January 2002

Educational policy analysis archives

Arizona State University

University of South Florida

Education Policy Analysis Archives

Volume 10 Number 7 **January 25, 2002** **ISSN 1068-2341**

A peer-reviewed scholarly journal
Editor: Gene V Glass
College of Education
Arizona State University

Articles appearing in **EPAA** are abstracted in the *Current Index to Journals in Education* by the ERIC Clearinghouse on Assessment and Evaluation and are permanently archived in *Resources in Education*.

¿Exito en California?
A Validity Critique of Language Program Evaluations and Analysis of English Learner Test Scores

Marilyn S. Thompson
Kristen E. DiCerbo
Kate Mahoney
Jeff MacSwan
Arizona State University

Citation: Thompson, M.S., DiCerbo, K.E., Mahoney, K. and MacSwan, J. (2002, January 25). *¿Exito en California?* A validity critique of language program evaluations and analysis of English learner test scores. *Education Policy Analysis Archives*, *10*(7). Retrieved [date] from http://epaa.asu.edu/epaa/v10n7/.

Abstract

Several states have recently faced ballot initiatives that propose to functionally eliminate bilingual education in favor of English-only approaches. Proponents of these initiatives have argued an overall rise in standardized achievement scores of California's limited English proficient (LEP) students is largely due to the implementation of English immersion programs mandated by Proposition 227 in 1998, hence, they claim *Exito en California* (Success in California). However, many such arguments presented in the media were based on flawed summaries of

these data. We first discuss the background, media coverage, and previous research associated with California's Proposition 227. We then present a series of validity concerns regarding use of Stanford-9 achievement data to address policy for educating LEP students; these concerns include the language of the test, alternative explanations, sample selection, and data analysis decisions. Finally, we present a comprehensive summary of scaled-score achievement means and trajectories for California's LEP and non-LEP students for 1998-2000. Our analyses indicate that although scores have risen overall, the achievement gap between LEP and EP students does not appear to be narrowing.

Education policy concerning the instruction of limited English proficient (LEP) students in the United States has been debated for a number of decades and considerable attention has been given to the best method of instruction for these students. In recent years, the controversy regarding how to best educate LEP students has surfaced in the form of political legislation. According to the United States Department of Education (1994) the term "limited English proficient" refers to individuals who (1) were not born in the U. S. and whose native language is other than English, or (2) come from environments in which a language other than English is dominant. The education of LEP students is important, especially given that during the 1996-1997 academic year U. S. school districts reported an enrollment of approximately 3.5 million limited-English proficient (LEP) students, accounting for 7.4% of the total reported enrollment (Macías, Nishikawa, & Venegas, 1998). The proportional rate of increase in LEP students from 1995 to 2020 is projected to be 96%, compared to an expected increase of 22% for native-English speakers (Campbell, 1994).

The controversy regarding the education of LEP students generally focuses on the amount of instruction provided in the children's native language. English immersion programs provide instruction almost exclusively in English, which the teacher attempts to make accessible to LEP students. Bilingual education programs provide a substantial amount of content area instruction in the students' native language, while some time each day is spent developing English skills. It should be noted, however, that the actual implementation of programs varies across states, districts, schools, and even classrooms (August & Hakuta, 1998; Berliner, 1988).

Proponents of bilingual education argue that without support in their native language, LEP students will fall behind academically while they are learning English (Crawford, 1999; Krashen, 1996). They also argue that if students first learn to read in the language in which they are fluent, they can then transfer those skills to reading in English (Krashen, 1996). Proponents of English immersion argue that instructional time devoted to intensive learning of English will more likely benefit children's academic achievement in a second language environment (Rossell & Baker, 1996).

Recently, the debate surrounding the education of LEP students has shifted to the political arena. Several states, including California, Colorado, and Arizona, have faced ballot initiatives that propose to restrict the types of educational methods and programs that may be used to instruct LEP students. Specifically, these restrictions functionally eliminate bilingual education programs in favor of an English immersion approach. The first such proposition was California's Proposition 227, which passed by a majority vote in 1998. In November of 2000, voters in Arizona approved a similar, but even more

restrictive measure, Proposition 203. In early 2001, measures similar to these propositions were introduced in the state legislatures of Massachusetts, Oregon, and Rhode Island.

In this article, we first provide an overview of the media coverage surrounding the implementation and evaluation of California's Proposition 227 and then review scholarly analyses related to its initiation, including both qualitative studies of implementation and quantitative evaluations of student achievement scores. We then discuss several methodological problems we have observed with the use of California Department of Education's Standardized Testing and Reporting (STAR; California Standardized Testing and Reporting, 2000) data to support arguments about the effects of Proposition 227. We limit our discussion to Stanford-9 scores released in 1998, 1999 and 2000 because we are concerned with the validity of claims about the success of Proposition 227 which purport to derive from these specific data. See Hakuta (2001) for remarks on the 2001 Stanford-9 scores of California's English learners. We frame these problems in the context of specific threats to validity and inappropriate approaches to data analysis. Finally, we present a comprehensive reanalysis of the STAR data and summarize achievement trajectories for LEP and non-LEP students. In interpreting our reanalysis, we discuss policy issues we feel cannot be adequately examined based on these data.

All the News That's Fit to Print?
Media Accounts of California's Stanford-9 Scores and Proposition 227

California implemented Proposition 227 during the 1998-1999 school year. During the previous year, California also began statewide administration of the Stanford Achievement Test, 9th edition (Stanford-9). These test results are publicly available, aggregated by grade level for each school, through the STAR system (California Standardized Testing and Reporting, 2000). In the past two years, many educators, media sources, and political stakeholders have reported summaries of these data as evidence of the effectiveness of Proposition 227.

The *New York Times*, which is widely recognized as one of the most influential newspapers in the United States, published a news story focusing on Stanford-9 achievement scores of LEP children in California on August 20, 2000. *Times* reporter Jacque Steinberg claimed that the increase in scores "at the very least" represented "a tentative affirmation" of the vision of Ron Unz (Steinberg, 2000, A1), who had sponsored the California initiative that banned bilingual education two years earlier. The *Times* story appeared as front-page news, running 1,744 words in length, and opened with the following statement:

> Two years after Californians voted to end bilingual education and force a million Spanish-speaking students to immerse themselves in English as if it were a cold bath, those students are improving in reading and other subjects at often striking rates, according to standardized test scores released this week. (p. A1)

Steinberg concluded the test results provide tentative evidence that Proposition 227's prescribed "cold bath" of English immersion is responsible for the increase in scores, and characterized the results as "remarkable." The *Times* piece also included an extensive anecdote of a school superintendent from the Oceanside district who

converted from an advocate of bilingual education to a proponent of structured English immersion.

To present a contrasting view, Steinberg included a 59-word paragraph in which he suggested alternative explanations for the increase in scores, citing class-size reduction in particular. However, these alternatives were introduced by the suggestion that Proposition 227 was at least in part responsible for the increase, which Steinberg found to be "remarkable given predictions that scores of Spanish-speaking children would plummet" (p. A1). Steinberg also quoted Stanford Professor Kenji Hakuta, who had conducted an analysis of the test scores and posted them on the World Wide Web the same day they were released. Rather than discussing Hakuta's study, Steinberg briefly summarized that it was Hakuta's view that "few conclusions could be drawn from the results, other than that 'the numbers didn't turn negative,' as many had feared" (p. A1). Steinberg appeared to use Hakuta's quote essentially to sustain his main point, namely, that the increase in test scores is a tentative affirmation of Proposition 227, and a clear indication that educators were wrong to predict children would suffer.

The *Times* story was syndicated in the *Milwaukee Journal Sentinel* (Steinberg, 2000b, p. 3A), where the opposing viewpoint was cut by half, and in the *Baltimore Sun* (New York Times New Service, 2000, p. 3A) and Cleveland's *Plain Dealer* (Steinberg, 2000c, p. 21A), where it was entirely eliminated. After the *New York Times* story appeared, the idea that California's Stanford-9 gains for LEP students resulted from the implementation of Proposition 227 was cited in 24 major U. S. newspapers, frequently without any question of the accuracy of the claim. Of these, 17 (or 71%) gave no voice to opposing viewpoint at all (see Table 1). (Note 1)

Table 1
News Stories in Major Newspapers (Aug. 20, '00—June 9, '01) Mentioning the Increase in California's Stanford-9 Test Scores as Evidence of the Success of Structured English Immersion (Proposition 227)

Newspaper	Date of publication	Length of article (in words)	Length of opposing view (in words)
The Plain Dealer	8/20/00	712	0
Milwaukee Journal Sentinel	8/20/00	839	41
The Baltimore Sun	8/20/00	848	0
New York Times	8/20/00	1744	78
The Arizona Republic	8/22/00	679	46
The Christian Science Monitor	8/23/00	1172	0
The Houston Chronicle	8/28/00	261	105
Star Tribune	8/28/00	540	0
USA Today	8/28/00	996	0

Newsday	9/09/00	450	0
The Arizona Republic	9/22/00	844	0
The Christian Science Monitor	9/27/00	862	0
The San Diego Union Tribune	10/06/00	574	0
The Arizona Republic	10/29/00	1283	0
Los Angeles Times	11/07/00	98	46
The Arizona Republic	11/08/00	678	0
New York Times	11/15/00	600	0
The Boston Globe	12/31/00	1125	0
The Atlanta Journ. & Constitution	1/04/01	647	0
The Boston Globe	1/14/01	436	0
The Arizona Republic	3/01/01	431	0
The Arizona Republic	3/02/01	448	40
The Denver Post	3/28/01	811	29
New York Times	4/01/01	228	0
	Averages:	721.1	16.0

Interestingly, the Associated Press (AP), which writes stories circulated to its 1,550 clients, including the *New York Times*, wrote a considerably more balanced story a week before the publication of the *Times* piece. AP reporter Jennifer Kerr's story opened as follows: "Two years after voters ended most bilingual education in California, statewide test scores for non-English speakers jumped about as much as scores for their fluent fellow students" (Kerr, 2000). Kerr's story noted that test scores had risen for all students in the state about equally, that the Stanford-9 was not written for English learners and is arguably inappropriate, and provided much stronger objections from the research community.

Only three news stories appeared in major U.S. newspapers before the *New York Times* story, one in the *Los Angeles Times* (Groves, 2000, p. A3) and two in the *San Diego Union-Tribune* (Moran & Spielvogel, 2000, p. B1). Like the AP story, these papers presented a much more balanced account. Groves' *Los Angeles Times* story began,

> California students who are not proficient in English improved their scores on the Stanford 9 standardized test at about the same rate as their fluent classmates, but new state data released Monday continue to show an immense disparity between the two groups. (p. A3)

The main *San Diego Union-Tribune* story (Moran & Spielvogel, 2000) opened this way:

> Celebrated gains in student state *test scores* are spread among all students—whether advantaged or disadvantaged, whether they speak English or

not—according to data released today. (p. B1)

However, the view appearing in the *New York Times,* due to the paper's enormous influence on the national press, strongly predominated. The *Times* was cited as an authority on the issue in 56 published letters and editorials, and in one story appearing in the *Arizona Republic* (Gonzalez, 2000, p. EX1). Following the appearance of the *Times* article, numerous television and radio news shows, including those of the major television networks, broadcasted the story that rising scores in California indicated Proposition 227 was a success in that state. A story in *Newsday* (Willen & Kowal, 2000, p. A10) said that the conclusion followed from "a recent California study."

