Current Research in Psychology 1 (2): 82-101, 2010
ISSN 1949-0178
© 2010 Science Publications

Locally-Chosen versus State-Mandated Success Criteria: A Retrospective Evaluation of the Effectiveness of the Success-For-All Program with Fourth and Seventh Grade Hispanic Students

James Carifio and Dalis Dominguez
Graduate School of Education, University of Massachusetts-Lowell, USA

Abstract: Problem statement: This study compared the growth in reading competencies of Hispanic female and male students in the SFA program as measured by the Massachusetts Comprehensive Assessment System (MCAS) and the Scholastic Reading Inventory (SRI). **Approach:** The study was a retrospective study with several a priori hypotheses and examined logical comparison groups of 4th and 7th graders, who received 1 and 3 years of SFA reading instruction respectively. **Results:** This study found that the SFA recommended SRI test presented an over-rosy picture of students reading achievement levels and gains (as compared to the MCAS test) and masked important differential effects of the program. ANOVA results showed that female Hispanic students obtained significantly higher average scores on the MCAS than did male Hispanic students and that these differences between Hispanic female and male students were significantly larger in the seventh grade than in the fourth grade. These findings showed that the SFA program was, most probably, not very successful in developing the reading competencies of Hispanic male students beyond grade 4. **Conclusion:** The critical and unexamined issue of locally-selected versus state-mandated success criteria for evaluating high stakes programs is also discussed in this article.

Key words: Massachusetts Comprehensive Assessment System (MCAS), Scholastic Reading Inventory (SRI), Test of Basic Skills (CTBS), Pinellas Instructional Assessment Program (PIAP), English Language Arts (ELA), Hispanic students, retrospective evaluation, ANOVA results, state-mandated success criteria

INTRODUCTION

The Success for All (SFA) program is one of the most widely used whole-school reading programs that have been developed in the past decade and it is currently being used in over 1500 school districts nationally according to the Success-for-All Foundation. Success for All (SFA) is a school-wide research-based reform model program developed by Robert Slavin and his associates at Johns Hopkins University and the program is based on the assertion that all students can and must succeed in the early grades and succeed in reading in particular (Slavin, 1996). There has been recent on-going controversy about the Success-for-All Program and its primary competitor, Reading First, concerning the effectiveness of either program, as well as the comparative effectiveness of these two rivals in the public school marketplace (Glen, 2007). This study examined selected but critical aspects of this controversy and debate about the SFA Program, but its findings are relevant to the Reading First Program also.

This study also addressed a major issue in educational research, program evaluation and the evidence-based practice movement since the implementation of legislated educational reforms in the late 1990's. This core and critical issue is, "Which criteria should be used in designing and assessing the effectiveness of educational programs: Locally-or-research chosen criteria or state-mandated standards?" This thorny but quite important issue has been more or less avoided, if not ignored, by the educational and research communities, but it is a core and critical issue that needs to be addressed by all educational research and policy setting and evaluation efforts today in the current context of educational reform. This study addresses this latter issue in a formative manner that should help initiate further needed discussions and studies of this important issue as well.

The River City school district, the focus of the current study, completed the fifth year of implementing the SFA Reform Model in the 2005-2006 school year, making it a "rich data set' for examining several claims of the SFA program and a

Corresponding Author: James Carifio, Graduate School of Education, University of Massachusetts-Lowell, 87 Putnam Street, Watertown, MA 02472 Tel: 617-513-6279

number of rival hypotheses as well. The River City school system is a poor, urban and predominantly Hispanic school system in eastern Massachusetts.

Purpose: The primary purpose of this study was to compare the growth in reading competencies of Hispanic female and male students in the SFA program implemented in the River City schools as measured by the Massachusetts Comprehensive Assessment System (MCAS) and the Scholastic Reading Inventory (SRI). The SRI was the measure of reading competency recommended by SFA to evaluate the success of the SFA program. The SRI, therefore, is the locally-chosen (and sanctioned) success criterion for the program. The MCAS, on the other hand, is the annual state-mandated test of reading competencies, which measures the State's definition of reading competencies (the state-mandated and sanctioned success criterion for the program). These particular comparisons were chosen as foci, as the evaluation of the SFA program's effectiveness has only been done in terms of overall sample averages with no assessment of differential effects and only using locally-chosen (and recommended) success criterion such as the SRI and the Woodcock, as opposed to reform and standards based state-mandated definitions and assessments of reading competencies and achievement. The differential effects of the SFA program, therefore, needed to be assessed (was it really achieving success for all), as well as the effectiveness of the SFA program relative to state-mandate success criteria such as the MCAS. This point is not only true of the SFA programs but all programs in this age of reform and to some degree all educational research and educational research studies as questions regarding the external validities of such studies cannot be ignored now that there are state standards and evidence-based practice movements around them.

The community and subjects: Table 1 presents a profile of selected social and economic characteristics for the residents of River City for the year 2000 as reported by the U.S. Census Bureau. As can be seen from Table 1 with a median family income of $28,000, River City remains the twenty-third poorest city of 50,000 or more residents in the United States with unemployment rates that consistently remain above those experienced by the rest of the Commonwealth of Massachusetts. These factors translate into a high percentage of children living in poverty (42.1%) within the community. In addition, nearly 80% of River City students qualify for free or reduce-priced lunch and nearly one-third receive some form of public assistance.

Table 1: Profile of selected social and economic characteristics

Educational attainment	Number	Percent
Population 25 years and over	40.940	100.0
Less than 9th grade	8.093	19.8
9-12th, no diploma	9.021	22.0
High school graduate (includes equivalency)	12.121	29.6
Some college, no degree	5.878	14.4
Associate degree	1.749	4.3
Bachelor's degree	2.391	5.8
Graduate or professional degree	1.687	4.1
Income in 1999 Households		
Less than $10,000	4.643	19.0
$10,000-$14,999	2.453	10.0
15,000-24,999	3.893	15.9
25,000-34,999	3.484	14.2
$35,000-$49,999	3.699	15.1
$50,000-$74,999	3.640	14.9
$75,000-$99,999	1.439	5.9
$100,000-$149,999	923.000	3.8
$150,000-$199,999	82.000	0.3
$200,000 or more	220.000	0.9

Despite strong community advocacy for quality public schools, support for educational achievement in many River City households is limited by crime, violence and the low educational achievement levels of parents-43% of River City adults have not completed high school. Of these factors, the greatest obstacle(s) for the majority of students in River City is the struggle with the language barrier. Over eighty percent (80%) of River City students identify Spanish as the primary language spoken in the homes.

The River City Public School District is comprised of sixteen schools serving students in Grades Pre-Kindergarten through Twelve. River City's Pre-K-12th grade public school population of 12,573 reflects the demographics of the City. Table 2 presents the ethnic/racial breakdown of the student population from 2001-2004. As can be seen from Table 2, the number of students of each ethnic racial type has been stable in the River City Schools over a four year period with approximately 92% of the students being minority students and roughly 85% of the student population being Hispanic.

The configuration of the 16 school in the River City School System is as follows: three (3) Early Childhood Centers, which entered into their fourth year of full-day kindergarten and half-day pre-school programs; four (4) K-8 Elementary Schools; five (5) 1-8 Elementary Schools; two (2) K-5 Elementary schools; one (1) 6-8 Middle School; one (1) 9-12 High School. Table 3 presents enrollments by grade level, gender, low income and limited English proficient students and teachers. As can be seen from Table 3, low income students range from 69% percent of the students in the school to a high of 90% in the school.

Table 2: Ethnic-racial break down of river city student population 2001-2004

	2001		2002		2003		2004	
Race/ ethnicity	(#)	(%)	(#)	(%)	(#)	(%)	(#)	(%)
Black	344.000	2.8	303.000	2.4	291.000	2.32	406.000	2.99
Hispanic	10.566	82.7	10.593	83.9	10,608.000	4.76	11.574	85.23
Asian/Pacific Islander	417.000	3.2	390.000	3.09	374.000	2.99	331.000	2.44
White	1.427	11.3	1.332	10.55	1.228	9.81	1.260	9.28

Table 3: Enrollments by school, grade level, gender, low income and limited English proficient students and teachers for river city public school system-school year 2004-2005

		Enrollment by gender			Low income students		Limited	
School	Grades	Male	Female	Total	(#)	(%)	English proficient	Number of teachers
A	K-8	514.000	498.000	1,012.000	912.000	90.12	223.000	82
B	PreK-K	188.000	158.000	346.000	240.000	68.97	157.000	16
C	1-8	355.000	344.000	699.000	628.000	89.84	112.000	53
D	K-8	422.000	443.000	865.000	647.000	74.71	80.000	72
E	1-8	572.000	540.000	1.112	1,016.000	91.37	194.000	81
F	PreK-K	186.000	153.000	339.000	240.000	70.38	196.000	25
G	PreK-K	92.000	76.000	168.000	144.000	85.71	89.000	10
H	9-12	1,262.000	1.233	2.495	1.883	74.75	478.000	169
I	K-5	302.000	277.000	579.000	540.000	93.26	149.000	39
J	6-8	211.000	183.000	394.000	378.000	95.94	83.000	38
K	1-8	345.000	348.000	693.000	610.000	88.02	144.000	49
L	K-8	634.000	615.000	1.249	991.000	79.28	219.000	86
M	1-8	110.000	22.000	132.000	125.000	93.28	14.000	34
N	1-8	562.000	557.000	1.119	998.000	88.95	153.000	84
O	K-5	170.000	131.000	301.000	264.000	87.71	84.000	26
P	K-8	376.000	333.000	709.000	600.000	84.27	118.000	57
Total		6.301	5.911	12.212	10.216	84.79	2.493	921

Similar variance was seen in the variation percentages of limited English speaking students in these 17 schools. Therefore, although there was some natural variation in these schools of these various factors, these variations reflected “fine-grained” differences as well as “section of the city differences” rather than truly major significant differences between these schools that were substantive as opposed to statistical.

