Reform-Related Changes in Educational Practices in SSI and Non-SSI States

Janet H. Kane and B. Darwin Kaufman

Wisconsin Center for Education Research

University of Wisconsin - Madison

Presented as part of a symposium on:

Analyzing Statewide Change in Mathematics through

NAEP Analyses Linked with Qualitative Analyses

American Educational Research Association Annual Convention

New Orleans, LA

April, 2002

The research presented in this paper was supported by the National Science Foundation under Grant No. REC-98874171 and by the Wisconsin Center for Education Research, School of Education, University of Wisconsin-Madison. Any opinions, findings, or conclusions are those of the authors and do not necessarily reflect the views of the supporting agencies.

Reform-Related Changes in Educational Practices in SSI and Non-SSI States

Introduction

The National Science Foundation (NSF) instituted the Statewide Systemic Initiatives (SSIs) in 1991 to promote systemic educational change based on high academic standards. The *Curriculum and Evaluation Standards for School Mathematics* (NCTM, 1989) defined the kind of standards-based curricula NSF encouraged. While states were not required to adopt the NCTM *Standards*, the expectation was that the policies and practices of the SSI states would at least be consistent with them.

Higher mathematics achievement for all students, including those historically underserved, is the ultimate measure of the success of an SSI. NSF labeled its achievement goals as the outcome drivers of educational system reform. In addition, NSF identified several process drivers, or policies and practices in support of high student achievement. This paper examines the process drivers, using items from the State NAEP teacher questionnaires to create indicators of the drivers.

By using items from the State NAEP, the same measures can be used with all states. In any NAEP year, all states that received SSI funding can be compared to all other states. Change over time can be assessed when the state NAEP teacher questionnaires includes the same or similar items from one year to the next.

The method section of this paper describes the development of reform-related indicators and the samples used to study the effects of the SSI program. Results of three different kinds of analyses are presented. First, cross-sectional and longitudinal designs are used to compare SSI and non-SSI states on the reform indicators. Second, multiple regression modeling is used to describe relationships among the indicators and student achievement. Third, differences among the SSI states are described, with reference to student achievement gains and accompanying changes in the reform indicators.

Method

Developing Indicators from NAEP Teacher Questionnaires

The state NAEP, begun in 1990 at grade 8, is designed to estimate parameters for an individual state. Besides the achievement test items, state NAEP includes teacher, student, and school questionnaires. Starting in 1992, the mathematics portion of the state NAEP has been administered every four years.

The research reported here uses items from the state NAEP teacher questionnaires to describe characteristics of the SSI and non-SSI states. The questionnaires requested information about the teachers' backgrounds, general training, and their instructional practices.

Teacher questionnaire item results are often reported in terms of the proportion of students whose teachers selected a specific response option. (See, for example, Shaughnessy, Nelson, & Norris, 1998.) Analyses are limited to nonparametric approaches that compare two or more groups on the proportion of responses in each category. Statistical models using questionnaire items frequently create dummy variables, collapsing the response categories into a dichotomous variable and, consequently, reducing the information content of the measure. As an alternative approach, we created scales by combining responses to related items. With a scale, random error is reduced and true score variability is increased. A scale simplifies reporting because the responses to several items are combined into a single measure. Scale scores allow the use of parametric statistics when the distribution of scale scores approximates a normal distribution.

We began with an examination of the teacher questionnaires in order to identify items indicative of the goals of the Statewide Systemic Initiatives. We used a model of systemic reform (Clune, 1998) to categorize the items and then examined the responses to each selected item. As a result of this review, some selected items were eliminated because almost all respondents chose the same response option, usually the highest or lowest. Either there is an extremely high level of teacher agreement on these items, or the items are not sensitive to differences among teachers.

We then reviewed the individual items of the 1996 grade 8 teacher questionnaire and discussed the "best" answer, from the perspective of mathematics reform. Most response options ranged from a low of "Never" or "None" to a high of "Almost Every Day" or "A lot." For most items, responses in the NAEP data set were coded from 1 to N, with N as the number of response options. In our analyses, we reversed the scales when necessary, so the highest value represented the most frequent occurrence. In discussions, project staff generally agreed that with successful statewide systemic initiatives, reform-related practices would increase, but that traditional practices focused on mastering facts, concepts, and routine procedures would also have a major role. We had concerns about a simple scale where "more" of something was considered to be "better" and explored assigning the greater number of points to response options that described a moderate frequency of occurrence. The alternative scales were evaluated using Cronbach's coefficient alpha, a measure of internal consistency (Cronbach, 1951). None of the proposed scoring systems improved on the original 1 to N coding, with 1 indicating the lowest frequency and N the highest.

The extensive review and analysis of the state NAEP 1996 teacher questionnaire items resulted in six indicators of mathematics reform (Webb, Kane, Kaufman, & Yang, 2001):

- ***I(RC), Relative Emphasis on Reasoning and Communication*** – how much reasoning and communication were addressed, relative to facts and procedures.
- ***I(MD), Mathematical Discourse*** – a scale of students' opportunities to discuss, present, and write about mathematical ideas.
- ***I(C), Calculator Use*** – a scale of the extent to which students used calculators in the classroom and on tests.
- ***I(S), NCTM Standards*** – a single item that asked about teachers' knowledge of the NCTM *Standards*.

I(PD), Last Year's Professional Development – a single item that asked how much time teachers spent in professional development in mathematics or mathematics education during the last year.

I(RT), Reform-Related Topics Studied – a count of the number of reform-related topics teachers have studied out of the seven topics listed in the NAEP questionnaire.

The 1996 state NAEP teacher questionnaire was not the same as the questionnaire administered in 1992. Several items were added, particularly items related to curricular reform. Wording of some items was modified, and for some items the number and labels of the response options was changed. Despite these differences, the similarities of the questionnaire items in 1992 and 1996 provided a means for comparing SSI and non-SSI states across time.

Standardizing the Reform Indicators

The ranges of the indicators were quite different, depending on the number of items that were used to create the indicator and how the responses were scored and combined. To facilitate comparisons, all measures were transformed to a standard scale with a mean of 0 and a standard deviation of 1. All states that participated in state NAEP in a given year served as the standardization sample. Because scales were not identical from one year to the next, standardization occurred within each year. Consequently, year-to-year comparisons of the standard scores provides information about the relative change of one group compared to another, but not the absolute change for both groups. In addition, if some state means increase from one year to the next on the standardized scale, other state means decrease.

Samples

Twenty-five states and Puerto Rico received funding through NSF's SSI program, and 25 states did not. NSF discontinued funding early for four states, resulting in 21 states with the full five years of funding. Under the SSI program, awards were made in three cohorts - the first in 1991, the second in 1992, and the third in 1993.

Not all states participated in state NAEP in any given year. Analyses and conclusions about the effects of the SSI program are limited to those states that chose to participate in state NAEP. While state NAEP also included data from the jurisdictions of Guam, Puerto Rico, the Virgin Islands, Washington, DC, and Department of Defense Schools, only state data were used for this study.

Yearly samples – 1992 and 1996. For each year of the state NAEP, comparisons can be made between all participating SSI states and non-SSI states, using all of the available data in a given year. Table 1 presents the number and percentage of SSI and non-SSI states participating each year at each grade.

Table 1
Number and Percentage of SSI and Non-SSI States Participating in State NAEP by Testing Year

	SSI States *n* = 21		Non-SSI States *n* = 25	
	n	%	*n*	%
1992				
Grade 8	18	86%	19	76%
Grade 4	18	86%	19	76%
1996				
Grade 8	18	86%	18	72%
Grade 4	19	90%	20	80%

Trend sample, 1992-2000. The trend sample includes those states that participated in three consecutive state NAEP administrations, 1992, 1996 and 2000. (See Table 2.) Fourteen SSI states (67% of all SSI states) and 13 non-SSI states (52% of all non-SSI states) are in the trend sample. By 1996, the first SSI cohort was completing its fifth year, and others were well into their third or fourth years. Data from the 2000 teacher questionnaire were not available for the analyses reported here.

While the 1992 measure provides a baseline for the 1996 measure, it is not necessarily independent of SSI. Since the first round of NSF funding started in 1991, some of the states had been funded for a time. More importantly, some of the states had extensive prior experience with reform initiatives, positioning them to be interested in and selected for NSF's Statewide Systemic Initiative program.

Table 2
Trend Sample States

SSI states *n* = 14	Non-SSI states *n* = 13
Arkansas	Alabama
California	Arizona
Connecticut	Hawaii
Georgia	Indiana
Kentucky	Maryland
Louisiana	Minnesota
Maine	Mississippi
Massachusetts	Missouri
Michigan	North Dakota
Nebraska	Tennessee
New Mexico	Utah
New York	West Virginia
South Carolina	Wyoming
Texas	

Unit of Analysis

State NAEP is designed to provide information about each state as a whole. The student is the sampling unit, and teachers' responses are matched with each of their students to define one record in the data file. Each student has an associated weight, based on the sampling plan, and state means are computed using weighted values (Allen, Jenkins, Kulick, & Zelenak, 1997). In this paper, the focus is on the state means and the variability among the means, rather than on the within-state variability.

With the state as the unit of analysis, SSI states are grouped together as replications receiving the treatment (e.g., the SSI program), and non-SSI states are grouped as replications not receiving the treatment. The SSI states used many and varied approaches to systemic reform. However, grouping the states together assumes that each is an instance of a general category, despite their differences. The statistical comparisons allow conclusions about whether something is more or less likely to occur in one group than another. There is no claim that all states in one group will share a characteristic that is not present in any of the states in another group.

Another caution in interpreting the results of these analyses is that, unlike experimental research, the SSI treatment was not randomly assigned to the states. States participating in the SSI program had to submit a proposal, and NSF selected the proposals to fund.

With states as the unit of analysis, the sample size is fairly small. In order to reject the null hypothesis, differences have to be fairly large. For comparisons between SSI and non-SSI states, we used an alpha level of .10 (Grissmer, Flanagan, Kawata, & Williamson, 2000).

Analytic Approaches and Results

Comparing SSI and Non-SSI States on Indicators of Mathematics Reform: Cross-sectional and Longitudinal Approaches

Cross-sectional comparisons. We examined the effect of the SSI program by comparing all SSI and non-SSI states in a given year. In this approach, all states that participated in state NAEP in a given year were included in the comparison. Table 3 reports the results.

In 1992, the multivariate F was not statistically significant at either grade (grade 8, $F = 0.66$, $p = 0.66$; grade 4, $F = 1.09$, $p = .39$). On most comparisons, the SSI states were slightly higher than the non-SSI states. In 1992, the largest difference was in two grade 4 indicators: Relative Emphasis on Reasoning and Communication and Use of Mathematical Discourse.

In 1996, there was a statistically significant multivariate F at both grade 8 ($F = 2.15$, $p < .10$) and grade 4 ($F = 3.58$, $p < .01$). At both grade levels, SSI states averaged higher than non-SSI states on I(RC), Relative Emphasis on Reasoning and Communication, and I(MD), Use of Mathematical Discourse. In addition, at grade 8, SSI states averaged significantly higher on I(S), Teachers' Knowledge of the NCTM *Standards*. At grade 4, SSI states averaged higher on I(PD), Time in Professional Development in the Last Year.