Inaccuracies in scientific and technical reporting are known to occur widely in journalistic writing (Simon, Fico, & Lacy, 1989; Singer & Endreny, 1993; Tankard & Ryan, 1974; Weiss & Singer, 1987). However, what is particularly disturbing about the *New York Times* story is that conclusions were drawn based on claims that disregarded basic principles of scientific research design and educational measurement. Undermining the credibility of the story were inadequate consideration of alternative explanations and improper interpretation and use of Stanford-9 scores. Further, the *Times* failed to discuss controlled studies comparing bilingual education to all-English instructional approaches (Ramirez et al., 1991; Willig, 1985) or recent comprehensive research syntheses prepared by the National Research Council (August & Hakuta, 1998; Meyer & Fienberg, 1992). These errors and exclusions are particularly grievous given the high-stakes nature of the inferences drawn regarding an extremely complex educational issue.

Brief Background on Proposition 227 Implementation

In this section, we provide some background on Proposition 227 and summarize briefly recent research addressing the implementation of the initiative. The full text of the California law can be reviewed online (English Language Education for Immigrant Children, 2001; http://www.leginfo.ca.gov/calaw.html), however the general mandate of Proposition 227 is the following: "Children who are English learners shall be educated through sheltered English immersion during a temporary transition period not normally intended to exceed one year" (Section 305, 2001). The law applies to English learners, defined as "a child who does not speak English or whose native language is not English and who is not currently able to perform ordinary classroom work in English, also known as a Limited English Proficiency or LEP child" (Section 306, 2001). The law also further defines sheltered English immersion as "an English language acquisition process for young children in which nearly all classroom instruction is in English but with the curriculum and presentation designed for children who are learning the language" (Section 306, 2001). The implementation of sheltered English immersion (SEI; equivalently referred to as 'structured' English immersion) as mandated by Proposition 227 has been addressed in the educational research literature. Most of these studies may be described as either qualitative studies of the implementation of SEI in California schools or quantitative summaries of standardized achievement scores pre- and post-implementation of Proposition 227.

Studies of Proposition 227 Implementation

Although districts, schools, and teachers did not ignore Proposition 227, there was not a

"sea of change" in programs for English learners apparent in the schools (García & Curry-Rodríguez, 2000). In fact, prior to implementation of Proposition 227, only 29% of English learners were in programs that included native language instruction, and 12% of students were still in those programs following implementation (Gándera et al., 2000). Maxwell-Jolly (2000) studied the interpretation and implementation of 227 in seven different school districts and found that although district interpretation of 227 set the tone, responses to and implementation of district policy regarding 227 varied widely. Further research indicated that when district administrators set a strong tone for eliminating native language instruction or providing alternatives to SEI, schools followed suit. However, when district leadership was lacking, implementation of the proposition varied across schools (Gándera et al., 2000).

Gándara (2000) documented the impact 227 had on instructional services, classroom pedagogy, and distribution of teachers, concluding the greatest impact of Proposition 227 was on classroom instruction. For example, teachers reported leaving out much of their normal literacy instruction, such as storytelling and story sequencing, to focus on English word recognition. Instructional challenges presented by Proposition 227 included having a lack of instructional materials and teaching students with a wider linguistic range (Schirling, Contreras, & Ayala, 2000). Teachers reported that even for programs in which parental waivers were obtained for native language instruction, they were required to include 30 days of English instruction before the waiver could take effect. Because schools did not know how many waivers they would receive, orders for instructional materials were delayed or made in insufficient quantities.

Hayes & Salazar (2001) evaluated instructional services offered to English learners enrolled in SEI in first, second, and third grade classes in Los Angeles Unified School District. They noted uneven implementation of SEI, with programs generally adopting one of two general approaches: use of primary language *for clarification* only and use of primary language for *concept development*. The effects of the proposition on teachers varied based on what the teachers had done prior to the passage of 227 and on teachers' education, skills, experience, and views on student learning (Gándera et al., 2000). For example, teachers who were certified to teach bilingual education were more likely to continue some level of native language support in their classrooms.

Studies offering various other perspectives on implementation of Proposition 227 have been published, many of them in a special issue of the *Bilingual Research Journal* devoted to the topic (e.g., Dixon, Green, Yeager, Baker, & Fránquiz, 2000; Palmer & Garcia, 2000; Paredes, 2000; Schirling, Contreras, & Ayala, 2000; Stritikus & Garcia, 2000). A California Research Bureau report by de Cos (1999) presented issues surrounding implementation of 227 in a historical context of language policy issues. The effects of STAR on English learners were also discussed and the author warned against using these publicly available test scores to evaluate SEI programs. We now review some published quantitative analyses of these STAR data that have been used to support arguments for or against Proposition 227.

Analyses of California Achievement Data

When the publicly available aggregated standardized test scores of California children were released following implementation of Proposition 227, they were quickly analyzed in an attempt to determine the effects of the initiative. It is well-documented in the

literature that LEP students made gains in test scores, as did all students in the state (Butler, Orr, Gutierrez, & Hakuta, 2000; Gándara, 2000; Garcia & Curry-Rodríguez, 2000). Butler et al. (2000) reported that schools maintaining strong bilingual programs had scores that equaled or exceeded those of schools that had dropped bilingual programs. In addition, Butler et al. (2000) noted that due to regression to the mean, scores of lower performing students are more likely to improve than those at the middle of the scale. Finally, they emphasized there was significant variation in test scores across schools in both the bilingual and English-only categories. García and Curry-Rodríguez (2000) studied a random sample of districts and found no specific patterns of test scores across schools with different 227 implementation strategies.

Amselle and Allison (2000) examined percentile rank increases for LEP students and found that LEP students made "significant gains in reading and writing in English as well as math" (p.1). They went on to examine percentile rank improvements in four school districts that reported to be in strict compliance with Proposition 227 and four school districts that reported maintenance of a bilingual program. They found greater score improvements in the select districts reporting compliance with the initiative. Finally, they pointed to Los Angeles Unified School District as a district that openly defied Proposition 227 and had percentile rank test scores below "the state average for LEP students" (p. 12). Unfortunately, Amselle and Allison (2000) failed to note the variability within districts. In addition, their focus on select districts did not allow them to examine the variability across districts that reported similar implementations of Proposition 227. Finally, they inappropriately used summaries of national percentile ranks to determine academic growth, a problem we discuss in greater depth later in this paper.

A noteworthy limitation of the publicly available data is the lack of student level information (Gándara, 2000). However, Gutiérrez, Asato, and Baquedano-Lopez (2000) acquired and utilized student-level data for LEP students from an urban unified school district. Over three years, they tracked student scores in this predominantly English-only, phonics-based literacy district. They found the percentage of LEP students scoring at or above the 50th percentile decreased dramatically over the three years. Disaggregation of these data by language group showed that the percentage of Spanish-speaking children reading at or above the 50th percentile dropped from 32% in the first grade to 30% in the second grade to 15% in the third grade. Other language groups (Cantonese, Russian, Hmong, and Mien) also experienced sharp declines between first and third grade (Gutiérrez et al., 2000). The specific causes of these declines were not explored.

In sum, published qualitative evaluation reports of Proposition 227 generally conclude the overall effect of the new law on the education of language minority students has been negative. Furthermore, with the exception of Amselle and Allison's (2000) report, quantitative analyses of the Stanford-9 data to date reveal comparable gains for English learners and their fluent English-speaking peers. Many arguments and quantitative summaries based on the STAR data have been replete with improper statistical analyses and fail to acknowledge the many limitations of these highly aggregated standardized achievement data (e.g., Amselle & Allison, 2000). We now discuss multiple validity concerns as they apply to use of the California Stanford-9 data for evaluating language policy.

Validity Issues

Of utmost concern when using assessment data in research should be the validity of the assessment for the intended purpose. A large literature exists addressing the conceptualization of validity in educational and psychological testing and research (see Messick, 1989, for a comprehensive discussion of validity); therefore, we do not attempt a comprehensive review of validity, but rather concentrate on those validity issues that appear to be most problematic in our research context. To help focus our discussion, we borrow from Messick (1989) a definition of validity as a unified concept with multiple facets: "Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment" (Messick, 1989, p. 13, emphasis added). As such, validity is not merely about the meaning of test scores. Validity encompasses "... the interpretability, relevance, and utility of scores, the import or value implications of scores as a basis for action, and the functional worth of scores in terms of social consequences of their use" (Messick, 1989, p. 13). The measurement context and the inferential context are both vitally important in forming validity judgments. We focus our discussion in the following section on issues that pose major threats to the validity of inferences based on the STAR data concerning LEP students: language of the test, alternative explanations, and sample selection.

Language of the Test

In this section, we consider the meaning of scores in the context of the assessment. Of particular concern is the administration of English-language standardized achievement tests to evaluate the academic achievement of students who are not proficient in English. Testing students in a language in which they are not yet proficient is problematic for multiple reasons. The *Standards for Educational and Psychological Testing* warn that when testing a non-native speaker in English, the test results may not reflect accurately the abilities and competencies being measured if test performance depends on the test takers' knowledge of English (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). The Stanford-9 is a test of academic achievement, not a test of language proficiency, and the test developers have not conducted any specific studies to establish validity of Stanford-9 scores for children who have limited ability in the language of the test (Harcourt Brace Educational Measurement, 1997c). Therefore, limited English proficiency should be regarded a likely source of measurement error in the Stanford-9 test scores intended to reflect academic achievement.

Language proficiency in general has been shown to influence performance on achievement tests (Ulibarri, Spencer, & Rivas, 1981). Pilkington, Piersel, and Ponterotto (1988) reported that the home language of a child influenced the predictive validity of kindergarten achievement measures. These studies suggest language proficiency plays a role in young children's performance on achievement tests. This relationship may continue in high school children, where LEP status was shown to be a significant predictor of both language arts and mathematics scores on the California Assessment of Progress, although a poverty measure was a stronger predictor (Wright & Michael, 1989).

Alternative Explanations

An important consideration in interpreting trajectories of achievement scores in California is the acknowledgement of potential confounding conditions and alternative explanations. In this section, we address five such considerations: simultaneous changes in educational policy and practice, inconsistent implementation of immersion programs, increasing test familiarity and preparation, the limitations of using aggregated data, and regression to the mean.

