Inadequate Conclusions from an Inadequate Assessment: What Can SAT-9 Scores Tell Us about the Impact of Proposition 227 in California?[1]

Yuko Goto Butler, Jennifer Evelyn Orr,
Michele Bousquet Gutiérrez, and Kenji Hakuta
Stanford University

Abstract

Proponents of Proposition 227 in California have argued for the effectiveness of English-only instruction over bilingual programs based on the increase in SAT-9 scores in the years since its implementation. Based on analyses of SAT-9 scores from 1998 to 2000, this article argues that: scores increased for all students, not just for English-learning students; scores increased for most districts regardless of the types of programs implemented; increases could be attributed to a number of possible factors, but it is not possible to separate out the impact of Proposition 227; and SAT-9 is not an appropriate measure for assessing English-learning students.

The scores that limited English proficient (LEP) students obtained on the Stanford 9 (SAT-9) test conducted in the spring of 2000 gained considerable media attention upon their release later that summer. The attention was in large part because the results presumably enabled an evaluation of the impact of Proposition 227, the "English for the Children" ballot initiative led by Ron Unz and passed by California voters in 1998.

The advocates of Proposition 227, highlighting the increases over the last three years of LEP students' SAT-9 scores, have argued that these validate the success of English immersion programs (e.g., Amselle & Alison, 2000). The results from Oceanside Unified School District were highlighted by anti-bilingual education advocates as well as by its superintendent, Ken Noonan, and received front page coverage in the *New York Times* (2000, August 20).[2]

In this article, we review a number of analyses of SAT-9 scores from 1998 to 2000. The results of these analyses clearly indicate that the SAT-9 scores of LEP students do not provide the basis for a resounding claim to victory for Proposition 227. We review below six factors that need to be taken into account in evaluating the SAT-9 scores to demonstrate why this is the case. Our results indicate how inadequate and misleading it could be to use SAT-9 results in order to evaluate the impact of Proposition 227.

SAT-9 scores rose for all students, not just for LEP students

The pattern of LEP student performance of any given school district should be considered in light of statewide patterns of performance by both LEP and native English speakers. Table 1 contains data on statewide reading scores specific to LEP students as well as scores for all students. The table contains data for 1998, 1999, and 2000. The numbers in parentheses indicate changes in percentile rank, and the last column shows the changes from 1998 to 2000. As one can see, there are virtually identical patterns of increases for both LEP students as well as for all students with the increases being most prominent in the earlier grades (especially grades 2 and 3). The same pattern holds for scores obtained in both math and language (see tables 2 and 3). Overall, the scores for all students as well as for LEP students rose, with large increases in the early grades, and performance tapering off in the fourth grade and beyond. It is clear from the pattern of increases seen herein that they cannot be explained simply by the effect of Proposition 227, but rather, there is something much more specific to the nature of SAT-9 causing such broad patterns of improvement.

Table 1
Statewide SAT-9 Scores (Reading) For LEP Students and For All Students

Grade	LEP Students				All Students			
	1998	1999	2000	Change 98-00	1998	1999	2000	Change 98-00
2	19	23(+4)	28(+5)	+9	39	43(+4)	48(+5)	+9
3	14	18(+4)	21(+3)	+7	36	40(+4)	44(+4)	+8
4	15	17(+2)	20(+3)	+5	40	42(+3)	45(+3)	+5
5	14	16(+2)	17(+1)	+3	40	41(+3)	44(+3)	+4
6	16	18(+2)	19(+1)	+3	43	45(+2)	47(+2)	+4
7	12	14(+2)	15(+1)	+3	41	43(+2)	45(+2)	+4
8	15	17(+2)	18(+1)	+3	44	46(+1)	47(+1)	+3