Table 3
Comparisons of all SSI and Non-SSI States that Participated in NAEP Each Year

	SSI states		Non-SSI States		
1992	*M*	*SD*	*M*	*SD*	*F*
Grade 8	*n* = 18		*n* = 19		
I(RC)	43.39	2.21	42.86	2.03	0.56
I(MD)	15.62	0.75	15.25	0.66	2.61
I(C)	11.35	0.97	11.21	1.13	0.16
I(PD)	3.26	0.23	3.26	0.21	0.00
I(RT)	4.77	0.29	4.69	0.39	0.41
Grade 4	*n* = 18		*n* = 19		
I(RC)	40.22	1.83	39.09	1.61	4.01
I(MD)	15.82	0.79	15.30	0.76	4.06
I(C)	7.28	0.64	7.21	0.59	0.12
I(PD)	2.58	0.21	2.59	0.22	0.02
I(RT)	4.72	0.26	4.73	0.24	0.02
1996					
Grade 8	*n* = 18		*n* = 18		
I(RC)	45.87	1.63	44.54	1.35	7.20*
I(MD)	23.29	1.21	22.33	1.01	6.82*
I(C)	9.92	0.68	9.85	0.72	0.08
I(S)	2.71	0.22	2.59	0.13	4.46*
I(PD)	3.49	0.35	3.30	0.24	3.60
I(RT)	5.20	0.33	5.02	0.28	3.22
Grade 4	*n* = 19		*n* = 20		
I(RC)	44.05	1.28	42.40	1.27	16.30*
I(MD)	23.89	1.04	22.98	1.22	6.09*
I(C)	8.11	0.51	7.91	0.48	1.51
I(S)	1.98	0.22	1.88	0.15	2.61
I(PD)	2.88	0.28	2.69	0.24	5.69*
I(RT)	4.87	0.24	4.78	0.30	1.02

*$p < .05$

In examining I(C), we also considered whether students were allowed to use calculators on the state's achievement tests. To identify states that allowed calculator use, we used the 1996 State Student Assessment Program Database (North Central Regional Educational Laboratory, 1996). Some of the states in the NAEP sample did not administer state tests in the 1994-95 school year, and others did not answer the question about the use of calculators on state tests. The subsample with testing information consisted of 27 states at grade 8 and 29 states at grade 4. At each grade level, 4 SSI states and 4 non-SSI states did not allow calculator use on their state assessments. At grade 8, a 2 x 2 ANOVA found that states allowing calculator use averaged significantly higher on I(C(96)). The mean for the states allowing calculators was 10.02 (*SD* = 0.64); states where calculators were not allowed had a mean of 9.30 (*SD* = .81). At grade 4, the

states allowing calculator use also averaged higher than the other states, but not significantly, with states that allowed calculators averaging 8.07 (*SD* = 0.56) and states not allowing them averaging 7.70 (*SD* = 0.38). This finding cautions researchers to interpret their findings in the context of state reform initiatives, in addition to the SSI.

The finding that SSI states averaged higher than the non-SSI states on several indicators of mathematics reform in 1996 and not in 1992 is evidence for the effectiveness of the SSI program. However, the cross-sectional results provide only weak evidence, because the states may have differed initially.

Longitudinal comparisons. In longitudinal comparisons, state means were compared across 1992 and 1996. A longitudinal design has the potential to provide stronger evidence for the effectiveness of the SSI program because change over time can be identified. However, the analyses have somewhat reduced power because of the smaller sample size: longitudinal comparisons were limited to those states that consistently participated in state NAEP—that is, the trend sample of 14 SSI states and 13 non-SSI states listed in Table 2.

The representativeness of the trend sample was examined by comparing the means of the trend sample with the means of the other states. The results are presented in Table 4. In general, the trend sample mean is not significantly different from the mean of other states participating in state NAEP.

Table 4
Representativeness of the Trend Sample

	Trend Sample		Other States		
1992	*M*	*SD*	*M*	*SD*	*F*
Grade 8	*n* = 27		*n* = 10		
I(RC)	42.78	1.98	44.03	2.28	2.71
I(MD)	15.42	0.70	15.47	0.81	0.47
I(C)	11.24	1.10	11.38	0.88	0.13
I(PD)	3.26	0.22	3.24	0.22	0.12
I(RT)	4.69	0.37	4.82	0.25	1.06
Grade 4	*n* = 27		*n* = 10		
I(RC)	39.62	1.99	39.70	1.18	0.02
I(MD)	15.59	0.88	15.44	0.57	0.25
I(C)	7.23	0.65	7.27	0.50	0.02
I(PD)	2.60	0.23	2.53	0.15	0.84
I(RT)	4.70	0.27	4.80	0.17	1.19
1996					
Grade 8	*n* = 27		*n* = 9		
I(RC)	45.02	1.55	45.87	1.86	1.86
I(MD)	22.69	1.30	23.16	0.98	0.94
I(C)	9.75	0.74	10.30	0.30	4.71*
I(S)	2.63	0.20	2.73	0.18	1.78
I(PD)	3.41	0.32	3.31	0.30	0.75
I(RT)	5.10	0.31	5.10	0.32	0.00
Grade 4	*n* = 27		*n* = 12		
I(RC)	43.19	1.42	42.22	1.75	0.00
I(MD)	23.38	1.28	23.51	1.08	0.09
I(C)	7.95	0.57	8.12	0.30	0.98
I(S)	1.90	0.18	2.00	0.20	2.39
I(PD)	2.80	0.27	2.74	0.27	0.48
I(RT)	4.83	0.26	4.82	0.31	0.00

*$p < .05$

Two different models were used to analyze change over time for the trend sample, depending on how the measures compared from one year to the next. When they were similar, but not exactly the same, a two-step hierarchical regression model was used. The prior year indicator was entered at Step 1, to assess the relationship between the two measures. At Step 2, SSI status was entered to assess the additional contribution of SSI status on the 1996 indicator. This approach can identify whether SSI states changed more than the non-SSI states. When the measures were the same, a repeated measures analysis of variance was used, with SSI as a between-subjects factor and time as a within-subjects factor. With this approach, absolute change over time can be examined as well as the interaction of time by SSI status. The following paragraphs present the results for each indicator at each grade level.

Grade 8

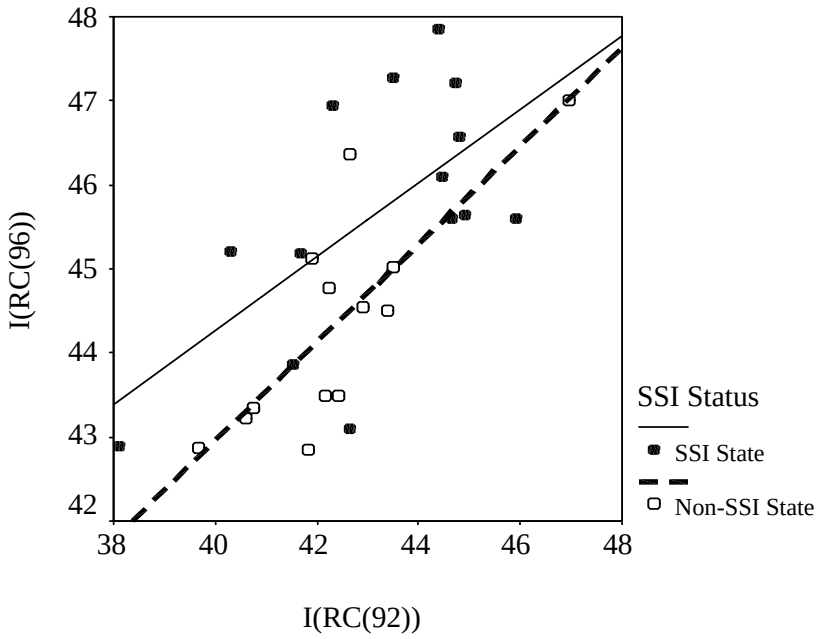
I(RC). I(RC), indicating the relative emphasis on reasoning and communication, is not directly comparable from 1992 to 1996. The regression analysis found that both the 1992 indicator and the state's SSI status were significantly related to the 1996 indicator. The results are summarized in Table 5 and Figure 1.

Table 5
Predicting I(RC(96)) from I(RC(92)) and SSI Status, Grade 8

	B	*SE B*	β	R^2	*F*	ΔR^2	F_Δ
Step 1							
$I_{RC(92)}$	0.54	0.11	.69	.48	22.78*		
Step 2							
$I_{RC(92)}$	0.49	0.17	.63				
SSI status	0.92	0.42	.30	.57	15.68*	.09	4.88*

*$p < .05$

Figure 1. Scatterplot of I(RC(92)) and I(RC(96)) for SSI and non-SSI states in the trend sample, grade 8.



As Figure 1 shows, the SSI states increased relatively more in I(RC) than the non-SSI states at grade 8, providing evidence that the SSI program resulted in states putting relatively more emphasis on reasoning and communication compared to facts, concepts, and routine problem-solving procedures. Since I(RC(96)) was significantly related to I(RC(92)), states

relatively high in 1992 were also relatively high in 1996, and those relatively low in 1992 were also relatively low in 1996.

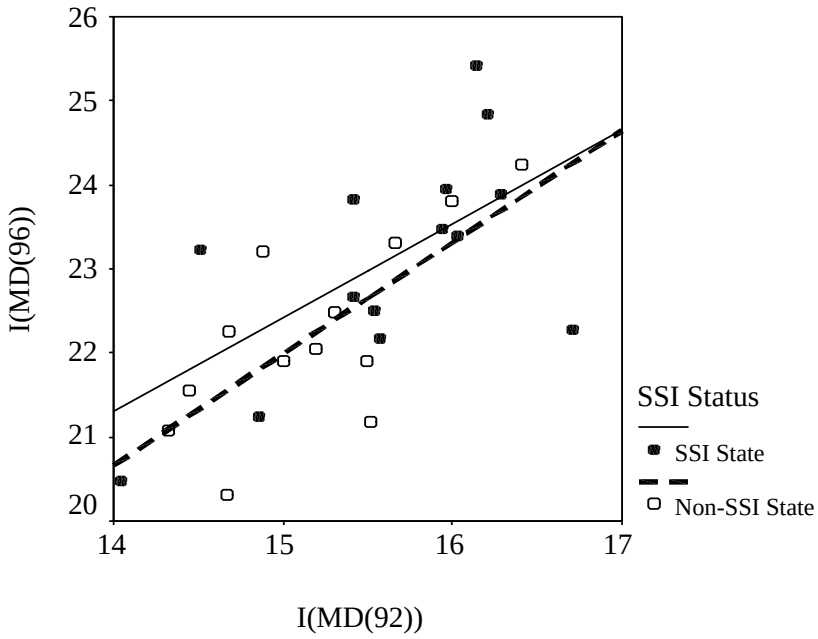
I(MD). This indicator of the use of mathematical discourse also cannot be directly compared from 1992 to 1996. Table 6 and Figure 2 present the results of the regression analysis.