Simultaneous Policy Implementations

Proposition 227 was introduced concurrently with other changes in educational policy and practice. In fact, Gándara (2000) explained Proposition 227 was enacted in what has been the most active period of education reform in recent times. Statewide initiatives include class size reductions from an average of 30 to 20 in early elementary classrooms and a switch from a whole language approach to a phonics-based method of reading instruction for poor readers. Gutiérrez et al. (2000) specifically noted that class size reduction, the new state standardized testing program, new reading and accountability initiatives, and the new language arts standards had all been implemented concurrently. In addition, other reforms have likely occurred at the district, school, and classroom level. Any or all of these may be important contributors to student gains.

Inconsistencies in Language Programs

Evaluating language program policy is further complicated by inconsistent implementation of English immersion programs, with instructional practices varying widely across districts, schools, and even classrooms (Berliner, 1988; Gándara et al., 2000). There are many different variations of English-only programs, as well as of bilingual programs. Therefore, it is not clear exactly which programs are being compared when simply examining changes in test scores from 1998 to 1999. The state education system is terrifically diverse and educational practices are far from uniform.

Test Familiarity and Coaching

Because the aggregated scores are publicly released, schools and districts feel pressure to achieve high test scores and thus encourage test preparation in varying degrees. As the stakes become higher, test preparation often becomes a high profit industry, as documented in Texas (McNeil, 2000; Sacks, 1999). The California test was tied to different, but motivating, rewards and sanctions for schools, teachers, and students. Schools and teachers could receive bonuses based on increased test scores. For example, each of the 1,000 certificated staff in underachieving schools with the largest growth in California receives $25,000 (Public Schools Accountability Act, 1999). On the other hand, schools that do not meet goals for academic improvement in 24 months may be taken over by the state Superintendent for Public Instruction. The Superintendent may then take a variety of actions, up to and including closing the school (Public Schools Accountability Act, 1999). Bilingual and English-only teachers alike, in the presence of so many rewards and sanctions, may feel pressure to specifically teach to the test or focus disproportionately on test preparation, as Moran (2000) has discussed.

Even without these extensive consequences, a dramatic and consistent rise in test scores is frequently observed in the first few years following implementation of a new testing program (Linn, Graue, & Sanders, 1990), such as occurred in California following implementation of Stanford-9 testing in 1998. There are several possible explanations for this trend. Coaching and teaching to the test, often at the expense of more desirable teaching and learning activities, can contribute to striking rises in test scores. A meta-analysis of 30 studies revealed that coaching for standardized tests increases test scores in the typical study by .25 standard deviations (Bangert-Drowns, Kulik, & Kulik, 1983). Coaching may refer to a variety of test preparation activities, including general test-taking strategies (e.g., guessing, underlining main ideas, time management), test-specific strategies (e.g., methods useful for quirks of a particular test), and academic instruction tied closely to the content and skills on the assessment (Anastasi, 1981; Bond, 1989).

As teachers and administrators become more familiar with the tests, coaching strategies may become more effective. Butler et al. (2000) suggested that since there is a trend for test scores to rise for all students in California, these broad patterns of improvement may result largely from "teaching to the test." Teachers in all types of classrooms, including bilingual and SEI classrooms, report having modified their teaching practices substantially, with a greater emphasis on preparing students to answer English standardized test-like questions (Alamillo & Viramontes, 2000; Gándara 2000).

Unknown Student and School Characteristics

Another factor limiting conclusions based on the STAR data is the nature of the data itself. We have noted the California STAR data are available to the public and the research community only in aggregated form—by grade level within school. The lack of student-level information makes these data insufficient for thoroughly exploring relations between student achievement and language program for LEP students. Although indicators of the dominant language program and enrollment numbers are available at the school level (California Language Census Data Files, 2000), this information cannot be tied to individual students or even grade-level averages. Further, relevant student-level information such as socioeconomic status, level of English proficiency, and the previous year's score cannot be used to control for potentially relevant individual differences.

A problem of particular relevance for studying the effects of language programs on LEP students is the variability among districts in the criteria for defining LEP students, as well as who will be tested. As noted by Gándara (2000) and Butler et al. (2000), redesignation of students with borderline English proficiency could have a profound effect on aggregated scores in LEP and non-LEP groups. Although the Stanford-9 is not a measure of English language proficiency, some districts redesignate LEP students achieving a certain score on the Stanford-9 to EP status for the following year. This skimming effect may result in depressed scores for the LEP group. In contrast, Gándara (2000) pointed out that some districts are not reclassifying students on the basis of high scores on the Stanford-9, perhaps because they have not performed well on language proficiency tests. Differences in redesignation policies and rates, in the absence of student-level data, blur the meaning ascribed to LEP and EP score means.

Additionally, use of these aggregated data induces two distinct problems related to school size. First, grade-level averages of achievement scores for each academic subject were included only if there were at least ten students in the summary category represented. Schools were therefore unable to report summaries of any subgroup, such as LEP students, if there were fewer than ten students in the grade. It follows that the scores of many students are not represented in these subgroup aggregates. If, for example, schools with fewer LEP students tended to be schools having higher socioeconomic status (SES), systematic omission of these schools due to insufficient numbers of LEP students may introduce bias in estimated LEP group means related to average school SES.

A second problem related to school size is that the oft-reported statewide averages of the grade-level means do not account for the drastically varying numbers of students represented by the available grade-level within-school score aggregates. Certainly, the numbers of students in each grade varies across schools overall and for particular subgroups, yet computation of unweighted averages gives equal influence to small and large schools. This presents a unit of analysis problem when trying to make inferences about achievement at the student level. Unweighted averages of the grade-level means do not provide appropriate estimates of the statewide student averages.

As we later address, weighted means might provide better estimates of student scores. However, even weighted means do not allow representation of students excluded from subgroup summaries resulting from too few students in a category. Further, they do not address problems associated with the lack of relevant student covariates or redesignation of language proficiency status. Using aggregated rather than student-level data severely limits the nature and strength of generalizations that can be made based on these data. It is imperative that researchers realize the implicit limitations and fallacies associated with using such grossly aggregated data.

Regression to the Mean

Another explanation for score gains we must briefly consider is regression to the mean, a topic discussed thoroughly in many classic statistical textbooks (e.g., Campbell & Stanley, 1966; Glass & Hopkins; 1996) but often ignored by researchers as a genuine threat to validity. Regression to the mean refers to the tendency for student scores that are extreme upon initial testing (relative to the overall mean) to drift toward the population mean upon subsequent testing. Regression to the mean is particularly important when gains of extreme groups are of interest—such as in the comparison of low-scoring LEP students to other groups. For all available years of Stanford-9 score reports, LEP student scores are markedly lower than those for non-LEP students. In a district such as Oceanside, whose mean scores for LEP students were extremely low in 1998, scores might be expected to rise upon retesting—even without intervention—due to regression to the mean. Butler et al. (2000) compared low-scoring schools with mostly non-LEP students to low-scoring schools with mostly LEP students and demonstrated that schools with both compositions increased similarly from 1998-2000.

Failure to acknowledge or correctly address regression to the mean has been an issue in other large-scale policy analyses. For example, Camilli and Bulkley (2001), in a critique of Greene's (2001) evaluation of Florida's A-Plus accountability system, argued

convincingly for policy analysts to be aware of regression to the mean and to use statistical models that take regression to the mean into account. They noted such approaches have been recently employed in North Carolina's development of growth standards for the state. A detailed discussion of regression artifacts, particularly regression to the mean, may be found in a recent book devoted to this subject by Campbell and Kenny (1999).

Sample Selection

Many, if not most, of the published reports of analyses of California's Stanford-9 scores have been based on the consideration of a small sample of schools. While it is perhaps more feasible to focus on a few schools when attempting a descriptive study of the policy implementation process, it is dangerous to make inferences based on quantitative differences in mean achievement across a small number of select schools. The presence of school and classroom effects on student achievement is well documented. Students sharing a common classroom and/or school environment tend to perform more similarly on achievement tests than students sampled from multiple sites (Muthén, 1991; Thompson, 2000). Demographic similarities, as well as collective experiences of students sharing an educational environment, contribute to these classroom and school effects.

Oceanside initially became the district held up as representative of schools that strictly implemented Proposition 227. Beginning with the emphasis on Oceanside's Stanford-9 gains in press releases from Ron Unz and English-only supporters (e.g., English for the Children, 2000), score gains in this district have been repeatedly cited as evidence of the success of the proposition (Amselle & Adams, 2000; Steinberg, 2000). In response to these claims of success due to SEI, opponents of Proposition 227 pointed out marked gains seen in specific districts maintaining bilingual education. For example, Butler et al. (2000) chose schools nominated by Californians Together, a bilingual advocacy group, for analysis (e.g., Fresno Unified School District; Californians Together, 2000, August 21). They compared these to English-only districts held up by advocates of Proposition 227 as the most successful English-immersion schools.

Schools that maintained bilingual programs likely had administrators and teachers who were highly committed to these programs, given the effort needed to obtain parental waivers for participating children. This degree of commitment to bilingual programs may also suggest exceptionally strong and effective programs. Similarly, it can be argued districts such as Oceanside had atypically strong SEI programs. While comparing what seem to be the most successful districts of each program type is informative, the results of such a comparison should not be used to suggest that the same outcomes would be observed in districts with different characteristics. A characteristic unique to a school or district may contribute substantially to a rise in test scores; this characteristic may or may not be related to the language program. We should not be surprised to see contradictory results and inferences regarding program effects from studies that employ selective sampling of a few specific schools and districts. While comparisons of select schools and districts are informative, we urge caution in making generalizations based on such samples.

Data Analysis Decisions

The remaining issues we address are data analytic problems in summarizing the STAR data and using these summaries to support inferences. Our focus here is on analyses that manipulate students' scores in manners incongruent with the intended purpose of the assessment, and therefore these data analysis problems should also be considered threats to the validity of judgments based on STAR data. First, we discuss the misinterpretation and misuse of scores reported in the form of percentile ranks. Problems in using percentile ranks as a basis for longitudinal inferences result from incongruent norm group compositions, unequal score intervals, and difficulties in computing gains. We then discuss unit of analysis issues associated with using aggregated data to make inferences at the student level.

Misinterpretation and Misuse of Percentile Ranks

Individual student reports of performance on standardized achievement tests, including the Stanford-9, frequently feature percentile ranks. National percentile rank (NPR) scores indicate percentile ranks for a subtest relative to the national within-grade norm group. The NPR scores reported for students taking the Stanford-9 are derived from distributions of scaled scores broken down by grade and subject (Harcourt Brace Educational Measurement, 1997c). For example, a 2nd-grade student estimated to be at the 56th percentile on the math test would score higher than 56% of the students in the 2nd-grade norm group. Similarly, he or she would score lower than 44% of the students in the 2nd-grade norm group. The popularity of percentile rank scores is likely due to the ease with which they are understood at a practical level—parents are comfortable with the notion of a percentile scale on which their child's relative standing can be located. However, there often are hidden validity problems in utilizing a "relative" comparison group in interpreting achievement scores.