Table 2

Statewide SAT-9 Scores (Math) For LEP Students and For All Students

Grade	LEP Students				All Students			
	1998	1999	2000	Change 98-00	1998	1999	2000	Change 98-00
2	27	34 (+7)	41 (+7)	+14	43	50 (+7)	58 (+8)	+15
3	25	32 (+7)	39 (+7)	+14	42	49 (+7)	57 (+8)	+15
4	21	25 (+4)	30 (+5)	+9	39	44 (+5)	51 (+7)	+12
5	21	24 (+3)	28 (+4)	+7	41	45 (+4)	51 (+6)	+10
6	24	28 (+4)	31 (+3)	+7	48	52 (+4)	57 (+5)	+9
7	22	24 (+2)	27 (+3)	+5	45	47 (+2)	51 (+4)	+6
8	23	25 (+2)	27 (+2)	+4	45	48 (+3)	50 (+2)	+5

Table 3

Statewide SAT-9 Scores (Language) For LEP Students and For All Students

Grade	LEP Students				All Students			
	1998	1999	2000	Change 98-00	1998	1999	2000	Change 98-00
2	19	23 (+4)	28 (+5)	+9	40	45 (+5)	50 (+5)	+10
3	19	24 (+5)	29 (+5)	+10	39	44 (+5)	50 (+6)	+11
4	23	26 (+3)	29 (+3)	+6	44	46 (+2)	50 (+4)	+6
5	21	23 (+2)	25 (+2)	+4	44	46 (+2)	49 (+3)	+5
6	22	24 (+2)	26 (+2)	+4	47	49 (+2)	52 (+3)	+5
7	19	21 (+2)	23 (+2)	+4	49	51 (+2)	54 (+3)	+5
8	19	21 (+2)	22 (+1)	+3	47	49 (+2)	51 (+2)	+4

Increases in SAT-9 scores could arise from a number of possible causes

There is relatively little information describing what specific changes in SAT-9 scores have come about as a result of Proposition 227. In fact, the increases in SAT-9 scores can be attributed to any number of possible causes. During the past few years, for instance, there has been an enormous focus on school reform in California besides Proposition 227, including class size reduction, increased school accountability, and increased focus on English language development. One potential explanation for the increased performance could simply be the effect of teachers "teaching to the test." In the past year, schools and districts in California have taken the SAT-9 much more seriously and have in many instances focused on teaching students how to perform better on the test. In addition, younger children's scores are probably more likely to benefit from increased attention by teachers and school officials to the importance of the test. We also know from experience with testing policies in other states that the first several years of a testing program show increases in scores as schools become familiar with the new test.

Moreover, districts seem to vary considerably in terms of both whom they included as LEP and non-LEP students as well as in the percentages of LEP students that they tested. The guidelines for "redesignation" for fluent English proficient (FEP) status vary from district to district, and the numbers and rates of redesignation vary from year to year, even within the same district (see Table 4, which was created based on a report by Amselle and Allison, 2000, at the READ Institute).

Table 4
Redesignation Rates For Selected School Districts (1996 to 2000)

District	Year				
	1996	1997	1998	1999	2000
Statewide average	6.5%	6.7%	7.0%	7.6%	7.8%
Oceanside	7.9%	5.7%	5.4%	6.6%	4.1%
Santa Barbara	1.7%	4.0%	3.1%	2.3%	2.9%
Ceres Unified	17.2%	7.6%	4.9%	6.2%	12.5%
Alameda City	4.6%	5.8%	5.4%	4.6%	7.6%
San Jose Unified	1.7%	3.4%	5.0%	7.7%	6.2%
Santa Ana	3.3%	4.2%	3.8%	4.0%	5.6%
Vista Unified	3.0%	1.5%	5.4%	4.3%	7.8%
Note: This table was made based on Amselle & Allison (2000)					

It goes without saying that the results of a given school or a district's LEP students would depend a great deal on who is considered LEP and which LEP students are tested. The average test scores for LEP students presumably would go down if more students were reclassified as FEP students and thus excluded from the LEP data. On the other hand, if a given district decided to set a higher standard for redesignation, more proficient students would remain under the LEP designation and this could pull up the scores for LEP students. In any event, any claims about the "successes" of LEP students need to be scrutinized. It is certainly premature to claim that the gains in SAT-9 scores as a whole are due to the implementation of Proposition 227.

