ .001, Adjusted R^2 .588) and accounted for approximately 59% of the variance in student performance on the NJ ASK8 in language arts. Statistically significant predictor variables, their betas, and p values were as follows: (a) NJ ASK8 Math, .363, p<.001, (b) interim LA Posttest, .275, p<.001, (c) interim LA Pretest, .248, p<.001, and (d) gender, .152, p=.004.

NJ ASK8 Math achievement was the best predictor of NJ ASK8 LA achievement. However, the predictive power of the NJ ASK8 Math is suspect. It stands to reason that the true direction of the relationship comes from a student's language arts skills and those skills influence his/her mathematics scores because of the amount of reading necessary to engage the NJ ASK8 mathematics test. The seemingly apparent influence of math achievement on language arts achievement might be a false finding and should be interpreted with caution.

Readers should instead focus on the almost identical strength of the interim LA pretests and posttests. We find it interesting that the interim LA pretest, given in September of a school year has similar predictive power as the posttest, usually given in March or April of the school year. Are both tests and the time taken from instruction necessary in this school?

7.2 School A Mathematics

A statistically significant model emerged (F = 31.866, p < .001, Adjusted R^2 .594.) accounted for approximately 59% of the variance in NJ ASK8 math scores. Two statistically significant predictors variables were, (a) interim Math Posttest with a beta of .383, (p<.001) and a t value of 5.296, and (b) NJ ASK8 LA (p<.001) with a beta of .358 with a t value of 4.904.

7.3 School B Language Arts

A statistically significant model emerged (F = 36.329, p < .001, Adjusted R^2 .639.) that accounted for approximately 64% of the variance in student performance on the NJ ASK8 LA. Statistically significant

predictors variables, their betas, p values, and t values were as follows: (a) ASI, - .270, p<.001, t value of 4.732, (b) gender, .242, p<.001, t value of 5.132, (c) NJ ASK8 Math, . 227, p<.001, t value 2.773, (d) interim LA Posttest, .206, p=.001, t value of 3.509 , and (e) student SES, -.132, p=.005, t value of -2.875. However, as stated earlier in the results for School A, the finding for the strong prediction value of NJ ASK8 Math for NJ ASK8 LA is suspect.

7.4 School B Mathematics

A statistically significant model emerged (F = 49.333, p < .001, Adjusted R^2 .707.) Approximately 71% of the variance in NJ ASK8 Math scores was explained by the variables in the model. Statistically significant predictors variables were as follows: (a) ASI,- .279, p<.001, t value of -5.539, (b) interim Math Posttest, .240, p=001, t value of 3.541, (c) interim Math Pretest, . 211, p=002, t value 3.093, (d) NJ ASK8 LA, .184, p=.007, t value of 2.733, and (e) interim LA Pretest,.115, p=.02, t value of 3.351. As with the interim LA pretests and posttests in School A, the interim math pretests and posttests in School B were similar in their prediction strength. Potentially, the interim pretest math scores from a test given in September could predict about as much of a student's chance of being proficient on the state mandated test administered in May.

7.5 School C Language Arts

A statistically significant model explained (F = 20.246, p < .001, Adjusted R^2 .614.) approximately 61% of the variance in student performance on the NJ ASK8 LA. Statistically significant predictor variables, their betas, p values, and t values were as follows: (a) NJ ASK8 Math .703, p<.001, t value of 7.310, (b) interim LA Pretest, .214, p=.003, t value of 3.067, (c) interim LA Posttest, .212, p=.005, t value of 2.884, and interim Math Posttest, -.189, p=.021, t value -2.347. However, as stated earlier in the results for School A, we question the finding for the strong prediction value of NJ ASK8 Math for NJ ASK8 LA.

The findings for the predictive strength on the interim LA pretests and posttests mirror those from School A. The beta and t values for the interim LA pretest and posttest in School C are almost identical, calling into the question the usefulness of the pretest / posttest assessment scheme in School C.

7.6 School C Mathematics

A statistically significant model emerged (F = 33.437, p < .001, Adjusted R^2 .728.) and it accounted for approximately 73% of the variance in student performance on the NJ ASK8 Math. Statistically significant predictor variables, their betas, p values, and t values were as follows: (a) NJ ASK8 LA, .495, p<.001, t value of 7.310, (b) interim Math Posttest, .319, p<.001, t value of 5.163, (c) gender, -.208, p<. 001, t value -3.992, and (d) ASI, -.179, t value -3.230. The betas suggested that being female, not eligible for ASI services and doing well on the NJ ASK8 LA and interim math posttest predicted positive results for the mathematics portion.