Consider a statement from the earlier-mentioned *New York Times* article describing increasing Stanford-9 achievement scores of LEP students in California: "In second grade, ... average score in reading of a student classified as limited in English increased 9 percentage points over the last two years, to the 28th percentile from the 19th percentile in national rankings, according to the state" (Steinberg, 2000, p. 1A). While this statement may seem quite clear on the surface, there are multiple assumptions about the meaning and comparison of percentile ranks that may cloud the perception of true academic growth. We briefly develop several points, well-known to psychometricians, that discourage the use of percentile ranks as measures for assessing academic gains for a collective group of students.

NPRs Are Norm-referenced Scores

For any one student, percentile ranks indicate only relative standing within a norm group. It follows that NPR scores should always be interpreted with the characteristics of the norm group in mind. The norm sample for the Stanford-9 was balanced to generally represent the U.S. population according to socioeconomic status, ethnicity, and urbanicity, with nonpublic schools oversampled to facilitate a separate norm group. Sampled schools were asked to test students who would typically be tested with other students in regular education classrooms, except those classified as trainable mentally handicapped or severely/profoundly mentally handicapped.

Individual districts and schools were therefore able to include or exclude LEP students

and some classifications of special education students according to local policy. The tested student population in California contains a much greater proportion of LEP students than does the Stanford-9 norm group. Specifically, the Stanford-9 spring norm sample contained only 1.8% LEP students (Harcourt Brace Educational Measurement, 1997b). In contrast, California estimates approximately 25% of its students are LEP (Macías et al., 1998). The incongruence between the makeup of the reference group and California with respect to LEP students calls into question the validity of generalizations based on NPR scores for LEP students.

Due to their normative nature, NPRs are not a measure of academic achievement as defined by a level of knowledge or skill. It is possible that a true academic gain may appear as a decline according to the change in NPR across years. For example, a student could display greater mastery than in the previous year, but have a lower percentile rank if students in the norm group scored proportionally higher than the tested student in the second year. It again follows that changes in percentile ranks across multiple years are not well suited for demonstrating improvements in academic knowledge or skill.

NPR Score Increments Represent Unequal Achievement Intervals

To understand more thoroughly the pitfalls of manipulating NPR scores, we consider how NPR scores are derived from the students' raw scores. A raw score is simply the total number of items a student answers correctly on a test. The test publisher determines NPRs through a two-step score conversion process (Harcourt Brace Educational Measurement, 1997b). On a specific test, such as the Stanford-9 5th-grade reading test, the original raw scores from the norm group are first transformed into *scaled scores* by applying item response theory (IRT). The IRT model employed for the Stanford-9 takes item difficulties into account to estimate a proficiency level, or scaled score, that is both independent of the specific items to which the student responds (i.e., the form and level of the subtest may vary) and independent of the group of students to whom the test is administered. These scaled scores are on a single scale for a subject area, so they can be compared across different test forms and grade levels. Scaled scores also have the convenient property of an equal-interval scale that supports comparisons of proficiency level across time for a specific subject test (i.e., a one-unit increase from 1998-1999 on a subject test represents the same amount of achievement growth as a one-unit increase from 1999-2000, regardless of grade level).

To convert scaled scores into NPRs, the cumulative distribution of scaled scores from the norm group is transformed into a roughly uniform distribution of percentile ranks ranging from 1 to 99. When a new group of students is administered the exam, their raw scores are first converted to scaled scores. NPRs are then determined for students in the testing group such that the percentile rank for a specific scaled score reflects the percentage of the norm group scoring at or below that level. Table 2 illustrates conversions among raw, scaled, and percentile rank scores on the 5th-grade reading subtest of the Stanford-9, using the spring national norm sample (Intermediate 2 Reading Scores, Form T; Harcourt Brace, 1997b). This reading subtest had 84 items. Because there are 4 choices, it is worth noting that we might expect merely guessing on the exam to yield correct answers to 21 of the 84 items, which would place the score at the 3rd percentile on the NPR scale.

Table 2

Conversions of Select Raw, Scaled, and National Percentile Rank Scores for the Stanford-9 5th-grade Reading Subtest

Raw score	Scaled scores (Form T)	National percentile rank
80	742	99
75	710	93
70	691	82
65	676	72
60	663	59
55	652	48
50	642	38
45	632	29
40	622	22
35	612	15
30	599	8
25	591	6
20	579	3
15	564	1

(Note: Adapted from Harcourt Brace Educational Measurement, 1997b. Intermediate 2 (grade 5) Total Reading, Form T, spring norm sample.)

This transformation of the distribution of scaled scores into percentile ranks is nonlinear, resulting in a loss of the equal-interval scale property of scaled scores. Equal differences in percentile ranks do not reflect equal differences in achievement or skill. Because the relative frequency of the original scaled scores is typically greater in the middle range than in the upper and lower ranges, conversion to percentile ranks results in spreading of the mid-range scaled scores and condensing of the upper- and lower-range scaled scores (the tails of the distribution).

This lack of an equal-interval scale has several important implications. An achievement gain of 1 scaled score point does not result in a consistent gain in percentile ranks throughout the range of the scale. Specifically, a gain of 1 percentile rank will reflect a greater achievement difference (reflected by the scaled score difference) in the upper or lower range than in the middle range. A difference in percentile ranks between 10 and 20 or between 80 and 90 may reflect a greater achievement gain than a difference between 50 and 60. This can be seen from the example in Table 3. Student 1 scored lower than most other students taking the test and Student 2 scored near the middle of students taking the test. Both students improved 50 scaled score points from 3rd grade to 4th grade, representing equivalent achievement gains. Student 1, at the low end of the scale, improved from the 1^{st} percentile to the 7^{th} percentile, an increase of 6 percentiles.

However, Student 2, in the middle of the scale, improved from the 41st to the 67th percentile, an increase of 26 percentiles. In addition, the accuracy of percentile ranks differs across the range of scores (Rogosa, 1999). Comparisons of percentile gains for students at different skill levels are not transparent and may be regarded as misleading at best.

Table 3
Example of Percentile Rank Gains at Different Points in the Distribution

	Student 1		Student 2	
	Grade 3	Grade 4	Grade 3	Grade 4
Scaled score	525	575	605	655
Percentile rank	1	7	41	67

(Note: Adapted from Harcourt Brace Educational Measurement, 1997b. Primary 3 (grade 3) & Intermediate 1 (grade 4) Total Reading, Form T, spring norm sample.)

NPRs Should Not Be Averaged or Used to Compute Gains

Another important implication of the lack of an equal-interval scale is that means and gain scores of NPRs should not be computed (Crocker & Algina, 1986; Cronbach, 1960). The California Stanford-9 data are only publicly available as means for each grade level within school. Additionally, these grade-level within-school means often are averaged further in an attempt to summarize subgroup means. Due to the unequal intervals created in deriving the NPR scale, such averaging of data may drastically propagate errors in estimating true achievement.

Although both the California STAR website (California Standardized Testing and Reporting, 2000) and Stanford-9 Technical Manual (Harcourt Brace Educational Measurement, 1997c) state explicitly that NPRs should not be used to determine true academic change across years, many reports have focused on changes in percentile ranks (e.g. Amselle & Allison, 2000; Butler et al., 2000; English for the Children, 8/14/2000; Steinberg, 2000, p. 1A). The major argument against using changes in NPRs to assess achievement gain for a student is that a gain of 1 scaled score point translates into different NPR intervals at different points on the scale, thereby obscuring measurement of true achievement gains. Further, because scores within each year have already been averaged, the collective gains of a group are even more problematic to assess and compare.

Gains have been computed from two perspectives: within-grade changes across years or cohort gains across years. *Within-grade changes*, which have been more commonly reported in the context of California STAR data, compare means for a grade level across subsequent years (e.g., 2nd-graders in 1998 to 2nd-graders in 1999). Even when using scaled scores, these within-grade changes are not true achievement gains in the sense that they are based on different groups of students. In contrast, *cohort gains* compare scores for a cohort of students across subsequent years (e.g., 2nd-graders in 1998 to

3rd-graders in 1999). However, we caution that when working with aggregated data, individual students cannot be tracked and therefore student mobility introduces some uncertainty to within-school cohort gains. Further, even if we assume the student group to be relatively consistent across years, the norm group used to determine the NPRs is different. As earlier described, it is therefore possible for slight gains in true achievement to appear as declines if the relative standing in the norm group is lower in the second year of testing, and vice versa.

In summary, the use of NPRs for tracking and comparing student achievement trajectories is problematic from multiple perspectives. The utility of percentile ranks is limited to informing judgments of how well a student does relative to the norm group for a subject and grade, gaining a view of a student's relative score profile across subjects, and evaluating whether a student has improved standing relative to the norm group from one year to the next. Such judgments are only valid if the norm group is an appropriate basis for comparison, which it quite clearly is not for LEP students. The characteristics of percentile ranks—including norm group inconsistencies and an unequal interval scale—make this score form unsuitable for large-scale longitudinal policy analysis. We now treat one additional data analysis problem we have observed in reports of California achievement results—inappropriate averaging of data.

Averaging Scores Across Subjects and Grades

Regardless of the score form reported, it is incorrect to average scores across different subjects and different grades (California Standardized Testing and Reporting, 2000; Harcourt Brace Educational Measurement, 1997c). Yet we have observed multiple citations of score improvements that involve averages across both grades and subjects. For example, consider the following statement, from a press release on the *English for the Children* website, that attempts to summarize the academic progress of English learners in California:

> From 1998 to 2000, California English learners in elementary grades (2-6) ... raised their mean percentile scores by 35% in reading, 43% in mathematics, 32% in language, and 44% in spelling, with an average increase of 39% across all subjects (English for the Children, 2000, August 14).

The same site also displays tables showing mean percentile ranks across elementary grades 2-6 and across multiple subjects. Even prior to computing percentage improvements, consider the layers of averages implied in these numbers:

1. LEP student scores are first aggregated to grade-level, within-school means (before release of data);
2. LEP grade-level, within-school means are averaged across grades (2 through 6) *and* schools;
3. LEP grade-level, within-school means are averaged across subjects (reading, mathematics, language, and spelling) *and* schools; and
4. LEP grade-level, within-school means are then simultaneously averaged across grades *and* subjects *and* schools.

We therefore see that "average increase of 39% across all subjects" relies on a notion of gains based on *means of means of means of means*. What do these *mean*?

These overall averages are not meaningful or defensible from a measurement perspective and, further, they may obscure important differences in means that exist across grades and subjects. Such data summaries are psychometric nightmares, and are particularly haunting when used to support arguments for educational policy that may strongly impact students' educational opportunities.

An Analysis of the California STAR Data

Motivated by the validity problems we have observed in other summaries of these data, we conducted a reanalysis of the California STAR data. Although we have argued that the Stanford-9 has significant limitations as a measure of achievement for LEP students and that these aggregated data lack information necessary to inform language program policy, it is apparent from our review of press and research reports that trends observed in these data will continue to be cited as evidence for arguments on both sides of the language policy spectrum. Here we attempt to analyze differences in score means and trends for California LEP and EP students and interpret them thoughtfully—without leaping to unwarranted inferences about language program effects. We present a comprehensive summary of means and gains for all grades and subjects tested; however, we focus our discussion on reading, language (Note 2), and mathematics scores for the elementary grades 2-6.