The SFA Foundation, under a 7 figure annual contract, trained the teachers in the SFA program, guided and monitored the implementation the SFA program in these 17 schools, carried out the implementation compliance evaluations and revisions as well as evaluated the success and progress of the program using the SRI and other qualitative and quantitative assessments they used, all of which were presented to teachers and administrators on summarized reports. Figure 1 present a general summary of the degree to which certain features and required characteristics of the SFA program were successfully implemented in the 17 River City schools according to and as reported by to which certain features and required characteristics of the SFA program were successfully implemented in the 17 River City schools according to and as reported by the SFA.

Foundation to the River City schools superintendent. The detailed definitions and characteristics of the six implementation criteria given in Fig. 1 are reported elsewhere. As can be seen from Fig. 1, the SFA program became successively better and more successfully implemented in the City Schools between 2001 and 2004 with all of its successful implementation percentage being above 80%. More will be said on this point below. However, it is clear from the SFA Foundations own evaluations and reports, it consider the SFA program well, properly and implemented correctly and successfully in the River City schools.

The SFA program:

SFA program components: Fully describing the components of the SFA program implemented in the River City schools is beyond the scope of this study and is done in detail elsewhere. The SFA program implemented in the River City schools, therefore, will be briefly characterized here to give a general description of the particular instantiation of the SFA program in the River City Schools. SFA program components are designed to meet the needs of students in a variety of developmental stages. The adopted instructional components for River City Public Schools are.

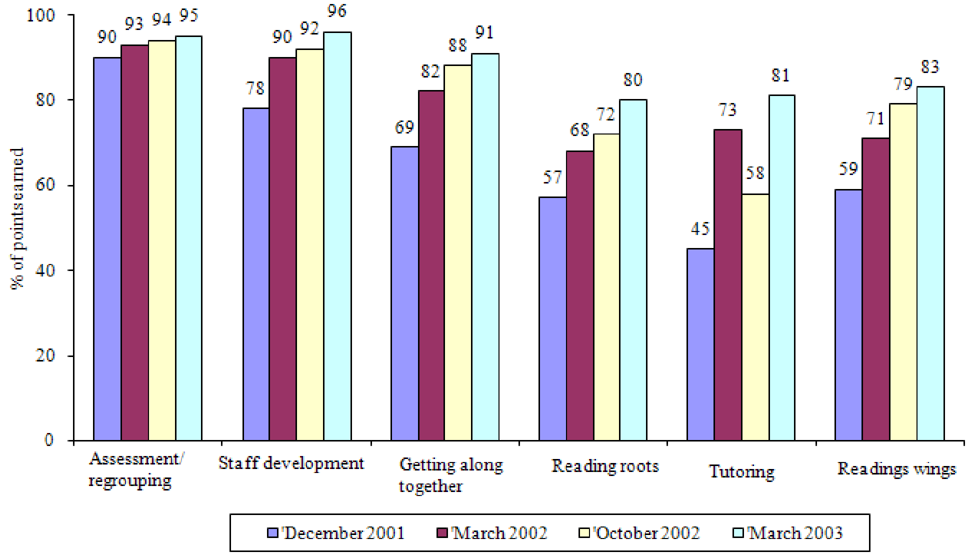

Fig. 1: Summary of SFA implementation reports in river city public schools, **Note:** From student achievement profile: school year 2002-2003 (p.13). river city public schools. adapted with permission

Roots 3rd edition: Beginning reading program for primary grade students (includes the addition of Fast Track Phonics 3). This component enables students to read books that use a phonetically controlled vocabulary. In these books, called "Shared Stories," the children know that with the exception of a few sight words, every word they uses letter sounds they know.

Older roots: Reading program for students in grades 2-5 at a beginning reading level.

Wings: Reading program for students in grades 1-5 beyond the beginning reading level. From grades 2 to 8, the Success for All reading program, Reading Wings, focuses on advanced phonics, continuing to develop fluency, as well as reading comprehension, higher order thinking skills and vocabulary building. The curriculum uses novels or trade books. Classroom materials, keyed to each novel or trade book, guide students to use strategies known to enhance comprehension, such as: Clarification, summarization, prediction and question generation-- and to represent their thinking using graphic organizers. Reading Wings lessons immerse students in high quality literature while focusing on critical skills.

Reading edge: Middle school reading program. It allows 6-8 grade students to strengthen their reading and writing skills and apply them to humanities and science units, where students experiment, investigate, solve problems and draw on resources beyond the classroom to expand their knowledge of the world. The Reading Edge program component has two stages designed to meet the needs of all students. These are: Stage I-reading instruction for students in grades 6-8 who need it at a beginning reading level (includes the addition of Fast Track Phonics) and Stage II- more advanced reading instruction for students who are grouped homogeneously by their reading levels.

Humanities: Humanities is offered to grade 6-8 students in a daily sixty minute block integrating reading, writing, school social studies and language arts program fro grades 6-8 students. It consists of a 60 minute block integrating reading, writing, social studies and language arts. Students learn by investigating real-world problems and topics in cooperative groups, connecting what they learn about the past with their own lives and present their findings in various forms of writing, extensive use of writing skills, reading of expository and narrative texts, math, science and fine arts. In addition students are asked to collect data, investigate, encouraged to ask questions and to predict the impact of actions or events.

Reading level grouping: From 1st grade on, children are grouped for reading according to reading level, not grade level and they are regrouped every quarter over the year based on their reading progress or not. It should be well-noted that this component and practice of the program makes the psychometric qualities of the assessment procedure (test and grouping score) used to group and regroup students critically important and critically important to the operations and functioning of the program.

Table 4: Summary of River City District's implementation of the Reading Edge Stage I program*

		October 2002		March 2003	
Reading edge stage I	Total points possible	avg. # of points	Stage	Avg. # of points	Stage
Active Instruction	30	17.0	Mechanical	23.6	Routine
Teamwork	24	13.0	Mechanical	16.2	Routine
Time for Reflection	12	7.3	Mechanical	8.2	Routine
Assessment	6	4.0	Routine	4.4	Routine
Classroom Environment	9	7.0	Routine	6.6	Routine

***Note:** From Student Achievement Profile: School Year 2002-2003. River city public schools. Adapted with permission

Table 5: Summary of river city district's implementation of the reading edge stage ii program*

		October 2002		March 2003	
Reading edge stage II	Total points possible	Avg. # of points	Stage	Avg. # of points	Stage
Active instruction	30	12.0	Mechanical	17.1	Mechanical
Teamwork	24	10.3	Mechanical	16.7	Routine
Time for Reflection	12	4.7	Mechanical	8.0	Routine
Assessment	6	3.3	Mechanical	3.9	Routine
Classroom environment	9	5.4	Mechanical	7.3	Routine

***Note:** From Student Achievement Profile: School Year 2002-2003 (p.15). River City Public Schools. Adapted with permission

Table 6: Summary of the district's implementation of the SFA humanities program*

		December 2001		October 2002		March 2003	
Humanities component	Total points possible	Avg. # of points	Stage	Avg. # of points	Stage	Avg. # of points	Stage
Setting the Stage	9	3.8	Mechanical	4.3	Mechanical	7.3	Routine
Active Instruction	21	8.8	Mechanical	10.0	Mechanical	17.1	Routine
Teamwork	30	11.2	Mechanical	15.1	Mechanical	23.4	Routine
Time for Reflection	12	3.9	Mechanical	5.3	Mechanical	8.6	Routine
Assessment	6	2.8	Mechanical	3.6	Routine	4.3	Routine
Classroom Environment	9	4.9	Mechanical	6.3	Routine	7.9	Routine

* Note: From Student Achievement Profile: School Year 2002-2003 (p.16). River City Public Schools. Adapted with permission

Table 7: Basic logical comparison group design for assessing the effects of SFA Program (all groups and independent groups)

	School year of group and group's data and number of participating students	
Grade	2001-2002 school year	2003-2004 school year
4th graders*	Received one year of SFA program (N = 890)	Received three years of SFA program (N = 531)
7th graders*	Received one year of SFA program (N = 832)	Received three years of SFA program (N = 558)

*: Each group is an independent group

During the fall and spring of each school year, consultants/coaches from the Success for All Foundation visit schools to evaluate their progress in the implementation of the SFA elementary reading program. As stated above (Fig. 1), proper and successful implementation of the components of the SFA program described about became better each year with the ratings on all components being high by 2002, particularly on the Roots (elementary school) component.

These same points were true for the Reading Edge and Humanities components of the SFA program as can be seen from Table 5-7. The data in Table 5-7 were again extracted from reports done by the SFA foundation for the River City school superintendent. Therefore, according to the SFA Foundation, the standard or vanilla version of the SFA program was properly and successfully implemented in the River City Schools and the degree of proper program implementation (theoretically) would not be an intervening factor in explaining or accounting from the outcomes observed in this study. This fact was confirmed statistically across the 17 River City schools using the implementation scores generated by SFA for each school (and summarized in Table 4-7), as well as other data such as attendance and age.

