Running head: HIGH STAKES AND HIGH EXPECTATIONS

High Stakes and High Expectations:

An Analysis of the Efficacy of High Stakes Testing as a School Reform Policy

Randy Hendricks, Ed.D.

Tarleton State University

ABSTRACT

The purpose of this study was to assess the efficacy of high stakes testing as a school reform policy by determining the impact that high stakes testing pressure has on academic achievement. High stakes testing pressure was operationalized by the High Stakes Testing Value (HSTV); a value determined by accumulating the number of testing stakes in each state's accountability policy. Academic achievement was operationalized by reading and mathematics National Assessment of Academic Progress (NAEP) scores.

NAEP achievement scores in reading and mathematics served as criterion variables for the multiple regression analyses while the following served as predictor variables; (a) the percentage of students identified as low socio economic status, (b) the percentage of students identified as English language learners, (c) the expenditures per student adjusted by the comparable wage index (CWI), and the HSTV.

The statistical data produced to address the six null hypotheses guiding this study showed that a lack of correlation existed between the HSTV variable and NAEP scores; therefore, all six null hypotheses were retained. However, interpretation of auxiliary statistical data suggested a link between the HSTV and higher than expected NAEP scores for states with high populations of low socio economic status students.

High Stakes and High Expectations:

An Analysis of the Efficacy of High Stakes Testing as a School Reform Policy

Introduction

The public schools of America are in the throes of a reform movement that has established accountability for educational outcomes as the driving force behind school improvement. High stakes testing is the centerpiece of this reform movement and has become the device by which schools and students are measured. This study assessed the efficacy of high stakes testing as a school reform policy by determining the impact that high stakes testing pressure has on academic achievement as measured by the National Assessment of Educational Progress (NAEP).

English and Steffy (2001) defined high stakes testing as the use of standardized tests to make important academic decisions about students. Madaus (1998) described high stakes testing as a reform strategy that coerces change by rewarding or punishing schools based on student performance on standardized tests. Generally, a high stakes test can be defined as any single assessment that results in rewards or sanctions for schools, educators, or students.

The need to critically assess high stakes testing as a school reform policy arises from the political context in which high stakes testing polices evolve. Accountability, as a lever for school improvement, is a part of the political lexicon; it is not a primary consideration of psychometrics, the science and practice of testing. The polis, or political community, is primarily motivated by the gain or loss of power and, therefore, depend upon negotiations, building alliances, employing tactics, and construct and deploy symbols as they jockey for political positioning. While the psychometric community is more concerned about statistical reliability and validity, the polis is motivated by how rhetoric and symbols can serve political objectives: “In the polis, concerns about reliability and validity of a testing instrument assume a lower priority than concerns about the political consequences of the test as a policy instrument” (Smith and Fey, 2000, p. 335).

Since the priorities and values of politics and psychometrics are inherently at odds, and since high stakes testing policies arise from the culture of politics, the debate over the efficacy of high stakes testing has taken on a decidedly political hue. The debate discourse seems to revolve around three primary issues. First, the possibility that high stakes testing is more symbolic than substantive has been raised and demands a response from the research community. If, as it has been suggested by some researchers, high stakes testing policies are intended as political symbols aimed at appeasing a public concerned about the quality of public education and are not necessarily intended to improve student learning then the unquestioned legitimacy of the current testing reform movement will only prolong more substantive public education policy responses (Smith and Fey, 2000; Amrein and Berliner, 2002a; Jones, Jones, and Hargrove, 2003; Pedroza, 1997).

Secondly, the cost of high stakes testing as a school reform policy extends beyond the financial expenditures invested in test construction, administration, grading, and reporting; it also extends to the cost of lost human capital if such policies do not yield their intended outcomes and students suffer the educational consequences (McNeil and Valenzuela, 2000; Pedroza, 1997). Lastly, political solutions to educational problems are historically implemented without sufficient evidence of policy efficacy in regards to the identified problem (Pedroza, 1997). Some researchers have suggested that the rapid spread of high stakes testing across the country exemplifies a premature and simplistic approach to an educational problem arising from complex social conditions and is an inherently flawed policy response (Popham, 2006; Pedroza, 1997; Haney, 2000).

Conceptual Framework

Various arguments have been used to promote the use of high stakes testing as a force to leverage failing public schools from the inertia of status quo. Among the most popular arguments are the following (Amrein and Berliner, 2002a):

1. High stakes tests serve to prioritize the curriculum for teachers, administrators, and school boards.

2. The sanctions associated with high stakes testing motivate educators and students to put more effort into teaching and learning.
3. High stakes tests aligned with state content standards provide all students, particularly those from historically underperforming student subgroups, with increased exposure to a high quality curriculum and, therefore, will lead to an elimination of the achievement gap between student subgroups.
4. School administrators use high stakes testing results to focus school improvement on specific content domains and student subgroups that have historically been underserved by public schools.
5. High stakes testing provides the general public with an understandable metric of school performance and arms them with information needed to meaningfully engage in the school reform dialogue.

The assumed efficacy of high stakes testing rests on the assumption that sanctions attached to governmental accountability policies will coerce improved teacher and student performance which, in turn, will translate into student gains on state assessments. However, if student gains on state assessments are not accompanied by gains on other legitimate measures of learning then one could conclude that the learning reflected on those state assessments is a very narrow form of learning that is restricted to the very specific context of the state assessment. For high stakes testing to provide evidence for improved student learning, there must be some other measure of learning besides state mandated tests; in other words, there must be some way to audit the effectiveness of high stakes testing policies independently from state assessments. In the context of this study, the National Assessment of Educational Progress (NAEP) will be utilized in lieu of state assessments as a way to audit the effectiveness of high stakes testing policies.

The conceptual framework underpinning this study is illustrated in Figure 1. In this schematic, states' testing polices (the independent variable) influence the following mediating variables; a) curriculum prioritization, b) educator and student motivation, c) the quality of the curriculum offered to

all students, d) educator focus on historically underperforming student subgroups, and e) public involvement in school curricular matters. In turn, positive influences on the mediating variables result in greater student learning as measured by selected NAEP tests (the dependent variable). Implicit in this conceptual framework is the assumption that increasing the number of testing stakes will generate greater influence on the mediating variables which will result in higher scores on NAEP assessments.