Table 6
Predicting I(MD(96)) from I(MD(92)) and SSI status, Grade 8

	B	*SE B*	β	R^2	*F*	ΔR^2	F_Δ
Step 1							
$I_{MD(92)}$	1.28	0.26	0.69	.48	23.20*		
Step 2							
$I_{MD(92)}$	1.20	0.28	.65				
SSI status	0.34	0.38	.14	.50	11.91*	.02	0.80

*$p < .05$

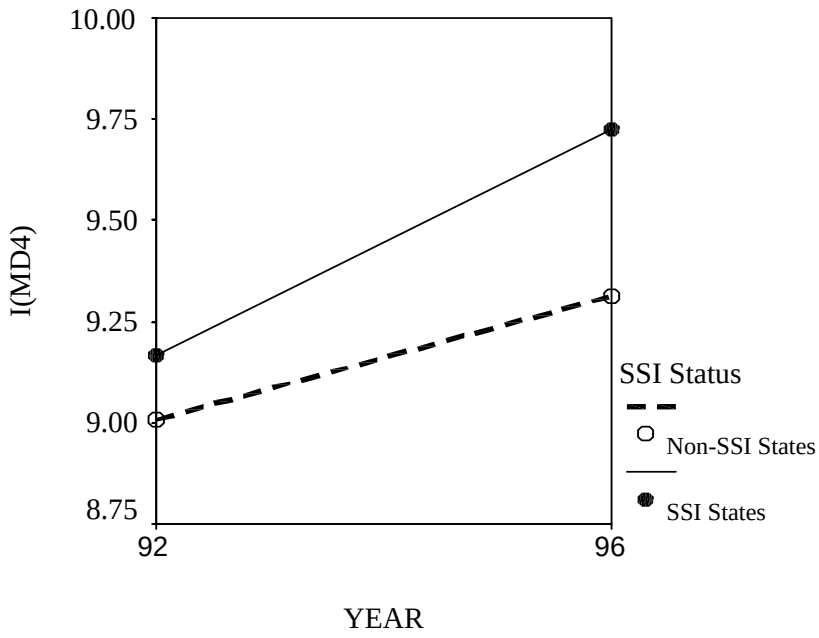
Figure 2. Scatterplot of I(MD(92)) and I(MD(96)) for SSI and non-SSI states in the trend sample, grade 8.



As the results show, states high in I(MD) in 1992 were also relatively high in 1996, and states low in 1992 were also low in 1996. For I(MD), the only predictor of the 1996 state mean was the 1992 state mean.

Four items on the mathematical discourse scale are exactly the same in 1992 and 1996. These items can be used to gain a sense of whether the overall use of mathematical discourse changed from 1992 to 1996. Figure 3 shows the results of a repeated measures ANOVA, using the four item scale, I(MD4). As the graph shows, I(MD4) increased significantly for both SSI and non-SSI states ($F = 28.65$, $p < .05$). The main effect for SSI was statistically significant ($F = 3.69$, $p < .10$), but the year-by-SSI-status interaction was not ($F = 2.47$, $p = .13$).

Figure 3. Change in the mean of I(MD4) for SSI and non-SSI states from 1992 to 1996, grade 8.



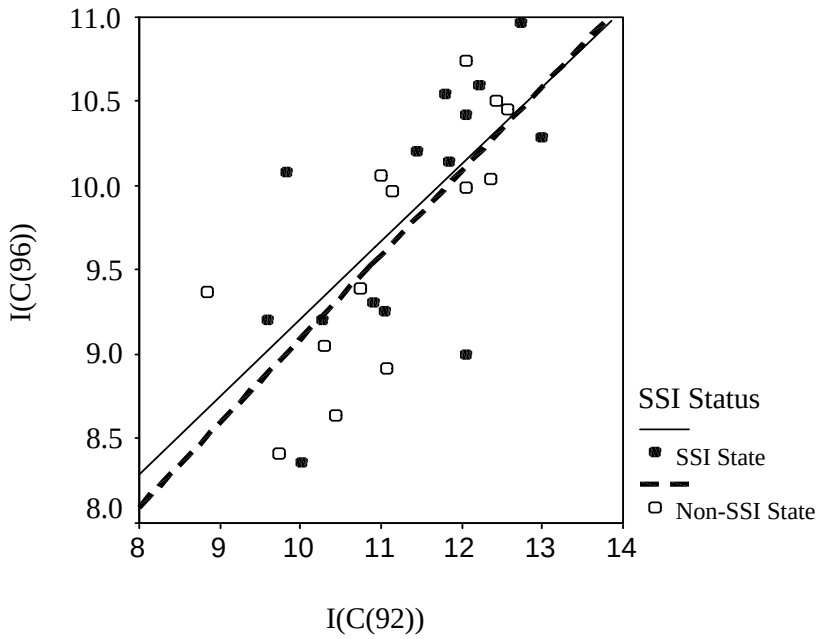
I(C). At grade 8, the calculator use indicator is scaled differently in 1992 and 1996. Table 6 and Figure 4 show the results. I(C(96)) is strongly related to I(C(92)), but SSI status does not add to the prediction of I(C(96)).

Table 6
Predicting I(C(96)) from I(C(92)) and SSI Status, Grade 8

	B	*SE B*	β	R^2	*F*	ΔR^2	F_Δ
Step 1							
$I_{C(92)}$	.48	.10	.71	.51	25.72*		
Step 2							
$I_{C(92)}$	.48	.10	.71				
SSI status	.07	.21	.05	.51	12.46*	.00	0.11

*$p < .05$

Figure 4. Scatterplot of I(C(92)) and IC(96)) for SSI and non-SSI states in the trend sample, grade 8.



A second regression model examined whether state policy about students using calculators on the state assessment was related to I(C(96)). Table 7 presents the results. In the reduced sample of 21 states, state policy was a significant predictor of I(C(96)).

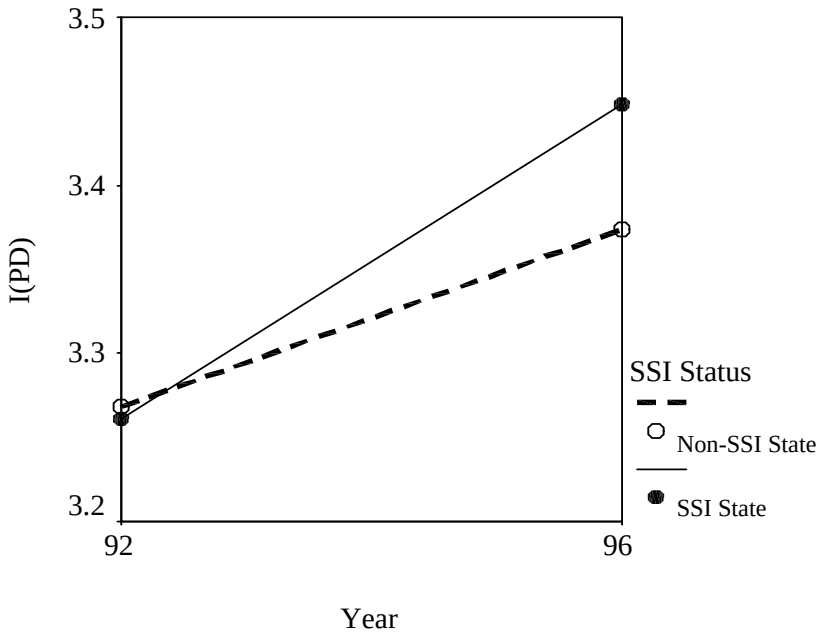
Table 7
Predicting I(C(96)) from I(C(92)), Calculator Use Policy, and SSI Status, Grade 8

	B	*SE B*	β	R^2	*F*	ΔR^2	F_Δ
Step 1							
$I_{C(92)}$	.40	.11	.59				
Calculator Use	.53	.27	.32	.58	12.65*		
Step 2							
$I_{C(92)}$	.40	.11	.59				
Calculator Use	.53	.28	.32				
SSI status	.04	.25	.02	.58	7.98*	.00	0.02

*$p < .05$

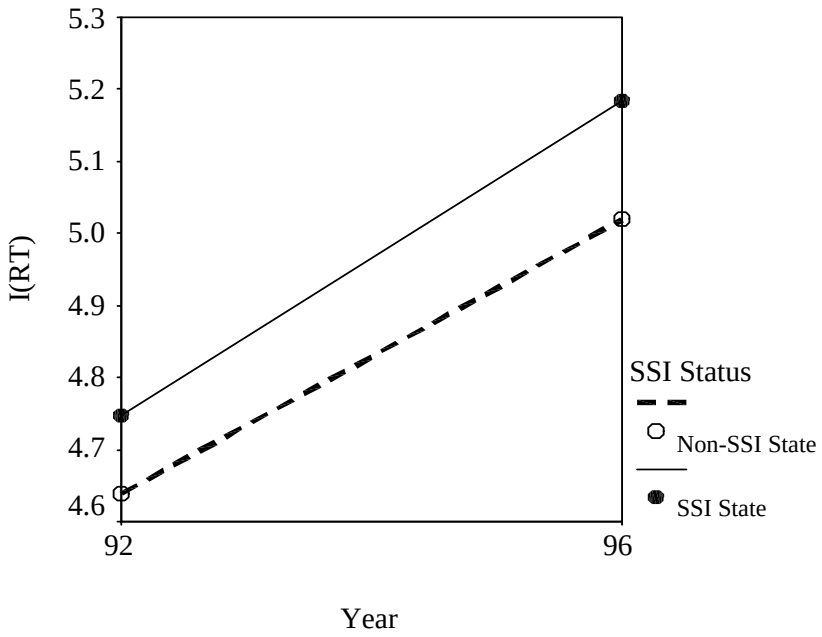
I(PD). I(PD), an indicator of the amount of time teachers spent in mathematics-related professional development last year, is scaled the same in 1992 and 1996. As Figure 5 shows, I(PD) increased from 1992 to 1996 ($F = 5.29$, $p < .05$). While SSI states increased slightly more than non-SSI states on this indicator, the difference was not significant ($F = 0.41$, $p = .53$).

Figure 5. Change in the mean of I(PD) for SSI and non-SSI states from 1992 to 1996, grade 8.



I(RT). I(RT) is an indicator of the number of reform topics teachers have studied out of the seven listed in the questionnaire. Figure 6 shows a significant increase in I(RT) from 1992 to 1996 ($F = 47.41$, $p < .05$). While the SSI states averaged higher than the non-SSI states in both 1992 and 1996, the main effect for SSI status was not statistically significant ($F = 1.33$, $p = .64$).

Figure 6. Change in the mean of I(RT) for SSI and non-SSI states from 1992 to 1996, grade 8.



Grade 4

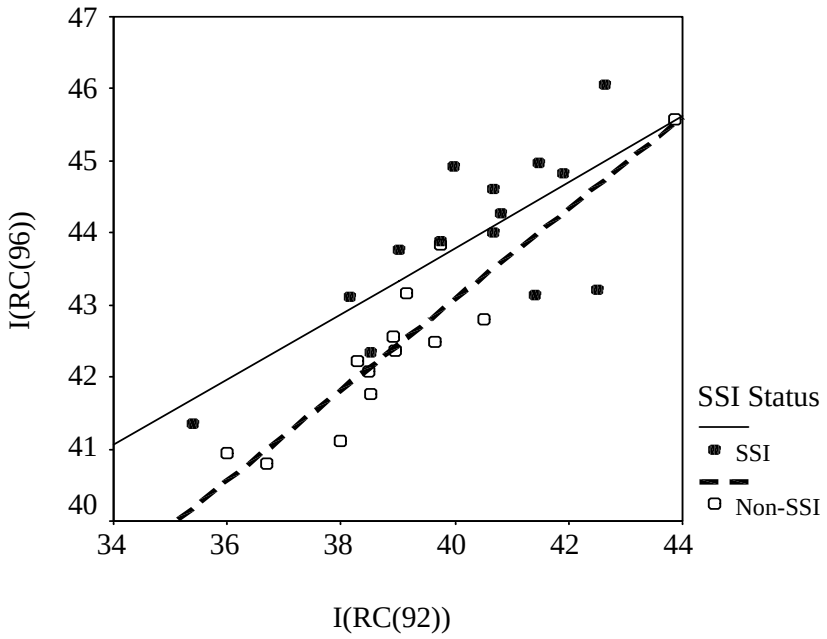
I(RC). At grade 4, the regression analysis found that both the 1992 indicator and the state's SSI status were significantly related to the 1996 indicator, replicating the findings from grade 8. The results are summarized in Table 8 and Figure 7. Figure 7 also shows that at grade 4 in 1992, five of the six states scoring highest on I(RC) were SSI states.