MATERIALS AND METHODS

This study was a retrospective study, which means that the data existed prior to the date that the study was designed. This fact posed a number of potential limitations, as data collected retrospectively is fixed in time and are not necessary the consequence of purposeful and systematic a priori manipulations by the researchers and must be analyzed and interpreted cautiously for a number of reasons including missing data points and various non-randomization factors. The present study examined comparison groups of fourth and seventh graders, who had received one and three years of SFA reading instruction respectively. The basic design of this study was a Grade-Level by Amount of SFA Program by Gender design (2×2×2) with several a priori hypotheses and comparisons concerning gender and other differences specified by the theoretical framework for this study. Fourth and seventh grade students were chosen as the focus of this study as these grades are one year before the termination points of the two major components or stages of the SFA program and correspond to the state testing points in its state-wide mandated readings assessment program. The subjects in this study (N = 1,651), therefore, were several purposely-selected (logical) comparison groups of fourth and seventh graders. The first comparison group in the study was fourth graders in the 2003-2004 school year who received three years of SFA reading instruction. The second group in the study was fourth graders in the 2001-2002 school year who received only one year of SFA reading instruction. These two groups of fourth graders were independent and uncorrelated groups. The third group in the study was seventh graders in the 2003-2004 school year who received three years of SFA reading instruction. The last group in the study was seventh graders in the 2001-2002 school year who received only one year of SFA reading instruction. Again, these two groups of seventh graders were independent and uncorrelated groups. This logically constructed retrospective design, therefore, allowed the studies a priori hypotheses to be tested in a more precise and controlled way than a simple “historical by successive years” design.

One of the major goals of this study was to determine the degree to which the SRI and the MCAS gave similar assessments of the reading competencies and reading levels of River City students and the success of the SFA program in developing their reading competencies and reading levels, particularly relative to what the state of Massachusetts defined and deemed to be the twenty-first century level. The working hypothesis of this study was that the SFA program was less successful in helping males to become competent readers (by the standards and on the measures the SFA program uses) and least successful in helping poor, ethnic males in becoming competent readers. These later males in particular lag behind other students in developing reading competencies and this lag becomes progressively greater with each grade.

The order and magnitude of differences in reading achievement between all of the comparison groups in this study, given the design described above, were predicted on an a priori basis using both individual difference and developmental learning theories and data from the literature. Table 8 presents the logical comparison group design of this study and Table 9 presents the a priori hypotheses of this study derived from its theoretical perspective (for full explication of this theoretical framework and derivation) and from review of prior evaluations done of the SFA program. All prior evaluation of the SFA program that we could locate only report aggregated data of program effects (one definition and view of “All”) and ignored Simpson’s Paradox (Bracey, 2006) relative to subgroup effects not necessarily or usually being the same as aggregated effects (a different definition and view of “All”). Further, the comparability of SFA-chosen success criteria (i.e., the SRI) and state-mandated success criteria (i.e., the MCAS) and the similarity of success and successfulness estimates each give of the program were not studies or reports that we were able to locate or that seemed to be available except for two highly flawed studies done with the state of Florida’s test. This particular problem is not unique to the SFA program, or the SRI instrument, but rather is endemic in education since educational reform and is a major problem in the research and evaluation literature and in determining or coming to some conclusions about “what is and is not working and for whom in what context,” never mind meta-analyses addressed at answering this question. Calibrating the SRI and the MCAS, therefore, had to be done to carry out this study.

Reliability and validity of measures: The Scholastic Reading Inventory (SRI) is a computer-adaptive assessment instrument for grades 1-12 that allows educators to assess reading comprehension and match students to books in a speedy and accurate manner (Scholastics for details). This assessment can be used to help place students at the best level in a reading program so they can read with success. It can also help teachers in monitoring student reading growth and differentiating instruction.

The SRI tests comprehension of written literature, not just vocabulary. According to Scholastic, all 3,000 questions in the item bank are based on passages from authentic children’s literature, both fiction and nonfiction, as well as excerpts from young adult and classic literature, newsstudys, magazines and periodicals. According to Scholastic, the SRI does not require prior knowledge of ideas outside of the assessment passages. Passages require students to make inferences, draw conclusions and demonstrate vocabulary knowledge in context, among other higher order thinking skills. The SRI test self-adjusts in response to the student’ s reading ability.

Table 8: A Priori Differences Expected in Basic Grade-Level by Amount of SFA Program by Gender (2x2x2) Design

	Grade Level	
Amount of SFA Program	Fourth	Seventh
One Year (2001-2002)	Smallest M/F differences F>M	Smaller M/F differences F>M
Three years (2001-2004)	Second largest M/F difference F>M	Largest M/F differences F>M

Table 9: Range of scores associated with each SRI interactive performance level

Grade	At-Risk (Significantly below grade level)	Basic (Below grade level)	Proficient (On grade level)	Advanced (Above grade level)
1	_	99 and Below	100-400	401 and Above
2	99 and Below	100-199	200-500	501 and Above
3	249 and Below	250-499	500-800	801 and Above
4	349 and Below	350-599	600-900	901 and Above
5	449 and Below	450-699	700-1000	1001 and Above
6	499 and Below	500-799	800-1050	1051 and Above
7	549 and Below	550-849	850-1100	1101 and Above
8	599 and Below	600-899	900-1150	1151 and Above
9	649 and Below	650-1049	1050-1300	1301 and Above
10	699 and Below	700-1099	1100-1350	1351 and Above
11	799 and Below	800-1149	1150-1400	1401 and Above

Students start the test; the test steps up or down according to their performance; and when the computer has enough information, the test stops. According to Scholastics, adaptive testing shortens test-taking time and increases testing accuracy by varying the number and difficulty of questions that students answer.

When taking the SRI, students read paragraphs from actual books (not text made up by test writers) and then answer one question about each paragraph. If the student answers correctly, the next question is just a little bit more difficult. If the student answers incorrectly, the next question is just a little bit easier. The test continues in this way until a Lexile level is established.

The main purpose of the Scholastic Reading Inventory (SRI) testing is to monitor individual and group progress in reading comprehension skills. The SRI measures reading comprehension by focusing on the skills readers use when studying written materials sampled from various content areas. The SRI is made up of five components: (a) Words in Isolation Checklist (b) Passage Comprehension, (c) Word Recognition Accuracy, (d) Vocabulary in Context and (e) Predictive Comprehension. These skills include referring to details in the passage, drawing conclusions and making comparisons and generalizations. SRI Interactive consists of embedded completion items. This format is similar to the fill-in-the-blank format. The passages are shorter for beginner readers and longer for more advanced readers. The passage is then response illustrated, which means that a statement is added at the end of the passage with a missing word or phrase followed by four options. The reader is then asked to select the best option that completes the statement.

This SRI Interactive was chosen by River City to monitor individual and group progress in reading comprehension skills. The data are used to drive placement and curricular decisions quarterly for each student as prescribed by the SFA reading program in the River City school district. The SRI is administered every 8-weeks to students in grades 2-8. In the River City school district the SRI test is administered via computer and takes approximately 30-40 minutes. At the end of the testing session, students are given a Lexile score. The lexile indicates the level at which the student reads. The lexile level is the measure of reading difficulty given to text, based in part on sentence difficulty and length and word frequency. The word Lexile is a blend of the root word “Lex”, which refers to words and echoes the word “percentile”, a comparative unit of measure. The Lexile Framework provides a single scale that can be used for targeting readers with text that provides an appropriate level of challenge. A lexile is a reading level determined by analyzing text and determining its difficulty level based on vocabulary and sentence structure. The lexile scale ranges from 200 for a beginning reader to 1700 for advanced texts. According to MetaMetrics, Inc., the company that developed the Lexile framework, the average middle school student grows about one to two lexiles per week. The average primary school student grows about three to four lexiles per week. Therefore, over a full, 36-week academic year, middle school students should expect to grow 90-140 lexiles.

Table 10: General MCAS performance level definitions

Performance level	Description
Advanced 260-280	Students at this level demonstrate a comprehensive and in-depth understanding of rigorous subject matter and provide sophisticated solutions to complex problems
Proficient 240-258	Students at this level demonstrate a solid understanding of challenging subject matter and solve a wide variety of problems
Needs improvement 220-238	Students at this level demonstrate a partial understanding of subject matter and solve some simple problems
Warning/failing 200-218	Students at this level demonstrate a minimal understanding of subject matter and do not solve simple problems

In the River City school district, The September and June SRI administrations are used to assess and report annual student progress and program success. A student Lexile score is then assigned a performance level in one of four categories: At-Risk, Basic, Proficient, or Advanced (Table 10). School level results are reported as the aggregate number and percent of students in each performance level on the September and June assessments and the aggregate differences between these two assessment points. For the purpose of this study, both Lexile scores and SRI performance level categories were used as unit of analysis.

The SRI Interactive normative information is based on a sample of 512, 224 students from a medium-large state. According to Scholastics, this state has shown similar means and standard deviations to the nation as a whole on national norming criteria. The SRI Interactive uses the same item format as the print version of SRI. During field testing, SRI Interactive results were correlated with SRI (Print), the Comprehensive Test of Basic Skills (CTBS), the North Carolina End-of-Grade Test of Reading Comprehension (NCEOG) and the Pinellas Instructional Assessment Program (PIAP). Correlations coefficients ranged from .56 (with CTBS) to .83 (with SRI-Print). From these results, Scholastics concluded that SRI Interactive measures a similar construct or trait as measured by other standardized tests designed to measure reading comprehension.

It should also be noted that the SRI Interactive was developed using the Rash one-parameter item response theory model to relate reader's ability and the difficulty of the items. Due to violation of model assumptions, there is a unique amount of measurement error which is associated with each score on SRI Interactive. The computer algorithm that controls the administration of the assessment uses a Bayesian procedure to estimate each student's reading comprehension ability. This procedure uses prior information about readers and students in order to control the selection of questions and the recalculation of each student's reading ability after responding to each question. Compared to a fixed-item test where all of the students take the same questions, with a computer-adaptive test every reader takes a different test. Also, all of the students receive approximately the same raw score or number of items correct. This outcome is due to the fact that all students are answering questions that are targeted for their unique ability and not questions that are too easy or too hard. Therefore, the error associated with any one score or student is unique, which is a characteristic that can pose several analytic difficulties.