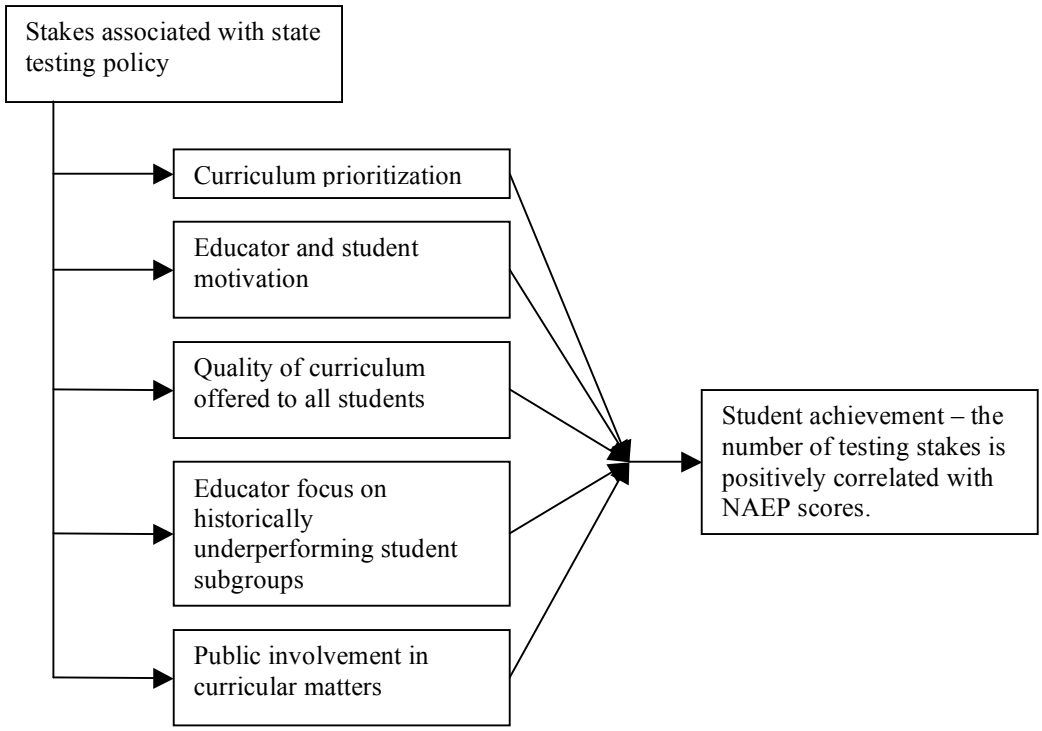

Figure 1. Conceptual framework of the study illustrating the relationships between state testing policies, five mediating variables, and student achievement.

Methodology

Research Question and Null Hypotheses

The purpose of this study was to assess the efficacy of high stakes testing as a school reform policy by determining the relationship between accumulated testing stakes and academic achievement as measured by the National Assessment of Educational Progress (NAEP). Accumulated testing stakes, a numerical value referred to as the High Stakes Testing Value (HSTV), was determined for each state by reviewing state accountability policies. The correlation between high stakes testing and academic

achievement was determined by establishing the potency of the HSTV as a variable in predicting student achievement on NAEP, a standardized assessment administered nation wide. In order to isolate the impact of the HSTV on student achievement, other variables known to influence student achievement were controlled. More specifically, the following research question and null hypotheses served to guide the study:

Research Question: Does the HSTV improve the accuracy of predicting student academic achievement when the mediating achievement variables of socio economic status, English language proficiency, and expenditures per student are controlled?

Null Hypothesis 1: The HSTV will not add to the accuracy of predicting NAEP reading scale scores when the mediating achievement variables of socio economic status, English language proficiency, and expenditures per student are controlled.

Null Hypothesis 2: The HSTV will not add to the accuracy of predicting NAEP mathematics scale scores when the mediating achievement variables of socio economic status, English language proficiency, and expenditures per student are controlled.

Null Hypothesis 3: The HSTV will not add to the accuracy of predicting the percentage of students scoring at the basic achievement level or higher on NAEP reading when the mediating achievement variables of socio economic status, English language proficiency, and expenditures per student are controlled.

Null Hypothesis 4: The HSTV will not add to the accuracy of predicting the percentage of students scoring at the basic achievement level or higher on NAEP mathematics when the mediating achievement variables of socio economic status, English language proficiency, and expenditures per student are controlled.

Null Hypothesis 5: The HSTV will not add to the accuracy of predicting gains in the percentage of students scoring at the basic achievement level or higher on NAEP reading when the mediating variables of socio economic status, English language proficiency, and expenditures per student are controlled.

Null Hypothesis 6: The HSTV will not add to the accuracy of predicting gains in the percentage of students scoring at the basic achievement level or higher on NAEP mathematics when the mediating achievement variables of socio economic status, English language proficiency, and expenditures per student are controlled.

Description of the Subjects

State level NAEP data served as the unit of analysis for this study ($N = 44$). Only states that reported the following to the National Center for Educational Statistics were used in the data analysis; (a) percentage of students eligible for free or reduced lunches, (b) percentage of students placed in limited-English proficiency programs, and (c) the state wide average of expenditures per pupil.

Procedures

The researcher collected the following data from the National Center for Educational Statistics (NCES) website: (a) the average scale score and percentage of student scoring at the basic achievement level or above for each state on the 2005 NAEP fourth grade reading test, the 2005 NAEP eighth grade reading test, the 2005 NAEP fourth grade mathematics test, and the 2005 NAEP eighth grade mathematics test; (b) the percentage of students scoring at the basic achievement level or above on the 2003 NAEP mathematics and reading tests for both fourth and eighth grades; and (c) the percentage of students eligible for free or reduced lunch, the percentage of students identified as limited English proficient, and the state wide average of expenditures per student for each state as reported for the 2005 NAEP administration. In addition, the nature of the testing stakes associated with each state's assessment program as of 1999, 2002, and 2005 was collected from the annual Quality Counts report compiled by *Education Week* and the NCES website.

Since the nature of accountability measures reported by the states have evolved over the span of the six years included in the study, the accountability stakes used for each of the benchmark years varied somewhat. For the year 1999, the following state imposed stakes were utilized to define accountability pressure; (a) high school graduation was contingent upon performance on state assessment, (b) high

performing districts were rewarded, (c) high performing schools were rewarded, (d) low performing districts were sanctioned, (e) low performing schools were sanctioned, (f) sanctions included written warnings, (g) sanctions included probationary status for schools or districts, (h) sanctions included loss of accreditation for school or districts, (i) sanctions included loss of funding, (j) sanctions included school reconstitution, (k) sanctions included school closure, and (l) sanctions included school take over by the state.

For the year 2002, the following state imposed stakes were utilized to define accountability pressure; (a) state required school-level report cards, (b) student performance data was disaggregated by race, (c) student performance data was disaggregated by poverty, (d) student performance data was disaggregated by limited English proficiency, (e) student performance data was disaggregated by special education status, (f) the state assigned ratings to all schools or identified low performing schools, (g) sanctions included school closure, (h) sanctions included school reconstitution, (i) sanctions included permitting student transfers, (j) sanctions included turning low performing school over to private management, (k) sanctions included loss of funding, (l) high performing or improved schools were rewarded, (m) at least one grade promotion was contingent upon performance on state assessment, and (n) high school graduation was contingent upon performance on state assessment.