Both bilingual districts and English-only districts showed increases in their SAT-9 scores

District level analyses of SAT-9 data show across-the-board increases in scores across districts that have different types of programs. We examined the SAT-9 performance from 1998 to 2000 for school districts that were reported to be:

1. Schools that never had bilingual programs and therefore were not impacted by Proposition 227 (schools in the Evergreen, Magnolia, Westminster, and Orange Unified districts);
2. Schools that had bilingual programs, but which dropped such programs as a result of Proposition 227 (Oceanside); and
3. Schools that retained bilingual programs to varying degrees (Santa Ana, Vista, Ocean View).

The data for each of the above groups show a pattern of score increases very similar to that observed in the statewide analysis: there were overall increases in scores, and these were particularly noticeable in grades 2 and 3. Tables 5 and 6 show the results for reading among second and third graders.

Table 5

Second and Third Grade LEP Students' Percentile Scores in Reading For Selected Districts That Did Not Have Bilingual Education Prior to Proposition 227

	2nd grade			3rd grade		
	1998	1999	2000	1998	1999	2000
Evergreen	50	54	54	30	42	44
Magnolia	18	21	21	12	16	17
Westminster	25	33	38	17	20	27
Orange Unified	16	23	26	15	16	17

Table 6

Second and Third Grade LEP Students' Percentile Scores in Reading For Oceanside School District and Selected Districts Maintaining Bilingual Education

	2nd grade			3rd grade		
	1998	1999	2000	1998	1999	2000
Oceanside City Unified	12	26	32	9	15	22
Santa Ana	17	23	22	14	17	17
Vista	18	21	25	13	16	18
Ocean View	17	27	27	23	17	20

It is also noteworthy that the increases were visible just as much in math as they were in reading and language.

In interpreting these data, one has to keep in mind that the scores for schools and districts are characterized by a substantial amount of unexplained sources of variability. For example, we have little information on exactly what types of programs have been implemented in each school and district. Among schools that reported using "bilingual instruction," bilingual programs vary tremendously from school to school and district to district in terms of the extent to which different languages are used, the types of materials they use, teacher qualifications, and so forth. The same holds true for schools with structured English programs. The types of English programs used therein differ substantially from one school to another. Despite such uncontrolled factors, the SAT-9 scores rose as a whole, however imperfect this measure may be. And yet it is important to note that while the overall scores did rise, the margins of errors would be so large that it would not be possible to distinguish between different types of language programs.

There was significant variation in performance within both English immersion programs and bilingual programs

A school-level analysis also revealed that there was significant variation in performance within both English immersion programs and bilingual programs. We examined the SAT-9 scores for 13 elementary schools that were selected in a report by Californians Together, an advocacy group for bilingual education. The Californians Together report highlighted the SAT-9 scores for 10 elementary schools that had large enrollments of LEP students and that "offer substantial bilingual instruction with adequate materials and with qualified teachers for Spanish speaking English learners." The report further compared those results with "three schools that have been highlighted by proponents

of Proposition 227 as good examples where school district and community preferences have limited instruction exclusively to structured English immersion." The report concludes that bilingual schools "can equal or exceed the performance in English of schools providing instruction only in English (Californians Together, 2000, August 21."

The Californians Together report does not provide data on percentile ranks that are consistent with the way in which we have chosen to report data in our analyses. We therefore went back to the state data reported for the schools named in their analyses: 10 schools that are designated as exemplary bilingual, and three schools that had been included because they have been the focus of pro-English immersion advocacy (including one in Oceanside with a high proportion of LEP students). We downloaded data from the California State Department of Education web site on percentile ranks for the mean NCE for each school (as in the previous analyses we performed noted above). Figures 1 and 2 show the results in reading among English-learning second and third graders. Since Oceanside has been the focus of such intense interest as a key success by Proposition 227 advocates, we examined the Oceanside pattern in the context of the bilingual school data provided by Californians Together. Namely, we added the average scores for the Oceanside district in those figures. The dotted lines indicate the three schools which use only structured English immersion programs. The regular lines show the 10 schools that were reported to "provide substantial bilingual instruction." The bold line indicates the districtwide performance at Oceanside (namely, the average scores for all elementary schools in the Oceanside school district).