The Massachusetts Comprehensive Assessment System (MCAS) is the mandated statewide assessment program used to measure the performance of students, schools and districts on the academic learning standards contained in the Massachusetts Curriculum Frameworks, fulfilling requirements of the Education Reform Law of 1993. The grade 3 Reading Test and grades 4, 7 and 10 English Language Arts Composition tests are administered every spring to all students in Massachusetts educated in publicly funded schools.

The MCAS tests involve a "common/matrix-sampled item" test design. With the exception of English Language Arts Composition, each test contains both common and matrix-sampled items. Individual student test scores are based exclusively on common items, which comprise roughly 80 percent of items in a test booklet. All students in a grade level are tested on the same set of common items. These common items are released to the public after testing is completed (http://www.doe.mass.edu/mcas for prior year items). Approximately 20 percent of each test booklet is dedicated to matrix-sampled items. These items differ across test forms and are used to equate tests across administrations as well as to field test new items for future use. Results from these items are also combined with those from common items for reporting to schools and districts based on the major strand levels of the Curriculum Frameworks.

The MCAS English Language Arts (ELA) tests measure three ELA Curriculum Framework content strands. These are: Language, Literature and Composition. These content strands reflect all of the standards of the Massachusetts English Language Arts Curriculum Frameworks that were feasible to incorporate into a large-scale state assessment program such as MCAS, as well as the number of common items per form for each Spring 2003 ELA test by grade level, strand and item type.

Table 11: Comparison of SRI and MCAS Performance Levels

SRI	MCAS
At-Risk: Students at this level do not demonstrate Minimally competent performance when reading grade-level appropriate text and can be considered as reading significantly “Significantly Below grade level.”	Warning/Failing: Students at this level demonstrate a minimal understanding of subject matter and do not solve simple problems (Significantly Below grade level)
Basic: Students scoring in this range demonstrate minimally competent performance when reading grade-level appropriate text and can be considered as reading “Below grade level.”	Needs Improvement: Students at this level demonstrate a partial understanding of subject matter and solve some simple problems. (Below grade level)
Proficient: Students scoring within this performance level demonstrate competent academic performance when reading grade-level text and can be considered as reading “On grade level.”	Proficient: Students at this level demonstrate a solid understanding of challenging subject matter and solve a wide variety of problems. (On grade level
Advanced: Students at this level exhibit superior performance when reading grade-level text and can be considered as reading “Above grade level	Advanced: Students at this level demonstrate a comprehensive and in-depth understanding of rigorous subject matter and provide sophisticated solutions to complex problems. (Above grade level)

Since the MCAS tests are designed to measure student performance on the learning standards contained in the Massachusetts Curriculum Frameworks, results are reported in terms of scale scores and performance levels that describe student performance in relation to established state standards. Individual student raw scores are translated into scaled scores and performance levels through a process called scaling. The 2003 MCAS Technical Report defines scaling as “a means by which numbers are converted from one numerical representation into another such that rank ordering and performance classification remain unchanged” (p.10). The reason for changing from one scale to another is to make the process of understanding scores more natural. School and district level results are reported as the number and percentage of student attaining each performance level at each grade in each subject area tested.

MCAS results for each student in grades 4 through 10 are reported in terms of four performance levels: Advanced, Proficient, Needs Improvement and Warning/Failing (categories similar in name to those of the SRI). Results on the grade 3 Reading test are reported in terms of three performance levels: Proficient, Needs Improvement and Warning.

MCAS test questions (items) focus on general Framework learning standards and/or their corresponding grade-specific standards. Some items incorporate standards identified for preceding grade levels; consequently, students are often required to demonstrate cumulative content knowledge and skills (e.g., grade 10 students may be tested on learning standards identified in a Framework from pre-Kindergarten through grade 10). Four types of response formats are used on MCAS tests. These four types are: Multiple-choice questions, open-response questions, short-answer questions and writing prompts. The details of these various items type and how they are scored are given at http://www.doe.mass.edu/mcas.

The MCAS ELA assessment has two components: Reading (Language and Literature) and writing (Composition). The reading test is composed of 36 Multiple-Choice (MC) items and 4 Open-Response (OR) items scored on a scale of 0-4 by trained raters at the state level. Thus, the maximal possible score students can achieve on the reading test is 52 points (36 points from MC items and 16 points from OR items). The writing component of the test consists of a single writing prompt that is scored using rubrics on two scales: Topic development (using a scale range of 1-6) and English conventions (using a scale range of 1-4). Each composition is scored by two readers whose scores are summed so that the final score range on topic development is 2-12 points and on English conventions 2-8 points, yielding a total score range of 4-20 points. Hence, the maximal possible total score on the ELA assessment is 72 points, 20 of which come from writing and 52 from reading component. When broken down by item type, 50 percent of maximal possible points on the ELA assessment are derived from selected-response questions (36 MC items) and 50 percent are derived from constructed-response questions (16 OR items and the composition).

In addition to performance levels, MCAS results are reported as scaled scores. Scaled scores in each content area range from 200 to 280. As previously stated, there are four performance levels: Advanced, Proficient, Needs Improvement and Failing (Table 11 for details).

MCAS tests are based on a common plus matrix-sampled design. Each MCAS test booklet contains both common and matrix-sampled questions. Common questions are those that are taken by all students at a grade level and comprise approximately 80 percent of a student’s test booklet. All performance level and scaled score results for students, schools and districts are based exclusively on the common items. All common items are publicly released following each year's test administration to inform local curriculum and to support public understanding of MCAS and the standards contained in the Curriculum Frameworks. The remaining 20 percent of the test questions in each student's test booklet are matrix-sampled questions, which differ across the multiple test forms at each grade level tested. Matrix-sampled items serve three primary purposes. First, they serve as the basis for equating MCAS tests from year to year. This equating allows for comparisons of performance at the school and district level over time. Second, the use of matrix-sampled questions allows for reporting of school and district results in greater depth and detail for a broader range of the curriculum than is possible with common items only. Results from the common and matrix-sampled items are aggregated at the school and district levels to produce additional information regarding performance on the major strands measured on each subject area test. Third, the use of matrix-sampled items allows for the field-testing of new MCAS items in an operational setting before they are used to generate individual student or school scores (Massachusetts Department of Education, 2007).

MCAS utilizes a raw-score-to-theta equating system in which test forms are equated every year to the theta scale of the reference test forms. According to Kerlinger and Lee (2000), equating is a statistical procedure by which scores on test forms designed to measure the same constructs are made interchangeable. This interchangeability of test scores obtained by alternate test forms allows the application of evaluation criteria established on a reference form to any new form that measures the same construct. For MCAS, equating enables the performance standards set on the reference forms to be applied on subsequent tests and for growth to be measured. This equating system is established through the chained linking design, which means that every new form is equated to the theta scale of the previous year's test form. Since the chain originates from the reference form, it can be assumed that the theta scale of every new test form is also the same as the theta scale of the reference form.

In view of the fact that the conversion of raw scores to scaled scores is mediated by theta values, the same theta scale must be maintained throughout different forms of the same grade/content test. Therefore, for each MCAS administration, items are calibrated to the same theta scale using the linking design mentioned in previous paragraph, so that raw scores can be mapped to the theta scale and theta scores can be transformed to scaled scores using the linear functions established on the reference forms.

The MCAS is a carefully administered test. Teachers and principals are provided with detailed administration manuals in order to ensure that test administration is uniform. The 2003 MCAS Technical Report presents evidence of the reliability and validity the MCAS. For instance, 72,480 fourth graders and 76,782 seventh graders took the English Language Arts (ELA) portion of the test. Cronbach alpha reliability was .88 for grade 4 and .90 for grade 7 and test-retest reliability coefficients above .90. The best external empirical evidence of the construct validity of the MCAS test is comparisons of students' performance on MCAS with their performance on commercial standardized tests. Results of Tacker and Hoffman's (1999) and Rump and Lesauxs (2006) studies provide support for the construct validity of the MCAS tests. For example, the ELA portion of the test has a .80 correlation with the Stanford Achievement Test and similar correlations to several other tests, even though content validity (and the test's congruence with the Massachusetts content frameworks) is the primary criterion for the validity of the MCAS. With respect to this later criterion, the MCAS was rated by a blue ribbon federal review panel as one of the top state assessment tests in the country and the state assessment test most in alignment with the NAEP scales (NCES 2007 for details). The content of the MCAS was also rated in a study done by Brown and Conley (2007) as the state assessment test most aligned to entry-level university courses. The various psychometric data available on the MCAS supports that it is a high quality test that measures a high quality content standard unlike many other state assessment tests (NCES, 2007). The MCAS has one of the most comprehensive web-sites in the nation on the content and development of the MCAS and its wide variety of reliability and validity information, as well as state-wide item performance information that may be downloaded and analyzed by anyone interested (http://www.doe.mass.edu/mcas).

Calibrating the SRI and the MCAS: Table 10 presents the range of scores associated with each SRI performance level, which in this table is grade level. This particular table is and will be very important in interpreting many of the findings about the SRI presented in the results section in terms of the observed SRI scores of fourth and seventh grade students and the observed fall to spring SRI difference or change scores. In general, the reader should note that Table 10 allows one to state what grade level a student at a given grade level is reading at in the fall or spring and that roughly 150 SRI points is a grade level gain or loss.