For the year 2005, the following state imposed stakes were utilized to define accountability pressure; (a) the state sanctioned low performing schools, (b) the state rewarded high performing or improving schools, (c) sanctions included school closure, (d) sanctions included school reconstitution, (e) sanctions included permitting student transfers, (f) sanctions included turning school over to private management, (g) sanctions included loss of funding, (h) the state assigned ratings based on state developed criteria in addition to adequate yearly progress, (i) student performance data was disaggregated by race, (j) student performance data was disaggregated by poverty, (k) student performance data was disaggregated by limited English proficiency, (l) student performance data was disaggregated by special

education status, (m) at least one grade promotion was contingent upon performance on state assessment, and (n) high school graduation was contingent upon performance on state assessment.

An electronic spreadsheet was used to total the values for each of the study variables. The variable totals were entered into the Statistical Package for the Social Sciences (SPSS) software to conduct the statistical tests with alpha set at .05.

Study Design

The study utilized a correlational design with multiple regression used as the method of statistical analysis. The various NAEP achievement measures served as the criterion variable while the following served as the predictor variables; (a) the percentage of students living in poverty, (b) the percentage of students identified as English language learners, (c) the expenditures per student adjusted by the comparable wage index (CWI), and (d) the HSTV. The adjusted expenditures per student were reduced to expenditures per hundred dollars to facilitate data reporting. Student poverty, as measured by the percentage of students on free and reduced lunch, English language learners, and expenditures per student were included as control variables.

The following multiple regression equation was utilized to test the research question and null hypotheses: $Y' = a + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4$ where Y' is the predicted value of NAEP achievement, a is the Y intercept, b_k is the unstandardized coefficient for the predictor variable obtained from the regression analysis, and X_k is the raw value for a predictor variable. A hierarchical multiple regression was conducted in order to determine the statistical significance of each predictor variable in the linear equation and the predictive potency that the HSTV adds to the regression model (Mertler and Vannatta, 2005). The HSTV was added as the fourth predictor variable in the full model in order to isolate the impact it had on R^2. Two separate hierarchical multiple regressions analyses were conducted for each of the null hypotheses with each multiple regression analysis utilizing a different grade level NAEP assessment as the criterion variable.

The expenditures per student for each state was adjusted by the Comparable Wage Index (CWI) in recognition of the wide geographic variation in the costs of educational resources. The adjustment to the expenditure per student variable was made by dividing the state's reported expenditures by the state's most recent (1999) CWI value and dividing that value by one hundred. For example, Texas reported state expenditures per student of $7,271. With a reported CWI value of 1.05, the adjusted expenditures per student for Texas becomes $6,924 (7,271 / 1.05) divided by 100 or 69.24.

Results

Null Hypothesis One

Null hypothesis one stated that the HSTV, as a predictor variable, will not add to the potency of predicting NAEP reading scale scores when the other predictor variables are controlled. The descriptive statistics for all the variables used in null hypothesis one are detailed in table 1.

Table 1

Descriptive Statistics for Null Hypothesis One Variables

Variable	Mean	*S D*	*N*
Fourth grade reading NAEP scale score	218.45	6.74	44
Eighth grade reading NAEP scale score	261.80	6.51	44
Percent of ELL students (ELL)	6.54	5.71	44
Percent of low socio-economic students (SES)	38.52	10.75	44
Adjusted expenditures per student ($)	88.76	17.03	44
High Stakes Testing Value (HSTV)	15.61	7.57	44

Fourth grade NAEP reading. Analysis of the fourth grade NAEP reading data revealed that both models significantly predicted the states' scale scores. The results for models one and two respectively were as follows: $R^2 = .644$, $R^2_{adj} = .617$, $F(3, 40,) = 24.120$, $p<.001$; $R^2 = .650$, $R^2_{adj} = .614$, $F(4, 39,) = 18.127$, $p<.001$. However, the analysis also determined that model two failed to significantly increase R^2 beyond model one, $\Delta R^2 = .006$, $F_{change}(1, 39,) = .698$, $p = .409$. Additionally, the regression analysis revealed that the ELL and SES variables were the only two significant predictor variables. Summaries for regression models one and two are shown in table 2; the change statistics summary is delineated in table

3. Bivariate and partial correlation coefficients between each predictor variable and the criterion variable are shown in tables 4 and 5 for models one and two respectively.

Table 2

Regression Summary for Models One and Two

Model	R	R^2	R^2_{adj}	F	df_1	df_2
One	.802	.644	.617	24.120**	3	40
Two	.806	.650	.614	18.325**	4	39

** Indicates significance at $p<.001$

Table 3

Model One to Model Two Change Statistics

ΔR^2	F_{change}	df_1	df_2
.006	.978+	1	39

+ Indicates lack of significance

Table 4

Regression Coefficients for Model One

Predictor Variable	B	β	t	Bivariate r	Partial r
ELL	-.403	-.342	-3.369*	-.443	-.470
SES	-.425	-.678	-6.214**	-.728	-.701
$	-.001	-.003	-.024+	.462	-.004

+ Indicates lack of significance
* Indicates significance at $p<.05$
** Indicates significance at $p<.001$

Table 5

Regression Coefficients for Model Two

Predictor Variable	*B*	*β*	*t*	Bivariate *r*	Partial *r*
ELL	-.411	-.348	-3.410*	-.443	-.479
SES	-.459	-.732	-5.758**	-.728	-.678
$	-.005	-.013	-.111+	.462	-.018
HSTV	.084	.094	.989+	-.332	.133

+ Indicates lack of significance
* Indicates significance at $p<.05$
** Indicates significance at $p<.001$

Eighth grade NAEP reading. Analysis of the eighth grade NAEP reading scale score data yielded results similar to the fourth grade NAEP reading analysis. Both regression models significantly predicted the states' scale scores. Model one and model two respectively produced the following: $R^2 = .674$, $R^2_{adj} = .650$, $F(3, 40,) = 27.600$, $p<.001$; $R^2 = .686$, $R^2_{adj} = .653$, $F(4, 39,) = 21.272$, $p<.001$. Model two failed to significantly increase R^2 beyond model one, $\Delta R^2 = .011$, $F_{change}(1, 39,) = 1.421$, $p = .241$. Additionally, just as the in the fourth grade analysis, the ELL and SES variables were the only two significant predictor variables. Summaries for regression models one and two are shown in table 6; the change statistics summary is delineated in table 7. Bivariate and partial correlation coefficients between each predictor variable and the criterion variable are shown in tables 8 and 9 for models one and two respectively. Based on the failure of the HSTV to significantly increase R^2 in either grade level analysis, null hypothesis one was not rejected.