7.7 School D Language Arts

A statistically significant model emerged (F = 30.316, p < .001, Adjusted R^2 .559) and accounted for approximately 56% of the variance in NJ ASK8 LA scores. Statistically significant predictor variables, their betas, p values, and t values were as follows: (a) NJ ASK8 Math, .443, p<.001, t value of 5.956, (b) interim LA Pretest, .187, p<.001, t value of 3.750, (c) ASI, -.119, p=.033, t value -2.141, (d) interim LA Posttest, .113, p=.044, t value of 2.023, and (e) gender,.102, p=.033, t value -2.147. The betas suggested that not being in need of ASI services, being female, and doing well on the interim LA pretest and posttests accounted for the most variance in performance on the NJ ASK8 LA test. Interestingly, the interim LA pretest was a stronger predictor of achievement on the state LA test than the posttest.

7.8 School D Mathematics

A statistically significant model emerged (F = 45.910, p < .001, Adjusted R^2 .660.) that indicated approximately 66% of the variance in NJ ASK8 Math scores was explained by the variables in the model. Statistically significant predictor variables, their betas, p values, and t values were as follows: (a) NJ ASK8 LA, .341, p<.001, t value of 5.956, (b) interim Math Pretest, .336, p<.001, t value of 5.999, (c) ASI, -.203, t value -4.301, (d) interim Math Posttest, .104, p=.056, t value 1.919. As we found in School B, the interim math pretest was a stronger predictor of student achievement on the state mandated math test than the interim posttest.

8 Conclusions and Recommendations

The results of the study revealed each school produced a combination of site-specific results and results common across sites regarding the strength of each independent variable to predict student achievement. The strength of the relationships between the various independent variables and the dependent variables varied greatly across sites. This suggests that the influence of interventions differ across sites due to contextual factors such as demographics of the community and student body. One size does not fit all in terms of interventions and influence on achievement.

8.1 Interim Assessment Variable

The interim pre and posttests were statistically significant predictors of achievement across all schools for LA and Math, with the exception of School C where the math pretest was not a statistically significant predictor of achievement. Seemingly, in all cases, the interim pretest accounted for the same or approximately the same amount of variance in student output on the state mandated tests as did the posttest. If the pretest predicts as much or almost as much of the variance in student achievement as the posttest, of what value is the posttest, and pretesting/posttesting cycle for that matter? The *Learnia* interim assessment tool is marketed to school and district administrators as being able to provide high quality feedback to teachers from the pretest so that they can adjust their instruction before the posttest, and before the gran cru of summative assessments, the state mandated NJASK8. In this case, similar to Thorndike's (1901; 1924) findings so many years ago, those who came to Grade 8 with the most academic achievement left with the most and *Learnia* testing could not influence the performance trajectory of those labeled as not-proficient on the pretest in order to become proficient on the posttest or the NJ ASK8.

Pearson recommends using *Learnia* as an important intervention to improve student achievement on state mandated tests. However, it is important to note that the reviewed research regarding interim assessments indicated that there is not a statistically significant relationship between the use of interim assessments and informed instructional change in the classroom (Goertz, Olah, & Riggan, 2009) nor is there a relationship between the types of assessment structures that are potentially available from the *Learnia* product and increased student achievement. More of an effort must be made to ensure that interim assessment products marketed to schools and interim assessment practices used in schools are vetted empirically by the corporation, not a common practice, and also vetted locally on a pilot basis to determine efficacy at various contextually unique school sites. The context in which the product is developed might not match the context in which it is deployed, and thus the results might not match those marketed by the company vending the product. Context matters.

The empirical literature revealed that the most effective classroom assessment in terms of influence on student achievement is formative assessment that is structured to provide opportunities for the students to practice the technique of self-evaluation, and reflection to help students learn to monitor and adjust their learning. For example, an experimental study conducted by Schunk (1996) revealed that students who had structured opportunities to practice the formative assessment technique of self-evaluation frequently, had greater motivation and increased achievement outcomes in comparison to those who did not participate in the practice of self-evaluation frequently. Generally, interim assessments do not provide those types of opportunities.

The existing empirical literature and the results from this study seem to suggest that the more proximal (closer to the student) the formative assessment activity is (i.e., self-evaluation), the greater the influence it has on learning (Sadler, 1989; Schunk, 1996) whereas the further the assessment practices, either formative or interim, are from the student (distal), the less influence they have on learning. The assessment product, *Learnia*, and products like it, are distal from the student. They are not formative assessments and the student is not actively involved in self-monitoring or self-assessing. It is unclear how the product, as currently used in the four schools in this study and marketed by the corporation facilitates reflection beyond superficial error identification.