Table 12: SRI and MCAS recoded categories for cross-classification analyses

SRI	MCAS	New category
At-Risk	Warning/failing	Fail
Basic	Needs improvement	
Proficient	Proficient	
Advanced	Advanced	Pass

Table 11 presents the score ranges on the state-mandated MCAS that are associated with each of the four proficiency levels for the test, one of which is at grade level with one above grade level and two being below grade level or significantly below grade level.

The MCAS significantly reduced classification problems and errors and is a high reliable test with test-retest coefficients in the .90's. The MCAS does not give grade level scores or ranges. The SRI also has a four category classification system like the MCAS and Table 12 presents these two four category systems side-by-side so that they can be compared. As can be seen from Table 12, although their labeling and terms are somewhat different, the four categories are essentially the same and comparable categories. To further simplify these categories, misclassifications and the analyses done with them, these categories were reduced to dichotomous pass-fall categories, thus allowing very simple 2x2 cross-classification analyses to be done to assess if the SRI and the MCAS classified the same students the same way in terms of these pass-fail categories and thus produced the same assessments of program effectiveness in terms of the percentage of female, male and all students who were classified as reading at grade level or above. In theory, the two tests should classify students reading levels similarly and should produce similar estimates of program effectiveness. Further details of this design and calibration process are given in as well as a fine-grained substantive analysis of the content of each of these two measures.

RESULTS

Large scale and critical problems with data collection and the quality of the locally chosen SRI success measure changed the meaning of this study from its original intention to some degree. Results of the preliminary analyses showed that 20% of the students were missing fall and/or spring scores on the (interactive) Scholastic Reading Inventory (SRI).

The percentage missing both fall and spring SRI scores consisted mainly of special education students (58.1%) or English Language Learner (44.0%) status, as compared to 13% of the Regular Education subjects. This finding represents a bias in the (missing) data, as it seems like most of the ELLs and SPED students were excluded from the testing. However, all students required to participate in the Massachusetts Comprehensive Assessment System (MCAS) participated. Therefore, the findings of this research suggest the need of further quality control by an external independent party in evaluating high-stakes and high cost programs such as the SFA program, which means an externally administered and provided assessment of high stake skills such as the MCAS, as some of the major changes in accountability brought about by the Education Reform Act of 1993 included independent assessments of high stake skills such as reading. This lack of internal data management and internal data quality control, particularly relative to high stakes data and programs, is not atypical in schools today, in the experience of one of us (i.e., Carifio) and appropriate data quality management (a key responsibility of any executive) is not an evaluation criterion today of school managers or an evaluation criterion that the SFA program seems to use in the evaluation of the implementation of its program or the programs effectiveness. It will be recalled that the SFA staff provided the superintendent of the River City schools with reports on the success the SFA program was having on improving the reading achievement of River City students. More will be said on this particular point below.

Since the most complete data available were on Regular Education Hispanic students, who were also the most dominant subgroup in the sample, only the results for Hispanic subjects enrolled in Regular Education (N = 1651) were used in the statistical analyses and assessment of effects, hypotheses and the effectiveness of the SFA program reported below.

Another major finding concerning the SRI test was the disparities between score distributions (shapes, skewness and kurtosis) and mean levels for the different subjects in the study. Figure 2-5 present graphic representations of the frequency distribution of the SRI fall scores for fourth and seventh grade students and the Fall-to- Spring SRI difference scores for the same fourth and seventh grade students. As can be seen from Fig. 2-4, some fourth graders were advanced readers (the SRI score test norms given in Table 10 for exact values), reading well above grade level (a Lexile score of 600-900 for fourth graders) in the fall and then became at-risk readers who were reading significantly below grade level in the spring. Conversely, some students who were reading significantly below grade level gained over 550 Lexiles from fall to spring (roughly 150 Lexiles is a whole grade gain).

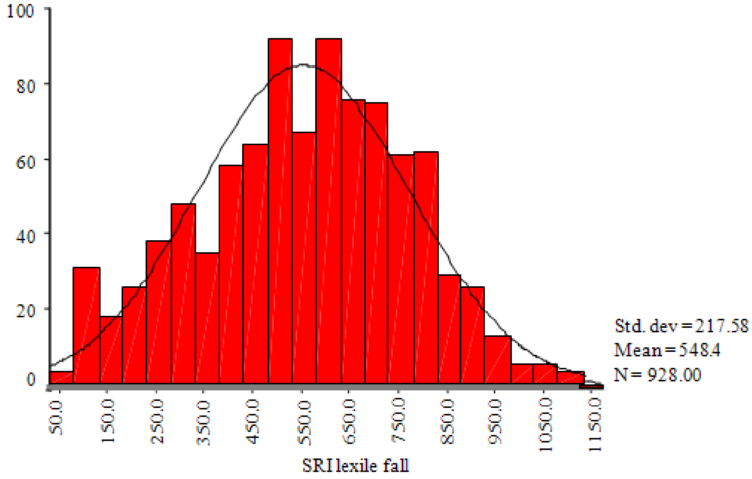

Fig. 2: Histogram of grade 4 SRI lexile fall scores

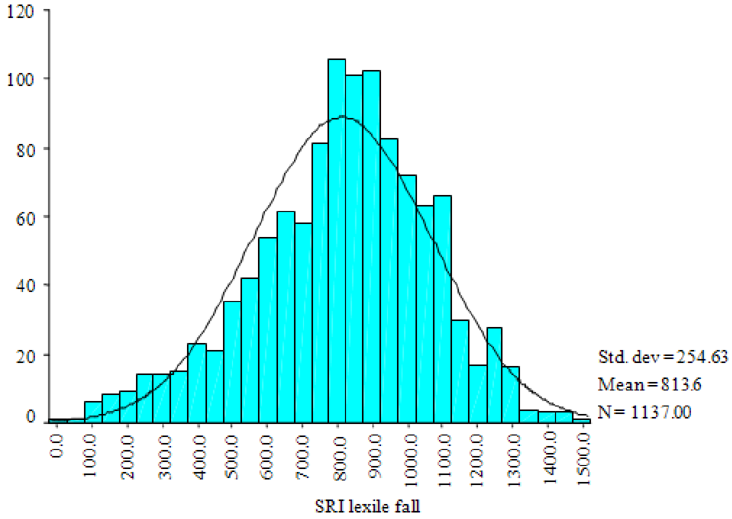

Fig. 3: Histogram of grade 7 SRI lexile fall scores

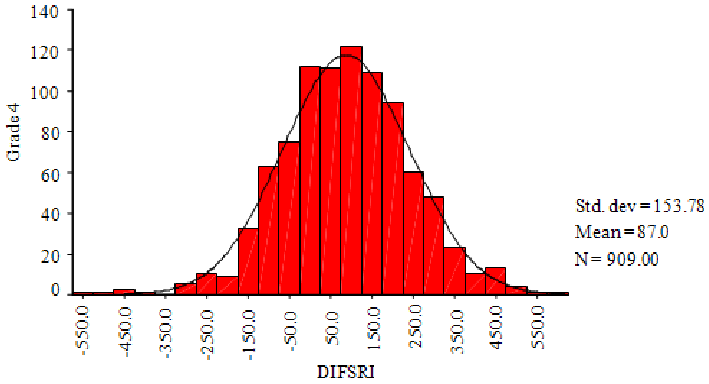

Fig. 4: Histogram of grade 4 SRI lexile difference scores

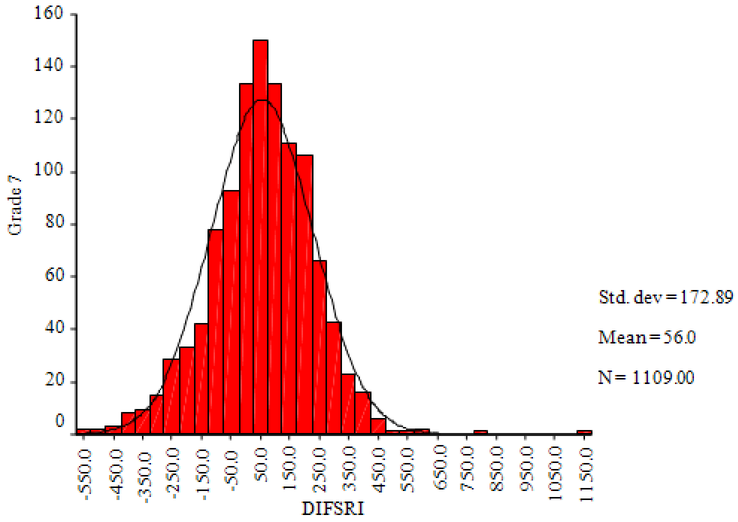

Fig. 5: Histogram of grade 7 SRI lexile difference scores

Likewise, many seventh graders (Fig. 3-5) were reading at the 11-th grad level in the fall (a Lexile score of 850-1100 was at grade level for seventh graders) and almost one third of the grade 7 students lost over 50 Lexiles from fall to spring, whereas others gained over 500 Lexiles from fall to spring. Something, obviously, is and was radically and very wrong with the SRI test, despite all of the rhetoric and alleged reliability and validity data in its test manual. However, this is the test recommended by the SFA foundation for use in evaluating the SFA program and one has to enquire as to how this test came to be recommended by the SFA program, which we did, but we did not receive any cogent answers other than the non-psychometric comment that it was computer administered and computer scored quickly and thus could efficiently regroup students, as the program required, as the end of each quarter. Perhaps administrative ease and speed were some (secondary) criterion of importance, but at what price.