Table 6

Regression Summary for Models One and Two

Model	*R*	R^2	R^2_{adj}	*F*	df_1	df_2
One	.821	.674	.650	27.600**	3	40
Two	.828	.686	.653	21.272**	4	39

** Indicates significance at $p<.001$

Table 7

Model One to Model Two Change Statistics

ΔR^2	F_{change}	df_1	df_2
.011	1.421+	1	39

+ Indicates lack of significance

Table 8

Regression Coefficients for Model One

Predictor Variable	*B*	β	*t*	Bivariate *r*	Partial *r*
ELL	-.356	-.312	-3.218*	-.432	-.453
SES	-.415	-.685	-6.558**	-.755	-.720
$	.017	.045	.409+	.503	.064

+ Indicates lack of significance
* Indicates significance at $p<.05$
** Indicates significance at $p<.001$

Table 9

Regression Coefficients for Model Two

Predictor Variable	*B*	β	*t*	Bivariate *r*	Partial *r*
ELL	-.346	-.303	-3.135*	-.432	-.449
SES	-.370	-.612	-5.076**	-.755	-.631
$	.023	.059	.532+	.503	.085
HSTV	-.109	-.127	-1.192+	-.499	-.187

+ Indicates lack of significance
* Indicates significance at $p<.05$
** Indicates significance at $p<.001$

Null Hypothesis Two

Null hypothesis two stated that the HSTV as a predictor variable will not add to the potency of predicting NAEP mathematics scale scores when the other predictor variables are controlled. The descriptive statistics for all the variables used in null hypothesis two are presented in table 10.

Table 10

Descriptive Statistics for Null Hypothesis Two Variables

Variable	Mean	*S D*	*N*
Fourth grade math NAEP scale score	237.80	5.78	44
Eighth grade math NAEP scale score	278.43	7.60	44
Percent of ELL students (ELL)	6.54	5.71	44
Percent of low socio-economic students (SES)	38.52	10.75	44
Adjusted expenditures per student ($)	88.76	17.03	44
High Stakes Testing Value (HSTV)	15.61	7.57	44

Fourth grade NAEP mathematics. The analysis revealed that model one significantly predicted the states' fourth grade NAEP mathematics scale scores, $R^2 = .543$, $R^2_{adj} = .508$, $F(3, 40,) = 15.821$, $p<.001$; model two was also found to be statistically significant, $R^2 = .558$, $R^2_{adj} = .513$, $F(4, 39,) = 12.324$, $p<.001$. Model two, however, failed to significantly increase R^2 beyond model one, $\Delta R^2 = .016$, $F_{change}(1, 39,) = 1.380$, $p = .247$. The analysis also revealed that the ELL and SES variables were the only two significant predictor variables. Summaries for regression models one and two are shown in table 11 and the change statistics summary is delineated in table 12. Bivariate and partial correlation coefficients between each predictor variable and the criterion variable are shown in tables 13 and 14 for models one and two respectively.

Table 11

Regression Summary for Models One and Two

Model	*R*	R^2	R^2_{adj}	*F*	df_1	df_2
One	.737	.543	.508	15.821**	3	40
Two	.747	.558	.513	12.324**	4	39

** Indicates significance at $p<.001$

Table 12

Model One to Model Two Change Statistics

ΔR^2	F_{change}	df_1	df_2
.016	1.380+	1	39

+ Indicates lack of significance

Table 13

Regression Coefficients for Model One

Predictor Variable	*B*	β	*t*	Bivariate *r*	Partial *r*
ELL	-.271	-.267	-2.327*	-.338	-.345
SES	-.377	-.701	-5.668**	-.693	-.667
$	-.033	-.096	-.732+	.353	-.115

+ Indicates lack of significance

Table 14

Regression Coefficients for Model Two

Predictor Variable	*B*	β	*t*	Bivariate *r*	Partial *r*
ELL	-.281	-.278	-2.420*	-.338	-.361
SES	-.423	-.786	-5.503**	-.693	-.661
$	-.038	-.112	-.853+	.353	-.135
HSTV	.113	.148	1.175+	-.278	.185

+ Indicates lack of significance
* Indicates significance at $p<.05$
** Indicates significance at $p<.001$

Eighth grade NAEP mathematics. Analysis of the eighth grade NAEP mathematics data revealed that both the restricted and full models significantly predicted the states' scale scores. The results for model one and two respectively were as follows: $R^2 = .560$, $R^2_{adj} = .527$, $F(3, 40,) = 16.979$, $p<.001$; $R^2 = .561$, $R^2_{adj} = .516$, $F(4, 39,) = 12.451$, $p<.001$. The analysis also determined that the full model failed to significantly increase R^2 beyond model one, $\Delta R^2 = .001$, $F_{change}(1, 39,) = .061$, $p = .807$. Additionally, the SES variable was found to be the only significant predictor variable. Summaries for regression models

one and two are shown in table 15; the change statistics summary is delineated in table 16. Bivariate and partial correlation coefficients between each predictor variable and the criterion variable are shown in tables 17 and 18 for models one and two respectively. Based on the failure of the HSTV to significantly increase R^2 in either grade level analysis, null hypothesis two was not rejected.

Table 15

Regression Summary for Models One and Two

Model	R	R^2	R^2_{adj}	F	df_1	df_2
One	.748	.560	.527	16.979**	3	40
Two	.749	.561	.516	12.451**	4	39

** Indicates significance at $p<.001$

Table 16

Model One to Model Two Change Statistics

ΔR^2	F_{change}	df_1	df_2
.001	.061+	1	39

+ Indicates lack of significance

Table 17

Regression Coefficients for Model One

Predictor Variable	B	β	t	Bivariate r	Partial r
ELL	-.220	-.165	-1.466+	-.251	-.226
SES	-.526	-.744	-6.132**	-.732	-.696
$	-.032	-.073	-.564+	.361	-.089

+ Indicates lack of significance
** Indicates significance at p<.001

Table 18

Regression Coefficients for Model Two

Predictor Variable	B	β	t	Bivariate r	Partial r
ELL	-.223	-.167	-1.463+	-.251	-.228
SES	-.539	-.762	-5.347**	-.732	-.696
$	-.034	-.076	-.580+	.361	-.092
HSTV	.031	.031	.246+	-.377	.039

+ Indicates lack of significance
** Indicates significance at $p<.001$

Null Hypothesis Three

Null hypothesis three stated that the HSTV, as a predictor variable, will not add to the potency of predicting the percentage of students scoring at the basic achievement level or higher on NAEP reading assessments. The descriptive statistics for all the variables used in null hypothesis three are illustrated in table 19.