Methods

Data

We compiled data from three publicly available data sources. First, Stanford-9 scores were obtained from the California STAR website (California Standardized Testing and Reporting, 2000). This dataset provided within-school grade-level means on subtests of the Stanford-9 for reading, mathematics, and language for grades 2 to 11, spelling for grades 2 through 8, and science and social studies for grades 9 through 11. In addition to the within-school grade-level means, STAR also reports subgroup means for EP and LEP students. However, as noted previously, data were not reported for groups of less than 10 students for reasons of confidentiality. For example, the grade-level aggregate scores for the LEP subgroup were not included in the data report if a grade had less than 10 LEP students. Additionally, we obtained supplemental demographic information from the language census website (California Language Census Data Files, 2000) and from the academic performance index data website (California Academic Performance Index Data Files, 2000).

Statistical Procedures

The outcome scores used in these analyses were in the form of subject-area scaled scores. Recall that scaled scores are academic proficiency estimates that can be compared across time and across different levels of a subject-area test. Weighted means were computed for each subject area and grade level for three groups. With weighted means, schools with more students are weighted more heavily than schools with fewer students in computing the overall mean, providing a closer approximation to student-level mean. The three groups for which means were computed were: all

students, LEP students, and EP students. Schools reporting overall scores were used in computing weighted means for the group of all students. For LEP and EP subgroup means, we included all schools that reported subgroup means for *both* LEP and EP students. In 1998, however, the STAR dataset did not include aggregate scores for EP students separately, so we were unable to compare these groups in 1998.

In order to determine changes in scores across years, we computed both within-grade changes and cohort gains. First, within-grade changes were computed by subtracting within-school grade-level means from one year to the next for a single grade; an example is the difference between 4th-grade reading scores in 1998 from 4th-grade reading scores in 1999. The weighted means of these within-grade changes were then computed for each grade level in each subject area. Second, cohort gains were computed by subtracting within-school grade-level scores from one year to the next in consecutive grades; an example is subtracting 3rd-grade reading scores in 1998 from 4th-grade reading scores in 1999. We regard this as a *loose* cohort because we do not have evidence regarding which students remained in the same school from one year to the next. Again, the weighted mean of these gains was computed in each subject area.

Results

Descriptive Statistics

In order to examine how grade level means change for each academic subject over the years 1998, 1999, and 2000, we computed weighted means and standard deviations for each grade in each subject (see Appendix A). Our main finding from these means is that over the three-year period, scores for LEP students remain substantially below the scores for EP students in schools that reported aggregate scores for both LEP and EP students. And, with few exceptions, the gap in LEP and EP students' scores does not appear to be narrowing. We describe the score trends for 2nd through 6th grades in reading, mathematics, and language in greater depth in the following section. We summarize mean score differences by examining within-grade changes, followed by cohort gains.

Within-grade changes

Reading. Within-grade changes for each grade in reading across consecutive years are shown in Table 4 for grades 2-6 and in Appendix B for all grades. Figure 1 shows that for 2nd grade, all student groups improved substantially from 1998 to 2000. For example, from 1999 to 2000, 2nd-grade LEP students gained an average of 4.20 scaled score points, EP students gained an average of 4.40 scaled score points, and all students gained an average of 5.39 scaled score points. (Recall that the group termed *all* students consists of a larger number of schools; LEP and EP means are based on schools reporting scores in both of these subgroups.) It is also interesting to note that although the overall means increased, gains varied considerably among schools, and not all schools experienced improvement. For example, from 1999 to 2000, 27.2% of schools experienced declines in mean second-grade reading for LEP students and 28.6% experienced declines for EP students.

Table 4
Weighted Mean Within-Grade Gains in

Reading for Grades 2-6

		1998-1999		1999-2000			1998-2000	
Grade		**LEP**	**ALL**	**LEP**	**EP**	**ALL**	**LEP**	**ALL**
2	M	7.84	5.88	4.20	4.40	5.39	12.90	11.21
	SD	9.82	8.89	10.54	9.67	8.31	10.88	9.66
3	M	8.30	5.02	3.12	4.72	4.57	11.58	9.60
	SD	10.87	8.36	10.98	9.41	7.80	10.24	8.71
4	M	6.47	2.69	2.16	3.58	3.79	7.60	6.48
	SD	10.27	8.02	10.07	8.69	7.48	9.65	8.44
5	M	3.81	1.65	1.27	1.88	1.83	4.54	3.44
	SD	7.68	7.09	7.92	7.75	6.90	8.88	7.56
6	M	3.56	2.12	1.96	1.97	1.51	4.18	3.68
	SD	8.01	5.93	6.84	6.52	5.52	7.35	6.23

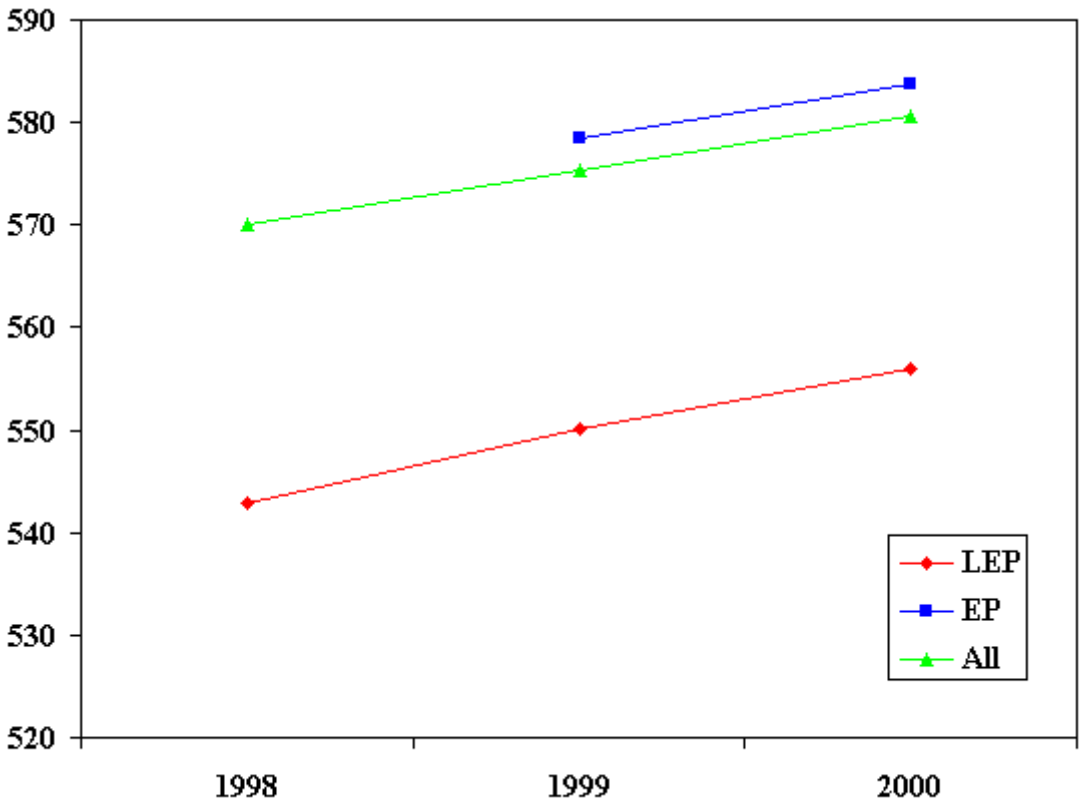

Figure 1. Mean within-grade changes for 2nd-grade reading.

A comparison of reading scores for LEP and EP students from 1999 to 2000 (EP students were not reported separately in 1998) across grades 2-6 indicated that EP students made slightly larger gains than LEP students in all 5 grades. In no instance,

however, was the difference in gains more than two scaled score points. This pattern is illustrated in graphs of reading means for grades 1 and 4 (see Figures 1 and 2, respectively), which show nearly parallel lines for LEP, EP, and all students.

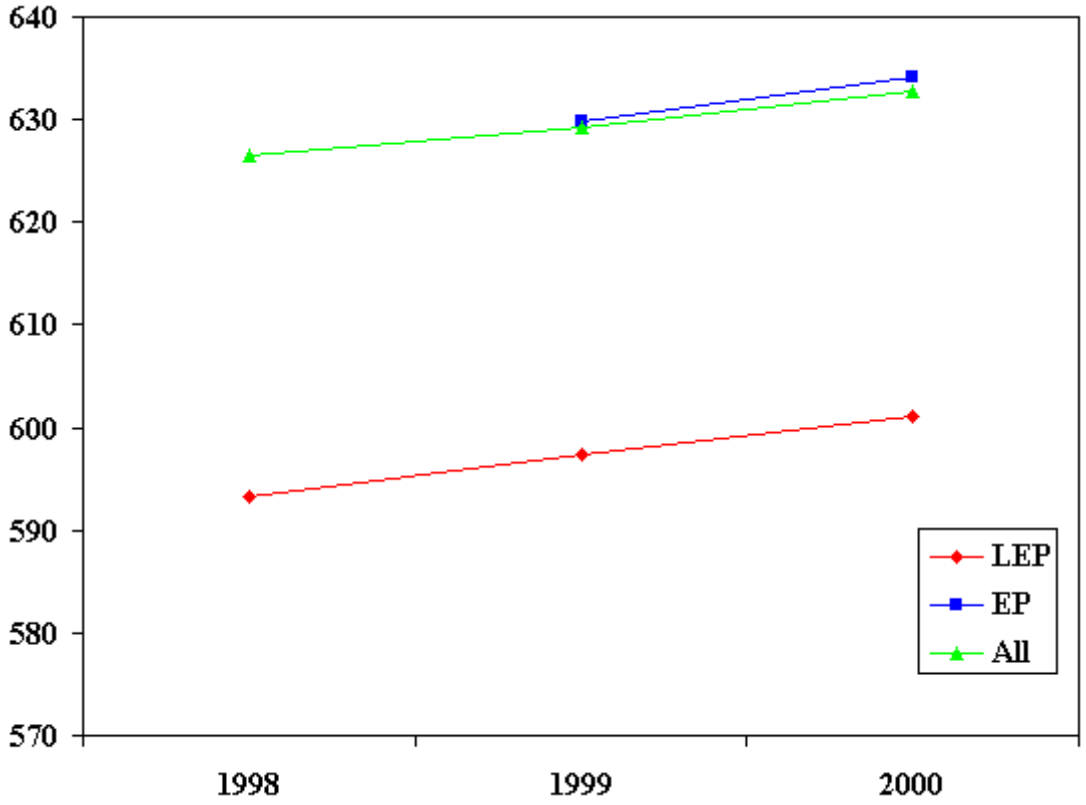

Figure 2. Mean within-grade changes for 4th-grade reading.