Given the easily observed inaccurate and misclassification of students by the SRI (i.e., a goodly number of fourth grade students in a very poor community reading at the eleventh grad level and then not), the classification unreliability of this test must have brought about a good deal of re-grouping havoc each quarter when it was used to regroup students as required by the SFA program and instructional model, with many students being fairly randomly over-grouped and under-grouped each quarter across each school year with measurable effects that would be observed somewhere and somehow and on some measure other than the SRI (i.e., the state-mandated exam). Couple this factor, which is bad enough on its own, with a possible cultural interaction where certain students would be more likely to be vocal and express concerns about such misclassification (i.e., female Hispanic students) than other more "laconic" students (i.e., male Hispanic students) who would be more like to passively and silently endure the "schooling hand they had been dealt," and you have a fairly chaotic and difficult silent and unobserved "submarine effect" occurring in the program that no one was attending too including the SFA monitors. Homogeneous grouping and regrouping may be a critically important characteristic and factor, theoretically, for the SFA program, but to implement this factor, one must have a test that classifies students with an extremely high degree of accuracy and reliability as well as validity. But this problem was not the only problem with the recommended and locally chosen SFA test.

Examination of Fig. 2-5 against the norms presented in Table 10 will quickly show the reader that the SRI presents an over-rosy picture of students reading achievement levels and gains, which means that this test is biased towards making SFA look successful. This locally determined test-choice bias is due to a number of factors, some of which (to be fair) are associated with parallel developments of new programs and new measures (including state-mandated ones), but the problem is also an old one dating back to the state-wide and national program evaluation efforts in the 1970's. This bias-in-success-criterion-choice point, however, also underscores both the wisdom and benefits of having external assessments, such as the MCAS, that are externally administered and have high and continuous public scrutiny of their quality, validity and meaningfulness by a broad array of experts.

The fall-to-spring difference score distributions (Fig. 2-4) clearly show that no meaningful (or logically consistent) gain or change score analyses could be done with the SRI data, as the number of students who became poorer readers by one or two grade-levels or more was only slightly less than the number of students who became better readers by one or two grade levels or more. Such a data pattern could indicate that there were possible one or more large student-type/program interactions submerged in the data (Simpson's Paradox again), but the data pattern clearly indicates that any aggregate means or aggregate gains are meaningless and uninterruptible and could be due to any number of factors other than the SFA program and any effects it was producing. The SFA foundation reported only aggregate gains to the River City school administrators who in turn reported them to all of the various constituencies they report to as well as the press. It is clear from Fig. 2-4 that the positive (gain) results reported were both spurious and meaningless, as was confirmed by various repeated measure ANOVA of the fall-to-spring SRI scores which had numerous interactions between all factors included in the analyses. Meaningful and interpretable program effects and gains in reading achievement could not be assessed for the SFA program in River City using the SRI tests. The point here is that psychometric knowledge, psychometric expertise and psychometric analyses of measures counts and is important and such knowledge and expertise needs to be part of the professional repertoire of both program providers (i.e., SFA) and program purchasers (i.e., River City administrators). The paradox is that in this era of standards, testing and accountability, this knowledge and expertise is close to an all-time low historically and one need only look at degree programs in education from certification to the doctorate to understand why (Fensham, 2004 and Shulman *et al*., 2006).

Cross-classification analyses: A commonly presented misanalysis of the relationship between a local criteria of success (such as the SRI) and the state criteria of success (the MCAS) is to state the correlation between the two tests and if it is substantial to say that the two tests are measuring the same thing and are equivalent in their descriptions of the performance levels of students and the successes of programs. Such a claim can be and often is fallacious (and sometimes purposefully so) as an easy test on which mean levels are (spuriously) high can be highly correlated with a more difficult test or higher standard level on which mean levels are much lower and which gives a very different assessment of the actual achievement levels of students. Correlations alone do not show equivalency and can be extremely misleading and misrepresenting as will be seen below.

The Pearson correlation that was observed between the spring SRI and the MCAS test for Regular Education Hispanic students in the study was r= +.69 (N= 1651). This latter correlation was significantly lower at the <.001 level (Z= 3.4) than was observed between the SRI and the state of Florida Comprehensive Assessment Test (FCAT) (Knutson, 2006), which was r= +.79 (N= 82,954). This large difference between the two correlation is due to several factors as well as sample composition, which is discussed elsewhere, but the point is again emphasized here that just because two tests highly correlate does not mean that they are equivalent, measuring the same thing or are equally difficult, as will be seen in the next analyses.

Table 11 presents the cross tabulation results of Pass-Fail frequencies for the Spring SRI and MCAS tests. As table be seen from Table 11, "cross-tabulation" results of the way both tests categorize students as to their reading achievement level showed that 68.1% of the students in the sample failed the mandated-state test MCAS (31.9% passed) as compared to 59.6% of the same students who passed the spring SRI (twice as many as the MCAS) within the next two weeks! Moreover, 29.8% (about 3 in 10) of the students who passed the SRI failed the MCAS and at the same time, 38.2% of the students who failed the spring SRI also failed the MCAS and only 2.2% of the students who failed the SRI passed the MCAS. All of the percentage differences above are statistically significant at the .001 level or greater. These results clearly showed that the SRI is not a valid criterion to predict MCAS reading scores at varying performance levels. It is also obvious from Table 11 that these two tests are giving two very different (and substantively very different) pictures of reading capabilities and SFA program success. The SRI is giving a very over-rosy picture of SFA program success and students reading levels as compared to the state-mandated MCAS test, which reflects the state's view and definition of what reading competencies to what level of attainment ALL students must have to be successful in the twenty-first century economy. The state's definition and view was established by a broad array of participants and experts and through numerous (public) screening processes including the state's development and validation of its criterion measure. It's view and measure, therefore, is more of a broad consensus view of reading competencies and levels than those reflected in either the SRI (which was not developed by the SFA program) or the SFA program (which was not developed by broad public consensus). The Massachusetts State Exam (MCAS), moreover, has been judged (as previously stated) by the U.S. Department of Education's Institute of Educational Sciences to be one of the top 5 State tests in the country and to be both a difficult and highly excellent test and standard (NCES, 2007). Given these points, is it really not very difficult to understand, given the clear and simple results presented in Table 11, why students, parents and administrators in the River City school system were and are stunned when they are informed of their MCAS test results after being told how well they are doing and progressing in reading for three consecutive quarters using the SRI and then so for three consecutive years. Is there really any puzzlement, given the clear and simple results given in Table 11, why there is such an outcry by all stake holders in the River City school system against the MCAS tests including all of the erroneous reasons typically given about why these highly dissonant results occurred? But the further and more important questions of note here are, "Who is responsible for this shock and why," and might this factor be related to the very high degree of school alienation observed among Hispanic students, but particularly Hispanic males. Perhaps the "regrouping misclassification effects" of the locally-chosen SRI and the over-rosy picture of performance and gain and then the shock of the MCAS results just might be factors in the school attitudes and alienation observed in this community.

Table 12 presents the cross-tabulation of Pass-Fail frequencies for the Spring SRI and MCAS Tests by Years of SFA Instruction for Regular Education Hispanic Students.

As can be seen from Table 12, the analyses of the cross-tabulated pass-fail frequencies for the SRI and MCAS by year of SFA instruction revealed the following:

- A total of 71.5% (632 of 884) of the students who received one year of SFA failed the MCAS as compared to 55.0% (486 of 884) who passed the SRI
- 71.5% (632 of 884) of the students who received one year of SFA failed the MCAS as compared to 62.9% (370 of 588) of the students who received three years of SFA. This 8.6% reduction in failure rates on the MCAS for students with three years of exposure to SFA instruction as oppose to one year is statistically significant at the .001 level (Z = 2.6) and indicates that increased exposure to SFA program is having some positive effects and produce gains in reading achievement even on the "more difficult" MCAS test, even taking into account that non-random and to some degree incomparable samples or groups are being compared here. An 8.6% difference is well beyond what would (most probably) be observed if the difference were due solely to uncontrolled marginal differences between the variables and samples being compared
- 42.6% (377 of 884) of the students who received one year of SFA instruction failed the SRI and the MCAS, whereas 31.6% (186 of 588) of the students who received three years of SFA failed the SRI and the MCAS. This 11.0% reduction in failure rates was significant at the .001 level of significance (Z= 3.7) and is additional evidence that increased exposure to the SFA program reduced failure rates on both the SRI and the MCAS
- The percent of students earning passing scores on the MCAS increased from 28.5% for students who received one year of SFA to 37.1% for students who received three years of SFA instruction. Likewise, the percentage of students passing both SRI and MCAS increased from 26.1% for students who received one year of SFA to 35.2% for students who received three years of SFA. This 9.1% difference in "double pass" rates with more exposure to the SFA program is highly significant at the .0001 level (Z= 7.1) and is most probably an upper case or best-estimate of the SFA program's cumulative effect

Although a 9.1% improvement in MCAS pass rates cumulatively over a three year period is very statistically significant and not to be un- or under-appreciated, it is not the kind of significant improvement in pass rates that one would call highly significant qualitatively over a three year period for such an extensive and well funded program. The view of the success of the SFA program using the MCAS is very different than the view of the success of the SFA program using the SRI, as is reasonably clear from these analyses. However, it should be clearly noted that program effects of a certain kind were found for the SFA program on the state-mandated MCAS. The question, of course, was how uniform were the effects found.

Table 13 presents the cross-tabulation results between SRI and MCAS for females and males subjects in the study. As can be seen from Table 13 the cross-tabulation analyses of the variables SRI and MCAS by gender revealed the following:

Table 13: Cross-tabulation of pass-fail frequencies for spring SRI and MCAS tests

	MCAS		
SRI	Fail	Pass	Total
Fail	563 (38.2%)*	32(2.2%)	595 (40.4%)
Pass	439(29.8%)	438 (29.8%)	877 (59.6%)
Total	1002 (68.1%)	470 (31.9%)	1472 (100.0%)

*: Table percentages are percentages of total subjects in table

A total of 63.5% (484 of 762) of the female students failed the MCAS as compared to 62.9% (479 of 762) who passed the SRI. Results for females between the two tests were the same.