Table 19

Descriptive Statistics for Null Hypothesis Three Variables

Variable	Mean	*S D*	*N*
Percent of students scoring at basic level or above on NAEP fourth grade reading test	64.14	7.56	44
Percent of students scoring at basic level or above on NAEP eighth grade reading test	72.75	7.05	44
Percent of ELL students (ELL)	6.54	5.71	44
Percent of low socio-economic students (SES)	38.52	10.75	44
Adjusted expenditures per student ($)	88.76	17.03	44
High Stakes Testing Value (HSTV)	15.61	7.57	44

Fourth grade NAEP reading. Analysis of the fourth grade NAEP reading data revealed that both models significantly predicted the percentage of students scoring at the basic achievement level or higher. The results for models one and two respectively were as follows: $R^2 = .657$, $R^2_{adj} = .631$, $F(3, 40,) = 25.529$, $p<.001$; $R^2 = .658$, $R^2_{adj} = .623$, $F(4, 39,) = 18.762$, $p<.001$. The analysis also determined that model two failed to significantly increase R^2 beyond model one, $\Delta R^2 = .001$, $F_{change}(1, 39,) = .129$, $p =$

.722. Additionally, the regression analysis revealed that the ELL and SES variables were the only two significant predictor variables. Summaries for regression models one and two are shown in table 20 and the change statistics summary is delineated in table 21. Bivariate and partial correlation coefficients between each predictor variable and the criterion variable are shown in tables 22 and 23 for models one and two respectively.

Table 20

Regression Summary for Models One and Two

Model	R	R^2	R^2_{adj}	F	df_1	df_2
One	.810	.657	.631	25.529**	3	40
Two	.811	.658	.623	18.762**	4	39

** Indicates significance at $p<.001$

Table 21

Model One to Model Two Change Statistics

ΔR^2	F_{change}	df_1	df_2
.001	.129+	1	39

+ Indicates lack of significance

Table 22

Regression Coefficients for Model One

Predictor Variable	B	β	t	Bivariate r	Partial r
ELL	-.457	-.346	-3.472*	-.442	-.481
SES	-.490	-.698	-6.511**	-.738	-.717
$	-.011	-.024	-.208+	.452	-.033

+ Indicates lack of significance
* Indicates significance at $p<.05$
** Indicates significance at $p<.001$

Table 23

Regression Coefficients for Model Two

Predictor Variable	B	β	t	Bivariate r	Partial r
ELL	-.461	-.348	-3.451*	-.442	-.484
SES	-.507	-.720	-5.731**	-.738	-.679
$	-.013	-.030	-.261+	.452	-.039
HSTV	.040	.040	.359+	-.377	.057

+ Indicates lack of significance
* Indicates significance at $p<.05$
** Indicates significance at $p<.001$

Eighth grade NAEP reading. Analysis of the eighth grade NAEP reading data paralleled the fourth grade regression results. Both eighth grade regression models significantly predicted the percentage of students scoring at the basic achievement level or above. Model one and model two respectively yielded the following: $R^2 = .617$, $R^2_{adj} = .589$, $F(3, 40,) = 21.503$, $p<.001$; $R^2 = .642$, $R^2_{adj} = .605$, $F(4, 39,) = 17.458$, $p<.001$. The full model failed to significantly increase R^2 beyond model one, $\Delta R^2 = .024$, $F_{change}(1, 39,) = 2.655$, $p = .111$. Additionally, just as the in the fourth grade analyses, the eight grade analysis revealed that the ELL and SES variables were the only two significant predictor variables. Summaries for regression models one and two are shown in table 24; the change statistics summary is delineated in table 25. Bivariate and partial correlation coefficients between each predictor variable and the criterion variable are delineated in tables 26 and 27 for models one and two respectively. Based on the failure of the HSTV to significantly increase R^2 in either grade level analysis, null hypothesis three was not rejected.

Table 24

Regression Summary for Models One and Two

Model	R	R^2	R^2_{adj}	F	df_1	df_2
One	.786	.617	.589	21.503**	3	40
Two	.801	.642	.605	17.458**	4	39

** Indicates significance at $p<.001$

Table 25

Model One to Model Two Change Statistic

ΔR^2	F_{change}	df_1	df_2
.024	2.655+	1	39

+ Indicates lack of significance

Table 26

Regression Coefficients for Model One

Predictor Variable	*B*	β	*t*	Bivariate *r*	Partial *r*
ELL	-.389	-.315	-2.997*	-.429	-.428
SES	-.422	-.643	-5.685**	-.714	-.688
$	.019	.046	.387+	.484	.061

+ Indicates lack of significance
* Indicates significance at $p<.05$
** Indicates significance at $p<.001$

Table 27

Regression Coefficients for Model Two

Predictor Variable	*B*	β	*t*	Bivariate *r*	Partial *r*
ELL	-.373	-.302	-2.927*	-.429	-.424
SES	-.352	-.537	-4.172**	-.714	-.555
$	.028	.067	.562+	.484	.090
HSTV	-.173	-.185	-1.629+	-.519	-.252

+ Indicates lack of significance
* Indicates significance at $p<.05$
** Indicates significance at $p<.001$

Null Hypothesis Four

Null hypothesis four stated that the HSTV, as a predictor variable, will not enhance predicting the percentage of students scoring at the basic achievement level or higher or NAEP mathematics assessments. The descriptive statistics for all the variables used in null hypothesis four are illustrated in table 28.

Table 28

Descriptive Statistics for Null Hypothesis Four Variables

Variable	Mean	*S D*	*N*
Percent of students scoring at basic level or above on NAEP fourth grade mathematics test	80.61	6.39	44
Percent of students scoring at basic level or above on NAEP eighth grade mathematics test	69.34	8.01	44
Percent of ELL students (ELL)	6.54	5.71	44
Percent of low socio-economic students (SES)	38.52	10.75	44
Adjusted expenditures per student ($)	88.76	17.03	44
High Stakes Testing Value (HSTV)	15.61	7.57	44

Fourth grade NAEP mathematics. The multiple regression analysis showed that model one significantly predicted the percentage of students scoring at the basic achievement level or above, R^2 = .519, R^2_{adj} = .483, $F(3, 40,) = 14.384$, $p<.001$; model two was also found to be statistically significant, R^2 = .532, R^2_{adj} = .484, $F(4, 39,) = 11.068$, $p<.001$. However, the analysis also determined that model two failed to significantly increase R^2 beyond model one, $\Delta R^2 = .013$, $F_{change}(1, 39,) = 1.058$, $p = .310$. Additionally, the ELL and SES variables were found to be the only two significant predictor variables. Summaries for regression models one and two are shown in table 29 while table 30 delineates the change statistics summary. Bivariate and partial correlation coefficients between each predictor variable and the criterion variable are shown in tables 31 and 32 for models one and two respectively.