Table 8
Predicting I(RC(96)) from I(RC(92)) and SSI Status, Grade 4

	B	*SE B*	β	R^2	*F*	ΔR^2	F_Δ
Step 1							
$I_{RC(92)}$	.60	.08	.84	.70	59.14*		
Step 2							
$I_{RC(92)}$	.54	.08	.75				
SSI status	.80	.29	.29	.77	41.43*	.07	7.75*

*$p < .05$

Figure 7. Scatterplot of I(RC(92)) and I(RC(96)) for SSI and non-SSI states in the trend sample, grade 4.



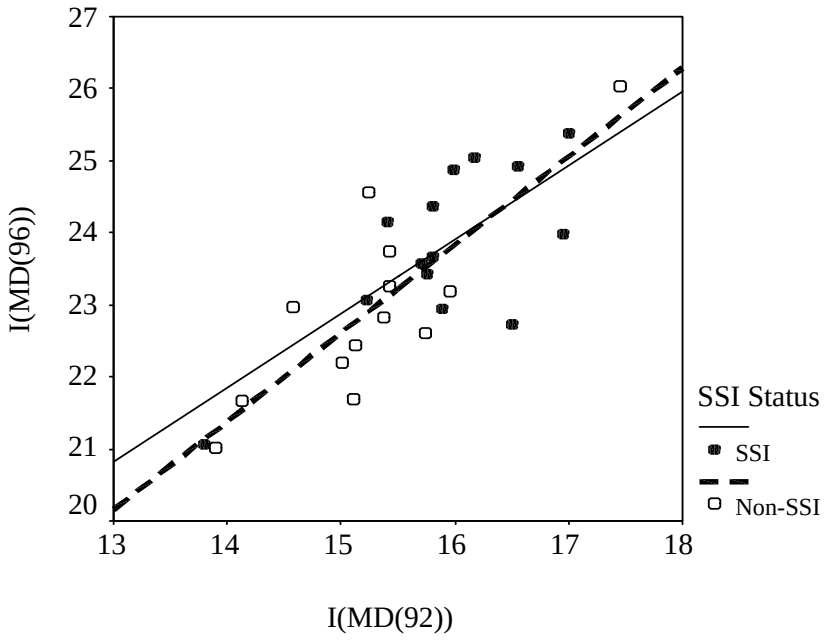
I(MD). Regression analyses found that I(MD(96)) was significantly related to I(MD(92)) and that SSI status was not. (See Table 9 and Figure 8.)

Table 9
Predicting I(RC(96)) from I(RC(92)) and SSI status, Grade 4

	B	*SE B*	β	R^2	*F*	ΔR^2	F_Δ
Step 1							
$I_{RC(92)}$	1.16	.17	.80	.64	44.74*		
Step 2							
$I_{RC(92)}$	1.13	.19	.78				
SSI status	0.15	.33	.06	.65	21.76*	.00	.20

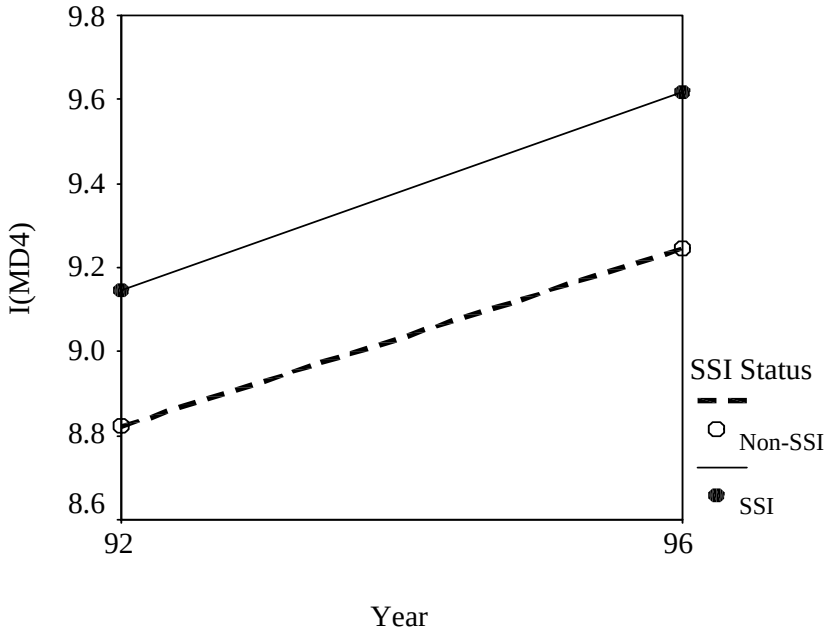
*$p < .05$

Figure 8. Scatterplot of I(MD(92)) and I(MD(96)) for SSI and non-SSI states in the trend sample, grade 4.



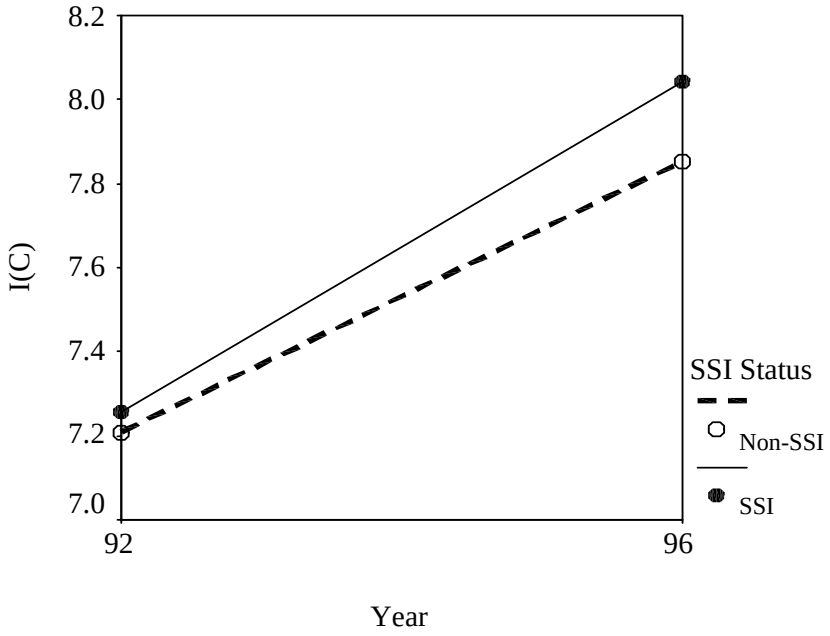
Analyses using I(MD4), a shorter scale that can be compared directly from 1992 to 1996, found that both SSI and non-SSI states increased from 1992 to 1996 ($F = 45.86$, $p < .05$), replicating the findings for grade 8. Results are graphed in Figure 9. At grade 4, the main effect for SSI status was also statistically significant ($F = 3.88$, $p < .10$).

Figure 9. Change in the mean of I(MD4) for SSI and non-SSI states from 1992 to 1996, grade 4.



I(C). At grade 4, the indicator of calculator use is directly comparable from 1992 to 1996. A repeated-measures ANOVA found a significant effect for year ($F = 83.50$, $p < .05$). As with the findings from grade 8, there was no significant effect for SSI.

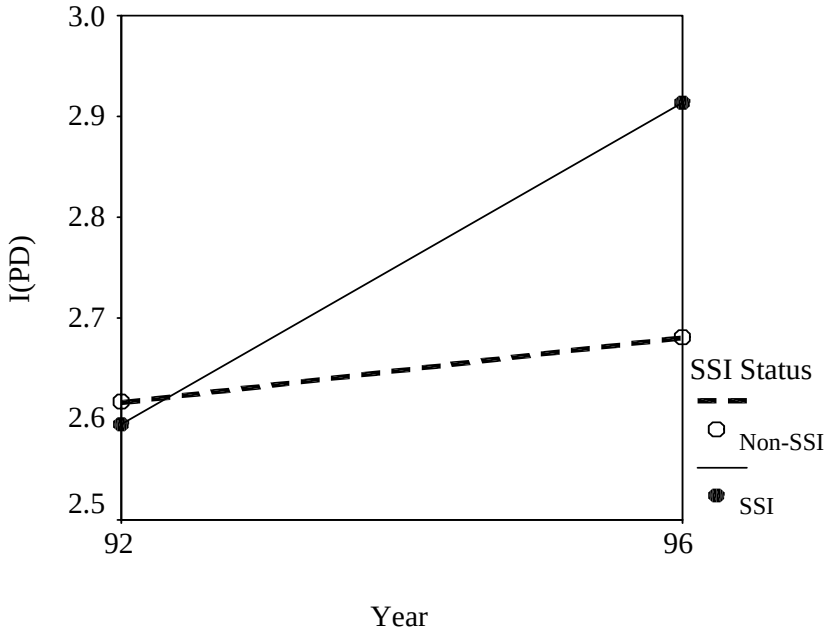
Figure 10. Change in the mean of I(C) for SSI and non-SSI states from 1992 to 1996, grade 4.



Supplementary analyses looked at the effect of state policy for using calculators on state assessments. For the subsample of 21 states with state assessment programs, there was no significant effect of the state's policy on I(C) evident at grade 4. This was unlike the findings for grade 8.

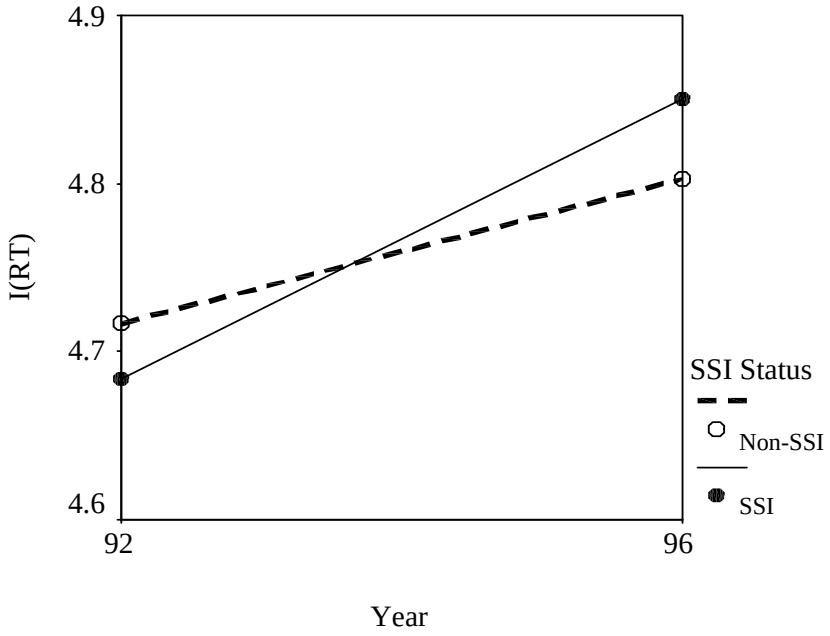
I(PD). Results for grade 4 are graphed in Figure 11. Both the main effect for year (F = 21.60, $p < .05$) and the interaction of year and SSI status (F = 9.61, $p < .05$) were statistically significant. In 1996, students in SSI states had teachers with more time in professional development related to mathematics and mathematics education than students in the non-SSI states. In grade 8, only the effect for year was statistically significant. Comparisons between Figure 5 and Figure 11 make it clear that grade 8 teachers averaged more time in mathematics-related professional development than grade 4 teachers.

Figure 11. Change in the mean of I(PD) for SSI and non-SSI states from 1992 to 1996, grade 4.



I(RT). Figure 12 shows that both SSI and non-SSI states increased in the number of reform-related topics grade 4 teachers had studied (F = 8.54, $p < .05$). The finding is the same as that at grade 8.