Figure 1. Performance in SAT-9 reading among schools with bilingual programs and English immersion programs (2nd grade).

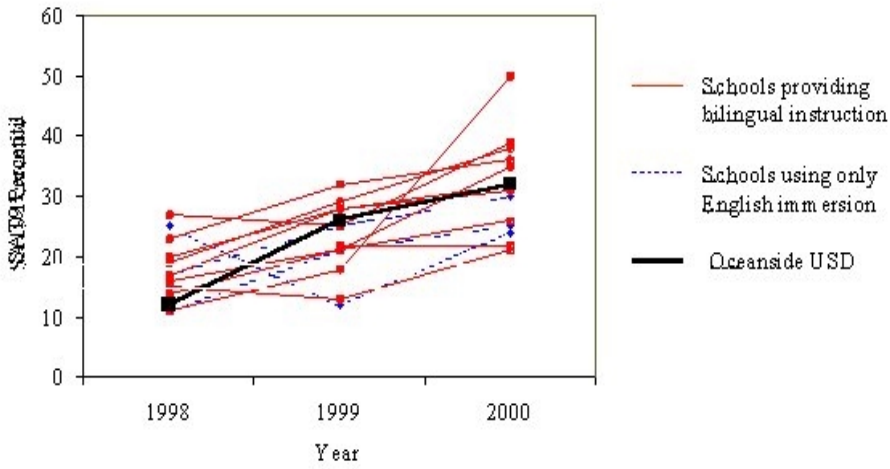
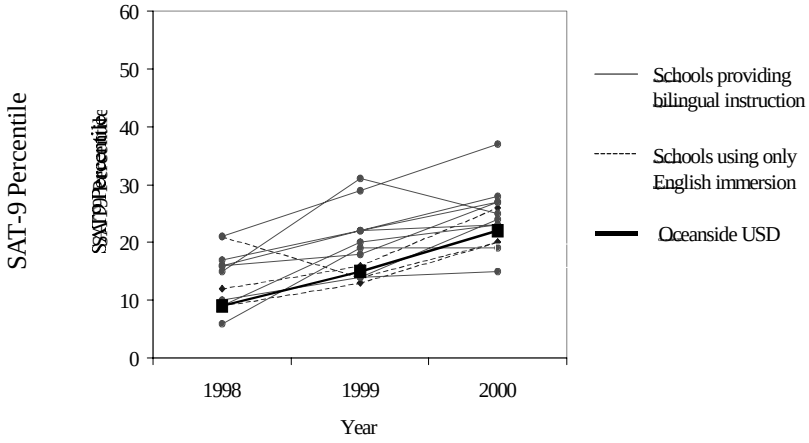


Figure 2. Performance in SAT-9 reading among schools with bilingual programs and English immersion programs (3rd grade).



In looking at the overall SAT-9 reading scores for those schools named in the Californians Together report, the following observations can be made:

1. The gains in those schools are clearly more evident in the second grade than in the third grade;
2. Of the schools chosen by Californians Together, the performance of those said to be providing substantial bilingual instruction on average seem to exceed the performance of the named English-only schools (as was also found by Californians Together); and
3. The scores clearly vary from school to school even within the program labels given (that is, schools that are "bilingual" vary, as do the schools that are "English-immersion").

One of the most important observations that can be made is a comparison of the performance of the schools chosen by Californians Together with the average performance of the Oceanside school district. Figures 1 and 2, which show reading scores for grades 2 and 3, show that the much-noted rise in Oceanside scores are indeed not that different from the patterns of increases that can be found in many bilingual schools. In this case, where comparisons are made with the schools highlighted by Californians Together that use bilingual education, there is relatively little that is remarkable about the score results for Oceanside.