Table 29

Regression Summary for Models One and Two

Model	*R*	R^2	R^2_{adj}	*F*	df_1	df_2
One	.720	.519	.483	14.385**	3	40
Two	.729	.532	.484	11.068**	4	39

** Indicates significance at $p<.001$

Table 30

Model One to Model Two Change Statistics

ΔR^2	F_{change}	df_1	df_2
.013	1.058+	1	39

+ Indicates lack of significance

Table 31

Regression Coefficients for Model One

Predictor Variable	*B*	*β*	*t*	Bivariate *r*	Partial *r*
ELL	-.363	-.324	-2.750*	-.384	-.399
SES	-.392	-.660	-5.206**	-.654	-.636
$	-.041	-.110	-.818+	.339	-.128

+ Indicates lack of significance
* Indicates significance at $p<.05$
** Indicates significance at $p<.001$

Table 32

Regression Coefficients for Model Two

Predictor Variable	*B*	*β*	*t*	Bivariate *r*	Partial *r*
ELL	-.373	-.333	-2.822*	-.384	-.412
SES	-.438	-.737	-5.011**	-.654	-.626
$	-.047	-.125	-.921+	.339	-.146
HSTV	.113	.134	1.029+	-.270	.163

+ Indicates lack of significance
* Indicates significance at $p<.05$
** Indicates significance at $p<.001$

Eighth grade NAEP mathematics. Analysis of the eighth grade NAEP mathematics data also revealed that both models significantly predicted the percentage of students scoring at the basic achievement level or above. The results for models one and two respectively were as follows: $R^2 = .559$, $R^2_{adj} = .526$, $F(3, 40,) = 16.882$, $p<.001$; $R^2 = .562$, $R^2_{adj} = .517$, $F(4, 39,) = 12.494$, $p<.001$. As in the fourth grade analysis, model two failed to significantly increase R^2 beyond model one, $\Delta R^2 = .003$,

$F_{change}(1, 39,) = .264$, $p = .610$. Unlike the fourth grade analysis, the SES variable was the only significant predictor variable. Summaries for regression models one and two are shown in table 33 while the change statistics summary is delineated in table 34. Bivariate and partial correlation coefficients between each predictor variable and the criterion variable are shown in tables 35 and 36 for models one and two respectively. Based on the failure of the HSTV to significantly increase R^2 in either grade level analysis, null hypothesis four was not rejected.

Table 33

Regression Summary for Models One and Two

Model	R	R^2	R^2_{adj}	F	df_1	df_2
One	.747	.559	.526	16.882**	3	40
Two	.749	.562	.517	12.494**	4	39

** Indicates significance at $p<.001$

Table 34

Model One to Model Two Change Statistics

ΔR^2	F_{change}	df_1	df_2
.003	.264+	1	39

+ Indicates lack of significance

Table 35

Regression Coefficients for Model One

Predictor Variable	B	β	t	Bivariate r	Partial r
ELL	-.320	-.228	-2.018+	-.309	-.304
SES	-.536	-.720	-5.922**	-.717	-.683
$	-.035	-.074	-.576+	.370	-.091

+ Indicates lack of significance
** Indicates significance at $p<.001$

Table 36

Regression Coefficients for Model Two

Predictor Variable	*B*	*β*	*t*	Bivariate *r*	Partial *r*
ELL	-.314	-.223	-1.955+	-.309	-.299
SES	-.509	-.682	-4.795**	-.717	-.609
$	-.032	-.067	-.514+	.370	-.082
HSTV	-.069	-.065	-.514+	-.439	-.082

+ Indicates lack of significance
* Indicates significance at $p<.05$
** Indicates significance at $p<.001$

Null Hypothesis Five

Null hypothesis five predicted that the HSTV, as a predictor variable, will not add to the potency of forecasting the gains in the percentage of students scoring at the basic achievement level or higher for NAEP reading assessments. The descriptive statistics for all the variables used in null hypothesis five are illustrated in table 37.

Table 37

Descriptive Statistics for Null Hypothesis Five Variables

Variable	Mean	*S D*	*N*
Gains in the percentage of students scoring at basic level or above on NAEP fourth grade reading test	.61	2.34	44
Gains in the percentage of students scoring at basic level or above on NAEP eighth grade reading test	-.84	2.05	44
Percent of ELL students (ELL)	6.54	5.71	44
Percent of low socio-economic students (SES)	38.52	10.75	44
Adjusted expenditures per student ($)	88.76	17.03	44
High Stakes Testing Value (HSTV)	15.61	7.57	44

Fourth grade NAEP reading. Analysis of the fourth grade NAEP reading data revealed that neither model significantly predicted the gains in the percentage of students scoring at the basic achievement level or higher. The results for models one and two respectively were as follows: $R^2 = .025$, $R^2_{adj} = -.048$, $F(3, 40,) = .341$, $p = .796$; $R^2 = .032$, $R^2_{adj} = -.067$, $F(4, 39,) = .327$, $p = .858$. Summaries for

regression models one and two are shown in table 38. Since neither model proved significant, the change statistics and regression coefficients were not relevant.

Table 38

Regression Summary for Models One and Two

Model	R	R^2	R^2_{adj}	F	df_1	df_2
One	.158	.025	-.048	.341+	3	40
Two	.180	.032	-.067	.327+	4	39

+ Indicates lack of significance

Eighth grade NAEP reading. Analysis of the eighth grade NAEP reading data departed from the fourth grade analysis with both regression models significantly predicting the gains in the percentage of students scoring at the basic achievement level or above. Model one and model two respectively yielded the following: $R^2 = .247$, $R^2_{adj} = .191$, $F(3, 40,) = 4.381$, $p = .009$; $R^2 = .285$, $R^2_{adj} = .211$, $F(4, 39,) = 3.882$, $p = .010$. However, model two failed to significantly increase R^2 beyond model one, $\Delta R^2 = .037$, $F_{change}(1, 39,) = 2.042$, $p = .161$. The regression analysis also revealed that the expenditures per student variable was the only significant predictor variable. Summaries for regression models one and two are shown in table 39; the change statistics summary is delineated in table 40. Bivariate and partial correlation coefficients between each predictor variable and the criterion variable are shown in tables 41 and 42 for models one and two respectively. Based on the failure of the HSTV to significantly increase R^2 in either grade level analysis, null hypothesis five was not rejected.