- A total of 73.0% (518 of 710) of the male students failed the MCAS as compared to 56.1% (398 of 710) who passed the SRI. Results for males on the two tests were not the same and in fact were every markedly different. The failure rate for males on the MCAS were 16.9 % higher than on the SRI (a roughly 1 in 5 difference!), which is significant at the .0001 level (Z= 7.2) using an adopted Z-test for correlated proportions
- 9.5% more male students failed the MCAS than female students which is a statistically significant difference at the .001 level (Z= 3.3), whereas only 5% more males failed the SRI than females which is also statistically significant at the .05 level (Z= 2.1). It should be clearly noted here that the gender differences in failure rates observed on the MCAS is twice as large as on the SRI and that the significance levels reported are for two-tailed (non-directional) tests rather than for one-tailed (directional) tests which would be appropriate for these and all other gender tests done. The results reported here on gender differences, therefore, are a very conservative statistically and this statistical testing strategy was employed to compensate for some of the imperfections in the data and study design
- 34.9% of the female students failed the SRI and the MCAS as compared to 41.8% of the males who failed both tests. The difference between these two percentages is statistically significant at the .0001 level (Z= 4.72). Roughly, 7% more Hispanic males failed both tests than Hispanic females

As can be seen from Table 15, the cross-tabulation results are consistent with our a priori hypotheses. The program effects found for the SFA program implemented in the River City School were due primarily to the performance of Hispanic Female students with Hispanic male students doing less well. It was, therefore, a differential and not “ALL” effect disguised in the overall (aggregate) frequencies and averages. So, the results of Success For All in the River City Schools systems was not quite due to success for all types of groups of students. Rather, it was due primarily to the success for some groups (types of learners) as opposed to others. Female scores, however, should be higher than male scores at each of these grade levels given the “developmental lead” females have during these two time periods. Therefore, to further examine these differences, more detailed analyses were conducted of the differences reported above, which included the independent variables grade, years of SFA instruction and gender and use of various modes of ANOVA. The 4x4 cross-classification results showed the same findings as given above and are reported.

Gender differences: A factorial (2×2×2) design was used to evaluate the effect of the independent variables gender, grade and years of SFA instruction. Results were analyzed on the basis of the possible interaction among the independent variables hypothesized before the study was done. The dependent variable was achievement performance of male and female students in grade 4 and 7 as measured by the MCAS, as this was both the psychometrically and substantively best as well as fairest and most stringent test available to test the hypotheses posed. The actual number of subjects in these analyses (N= 1645) only included Regular Education Hispanic students who took the MCAS test. This study found that none of the uncontrolled background variables, such as degree of appropriate program implementation, SES indirectly measured by the relative prosperity of the neighborhood (according to Census data) in which the school was located, attendance, age and percentage of students in the school receiving free lunch, were observed to have significant main effect or interaction effects with reading achievement level, individually or in combinations, accounting for more than 1% of the variance on MCAS scores, with this one percent effect, when observed, being observed for combinations (namely, interactions) of these aforementioned variables.

Table 16 presents the cell means, N’s and Standard deviations for a Grade by Gender by Years of the SFA program three-way of MCAS score for Regular Education Hispanic students (N=1,645) and Table 17 presents the three-way ANOVA results table for the data given in Table 14. The ANOVA results showed that a significant main effect was found for gender (F= 67.45, df1=1, df2= 1637, p < .0001). This result shows that female Hispanic students obtained significantly higher average scores on the MCAS (M= 44.41) than did male Hispanic students (M= 40.69) which was the main a priori hypothesis of this study. However, no significant difference was found for “number of years of exposure to the SFA program,” or the interaction between years of exposure and gender. These results showed that gains made by Hispanic female students in both years were statistically about the same, greater than those of the Hispanic male students, but that the gap was widening between females and males from the fourth to seventh grade level despite the amount of exposure to the program the males and females received.

A significant main effect was also found for grade effect (F= 4.28, df1=1, df2= 1637, p< .04). The results of the study showed that fourth grade Hispanic students had significantly higher average scores on the MCAS (M= 43.28) than seventh grade Hispanic students (M= 42.05). This difference or result was significantly larger, moreover, for fourth graders with three years of SFA instruction as compared to those fourth grade students with one year of SFA instruction. However, this three years as compared to one year of the SFA program difference was much less pronounced for seventh graders, which suggests that the SFA is not being as (marginally) successful after the fourth grade as it is before the fourth grade and that further research is needed on this point or finding. Furthermore, the group with the lowest MCAS mean performance levels (relative to all of the other comparisons) was seventh grade Hispanic males with one year of the SFA program (M= 39.14) and the second lowest was seventh grade Hispanic males with three years of the SFA program (M= 41.74). This finding not only supports our hypotheses about the gender difference in the SFA program’s success, but also the previously reported finding that the SFA program is, most probably, not being very successful at all with Hispanic male students beyond the fourth grade.

ANOVA results using MCAS scores in the study also showed a three-way interaction between gender, grade and years of exposure to the SFA program (F= 3.67, df1= 1, df2= 1637, p< .06). This three-way interaction was due primarily to the differences in reading achievement between Hispanic females and Hispanic males becoming more pronounced as they become older and get more exposure to the SFA program. The difference between male and female students with one year of SFA instruction gets bigger with more exposure to the program. Both grade four and seven Hispanic females scored higher in both years than did their Hispanic male counterparts with the gaps between the two becoming larger. However, no statistically significant differences were found for year or the interaction between year and gender, as the interaction was disordinal and disordinal interactions are often statistically insignificant or “silent” (due to the use of marginal averages for the testing), as in the present case. Although these differences can also be seen when using the spring SRI Lexile scores, the sizes of the differences are greatly lessened and very much “muted” to the point of seeming to be inconsequential perhaps because of the poor psychometric properties of the SRI test and it being a much easier test than the MCAS. ANOVA results using the SRI showed a much smaller difference between males and females on reading achievement. This smaller difference is due to the SRI being very unreliable and a much easier test than the MCAS and thus having a "ceiling effect," as it does not have the more difficult achievement levels as part of what it measures. These facts are just further evidence of the unsuitability and invalidity of the SRI as a fair and reasonable assessment of reading achievement gains and differences, as well as program success.

The findings of this study raise serious questions and far reaching questions about locally chosen success criteria for educational problems and program evaluations (even if they are chosen or recommended by university professors) versus state-mandated success criteria that need further research and study. The SRI (the locally-chosen/university and SFA recommended success criterion) over-estimated the effects and the success of the SFA program by 50% as compared to the state-mandated MCAS test. A difference of this order of magnitude is simply not acceptable (or a matter of academic or research freedom) today or in the current educational climate, particularly given the state-testing shock, fall out and collateral damage it induces on stake holders and particularly students. This same point holds for the misclassification and mis-grouping havoc the SRI caused in the SFA program in the River City schools and all of the negative impacts this factor, which happened each school quarter, had on students as well as teachers and parents. Research, program providers and school administrators have responsibilities and accountabilities and very clear responsibilities and accountabilities on all of these items and all of the points made in this study. It is not just students and classroom teachers who are responsible and accountable now and in the current educational context.

Little research has been done of the relationships between locally-chosen success criteria and state-mandated success criteria and a great deal more work and research needs to be done on this topic and issue to understand how successful schools are really being relative to their program and the mandates of the Educational Reform Act of 1993 and this same points holds for research studies on programs that are used or are going to be recommended for use in schools in the current age of educational reform.

Table 14: Cross-tabulation of pass-fail frequencies for the spring SRI and MCAS tests by years of SFA instruction for regular education hispanic students

			MCAS		
Years of SFA			Fail	Pass	Total
1 Yr.	SRI	Fail	377 (42.6%)*	21 (2.4%)	398 (45.0%)
		Pass	255 (28.8%)	231 (26.1%)	486 (55.0 %)
Subtotal			632 (71.5%)	252 (28.5%)	884 (100.0%)
3 Yrs.	SRI	Fail	186 (31.6%)	11 (1.9%)	197 (33.5%)
		Pass	184 (31.3%)	207 (35.2%)	391 (66.5%)
Subtotal			370 (62.9%)	218 (37.1%)	588 (100.0%)
Total			1002 (68.1%)	470 (31.9%)	1472 (100%)

*The percentages are percentages of all subjects (the Total N) in table

Table 15: Cross-tabulation of Pass-Fail Frequencies for the Spring SRI and MCAS Tests by Gender for Regular Education Hispanic Students (N= 1472)

			MCAS		
Years of SFA			Fail	Pass	Total
Female	SRI	Fail	266 (34.9%)*	17 (2.2%)	283 (37.1%)
		Pass	218 (28.6%)	261 (34.3%)	479 (62.9 %)
Subtotal			484 (63.5%)	278 (36.5%)	762 (100%)
Male	SRI	Fail	297 (41.8%)	15 (2.1%)	312 (43.9%)
		Pass	221 (31.1%)	177 (24.9%)	398 (56.1%)
Subtotal			518 (73.0%)	192 (27.0%)	710(100.0%)
Total			1002 (68.1%)	470 (31.9%)	1472 (100%)

*: The percentages are percentages of all subjects (the Total N) in table

Table 16: Cell Means, N's and SD's for ANOVA Results for the MCAS by Gender, Grade and Years of SFA Instruction (N = 1645)