Language. Within-grade changes in language were similar to those in reading (see Appendix B). Students in each group displayed gains in scores from 1998 to 1999, 1999 to 2000, and 1998 to 2000. A comparison of LEP and EP students from 1999 to 2000 again revealed that EP students made slightly larger gains than LEP students in grades 2-6, although the overall improvement was never greater than three scaled score points.

Mathematics. An examination of the within-grade mean increases in mathematics scores revealed that LEP students again improved slightly less than EP students in all grades from 1999 to 2000, as reported in Table 5 for grades 2-6 and Appendix B for all grades. For example, in 2nd grade, the mean change for LEP students was 6.91 scaled score points and the mean change for EP students was 7.63 scaled score points. In 4th grade, LEP students had an average increase of 4.95 scaled score points, while EP students had an average increase of 7.33 points. Figures 3 and 4 illustrate within-grade improvements for 2nd and 4th grades, respectively. A visual inspection suggests that increases were similar across groups; however, the cumulative effect of greater improvements for EP students over multiple years makes the trend worth noting. We again caution that these within-grade changes are average score improvements across years based on different groups of students.

Table 5
Weighted Mean Within-Grade Gains in Mathematics for Grades 2 through 6

		1998-1999		1999-2000			1998-2000	
Grade		**LEP**	**ALL**	**LEP**	**EP**	**ALL**	**LEP**	**ALL**
2	M	10.08	7.72	6.91	7.63	7.29	14.37	15.04
	SD	12.82	10.02	14.14	12.01	10.18	13.47	11.66
3	M	11.82	8.17	7.35	9.60	8.38	16.43	16.56
	SD	13.23	9.87	12.92	10.59	9.38	13.08	11.03
4	M	7.84	4.88	4.95	7.33	6.91	11.30	11.83
	SD	10.33	8.71	11.39	9.51	8.35	11.08	9.69
5	M	4.90	3.78	4.49	5.68	5.34	8.44	9.09
	SD	9.10	8.31	10.03	8.84	8.18	10.37	9.39
6	M	4.66	4.30	3.67	4.89	4.11	7.55	8.49
	SD	10.06	7.54	9.66	8.61	7.51	9.51	8.51

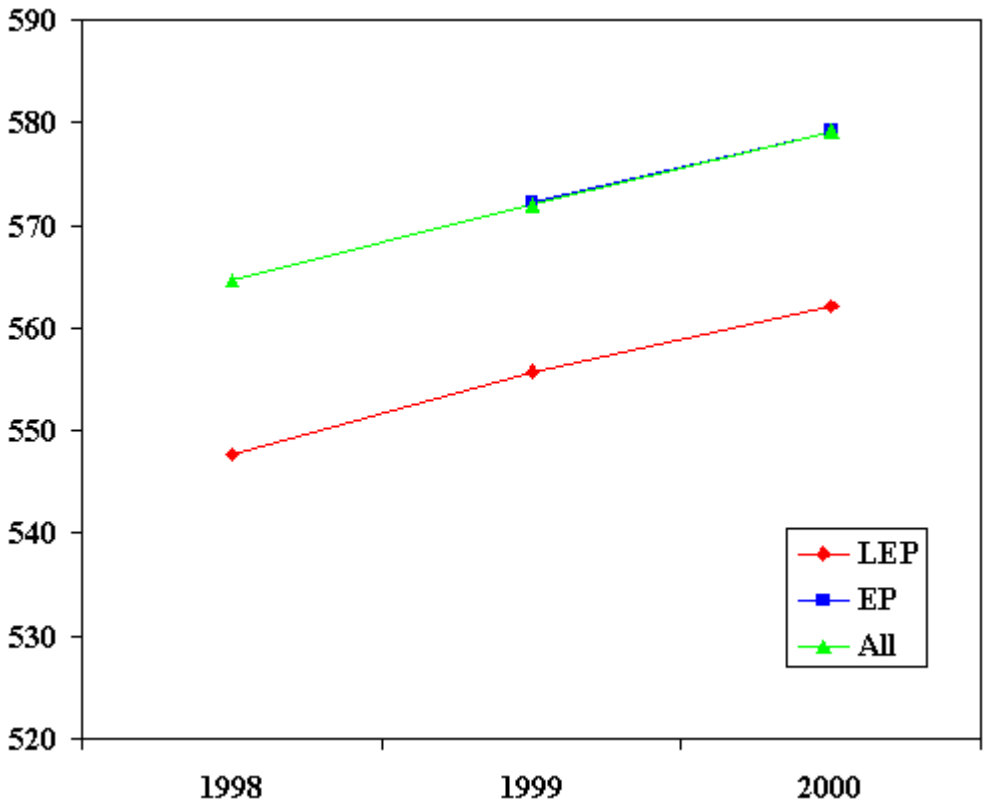

Figure 3. Mean within-grade changes for 2nd-grade mathematics.

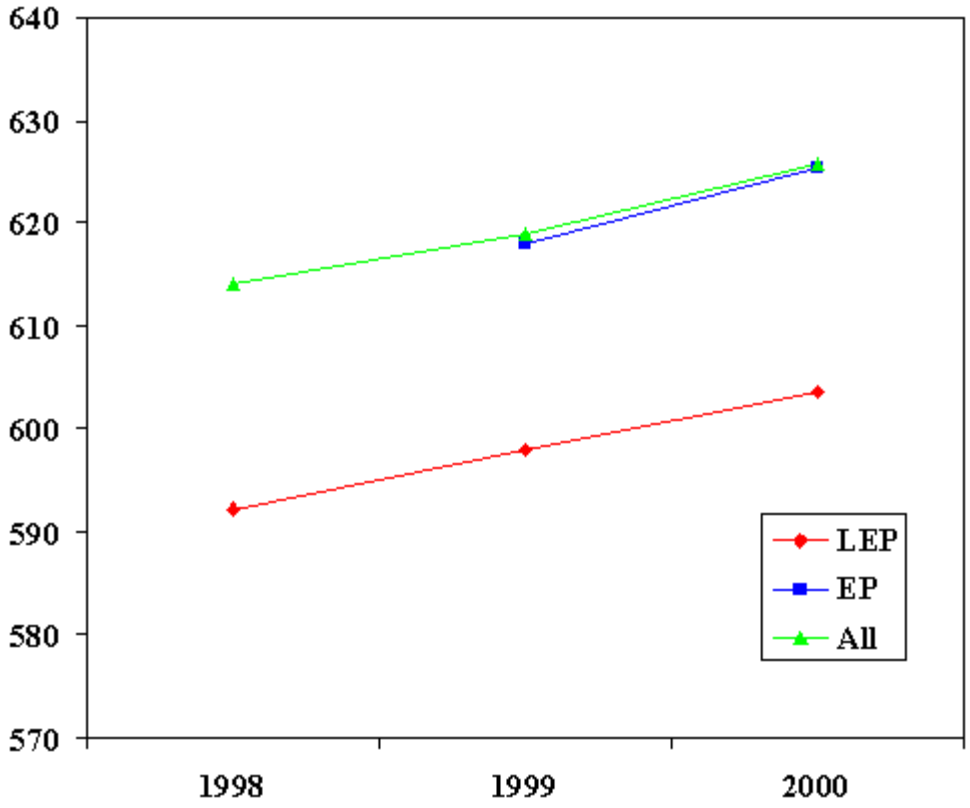

Figure 4. Mean within-grade changes for 4th-grade mathematics.

Cohort gains

Reading. We examined the gains made by cohorts of students across the three years of the test (see Table 6). Figures 5 and 6 show cohort gains for grades 2-4 and grades 4-6, respectively, in reading. Both LEP and EP cohorts improved substantially from 1999 to 2000; however, there was not a clear pattern with respect to which group gained more. From 2nd to 3rd grade, LEP students gained less (28.70 scaled score points) than EP students (34.21). The two groups' gains were similar to each other from 3rd to 4th grade and from 4th to 5th grade, while the LEP students gained more than EP students from 5th to 6th grade. Another interesting trend is that for all groups, cohort gains across grades are much greater in early elementary grades (2-4) than upper elementary grades (4-6).

Table 6
Weighted Mean Cohort Gains in Reading for Grades 2 through 6

		1998-1999		1999-2000				1998-2000	
Grades		**LEP**	**ALL**	**LEP**	**EP**	**ALL**	**Grades**	**LEP**	**ALL**
2 to 3	M	30.59	34.00	28.7	34.21	32.66	2 to 4	58.33	62.19
	SD	9.51	8.48	9.83	9.36	8.61		11.01	9.20

3 to 4	M	30.27	29.54	27.87	27.76	28.35	3 to 5	47.99	47.05
	SD	8.89	7.28	8.97	7.85	7.11		10.06	8.93
4 to 5	M	19.51	18.31	17.91	16.75	17.53	4 to 6	38.75	33.53
	SD	8.43	7.06	8.35	7.49	6.79		10.94	9.77
5 to 6	M	19.75	15.72	18.88	14.91	15.92			
	SD	7.41	7.24	8.23	7.27	7.00			

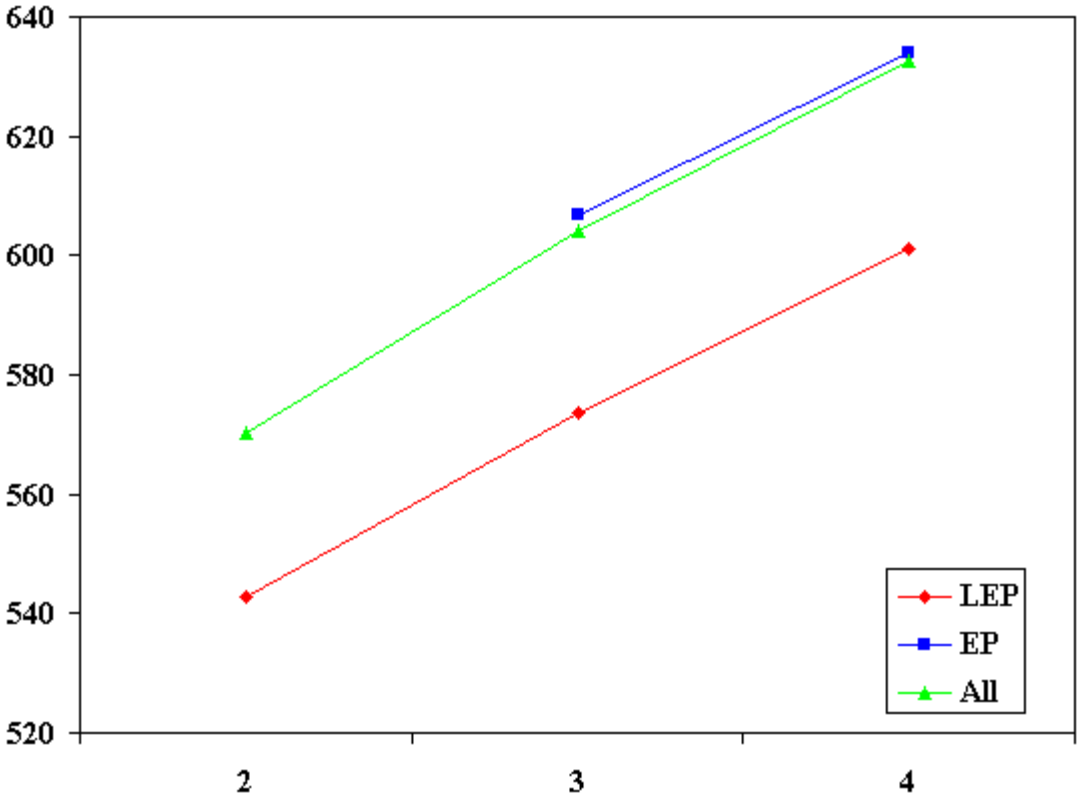

Figure 5. Mean cohort gains for 2nd through 4th graders in reading.