Table 39

Regression Summary for Models One and Two

Model	R	R^2	R^2_{adj}	F	df_1	df_2
One	.497	.247	.191	4.381*	3	40
Two	.543	.285	.211	3.882*	4	39

* Indicates significance at $p<.05$

Table 40

Model One to Model Two Change Statistic

ΔR^2	F_{change}	df_1	df_2
.037	2.042+	1	39

+ Indicates lack of significance

Table 41

Regression Coefficients for Model One

Predictor Variable	*B*	β	*t*	Bivariate *r*	Partial *r*
ELL	.091	.255	1.726+	.090	.263
SES	-.046	-.243	-1.533+	-.380	-.236
$	.042	.349	2.071*	.378	.311

+ Indicates lack of significance
* Indicates significance at $p<.05$

Table 42

Regression Coefficients for Model Two

Predictor Variable	*B*	β	*t*	Bivariate *r*	Partial *r*
ELL	.097	.270	1.852+	.090	.284
SES	-.021	-.111	-.613+	-.380	-.098
$	.045	.374	2.235*	.378	.337
HSTV	-.062	-.230	-1.429+	-.334	-.223

+ Indicates lack of significance
* Indicates significance at $p<.05$

Null Hypothesis Six

Null hypothesis six stated that the HSTV, as a predictor variable, will not add to the potency of predicting the gains in the percentage of students scoring at the basic achievement level or higher for NAEP mathematics assessments. The descriptive statistics for all the variables used in null hypothesis six are illustrated in table 43.

Table 43

Descriptive Statistics for Null Hypothesis Six Variables

Variable	Mean	*S D*	*N*
Gains in the percentage of students scoring at basic level or above on NAEP fourth grade mathematics test	3.14	2.28	44
Gains in the percentage of students scoring at basic level or above on NAEP eighth grade mathematics test	.64	2.28	44
Percent of ELL students (ELL)	6.54	5.71	44
Percent of low socio-economic students (SES)	38.52	10.75	44
Adjusted expenditures per student ($)	88.76	17.03	44
High Stakes Testing Value (HSTV)	15.61	7.57	44

Fourth grade NAEP mathematics. The analysis revealed that neither of the two models significantly predicted the gains in the percentage of students scoring at the basic achievement level or above, $R^2 = .046$, $R^2_{adj} = -.025$, $F(3, 40,) = .644$, $p = .591$ (model one); $R^2 = .101$, $R^2_{adj} = .009$, $F(4, 39,) = 1.099$, $p = .371$ (model two). Summaries for regression models one and two are shown in table 44. Since neither model proved significant, the change statistics and regression coefficients were not relevant.

Table 44

Regression Summary for Models One and Two

Model	*R*	R^2	R^2_{adj}	*F*	df_1	df_2
One	.215	.046	-.025	.644+	3	40
Two	.318	.101	.009	1.099+	4	39

+ Indicates lack of significance

Eighth grade NAEP mathematics. As in the fourth grade analysis, the eighth grade NAEP mathematics data revealed that neither of the two models significantly predicted the gains in the percentage of students scoring at the basic achievement level or above. The results for models one and two respectively were as follows: $R^2 = .154$, $R^2_{adj} = .090$, $F(3, 40,) = 2.421$, $p = .080$; $R^2 = .157$, $R^2_{adj} = .071$, $F(4, 39,) = 1.820$, $p = 1.45$. Summaries for regression models one and two are shown in table 45. Since neither model proved significant, the change statistics and regression coefficients were not relevant.

Based on the failure of the HSTV to significantly increase R^2 in either grade level analysis, null hypothesis six was not rejected.

Table 45

Regression Summary for Models One and Two

Model	R	R^2	R^2_{adj}	F	df_1	df_2
One	.392	.154	.090	2.421+	3	40
Two	.397	.157	.071	1.820+	4	39

+ Indicates lack of significance

Summary, Conclusions, and Recommendations

The original research question that led to the development of this study asked if the HSTV added to the potency of predicting NAEP reading and mathematics achievement when the effects of socio economic status, English language proficiency, and expenditures per student were controlled. The results of the study raise serious questions about the efficacy of the conceptual framework supporting it and suggest that the number of testing stakes is not a significant factor in regards to student achievement.

By extension, the results of this study challenge the efficacy of high stakes testing as a school reform policy. If accountability pressure results in genuine student learning, learning that transcends the narrow confines of a state's high stakes assessment, higher number of testing stakes could reasonably be expected to positively correlate with NAEP achievement. However, the results of this study provided no evidence of any such correlation.

Numerous researchers have concluded that students' socio economic status and English-language proficiency are two of the most significant factors influencing student achievement on standardized tests (English and Steffy, 2001; Jones, Jones, and Hargrove, 2003; Payne and Biddle, 1999; Petterson, Kupersmidt, and Vaden, 1990; Gitomer, Anadal, and Davision, 2005; Abedi and Leon, 1999; and Abedi and Dietel, 2004). The fact that socio economic status was a significant predictor variable in eight of the nine significant full regression models and that English-language proficiency was significant as a

predictor variable in six of the nine significant full regression models provides support for those conclusions.

Perhaps more revealing was the fact that the HSTV failed to produce significant predictor variable status in any of the full regression models while expenditures per student produced significant predictor variable status in only one of the full regression models. Of course, the failure of the HSTV to produce significant predictor variable status in any of the full regression models also resulted in all six null hypotheses being retained since the change in R^2 from the restricted model to the full model was also not significant. In other words, the HSTV failed to add to the predictive potency of any of the 12 hierarchal multiple regression analyses suggesting that states with low HSTVs and states with high HSTVs scored equally well on all 12 of the NAEP achievement measures.

In addition, a review of the correlation matrices for the 12 hierarchical regressions revealed the following bivariate correlations: The SES predictor variable showed a significant negative correlation with NAEP achievement ($p < .05$, one-tailed) in 9 of 12 instances with the Pearson r value ranging from -.380 to -.755 and the ELL predictor variable yielded a significant negative correlation with NAEP achievement in 7 of 12 instances ($p < .05$, one-tailed) with Pearson r varying from -.309 to -.443. This combination of results suggest a strong link between low percentages of students from poverty, low percentages of students with limited English language proficiency, and increased achievement on NAEP assessments.

On the surface, the results of this study seem to suggest that accountability pressure has no impact on student achievement and appear to support the primary finding of Coleman's seminal report on student achievement penned in 1966; "The social composition of the student body is more highly related to achievement…than is any school factor" (cited in Kahlenberg, 2001, The 1966 Coleman Report section ¶ 3). High stakes testing dissenters, relying on such findings, argued that high stakes accountability schemes are unlikely to reverse the social reality suggested by decades of educational research; students come from various backgrounds offering some distinct educational advantages over others (Jones, Jones, and

Hargrove, 2003; English and Steffy, 2001). This social reality, according to dissenters, ensures that students will continue to enter America's classrooms at different points along the school readiness continuum and, therefore, exit those classrooms at different points along the content mastery continuum regardless of well intended accountability policies (Mathis, 2004).