Figure 12. Change in the mean of I(RT) for SSI and non-SSI states from 1992 to 1996, grade 4.



Summary. Six potential indicators of educational reform were developed from the state NAEP teacher questionnaire. Cross-sectional comparisons for 1996 found that SSI states averaged significantly higher than non-SSI states on three of the six indicators at each grade level, as shown in Table 10. Longitudinal analyses, using a smaller sample of 27 states, found that at both grade levels SSI states increased more than non-SSI states on I(RC), the indicator of the relative emphasis on reasoning and communication. At grade 4, SSI states also increased more in the time teachers' spent in mathematics-related staff development during the last year. While the indicator of calculator use was not related to the state's SSI status, it was related to whether students could use calculators on the state assessment at grade 8.

One indicator, I(RT), was not related to SSI status in any of the comparisons. This indicator has to do with whether teachers have ever studied each of the seven topics listed. The question refers to the teacher's entire career and does not distinguish between brief study and in-depth, extensive study.

At grade 8, three indicators could be compared directly from 1992 to 1996: I(MD4), I(PD) and I(RT). All three increased significantly from 1992 to 1996. At grade 4, four indicators could be compared directly: the three compared at grade 8 and I(C). All increased significantly from 1992 to 1996. This finding suggests that across the 27 states in the longitudinal sample, mathematics reform progressed steadily during the mid-90s.

Table 10
Summary of Findings Comparing SSI and Non-SSI States on Six Indicators of Mathematics Reform

	Cross-sectional		Longitudinal	
Grade Level	8	4	8	4
I(RC) – Relative Emphasis on Reasoning and Communication	SSI higher	SSI higher	SSI gained more	SSI gained more
I(MD) – Students' Use of Mathematical Discourse	SSI higher	SSI higher	SSI higher overall	SSI higher overall
I(C) – Students' Use of Calculators	Related to state policy	*ns*	Gain related to policy	*ns*
I(S) – Teachers' Knowledge of NCTM *Standards*	SSI higher	*ns*	-	-
I(PD) – Time in Professional Development Last Year	*ns*	SSI higher	*ns*	SSI gained more
I(RT) – Number of Reform-Related Topics Teachers Have Studied	*ns*	*ns*	*ns*	*ns*

Note: *ns* means that no statistically significant effects were found.

In comparing indicators from 1992 to 1996, we found that most of the 1992 measures were strongly related to those in 1996. Table 11 summarizes the correlations. The correlations suggest that changes from 1992 to 1996 were gradual, building on the states' practices in 1992. They also indicate that states differ in their educational practices and that these differences are enduring across four years. Correlations in grade 4 are slightly higher than those in grade 8, suggesting that the 1992 measures are a stronger influence on the 1996 measures for grade 4 compared to grade 8. I(PD) for grade 8 is the one exception to the strong correlations.

Table 11
Correlations Between 1992 and 1996 Indicators at Grade 8 and Grade 4

Indicator[a]	Grade 8	Grade 4
I(RC)	.67	.78
I(MD)	.66	.74
I(C)	.71	.77
I(PD)	.23	.56
I(RT)	.53	.62

[a]I(S) is not included in the table because it was not measured in 1992

Relationships Between the Indicators and Student Achievement: Multiple Linear Regression Modeling

The relationships of the indicators to student achievement were examined with multiple regression modeling. This work was exploratory and directed to model development. The goal was to develop hypotheses to be tested with the 2000 state NAEP data. SSI is not a variable in the model. Rather, it is considered to be one of many factors external to the model that influence the variables. The model incorporates just a few of the many influences on student achievement.

Figure 13 illustrates relationships among the six indicators. The indicators fall into three groups, related causally. The two indicators on the left, I(PD) and I(RT), are enclosed in a broken line box to represent time and topics teachers have studied. The box represents teachers' opportunities to learn, but it does not indicate what teachers know or are able to do. The next group in the model is represented by only one indicator, I(S). This represents the teachers' knowledge and skills—specifically, teachers' knowledge of the NCTM *Standards*. Three indicators represent what teachers actually do while teaching and include instructional goals as well as teaching practices. These indicators are expected to be most directly related to student achievement and to result from the education and training teachers have completed.

Figure 13. Indicators of mathematics curricular reform and their relationship to student achievement.

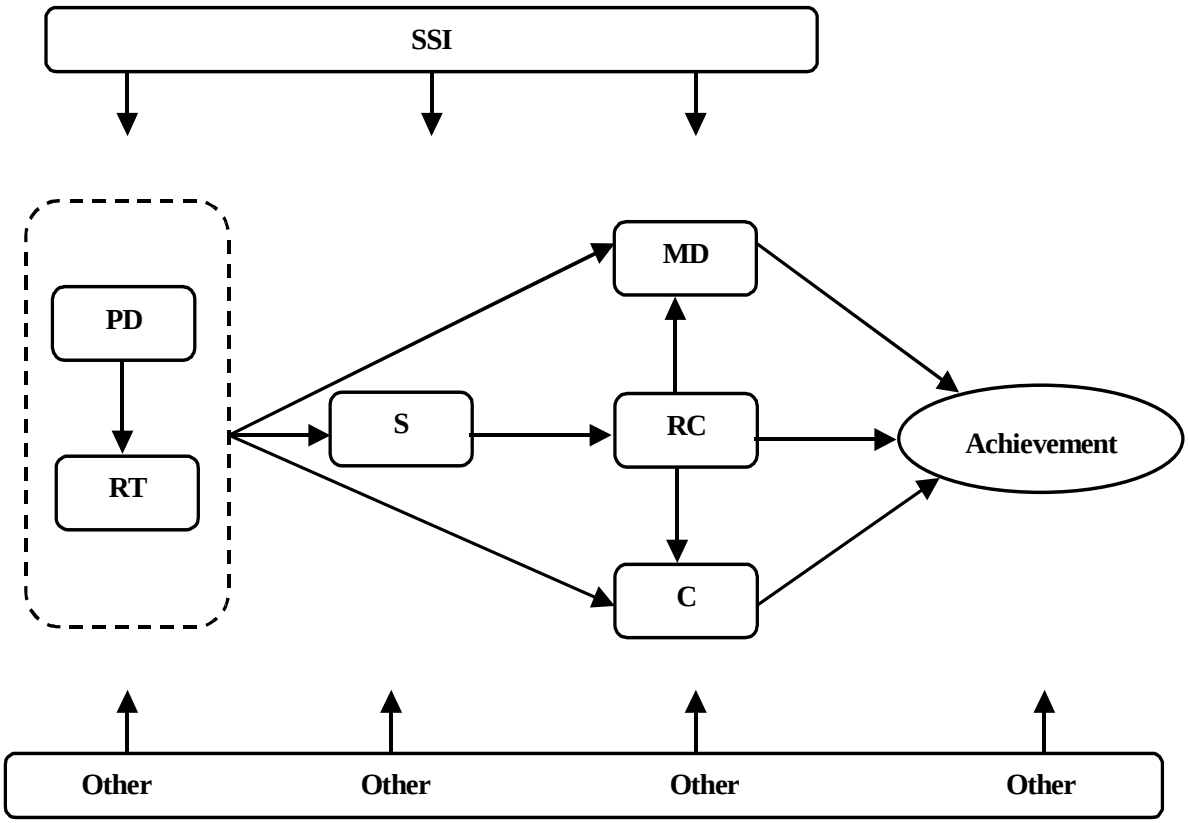


Table 12 presents the intercorrelations of the six indicators and the state mean mathematics composite scores for both grade 4 and grade 8 in 1996. As the table shows, the three classroom practice indicators are strongly interrelated, with the highest correlation between I(RC) and I(MD) at both grade 8 and grade 4. In addition, the teacher's knowledge of the NCTM *Standards* is correlated with the three classroom practice indicators at both grade 4 and grade 8. At grade 8, knowledge of the NCTM *Standards* is related to time spent in mathematics-related staff development during the last year, but not at grade 4. The two indicators of professional development are also significantly related at both grade 8 and grade 4.

The pattern of univariate correlations with the mathematics composite differs somewhat between grades 8 and 4. At both grades, the mean mathematics composite is significantly related to the calculator use indicator, and it is not significantly correlated with the mathematical discourse indicator. At grade 8, two additional indicators are also positively related to the mathematics composite: the relative emphasis on reasoning and communication and teachers' knowledge of the standards. At grade 4, two indicators are negatively related to achievement—amount of professional development during the last year and the number of reform topics studied. In interpreting the coefficients, it is important to remember that correlation does not imply causality.

Table 12
Intercorrelations Among the Six Indicators of Mathematics Reform and the State NAEP Mathematics Composite at Grade 8 and Grade 4, 1996

	Composite	I(RC)	I(MD)	I(C)	I(S)	I(PD)
Grade 8						
I(RC)	.32*					
I(MD)	-.05	.68*				
I(C)	.64*	.54*	.48*			
I(S)	.31*	.73*	.57*	.58*		
I(PD)	-.20	.41*	.48*	.18	.43*	
I(RT)	-.01	.28*	.49*	.33*	.18	.39*
Grade 4						
I(RC)	-.17					
I(MD)	-.17	.87*				
I(C)	.37*	.26	.47*			
I(S)	.09	.42*	.51*	.50*		
I(PD)	-.36*	.51*	.45*	.02	.24	
I(RT)	-.33*	.35*	.49*	.18	.17	.46*

* $p < .10$

Using multiple linear regression to assess the model in Figure 13 presents some problems because of the relatively high correlations among several predictors. Multiple linear regression assumes the predictors are independent. Relationships among predictors raise issues about how to estimate the model parameters. In part, this issue is solved by the model specification. If the model must include all predictors, an analytic method that will divide the shared variance among the predictors can be used.

According to the model, variation in I(S), I(PD), and I(RT) is expected to be reflected in variation in the classroom practice indicators, I(RC), I(MD), and I(C). Models with just the three classroom practice indicators were evaluated by checking whether adding additional indicators resulted in a better fit.

Grade 8, 1996. At grade 8, all three classroom practice indicators combined to predict the state's mean mathematics composite, Y, with a model R^2 of .63 ($F = 18.10$, $p < .01$). Adding the other three indicators did not add to the prediction, with an R^2 change of .04 ($F = 1.03$, $p = 0.39$). The prediction equation with standardized betas was:

Y = .34I(RC) - .65I(MD) + .77I(C)

All betas are significant. In the model, I(MD) is acting as a suppressor variable, since it has a negative sign. The beta represents the relationship between I(MD) and Y, the mean mathematics composite, with the effects of I(RC) and I(C) partialled out. Variability in I(MD) results from many factors. The coefficient in the model is for the variability in I(MD) that is unrelated to variability in the other indicators.