We also looked more closely at data from the Oceanside district. Specifically, we examined SAT-9 scores at the school level. We plotted SAT-9 reading data for LEP second and third graders at all elementary schools at the district, except the following four schools which had very low LEP enrollments: Stuart Mesa, Santa Margarita, North Terrace, and Ivey Ranch. Figures 3 and 4 show the results. While we have only scratched the surface of the data, so

to speak, it is clear that there is considerable variability from school to school within the Oceanside school district. In addition to the school variability in Oceanside, there are striking differences in the patterns for second and third graders (as was seen for both the state as a whole as well as in the Californians Together dataset). These observations simply serve to reinforce our initial observations about the limitations of SAT-9 data for drawing conclusions about the effects of Proposition 227.

Figure 3. Performance in SAT-9 reading among elementary schools in Oceanside school district (2nd grade).

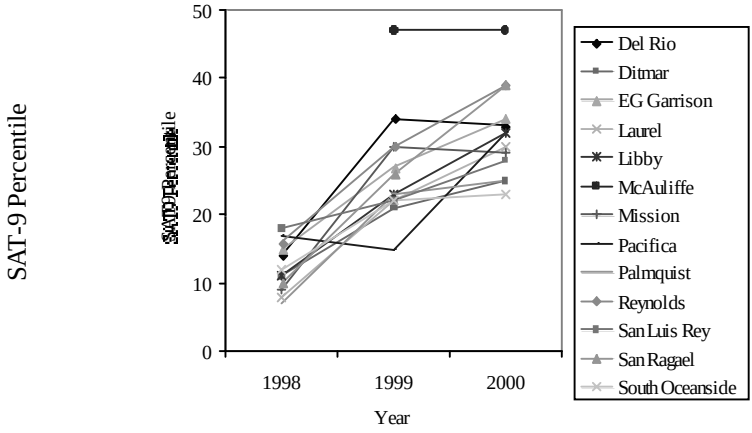
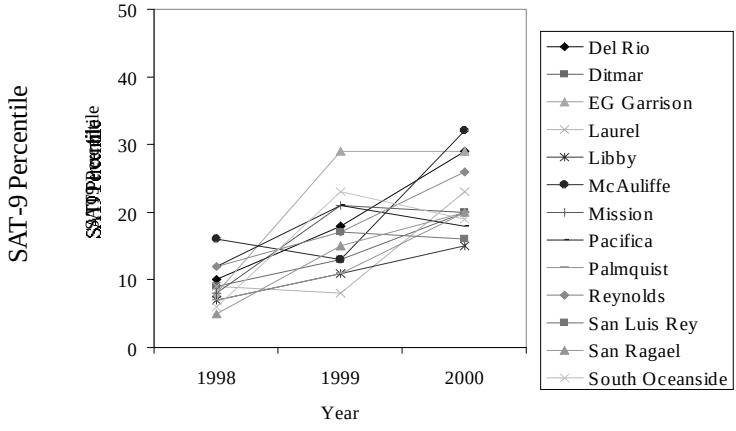


Figure 4. Performance in SAT-9 reading among elementary schools in Oceanside school district (3rd grade).



The results of the statistical effect known as "regression to the mean": Lower scores are more likely to move up

The large increase in SAT-9 scores at the Oceanside school district from 1998 to 2000 can also be explained by the law of statistics known as "regression to the mean" (Campbell & Kenny, 1999). One might notice that the performance of the Oceanside school district was particularly low in 1998. The district's average performance among LEP second graders was at the 12th percentile (whereas LEP students statewide performed at the 19th percentile), and the average performance among LEP third graders was at the 9th percentile (compared to the 14th percentile statewide). In other words, the LEP students in Oceanside started out among the lowest in a group of students whose score was low to begin with. The statistical phenomenon known as "regression to the mean" describes the effect whereby scores at the extreme ends of the statistical distribution tend to move toward the population average (mean). Simply stated, low scores tend to move higher and high scores tend to move lower. According to this law of statistics, the scores at the Oceanside school district should have been expected to rise upon retesting simply because their beginning scores were so low.