8.2 Recommendations for Policy and Practice

School and district administrators might want to consider redeploying the funds earmarked for the distally developed corporate assessment products toward building more internal assessment capacity through teacher developed formative and interim assessment practices based on prevailing evidence. Doing so could help to build local assessment capacity and allow for customization of practices to meet the needs of the district's diverse student population. Making formative and interim assessment activities more proximal to the student might help to increase the chances that those practices will actually lead to improved teaching and influence student achievement and more effective use of scarce public funds. At the very least. school administrators should know the efficacy of the products they bring into their buildings and they should understand the quality of the data from which they make decisions.

9 References

Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas accountability system. *American Education Research Journal,* 42(2), 231-268.

Buchanan, T. (2000). The efficacy of a World-Wide Web mediated formative assessment. *Journal of Computer AssistedLearning, 16*, 193-200.

Butler, R. (1988). Enhancing and undermining intrinsic motivation; the effects of task-involving and ego-involving evaluation on interest and performance. *British Journal of Educational Psychology, 58*, 1-14

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A.M., Weinfield, F.D., & York, R.L. (1966). *Equality of educational opportunity.* Washington, DC; U.S. Government Printing Office.

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

Davy, L. (2008). *Learnia formative assessment resources 2008-2009.* Retrieved August 15, 2009, from http://www.state.nj.us/education/assessment/formative/memo050908.pdf

Dunn, K.E., & Mulvenon, S.W. (2009). *Let's talk formative assessment...and evaluation?* Retrieved August 16, 2009, from http://eric.ed.gov (ERIC Document Reproduction Service No. ED505357)

Else-Quest, N.M., Hyde, J.S., & Linn, M.C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin*, *10*(1), 103-127.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). Los Angeles, CA: Sage Publishing.

Goertz, M. E., Olah, L. N., & Riggan, M. (December, 2009). *Can interim assessments be used for instructional change?* Consortium for Policy Research in Education Policy Brief.

Goldhaber, D. & Brewer, D. J. (2002). Does teacher certification matter? High school teacher certification status and student achievement. *Educational Evaluation and Policy Analysis, 22*(2), 129-145.

Green, S.B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research, 26,* 499 – 510.

Johnson, B. (2001). Toward a New classification of nonexperimental quantitative research. *Educational Researcher*, *30*(2), 3-13.

Koretz, D. (2009). *Measuring Up: What Educational Testing Really Tells Us.* Harvard University Press: MA.

Kuder, G. F. & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*, 151-160.
Michel, A.P. (2008). Variables from the New Jersey school report card that predict student achievement on the NJASK4. *New Jersey Journal of Supervision and Curriculum Development,* 52, 34-45.
New Jersey Department of Education (2009). *New Jersey Assessment of Skills and Knowledge 2008 technical report grades 5-8.* Retrieved July 28, 2009 from http://www.state.nj.us/education/assessment/ms/5-8/tech/2008TechReport.pdf
Perie, M., Marion, S., & Gong, B. (2007). *The role of interim assessments in a comprehensive assessment system: A policy brief.* Retrieved August 16, 2009, from http://www.achieve.org/files/TheRoleofInterimAssessmen
Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*, 119-140.
Schunk, D. H. (1996). Goal and self-evaluative influences during children's cognitive skill learning. *AmericanEducational Research Journal, 33*, 359-382.
Sirin, S.R. (2005). Socioeconomic Status and Academic Achievement: A Meta-Analytic Review of Research. *Review of Educational Research*, *75*(3), 417-453.
Sly, L. (1999). Practice tests as formative assessment improve student performance on computer-managed learning assessments. *Assessment and Evaluation in Higher Education,24* (3), 339-343.
Takacs, J.A. (2010). Using Formative Assessment in Professional Learning Communities to Advance Teaching and Learning. Unpublished doctoral dissertation. Walden University. AAT 3398977.
Thorndike, E. L. (1924). Mental discipline in high school studies. *Journal of Educational Psychology, 15*, 1-22, 98.
Thorndike, E. L., &Woodworth, R. S. (1901). The influence of improvement in one mental function upon efficiency of other functions. *Psychological Review, 8*, 247-261, 384-395, 553-564.
Tienken, C.H. (2008a). A descriptive study of the technical characteristics of the results from New Jersey's assessments of skills and knowledge in grades 3,4, and 8. *New Jersey Journal of Supervision and Curriculum Development*, *52*, 46-61.
Tienken, C.H. (2008b). The characteristics of state assessment results. *Academic Exchange Quarterly, 12*(3), 34-39.
Velan, G. M., Rakesh, K. K., Mark, D., & Wakefield, D. (2002). Web-based self-assessments in Pathology with Questionmark Perception. *Pathology*, *34*, 282-284.
Wang, T. H. (2007). What strategies are effective for formative assessment in an e-learning environment? *Journal of Computer Assisted Learning, 23*, 171-186.
White, B.Y., & Frederiksen, J.R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction, 16*(1), 3-118.
Willingham, W. W., & Cole, N. S. (1997). *Gender and Fair Assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.