Grade 8, 1992. For comparison purposes, the 1992 grade 8 data were analyzed with the same regression model. As in 1996, the three classroom practice indicators all contributed to the prediction of the state mean mathematics composite, Y ($F = 10.84$, $p < .01$). In addition, adding the other indicators did not improve the prediction ($F = 0.33$, $p = .72$). The 1992 regression equation was:

Y = .38I(RC) - .59I(MD) + .70I(C)

Grade 4, 1996. At grade 4, the three classroom practice indicators combined to predict the state's mean mathematics composite with a model R^2 of .30 ($F = 5.01$, $p < .01$). As was found at grade 8, adding the three other indicators did not add to the prediction, with an R^2 change of .08 ($F = 1.36$, $p = .27$). For grade 4, the regression equation was:

Y = .23I(RC) - .66I(MD) + .62I(C)

In this model, the coefficient for I(RC) was not statistically significant, leaving a reduced model of :

Y = - .44I(MD) + .58I(C)

Substituting other indicators in the model, I(PD) or I(RT) can take the place of I(MD). Both substitute indicators also have a negative coefficient, and predict the state's mean mathematics composite along with I(C). The equation with I(PD) is:

Y = .38I(C) - .36I(PD)

Grade 4, 1992. In 1992, the regression model, with the three classroom practice indicators as predictors of the states' mean mathematics composite score, was not statistically

significant (F = 2.13, p = .12). However, as in 1996, the model with the indicator of calculator use and the indicator of professional development was significant (F = 6.87 p < .01). As in 1996, the coefficient for I(PD) had a negative sign:

Y = .43I(C) - .50I(PD)

Summary and discussion. The grade 4 results are similar to those at grade 8, with the added finding that indicators of professional develop enter negatively in the model, as a replacement for I(MD). Differences between the grade 4 and grade 8 model may be related to differences in educational practices at the two grades. Grade levels are compared in Table 13.

Table 13
Comparisons of Grade 4 and Grade 8 Indicators in 1992 and 1996, and Correlations of the Measures at the Two Grade Levels

		Grade 4	Grade 8	*t*	*r*
1996 (n = 36)					
I(RC)	*M*	43.12	45.24	9.82*	.67*
	SD	1.52	1.65		
I(MD)	*M*	23.37	22.81	-4.62	.82*
	SD	1.21	1.22		
I(C)	*M*	8.00	9.89	22.11*	.69*
	SD	0.52	0.70		
I(S)	*M*	1.94	2.65	31.03*	.75*
	SD	0.20	0.19		
I(PD)	*M*	2.78	3.39	15.71*	.69*
	SD	0.26	0.31		
I(RT)	*M*	4.82	5.10	7.90*	.73*
	SD	0.25	0.31		
1992 (n = 37)					
I(RC)	*M*	39.64	43.12	13.07*	.67*
	SD	1.79	2.11		
I(MD)	*M*	15.55	15.43	-1.44	.78*
	SD	0.80	0.72		
I(PD)	*M*	2.58	3.25	20.46*	.56*
	SD	0.21	0.22		
I(RT)	*M*	4.73	4.73	-0.07	.69*
	SD	0.25	0.34		

*p < .05

In 1996, five of the six indicators were significantly higher at grade 8 than at grade 4. The one exception was I(MD). In 1992, grade 8 averaged significantly higher on I(RC) and I(PD), and again there was no difference in I(MD). The comparisons show that grade 8 teachers put relatively more emphasis on reasoning and communication than grade 4 teachers. Since students develop these skills through mathematical discourse, we would expect to find more discourse in grade 8. As was found in the regression models, results of I(MD) seem contrary to expectations.

Interpretation of the negative coefficients for I(MD) is, necessarily, post hoc. Many possibilities have been considered, and three seem the most plausible. One is that using mathematical discourse is not an end it itself, but provides the conditions under which students become more proficient at solving unique problems and communicating their mathematical ideas. From this perspective, opportunities for talking or writing will not lead to increased student achievement unless those opportunities are appropriately tied to specific instructional goals. Additionally, when basic skills are the main instructional goals, providing opportunities for discourse is expected to be less effective in promoting student achievement than other methods of instruction. This explanation centers on the match between discourse opportunities and instructional goals. An alternative explanation addresses the issue of causality. Instead of expecting that discourse opportunities lead to achievement, the relationship is reversed—that is, states with low achievement scores may be more likely to promote the use of mathematical discourse as a way to raise student achievement. In the long term, scores would be expected to increase as students have increased experience with mathematical discourse. However, in the short term, the negative sign may indicate that lower-achieving states are implementing mathematical discourse more than higher achieving states. A third possibility is that the discourse measure is flawed. As discussed previously, the measure assumes that greater opportunity for discourse is better. We would prefer a scale where moderate levels of discourse are rated the highest. While too little discourse can prevent students from developing mathematical reasoning and communication skills, we also expect that high levels of discourse may not be optimal.

The results reported here are based on cross-sectional analyses, which cannot provide information about causal relationships. Future longitudinal studies incorporating data from the 2000 state NAEP will be useful in differentiating among these alternative explanations. More work is needed to clarify the role of mathematical discourse in student learning. Perhaps the explanation of the negative sign for I(MD) is related to all three alternatives.

Individual Differences Among States on the Indicators: Grouping States Based on Achievement Gains

The preceding sections of this paper described the reform indicators developed from the state NAEP teacher questionnaire and showed that SSI states gained more on some of the indicators than non-SSI states. In cross-sectional analyses, regression analyses identified indicators that were related to student achievement at grade 4 and grade 8. This section explores differences among the 14 SSI states in the longitudinal sample.

The 14 SSI states were sorted into groups based on increases in their mean mathematics composite scores from 1992 to 1996 and from 1996 to 1998. Nationally, increases at both grade 4 and grade 8 averaged about 4 points, or 1 point a year. In the sample of 14 SSI states, three groups were identified:

Steady Progress – Four states
Average gain of at least 1 point a year in each 4-year interval
Less Steady Progress – Five states
Average gain of at least 1 point a year in one of the 4-year intervals
Little Change – Five states
Average gain of less than 1 point a year in both 4-year intervals

The four Steady Progress states in the longitudinal sample are Texas, New York, Kentucky, and Massachusetts.

After classifying the states, the reform-related indicators were examined to identify any patterns unique to the Steady Progress states. To facilitate comparisons between the indicators, all measures were standardized with a mean of 0 and a standard deviation of 1. Because scales for some indicators varied from year to year, standardization took place within each year. With these relative measures, the mean for each year is 0; consequently, the average change from one year to the next is 0. Comparisons across years describe relative growth. Increases from one year to the next in a standardized mean can happen in two ways—either the state increases relatively more than other states, or the state decreases relatively less than other states. Figures 14 through 17 presents the graphs of the indicators in grades 4 and 8 for the four Steady Progress states.

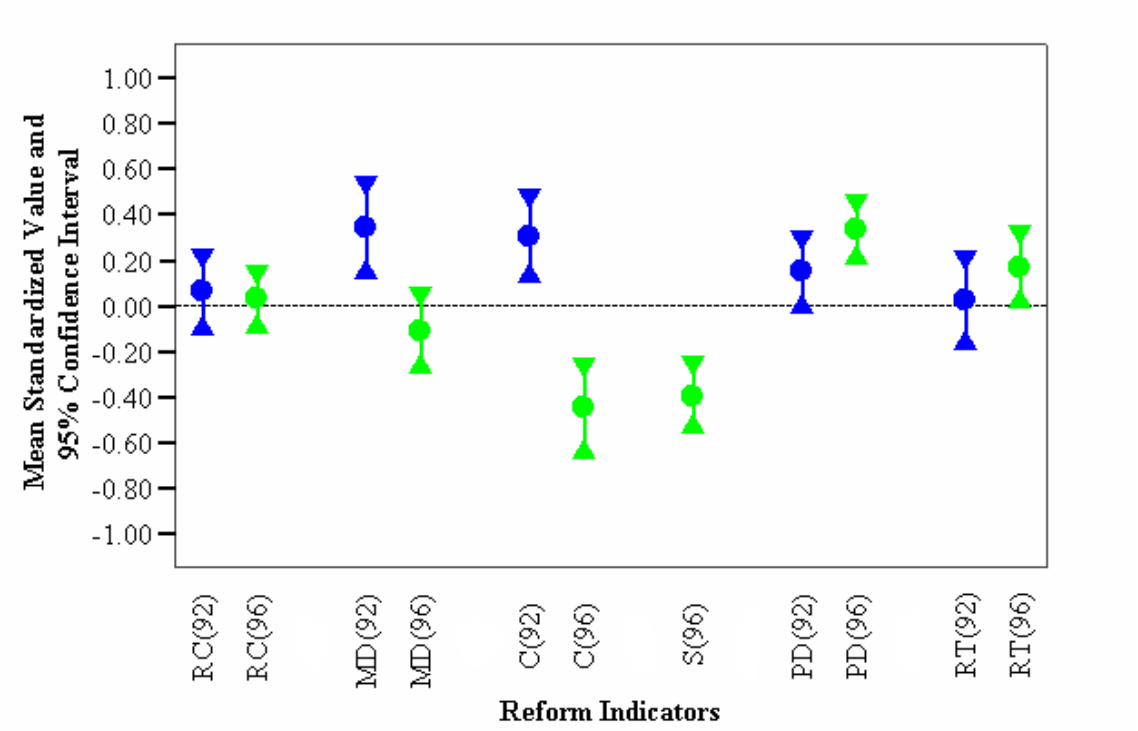
Three indicators describe classroom practices: I(RC), I(MD), and I(C). Among the 14 SSI states, few had significant changes in either I(RC) or I(MD) from 1992-1996, except for states in the Steady Progress group. At grade 8, both Kentucky and Massachusetts significantly increased in I(RC) relative to other states, and Massachusetts also increased in I(MD). In contrast, Texas, at grade 8 and grade 4, and New York, at grade 4, significantly decreased in I(MD), although neither had significant changes in I(RC). Three of the four Steady Progress states had significant changes in I(MD) for at least one grade level, although two decreased and one increased.

Only two of the ten other states had significant changes in I(RC) or I(MD). One state significantly increased in I(MD) at grade 8. The other state significantly decreased in both I(RC) and I(MD) at grade 4.

On the basis of these results, it seems that SSI states used at least two distinct approaches to increasing mathematics achievement. Massachusetts and Kentucky are generally above average in their emphasis on reasoning and communication, accompanied by above-average opportunities for mathematical discourse. In contrast, Texas and New York are generally near the average in their emphasis on reasoning and communication, accompanied by average opportunities for mathematical discourse. Another indicator, I(S), teachers' knowledge of the NCTM *Standards*, complements these findings. Massachusetts and Kentucky are above average in I(S), while Texas and New York are generally well below average. Future work will review state standards and assessments in light of this result.

Figure 14. Relative values of reform indicators in Texas at grades 8 and 4.