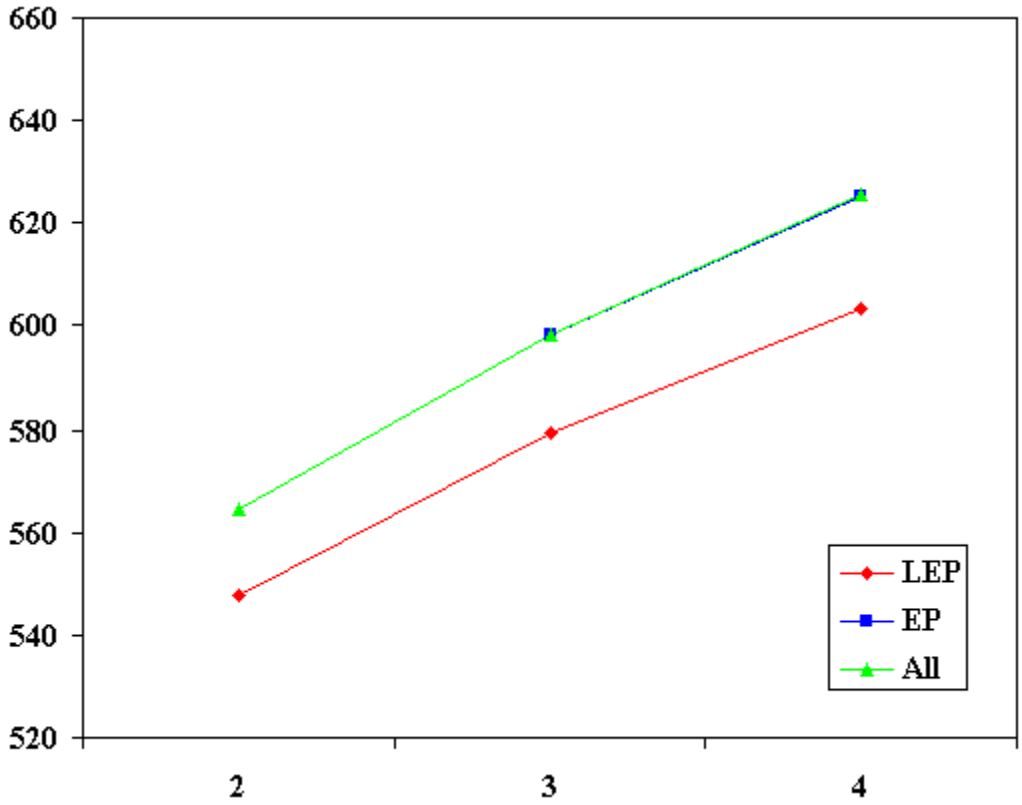

Figure 6. Mean cohort gains for 4th through 6th graders in reading.

Language. The cohort gains in language were similar to those in reading (see Appendix C). There was not a consistent pattern of gains for LEP, EP, and all students. From 1999 to 2000, EP students gained more (22.48 scaled score points) than LEP students (21.02) from 2nd to 3rd grade, although not by much. The two groups gained similarly across the other grade ranges, with EP students gaining slightly more then LEP students from 4th to 5th grade. LEP students gained slightly more than EP students from 3rd to 4th grade and 5th to 6th grade.

Mathematics. The cohort gains for mathematics, displayed in Table 7, reveal a pattern consistent with that seen in the within-group changes. From 1999 to 2000, LEP students gained less than EP students in every cohort. For example, from 2nd to 3rd grade, LEP students gained an average of 31.76 scaled score points, while EP students gained an average of 35.30 scaled score points. These patterns are suggested in Figures 7 and 8 as well, as the LEP line diverges slightly from both the EP and all lines.

Table 7
Weighted Mean Cohort Gains in Mathematics for Grades 2 through 6

		1998-1999		1999-2000				1998-2000	
Grades		**LEP**	**ALL**	**LEP**	**EP**	**ALL**	**Grades**	**LEP**	**ALL**
2 to 3	M	31.61	33.74	31.76	35.30	34.37	2 to 4	55.96	60.70
	SD	12.23	11.03	12.61	12.02	11.20		13.21	11.64

3 to 4	M	26.81	28.30	24.38	27.49	27.07	3 to 5	52.83	57.00
	SD	11.02	9.65	11.12	10.74	9.76		12.85	11.17
4 to 5	M	27.13	28.26	26.07	28.91	28.73	4 to 6	50.46	53.54
	SD	9.43	8.16	9.53	9.03	8.08		12.65	10.95
5 to 6	M	22.79	24.58	22.60	25.67	25.43			
	SD	9.36	8.96	9.64	9.32	8.80			

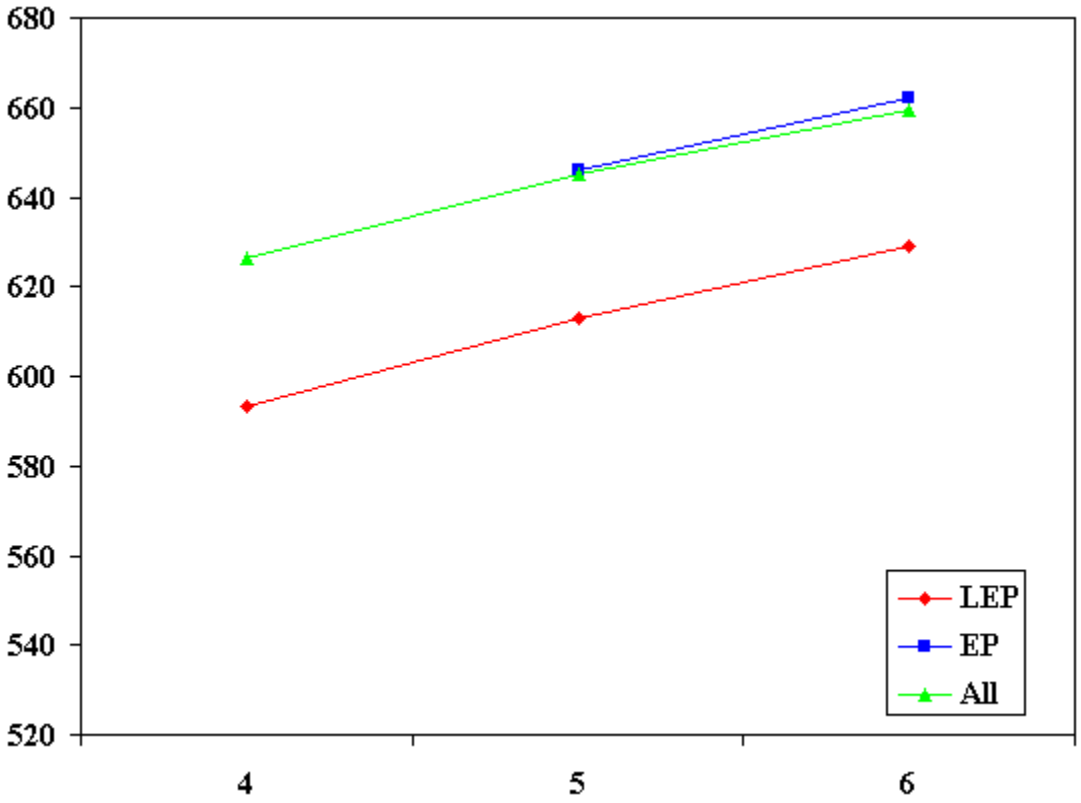

Figure 7. Mean cohort gains for 2nd through 4th graders in mathematics.

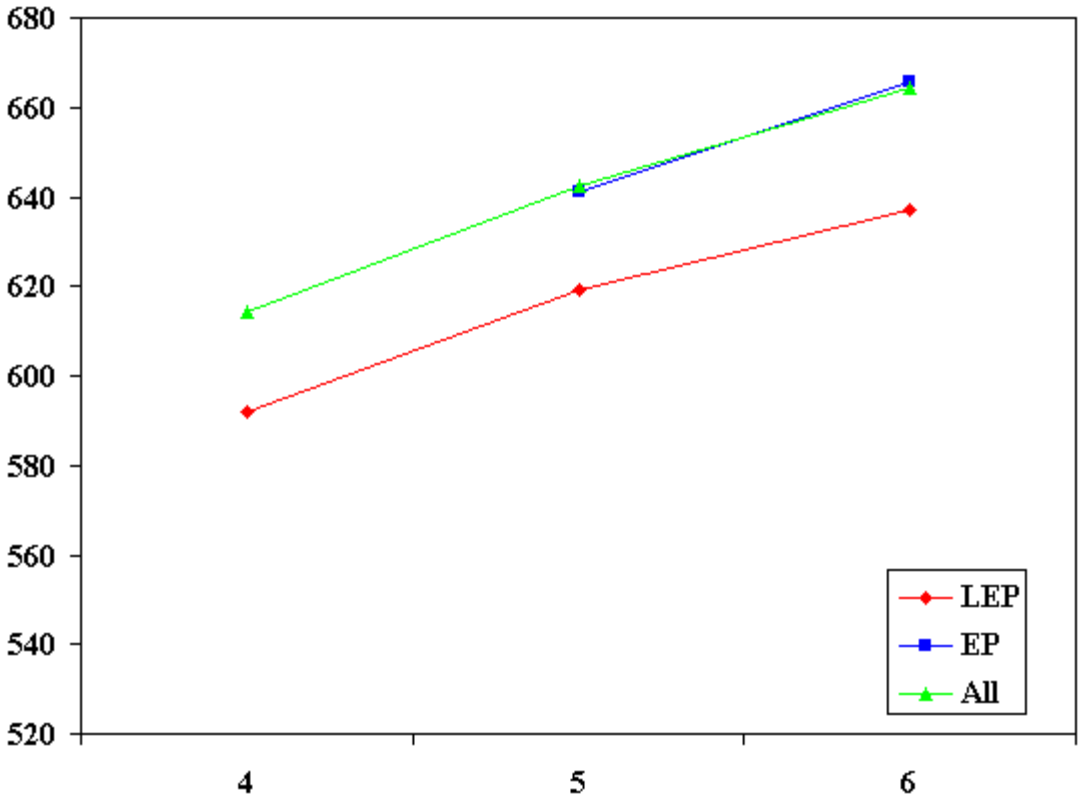

Figure 8. Mean cohort gains for 4th through 6th graders in mathematics.