However, when the totality of data produced by this study are assessed and interpreted beyond the "yes or no" response needed to answer the null hypotheses, a contradictory perspective begins to emerge. As was mentioned previously, both the SES and ELL predictor variables were shown to be negatively correlated with NAEP achievement. When the relative importance of the SES variable was compared to the relative importance of the ELL variable (calculated by taking the ratio of the squares of the respective beta weights), the SES variable was found to account for over three times as much of the variance in the criterion variable as the ELL variable in instances where both variables yielded significant predictor status (Kachigan, 1991). In short, the SES predictor variable accounted for the overwhelming majority of the variance in the criterion variable in eight of the nine significant full regression models.

A review of the correlation matrices also revealed that the SES predictor variable was positively correlated with the HSTV predictor variable ($r = .530$, $p = .017$, one-tailed), meaning that states with a high HSTV also tended to be states with a high population of low socio economic status students. In essence, states with high HSTVs tended to teach higher percentages of students from poverty and still produced comparable scores on NAEP achievement when compared to states with low HSTVs and, by implication, lower populations of low socio economic students.

Given that states with high HSTVs tended to educate high percentages of students from poverty and that NAEP scores from those states compared favorably to states with lower percentages of students from poverty, the argument can be made that accountability pressure made some positive impact for states with greater number of testing stakes. Put in more simplistic terms, high stakes testing states "tied" low stakes testing states in NAEP achievement even though states with low HSTVs were afforded a "head start" by virtue of the their higher performing student demographics. So, while the accountability pressure

associated with high stakes testing may not have resulted in higher achievement for low SES students when compared to high SES students, it may have resulted in increased achievement for students from impoverished home environments.

The overarching conclusion arising from the results of this study provides some fodder for those on either side of the high stakes testing debate. On one hand, states with more rigorous high stakes testing policies appeared to perform no better on NAEP than states with less rigorous high stakes testing policies. On the other hand, states with more rigorous high stakes testing policies appeared to perform just as well on NAEP as states with less rigorous accountability schemes even though states with higher HSTVs tended to serve harder to educate students.

The mix of conclusions arising from this study parallels the contradiction of conclusions found in the high stakes testing literature. Like the six blind men of Indostan who interpreted the nature of the elephant in differing yet accurate ways, conflicting conclusions regarding the effectiveness of high stakes testing are defensible depending upon the data collected and how it is interpreted (Saxe, n.d.). The complexity of the high stakes testing question transcends answers provided by a single study revealing only one facet of the issue. Just as the six blind men would have had a more accurate understanding of the elephant had they been willing to extend beyond their limited perspectives, educators and policy makers should continue to explore all aspects of the issue in an unbiased fashion before forming a simple yea or nay conclusion.

Is high stakes testing an effective school reform policy that warrants its prominent position as the centerpiece of today's school reform movement? The answer to that question, as addressed by the results of this study, appears to be inconclusive and dependent upon each state's student demographics; high poverty states seem to benefit from such policies while states with more wealthy demographics do not. Perhaps the true value of high stakes testing as a school reform policy lies somewhere between the bipolar extremes of curse or cure; producing beneficial results in some instances and negative unintended consequences in others (Amrein and Berliner, 2002b; Jones, Jones, and Hargrove, 2003).

References

Abedi, J, & Dietel, R. (2004). *Challenges in the No Child Left Behind Act for English language learners*. University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J. & Leon, S. (1999) *Impact of students' language background on content-based performance: Analyses of extant data.* Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Amrein, A. & Berliner, D. (2002a). High-stakes testing, uncertainty, and student learning. *Education Policy and Analysis Archives, 10*(18).

Amrein, A. & Berliner, D. (2002b). An analysis of some unintended and negative consequences of high-stakes testing. *Education Policy Studies Laboratory.* Retrieved November 15, 2006 from http://epsl.asu.edu/epru/documents/EPSL-0211-125-EPRU.pdf

English, F. & Steffy, B. (2001). *Deep curriculum alignment*: *Creating a level playing field for all children on high-stakes tests of educational accountability.* Lanham, MD: Scarecrow Press.

Gitomer, D., Andal, J., & Davison, D. (2005). *Using data to understand the academic performance of English language learners*. Retrieved January 2, 2007 from http://www.ncrel.org/policy/pubs/pdfs/pivol21.pdf

Haney, W. (2000). The myth of the Texas miracle in education. *Education Analysis and Policy Archives, 8*(41). Retrieved November 14, 2005 from http://epaa.edu/epaa/v8n41/

Jones, G., Jones, B., & Hargrove, T. (2003). *The unintended consequences of high-stakes testing.* Lanham, MD: Rowman & Littlefield Publishers, Inc.

Kachigan, S. (1991). *Multivariate statistical analysis: A conceptual introduction.* New York: Radius Press.

Kahlenberg, R. (2001). Learning from James Coleman. Retrieved February 21, 2007 from www.findarticles.com/p/articles/mi_m0377/is_2001_Summer/ai_76812255/print

Madaus, G. (1998). The influence of testing on the curriculum. *Yearbook (National Society for the Study of Education, 98*(2), 73-111.

Mathis, W. (2004). NCLB and high-stakes accountability: A cure? Or a symptom of the disease? *Educational Horizons, 82,* 143-152.

McNeil, L., & Valenzuela, A. (2000). *The harmful impact of the TAAS system on testing in Texas: Beneath the accountability rhetoric.* Retrieved November 29, 2005 from http://www.law.harvard.edu/civilrights/conferences/ testing98/drafts/mcneil_valenzuela.html.

Mertler, C. & Vannatta, R. (2005). *Advanced and multivariate statistical methods: Practical application and interpretation.* Glendale, CA: Pyrczak Publishing.

Patterson, C., Kupersmidt, J., & Vaden, N. (1990). Income level, gender, ethnicity, and household composition as predictors of children's school-based competencies. *Child Development, 61*(2), 485-494.

Payne, K. & Biddle, B. (1999). Poor school funding, child poverty, and mathematics achievement. *Educational Researcher, 28*(6), 4-13.

Pedroza, B. (1997). *Consequences of high-stakes testing on historically disenfranchised students: An analysis of student outcomes*. Unpublished doctoral dissertation, University of Texas, Austin.

Popham, W. J. (2006). *Assessment for educational leaders*. New York: Pearson Education, Inc.

Saxe, J. (n.d.) *The blind men and the elephant.* Retrieved June 1, 2007 from http://cs.wwc.edu/~aabyan/Poetry/blindmen.html.

Smith, M. & Fey, P. (2000). Validity and accountability in high-stakes testing. *Journal of Teacher Education, 51*(5), 334-344.