Grade 8



Grade 4

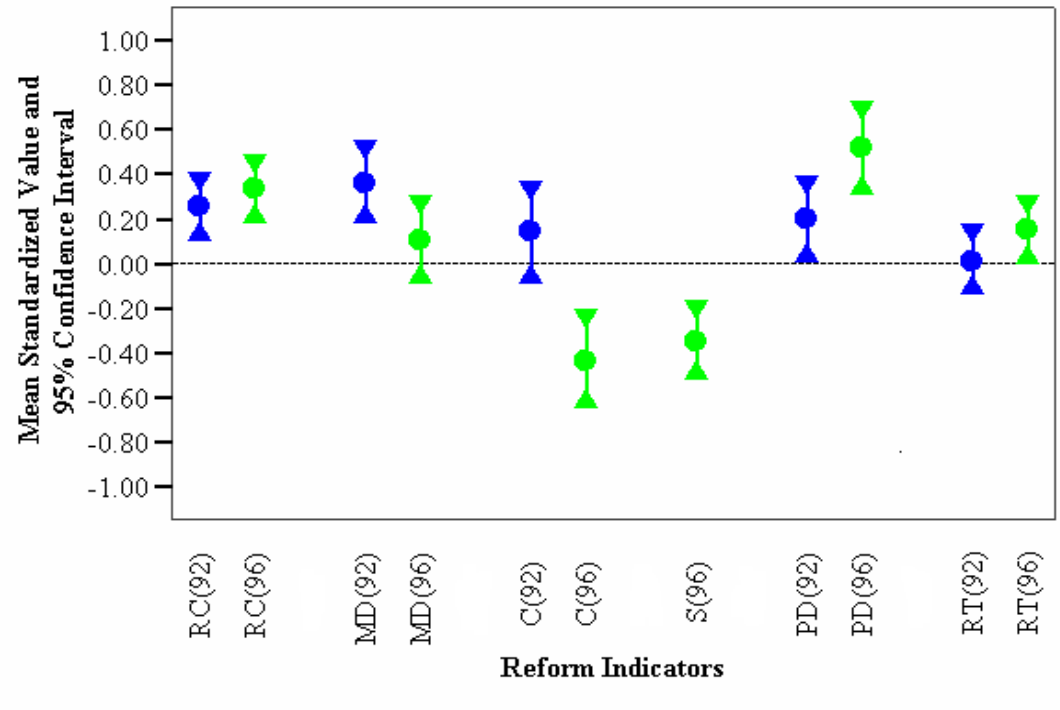
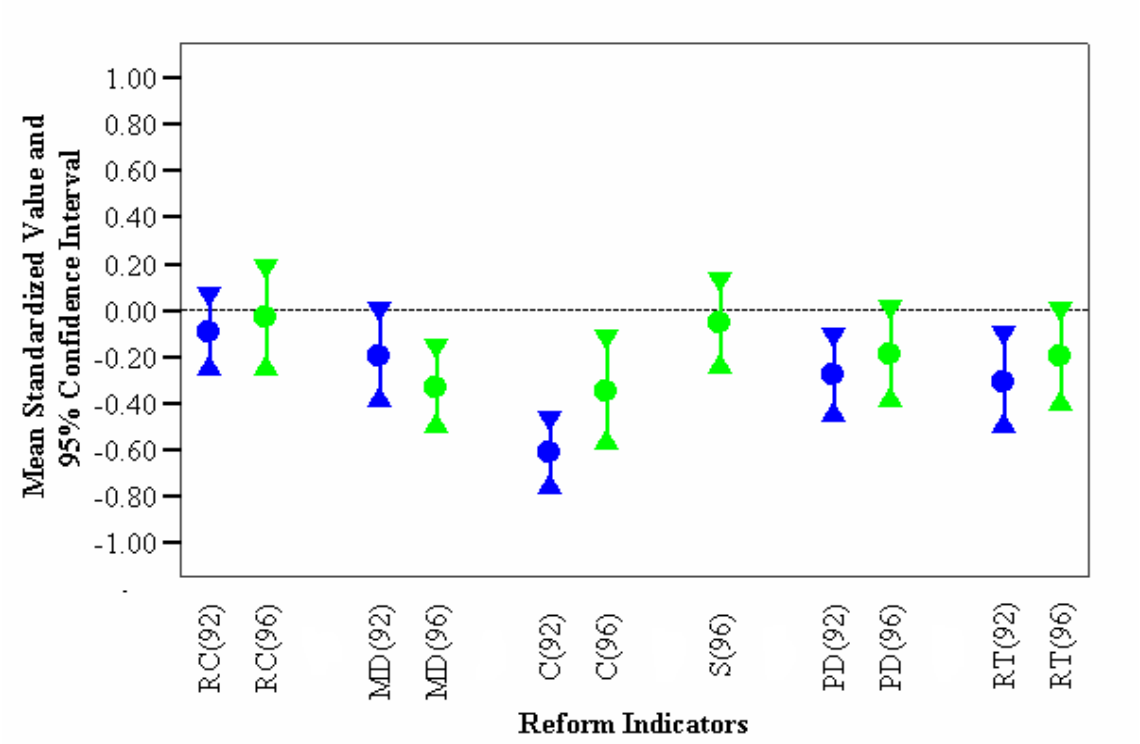


Figure 15. Relative values of reform indicators in New York at grades 8 and 4.

Grade 8



Grade 4

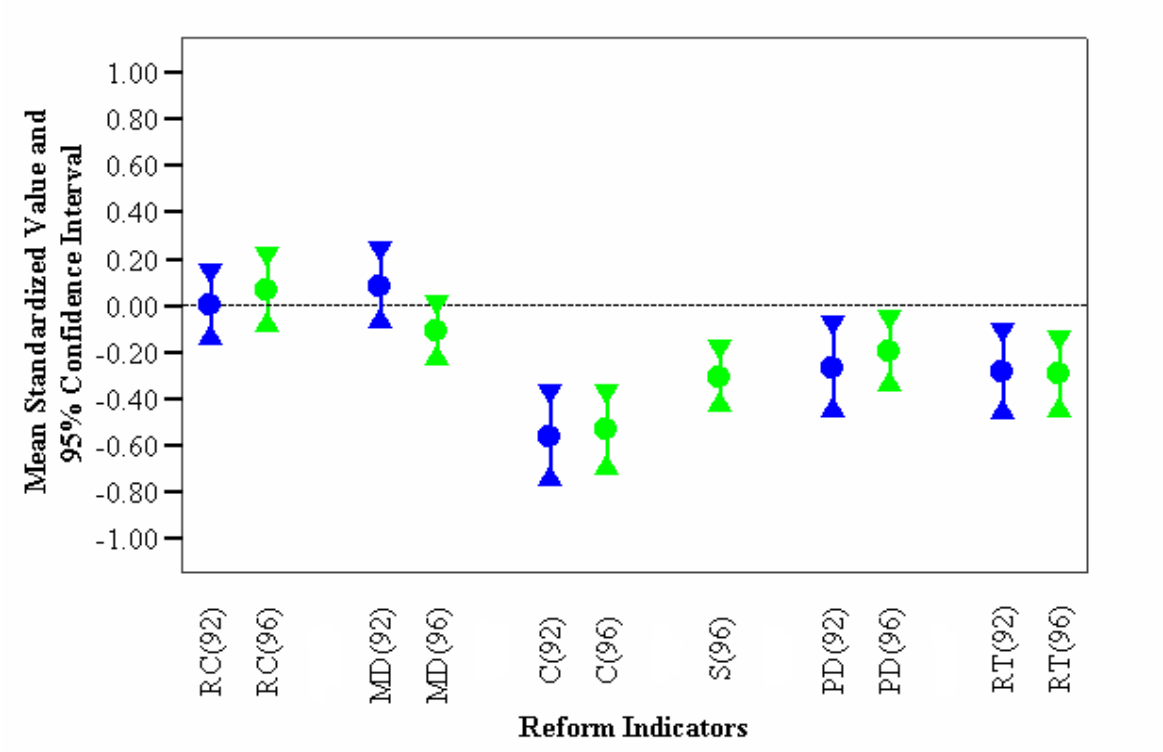
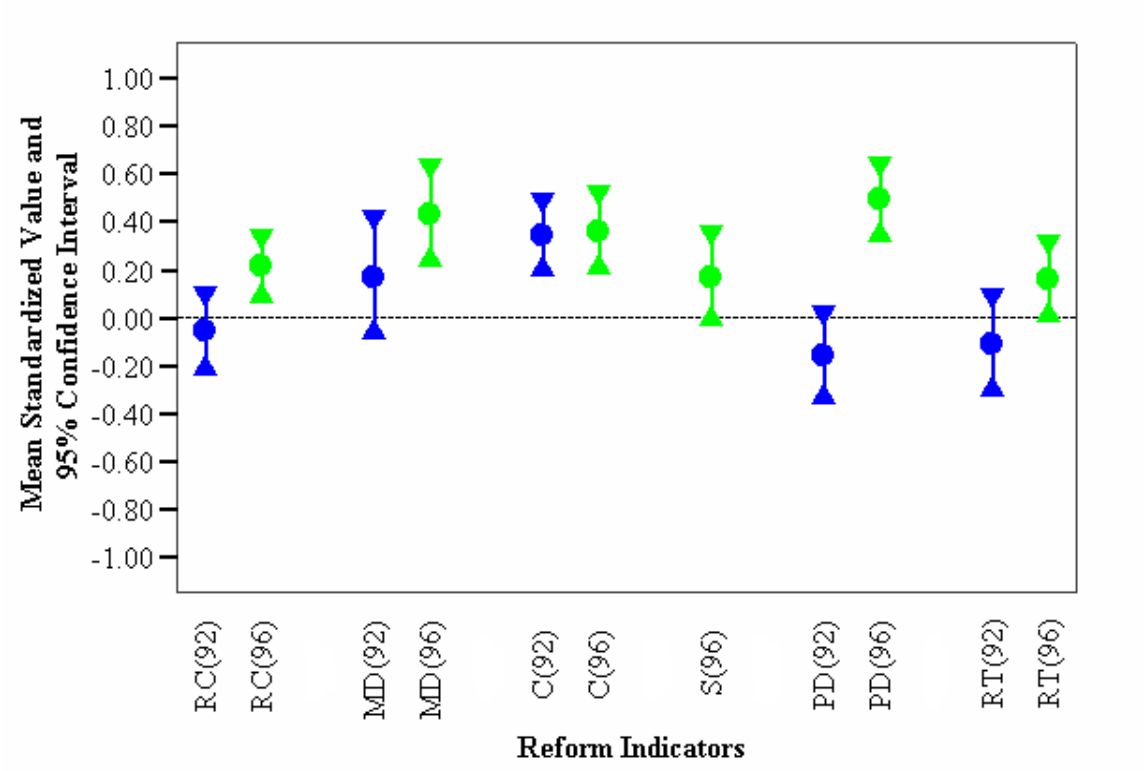


Figure 16. Relative values of reform indicators in Kentucky at grades 8 and 4.