		Gender								
		Female			Male			Total		
Years of SFA	Grade	N	M	SD	N	M	SD	N	M	SD
1 yr.	4	242	45.38	8.10	199	42.82	8.54	441	44.23	8.39
	7	229	42.70	8.91	238	39.14	9.05	467	40.89	9.15
	Total	471	44.08	8.60	437	40.82	9.00	908	42.51	8.94
3 yrs.	4	174	44.75	9.39	166	39.20	10.41	340	42.04	10.27
	7	213	44.85	8.88	184	41.74	9.40	397	43.41	9.24
	Total	387	44.81	9.10	350	40.54	9.96	737	42.78	9.75
Total	4	416	45.12	8.66	365	41.18	9.60	781	43.28	9.31
	7	442	43.74	8.95	422	40.27	9.28	864	42.05	9.27
	Total	858	44.41	8.83	787	40.69	9.43	1645	42.63	9.31

Table 17: Grade, Gender and Years of SFA Instruction ANOVA Table for MCAS Scores in Table 14 (N = 1645)

Source	df	Mean Square	F	Sig.
Year	1	6.6	0.08	0.78
Gender	1	5511.5	67.45	0.00
Grade	1	349.5	4.28	0.04
Year * gender	1	162.5	1.99	0.16
Year * grade	1	2045.9	25.04	0.00
Gender * grade	1	51.6	0.63	0.43
Year * gender * grade	1	299.8	3.67	0.06
error	1637	81.7		
Total	1645			
Corrected total	1644			

R Squared= 0.061 (Adjusted R Squared= 0.057)

DISCUSSION

In such studies researchers and program makers really are no longer free to use any criterion or test of their choosing and doing so without relating and calibrating the instrument chosen to the higher quality state standards. Until this type of research is done and the relationship between various measures are clearly established empirically and substantively, one would be wise to be extremely cautious about program claims of success and effects as well as about literature reviews, best practices summaries and meta-analyses that just indiscriminately pool and average effects from various and different criterion measures which have unknown relationships and who equivalency is just assumed as a matter of faith. Educational researchers, professionals and policies makers need to address these issues and the underlying quality and equivalency of the evidence base that is being used to make educational and policy decisions; namely, the very large and game changing elephant in the room. Psychometrics and psychometric knowledge, expertise and evidence counts and counts big time and should be the first question address by ALL and not the last question addressed my MOST in the research, development and evaluation of educational efforts, programs and policies. One of us (Carifio) has long advocated the development and use of standard models as an absolutely necessity in all areas of education (Carifio, 2005). Nowhere is this point more true and is the need greater than in the area of reading.

CONCLUSION

In sum, the SFA program was shown to have some limited and some selective effects on the acquisition of reading skills primarily for female as opposed to male Hispanic students as predicted. Also, as stated above, the findings of this research strongly suggest the need of data quality control by an external independent party and that data quality management training and responsibility be part of principalship training, as it is in business where it is view as a critical management function. This study also documented and clearly showed that school leaders must constantly monitor data collection and data analysis, as meaningful information can only be obtained from the thoughtful process of inquiry and analysis.

Moreover, as a sound educational investment, we strongly recommends that school districts develop policies that allow them to carefully analyze the implications, lasting educational benefits, as well as cost-effectiveness of implementing reform models such as SFA, particularly as “one size (or program) clearly does not fit all.” A wide variety of research tells us that males and females learn differently and differ from each other. Male underperformance in reading cannot be ignored. Reading and writing are basic skills required fro functioning in the 21st century. As educators in learning communities, we must consider gender difference when developing curriculum, implementing teaching methods and instructional practices, as well as analyzing data. There is no one silver bullet to solve the gender gap issue in schools. However, the development of intervention strategies, the right pedagogic approach and quality teaching paired with the closely monitoring of males progress can be effective when raising males’ reading achievement.

There is currently a hotly contested national controversy between the Successful all Program and the Reading First program (Glen, 2007). We are fully confident that if this study was replicated for the various implementations Reading First program approach, the results found would be highly similar if not exactly the same as the results found here. The Reading First program also uses locally-chosen/university recommended success criteria and not state-mandated success criteria and the evaluative evidence for its claims is just as suspect and perhaps even more suspect as that for the SFA program. We are also fully confident that if this study was replicated using the Woodcock-Johnson tests that are the criterion in many of the evaluations of the SFA program currently reported (Slavin and Madden, 2006), the results would be highly similar if not exactly the same as the results reported here, as the Woodcock has only been re-normed recently and not revamped and is still the same test as it was fifty years ago and thus not really a measure of the reading competencies and levels needed in the 21-st century. Again, the issue is the standard and the quality and validity of the standard in terms of public mandates, which are now the ineluctable modality of reality, as Stephen Daedalus was prone to say, to which we must all accommodate to some minimum degree, including researcher, program providers, educational managers and university professors. This last point, as well as the severity and the importance of the multiple and wide variations in success criteria being used to evaluate No Child Left Behind Programs and in State Standards themselves has been made manifestly explicit in a new research report by the Institute of Educational Sciences (NCES, 2007) on the relationship of the assessments of students’ reading achievements and levels in the state given by the State’s Mandated NCLB test and the assessment of the students’ reading achievement’s and levels given by the National Assessment of Educational Progress (NAEP) tests (NCES,2007). Very wide differences were seen in the assessments of students’ readings abilities and levels between these two measures with the level and percentages of the discrepancies being similar to the gaps seen in this study.

Three important points need to be noted about the NCES (2007) study reported above. First, the NCES study was done a year after this study was done. Second, reading scores for Massachusetts’ MCAS test was at the top of the list of states relative to correlating with scores from the NAEP test and in fact were almost in one-to-one correspondence. So the current study gives a very reasonable estimate of all of the problems it has identified and discussed and the levels of actual discrepancies from the NAEP standard as the success criterion right down to the level of the locally chosen success criterion, which in this case was the SRI for the SFA program. The third point of importance is that the data in the NCES (2007) study was not broken out separately for different minority groups so that the manner in which NCLB reading programs were being unsuccessful for Hispanic students could not be seen from the NCES data, nor could the manner in which these programs were being particularly unsuccessful for Hispanic male students was completely masked in the data as well as how bad this problem actually is from state to state.

The critical, practical and humane never mind political importance of these gaps and the findings of this study cannot be over-estimated. One must ask oneself how one would feel and how one would react if one had been told all year long that one was doing a good job in reading (on the SRI or similar test) and then told one was failing reading miserably (on the MCAS or similar above national state standards test) and then told one is doing a good job on reading (on the SRI or similar test) again two weeks later! What would one (and one's parents) possibly think and believe and believe about school and school as a serious place and success path for one's self (or one's children) and just who is responsible for and accountable for this kind of fairly baffling experience by young Hispanic males and their parents. Who is, in fact, creating the many problems we observed and the many difficulties that we are documenting that are "out there" and attributed to others. The answer was well-stated by Pogo a good while back and needs to be heard and reconsidered today. We need to reflect on US and critically so.

Last, it again needs to be stressed that this is a retrospective study and a study that needs to be replicated and that we are well aware and understanding of and sensitive to various facts about people and events happening and working in parallel and mandated and driven haste and what the results of such actions often are. However, we are also very clear about what questions are first questions and what questions need to be addressed first and not last, which is in part one of the central foci of this study.

There is not a first or a second author for this study; both authors are first authors of this study as this is an interdisciplinary and collaborative study.

REFERENCES

Bracey, G.W., 2006. Reading Educational Research: How to Avoid Getting Statistically Snookered. 1st Edn., Heinemann, USA., ISBN-10: 0325008582, pp: 188.

Brown, R.S. and D.T. Conley, 2007. Comparing state high school assessments to standards for success in entry-level university courses. Edu. Assessm., 12: 137-160. DOI: 10.1080/10627190701232811

Carifio, J., 2005. Towards a standard integrated information processing/ cognitive model of learning. University of Massachusetts Lowell. http://www.ihpst2005.leeds.ac.uk/papers/Carifio.pdf

Fensham, P.J., 2004. Defining an Identity: The Evolution of Science Education as a Field of Research. 1st Edn, Springer, USA., ISBN-10: 1402014686, pp: 247.

Glen, D., 2007. Reading for profit. Chronicle of Higher Education. http://chronicle.com/article/Reading-for-Profit/10194

Kerlinger, F.N. and H.B. Lee, 2000. Foundations of Behavioral Research. 4th Edn., Harcourt College Publishers, California, ISBN-10: 0155078976, pp: 890.

Massachusetts Department of Education, 2007. Progress Report on Students Attaining the Competency Determination Statewide and by School and District: Classes of 2007 and 2008. Massachusetts Department of Education. http://www.doe.mass.edu/mcas/2007/results/CD.pdf

NCES, 2007. Mapping State Proficiency Standards Onto the NAEP Scales 2005-2007. Institute of Education Sciences. http://inpathways.net/mapping%20state.pdf

Rump, A.A. and N.K. Lesaux, 2006. Meeting Expectations? An empirical investigation of a standards-based assessment of reading comprehension. Edu. Eval. Policy. Anal, 28: 315-333. DOI: 10.3102/01623737028004315

Shulman, L.S., C.M. Golde, A.C. Bueschel and K.J. Garabedian, 2006. Reclaiming education's doctorates: A critique and a proposal. Edu. Res, 35: 25-32. DOI: 10.3102/0013189X035003025

Slavin, R.E., 1996. Every Child, Every School: Success for All. 1st Edn., Corwin press, USA., ISBN-10: 0803964366, pp: 264.

Slavin, R.E. and N.A. Madden, 2006. Success for All/Roots and Wings: 2006 summary of research on achievement outcomes. Johns Hopkins University. http://www.successforall.net/_images/pdfs/SummaryofResearch-2003.pdf