In order to demonstrate this effect of regression to the mean, we examined the performance of third grade students who were low-achieving LEP and non-LEP students. We selected schools that had low overall performance in reading for both LEP and non-LEP students. Specifically, for non-LEP students, we identified all the schools in which there were fewer than 3% LEP students, but in which the average National Percentile Rank score in reading was low (< 27th percentile) for 1998. There were 30 such schools altogether. We introduced this method in order to compare the performance among non-LEP students because there were no separate data available for non-LEP students in 1998. We then tracked the changes in these schools for 1999 and 2000. For the LEP students, we identified 26 schools that had both high percentages of LEP students (> 80%) and low reading scores on the SAT-9 test in 1998 (< 10th percentile). We also traced the changes in these schools for 1999 and 2000. It is important to note that we have no information on what types of programs were implemented at these schools: Some schools would have employed some forms of bilingual programs while others would have employed English immersion programs. It is our belief, however, that tracking student performance in these schools provides arguably the cleanest comparison of LEP and non-LEP students in low-reading schools allowable by the data presently available.

The results in reading are shown graphically in figures 5 and 6. As the bold lines (indicating the median percentile rank) show, there are clear increases in reading scores across the three years for *both* LEP and non-LEP native English speakers in schools with low reading scores. While there was, on average, a 7% increase in scores among schools with mostly LEP students from 1998 to 2000, SAT-9 scores among schools with mostly non-LEP students

increased by 10% from 1998 to 2000. (Certainly, this increase among schools with mostly non-LEP students cannot be attributed to the effects of Proposition 227!) A similar increase was observed in math and language data as well. Again, as in the statewide statistics, the increased performance on the SAT-9 test seems to be across the board. Regardless of the types of programs and irrespective of LEP or non-LEP status, the schools that started from low scores showed improvement. One can observe herein clear indications of a regression to the mean in the SAT-9 data, as the schools were all selected for their initially low reading scores.

Figure 5. SAT-9 reading percentile scores from sample schools with mostly LEP students, 1998, 1999, and 2000 (N = 26). Data are for LEP students only.

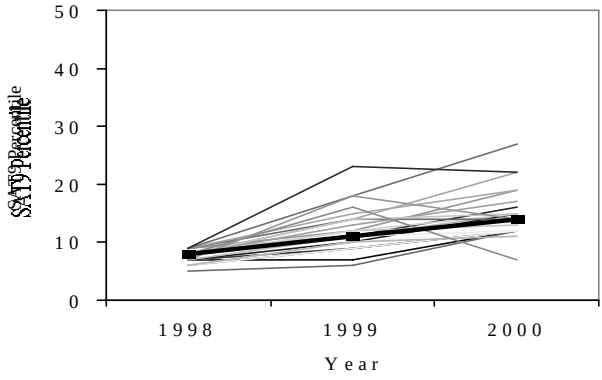
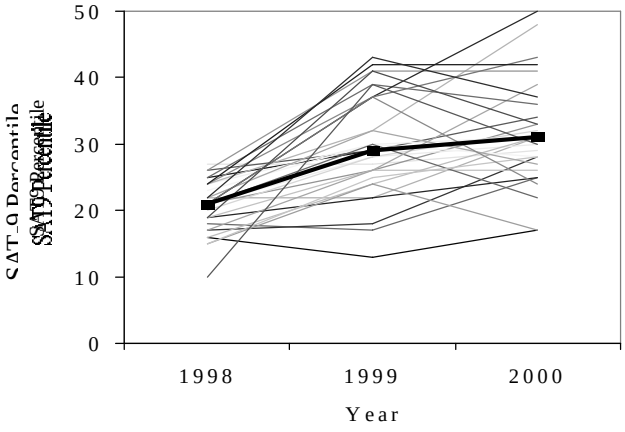


Figure 6. SAT-9 reading percentile scores from sample schools with mostly native English-speaking students (< 3% LEP), 1998, 1999, and 2000 (N = 30).