Grade 8



Grade 4

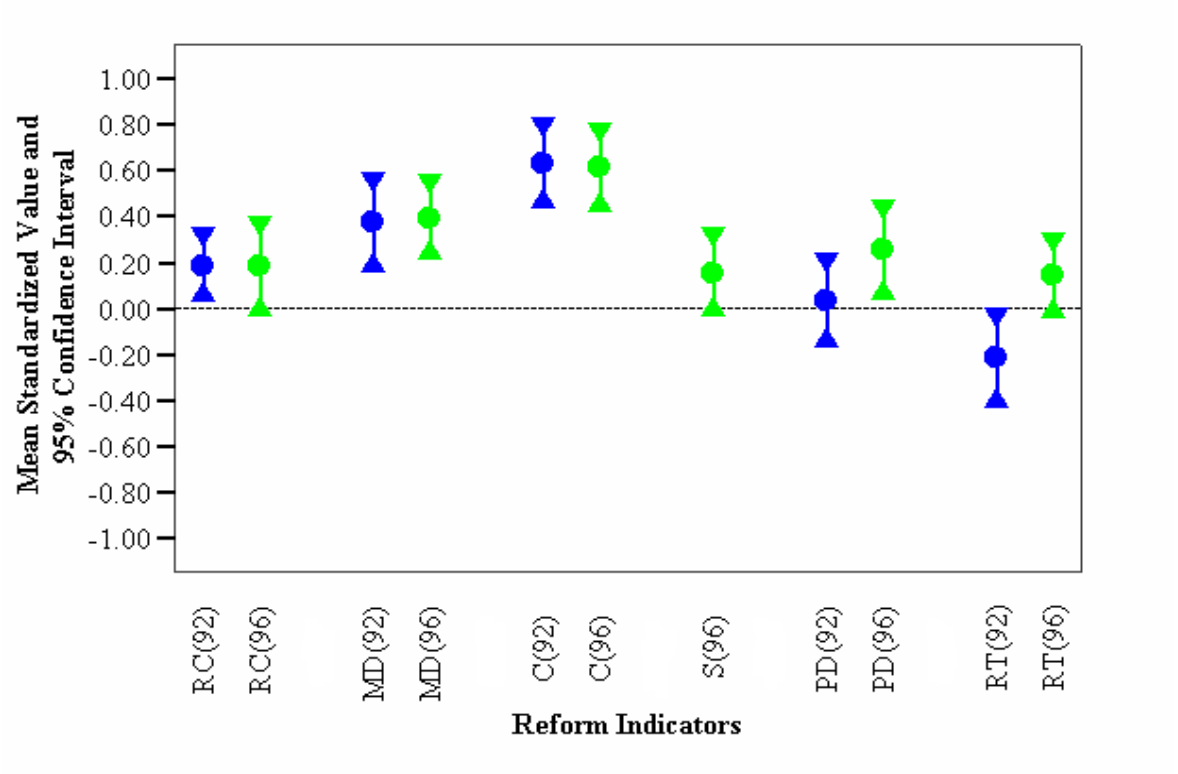
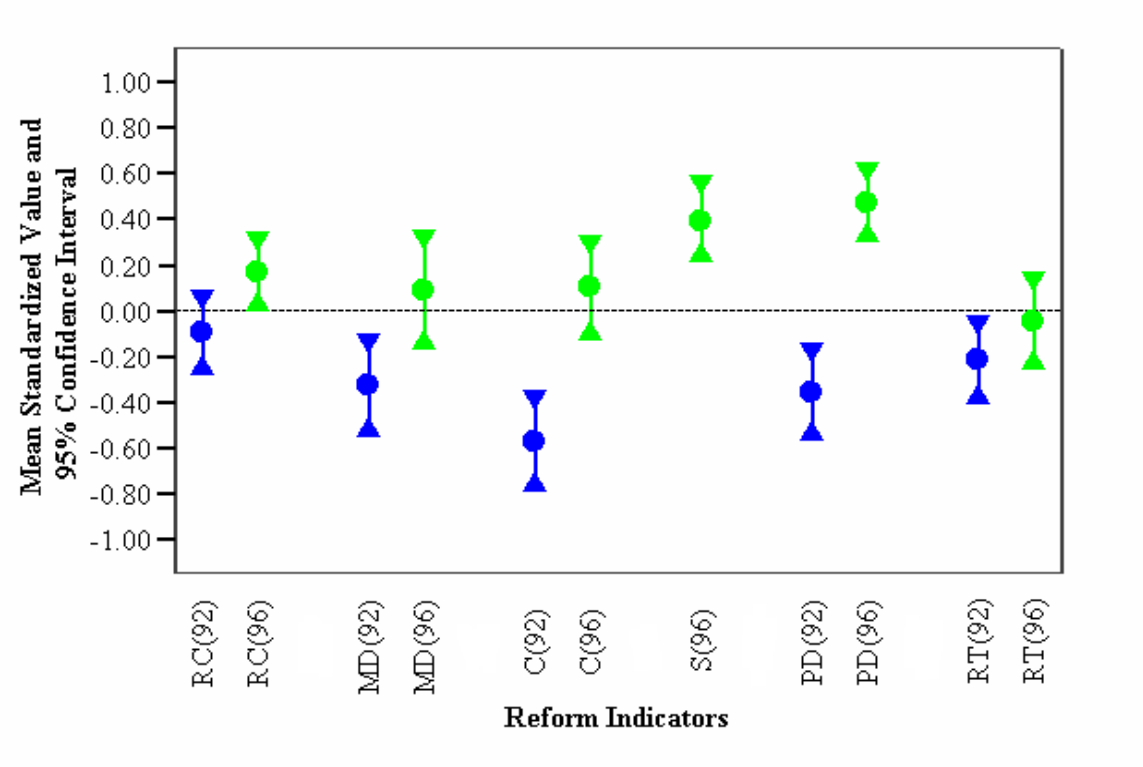
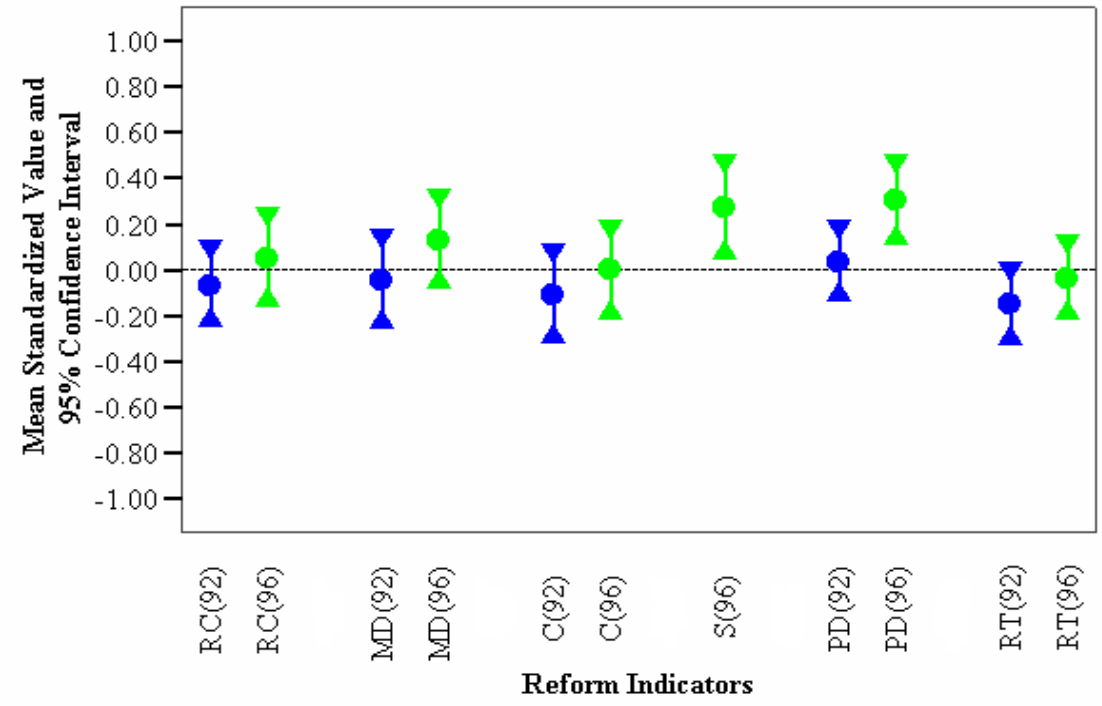


Figure 17. Relative values of reform indicators in Massachusetts at grades 8 and 4.

Grade 8



Grade 4



For the third classroom practice indicator I(C), seven SSI states had significant changes at grade 8 and five at grade 4. However, the direction of change varied, with some states increasing and others decreasing. Table 14 summarizes the changes in I(C) from 1992 to 1996.

Table 14
Changes in I(C) and I(PD) from 1992 to 1996 as a Function of the State's Achievement Gains from 1992 to 2000

	Significant Increase	Significant Decrease	No Change
I(C)			
Steady progress			
Grade 4	0	1	3
Grade 8	2	1	1
Less steady progress			
Grade 4	0	1	4
Grade 8	0	2	3
Little change			
Grade 4	3	0	2
Grade 8	0	2	3
I(PD)			
Steady progress			
Grade 4	2	0	2
Grade 8	3	0	1
Less steady progress			
Grade 4	2	1	2
Grade 8	0	0	5
Little change			
Grade 4	3	0	2
Grade 8	0	2	3

In the Steady Progress group, both Massachusetts and New York increased in I(C) at grade 8. Massachusetts increased to slightly above average, while New York was still below average despite the significant increase. Texas decreased from above average in 1992 to below average in 1996 at both grade 8 and grade 4. Kentucky did not change significantly, but it was substantially above average both years. As with the first two indicators, the Steady Progress group has two distinct subgroups. The two states above average on I(C) in 1996 are also above average on I(RC), I(MD), and I(S); the two states below average on I(C) are generally below average on the other indicators.

I(PD) is an indicator of the amount of time teachers spent in professional development over the last year. Massachusetts and Texas increased at both grades 4 and 8 and Kentucky increased at grade 8. New York did not change significantly, averaging below the mean both years. In the other groups, no states had significant increases in I(PD) at grade 8, though five did at grade 4. In the Little Change group, two states significantly decreased in I(PD) at grade 8. Three of the four states in the Steady Progress group were well above average on I(PD) in 1996.

Only one state had a significant increase in I(RT)—Kentucky increased at both grade 4 and grade 8. Since the indicator measures topics teachers have "ever" studied, it may be relatively insensitive to short-term influences.

While the sample size is not large, these results suggest that states adopted at least two different strategies for improving mathematics achievement. Both successful strategies involved aligning instructional goals and students' opportunities for mathematical discourse. In states making steady progress, students generally had teachers who devoted well above an average amount of time to mathematics-related professional development.

Summary and Conclusions

This paper describes six reform-related indicators developed from the state NAEP teacher questionnaire. The indicators were used to:

- compare SSI with non-SSI states;
- assess changes over time in SSI and non-SSI states;
- predict students' mathematics achievement;
- describe reform-related activities in individual states;
- differentiate among SSI states making steady achievement gains from 1992-2000.

Both cross-sectional and longitudinal analyses found that SSI states averaged significantly higher than non-SSI states on several indicators in 1996 and that there were no statistically significant differences on the other indicators.

Three indicators were related to classroom practices:

I(RC) – Relative Emphasis on Reasoning and Communication;
I(MD) – Students' Opportunities for Mathematical Discourse; and,
I(C) – Students' Use of Calculators.

I(C) and I(RC) were positively related to the states' mean mathematics composite scores at grade 8, but only I(C) predicted the mean mathematics composite scores at grade 4. For both grades, I(MD) alone was unrelated to mathematics achievement, and it had a negative sign in multiple regression models. A general conclusion is that the value of mathematical discourse may depend on its curricular purpose and its instructional relationship to other skills.

For states making steady progress in increasing students'mathematics achievement, changes in classroom practice indicators are evident. However, states seem to take different approaches. One approach puts average emphasis on reasoning and problem solving and provides average opportunities for mathematical discourse, accompanied by below-average use of calculators and teachers' knowledge of the NCTM *Standards*. The other approach is characterized by above-average emphasis on reasoning and communication, accompanied by above-average means on I(MD), I(C), and I(S). Regardless of the approach, states making steady gains in students' mean mathematics composite score generally increased in professional development from 1992 to 1996.

Next steps for this research will be to use information from the individual states about the specifics of a state's SSI, along with information about the state's standards, assessments, and accountability policies from 1990 to 2000 to understand the reasons for individual differences among the SSI states.

References

Allen, N.L., Jenkins, F. Kulick, E., & Zelenak, C. A. (1997). *Technical report of the NAEP 1996 State Assessment Program in Mathematics*. Washington, DC: National Center for Education Statistics.

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

Clune, W. H. (1998). *Toward a theory of systemic reform: The case of nine NSF Statewide Systemic Initiatives* (Research Monograph No 16). Madison: University of Wisconsin, National Institute for Science Education.

Grissmer, D. W., Flanagan, A., Kawata, J., & Williamson, S. (2000). *Improving student achievement: What NAEP state test scores tell us*. Santa Monica, CA: RAND, MR-924-EDU.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.

North Central Regional Educational Laboratory. (1996). *State student assessment programs database*. Oakbrook, IL: Author.

Shaughnessy, C. A., Nelson, J. E., & Norris, N. A. (1998). *NAEP 1996 mathematics cross-state data compendium for the grade 4 and grade 8 assessment.* Washington, DC: National Center for Education Statistics.

Webb, N.L., Kane, J., Kaufman, D., & Yang, J. (2001). *Study of the impact of the Statewide Systemic Initiatives Program*. Madison, WI: Wisconsin Center for Education Research.