The SAT-9 test is not designed to measure English development and academic achievement for LEP students

The last point that we would like to make is that the SAT-9 test is not an appropriate measure to assess English development and academic achievement for LEP students. The SAT-9 was developed to distinguish academic achievement among native speakers of English; it is not a measure of English language development for LEP students. Therefore, the test measures constructs that are qualitatively different from what would be expected of students learning English.

Even for native speakers of English, the low end of the score distribution is poorly measured by the test. Tests such as the SAT-9 are designed to make fine discriminations between students who score in the middle part of the bell-shaped distribution where most students perform, and far less to distinguish incremental performance among the top and bottom ends of the bell curve. Imagine a situation where you need to take a multiple-choice test on a subject that you know little about. You might need to guess most of the answers so that your scores might simply tell you how lucky you are, but not how much you know about the subject. Depending on your "luck," your scores would be totally different. The scores among LEP students appear to indicate an effect similar to this. Indeed, the scores themselves are subject to noisy measurements and the bouncing around of scores, as one can see in figures 5 and 6.

Given that the SAT-9 is a weak measure of English ability among LEP students, it can provide us with very limited, gross information. It is certainly not refined enough to tell us about differences between program labels, such as bilingual versus English immersion programs.

Conclusion

A substantial amount of attention has been given to the impact of Proposition 227. Proponents of this initiative reported the "effectiveness" of English immersion programs over bilingual programs based on an increase in SAT-9 scores among selected schools that were reported to have switched to English immersion from bilingual programs. In this article, based on a number of analyses of SAT-9 scores from 1998 to 2000, we have argued that there are serious limitations in using the SAT-9 data to make this claim. Our arguments against the use of SAT-9 data for this purpose include the following:

1. SAT-9 scores rose not only among LEP students but also for all the students statewide. Thus, the gains in scores cannot be attributed solely to the effects of Proposition 227;
2. Increases in SAT-9 scores could be the result of any number of possible causes that are not distinguishable from the effects of Proposition 227 per se;

3. Both bilingual districts and English-only districts showed similar gains in their SAT-9 scores;
4. One can see significant variation in performance within both bilingual schools and English-only schools, indicating limitations in using the SAT-9 data to make claims about improvements based on program type;
5. Regardless of the programs implemented, or whether or not these were for LEP or non-LEP students, the schools that started with low initial scores showed much improvement across the board. This is the result of regression toward the mean. The gain in scores for the Oceanside school district that has been reported as an example of the successful implementation of Proposition 227 was not exceptional among schools or districts which started from very low scores in 1998; and
6. The SAT-9 test is designed to distinguish academic achievement among native speakers of English and not for that among English learning students. The scores contain large amounts of unexplained variability for LEP students. The SAT-9 test is not fine enough to assess either LEP students' English development or their academic achievement, and thus one should not rely on such a blunt tool to judge the impact of Proposition 227 for LEP students.

As a final policy point, we would point out that the failure of schools and states to ensure accountability for academic achievement for LEP students has been an enduring problem. Thus, we should be happy that the educational system and the public are starting to pay more attention to the academic progress of these students. Yet as the above analyses indicate, the situation with SAT-9 should be considered an extremely poor implementation of the idea of including LEP students in accountability and school reform.

References

Amselle, J., & Allison, A. C. (2000, August). Two years of success: An analysis of California test scores after Proposition 227. *Read Abstract* [On-line]. Available at: http://www.ceousa.org/html/227rep.html

Californians Together. (2000, August 21). Schools with large enrollments of English learners and substantial bilingual instruction are effective in teaching English [On-line]. Available at: http://www.bilingualeducation.org/news.htm

Campbell, D., & Kenny, D. (1999). *A primer on regression artifacts.* New York: The Guilford Press.

Steinberg, J. (2000, August 20). Increase in test scores counters dire forecasts for bilingual ban. *New York Times, p. A1.*

Endnotes

[1] This research was funded in part through grants from the Spencer Foundation and the James S. McDonnell Foundation.

[2] Mr. Unz provides a comprehensive list of media coverage on the release of the 2000 scores at http://www.onenation.org/news.html