Comparison of Weighted and Unweighted Means

To determine whether the use of weighted means yields substantially different results than unweighted means, we computed a limited number of unweighted means. The unweighted scaled score means for reading, language, and mathematics for grades 2 through 4 are displayed in Appendix D. The results for reading are quite interesting. The unweighted means for the LEP students are higher than the weighted means for all three grades in 1999 and 2000. In contrast, unweighted means for the EP students are lower than the weighted means for all three grades in these years. This indicates that when using unweighted means to summarize reading scores, the gap between EP and LEP students appears to be less than it is when using weighted means to estimate the average score at the student level. In other words, the gap in reading scores between LEP and EP students is wider when taking into account the number of students in each subgroup for a grade-level within a school. This phenomenon is also present in the language scores, but to a lesser extent. For mathematics, the unweighted means were consistently higher for both LEP and EP groups.

Discussion

Our concern for basing educational policy on valid evidence of academic success motivated this commentary and analysis. We first sought to provide a summary and validity critique of writings citing Stanford-9 scores in arguments regarding the success of Proposition 227 in California. Multiple issues have threatened the validity of inferences based on the California data concerning LEP students: testing LEP students in English, failing to consider myriad alternative explanations for score trends, and generalizing from a limited and nonrandom sample of schools. In addition, we have observed errors in

quantitatively summarizing this large dataset of standardized scores. In the context of the California data, the misuse and misinterpretation of percentile rank scores, the inappropriate averaging of data across years and grades, and the failure to consider the unit of analysis when using aggregated data are common problems. As validity arguments must include judgments about the appropriateness of interpretations made from test scores (Messick, 1989), this critique should raise many doubts regarding conclusions that have previously been drawn from LEP students' scores on the Stanford-9. In addition, the topics discussed in this paper generalize readily to other applications in which large standardized datasets are cited in educational policy debates.

Our analysis of the STAR dataset differed in three important ways from previous summaries of these data: a) we used scaled scores to assess academic gain across years; b) we computed weighted means to account for the number of students represented by an aggregate score; and c) we were modest with respect to the meaning our results hold for informing language program policy. As previously reported (e.g., Butler et al., 2000) means improved for both LEP and EP students over the three-year period. Our examination of weighted means revealed that from 1998 to 2000, scores for LEP students remained substantially below the scores for EP students in schools that reported aggregate scores for both LEP and EP students, and that with few exceptions this gap is not narrowing.

A within-grade comparison of reading scores for LEP and EP students across grades 2-6 indicated that EP students made slightly larger gains than LEP students in all grades. Loose cohort gains for LEP and EP students in reading were similar; however, for some grade intervals LEP students gained slightly more, while for other intervals EP students gained slightly more. The results for language arts subtest scores were similar to those in reading. In mathematics, an examination of the within-grade mean gain scores revealed that LEP students gained slightly less than EP students in all grades from 1999 to 2000. For 1999-2000 cohort gains, LEP students gained less than EP students in every cohort.

Because it was impossible to follow an individual student's growth across multiple years, the comparison of groups from year to year undoubtedly involved the comparison of different students. This problem was complicated by redesignation of students from LEP status to EP status. Not only was it unclear how many students were redesignated, but redesignation criteria differed across districts. Finally, there are many factors that have been repeatedly shown to influence student achievement. For example, LEP students may differ, on average, from their EP peers with respect to socioeconomic status and mobility, but there was no way to control for such differences using these data.

These findings should be regarded as descriptive summaries of the California STAR data, and we caution that these must be interpreted in light of the substantial limitations of these data for research purposes. *Whatever the score differences between LEP and EP students, judgments of the effects of language program policy on LEP student achievement are not warranted by these data.* To further address the question of performance differences between language programs on a large scale, we attempted to use language census data (California Language Census Data Files, 2000) and academic performance index (API; California Academic Performance Index Data Files, 2000) data to tie schools to specific program types. Most schools reported having students in nearly every program type. Because we could not identify individual students, we could not parse the data from schools into program type. We then attempted to compare schools that reported 100% of their LEP students in bilingual programs in 1998, 1999, and 2000 with schools that

reported 100% of their LEP students in English immersion programs in those years; however, this resulted in such a drastic reduction in data that we did not feel quantitative comparisons were warranted (only six schools reported 100% of LEP students in bilingual programs over the three years).

Further investigation is also needed to explore the differences in score trends observed when using unweighted versus weighted means, most notably the underestimation of EP and LEP mean differences with unweighted means. Factors associated with school size may offer meaning to these patterns. We attempted to use API data to investigate the relationship between score trends and mobility, socioeconomic status, and class size. However, due to the aggregated nature of these data, we were only able to reach very general and well-known conclusions, such as that schools with lower average SES tended to have lower test scores.

The evaluation of policy outcomes is a high-stakes activity requiring more thoughtful and detailed analyses than computing overall group NPR means. In the context of school accountability systems, Camilli and Bulkley (2001) summarized that tying accountability to single achievement outcomes does not automatically shed light on *why* certain changes were noted. They also argued that appropriate and informative use of statistical models for evaluating policy outcomes requires appreciable technical sophistication. We concur with these notions and find them relevant for our context of evaluating language program effects for LEP students. There is a strong need for research that is well-planned and well-executed that seeks to evaluate language program effects with better controls.

We offer several conclusions based on our validity critique and analysis of the Stanford-9 data. First, the scores of LEP students are not catching up to those of their English-proficient peers in any consistent manner across grades and subjects. Second, the success or failure of programs to remedy the disparity between LEP and EP students should be judged by means other than a single academic achievement test administered in English. The construct of language on an achievement test is qualitatively different from language proficiency as measured on an assessment of English as a second language. Using test scores for any purpose requires that we consider the appropriateness of the scores for the intended use and provide evidence to justify this use. In all assessments, not only should the psychometric validity of the tests be considered, but the potential consequences of the test's use must also be judged. Given the changing demographics of the United States, educators, researchers, and policymakers must join forces to establish policy that will provide maximal opportunity for LEP students to learn.

Notes

[1]We conducted a full-text search of the NEXIS/LEXIS Academic Universe archive of major U.S. newspapers using the search terms "bilingual education, test scores, California." "Major U.S. newspapers" are defined by NEXIS/LEXIS as U.S. newspapers listed in the top 50 in circulation in *Editor & Publisher Year Book*. In a manual inspection of the results, we excluded any publication that did not mention the increase in Stanford-9 test scores in relation to Proposition 227 and also included a *Newsday* article (Willen & Kowal, 2000, p. A10) that refers to the event as "a recent California study."

[2]The "language" subtest of the Stanford-9 measures comprehensive language arts proficiency and is intended for use with English-proficient students; therefore it should not

be regarded as an assessment of English language proficiency.

References

Alamillo, L., & Viramontes, C. (2000). Reflections from the classroom: Teacher perspectives on the implementation of Proposition 227. *Bilingual Research Journal, 24*, 155-168.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

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

Anastasi, A. (1981). Coaching, test sophistication, and developed abilities. *American Psychologist, 36*, 1086-1093.

August, D., & Hakuta, K. (Eds.). (1998). *Educating language-minority children.* Washington, DC: National Academy Press.

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C. C. (1983). Effects of coaching programs on achievement test performance. *Review of Educational Research, 53*, 571-585.

Berliner, D. C. (1988). Meta-comment: A discussion of critiques of L. M. Dunn's monograph *Bilingual Hispanic Children on the U.S. Mainland*. *Hispanic Journal of Behavioral Sciences, 10*, 273-300.

Berliner, D., & Biddle, B. (1995). *The manufactured crisis: Myths, fraud, and the attack on America's schools.* Reading, MA: Addison-Wesley.

Bond, L. (1989). The effects of special preparation on measures of scholastic ability. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.) (pp. 13-103). New York: Macmillan Publishing Company.

Butler, Y. G., Orr, J. E., Gutierrez, M. B., & Hakuta, K. (2000). Inadequate conclusions from an inadequate assessment: What can SAT-9 scores tell us about the impact of Proposition 227 in California. *Bilingual Research Journal, 24*, 141-154.

California Academic Performance Index Data Files. (2000). Sacramento, CA: California Department of Education. [On-line database, 8/15/2000] Available: http://api.cde.ca.gov/datafiles.html

California Education Code (2001). *English language education for immigrant children*, Division 1, Part 1, Chapter 3. [On-line]. Available: http://www.leginfo.ca.gov/calaw.html

California Language Census Data Files. (2000). Sacramento, CA: California Department of Education. [On-line database, 8/15/2000] Available: http://www.cde.ca.gov/demographics/files/census.htm

California Standardized Testing and Reporting (2000). Sacramento, CA: California

Department of Education. [On-line database, 8/15/2000] Available: http://star.cde.ca.gov

California State Board of Education. (2000). *State Monetary Awards Programs Based on the Academic Performance Index.* [On-line] Available: http://www.cde.ca.gov/ope/ae/pages/certstaffact.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: http://www.bilingualeducation.org/news.htm

Camilli, G., & Bulkley, K. (2001, March 4). Critique of "An evaluation of the Florida A-Plus Accountability and School Choice Program." *Educational Policy Analysis Archives, 9* [On-line journal] Available: http://epaa.asu.edu/

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

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Campbell, P. R. (1994). *Population projections for states by age, race, and sex: 1993 to 2020* (U.S. Bureau of the Census - Current population reports, pp. 17-23). Washington, DC: Government Printing Office.

Crawford, J. (1999). *Bilingual education: History, politics, theory, and practice.* Los Angeles: Bilingual Education Services.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, FL: Holt, Rinehart, and Wilson.

Cronbach, L. J. (1960). *Essentials of psychological testing (2nd ed.)*. New York: Harper.

de Cos, P. (1999). *Educating California's immigrant children: Overview of bilingual education* (CRB-99-010). California Research Bureau.

Duran, R. P. (1989). Testing of linguistic minorities. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13-103). New York: Macmillan Publishing Company.

Eastin (2000). *Eastin releases additional STAR 2000 test results*. [On-line] Available: http://www.cde.ca.gov/statetests/star/pressrelease2000b.pdf

English for the Children. (2000, August 14). *After two years of Prop. 227 English immersion, a huge rise in California's immigrant scores; Pro-bilingual education districts lag behind.* [On-line]. Available: http://www.yeson227.org/releases.html

Gándera, P. (2000). In the aftermath of the storm: English learners in the post-227 era. *Bilingual Research Journal, 24*, 1-14.

Gándera, P., Maxwell-Jolly, J., Garcia, E., Asato, J., Gutierrez, K., Stritikus, T., & Curry, J. (2000, April). The initial impact of Proposition 227 on the instruction of English learners. [On-line]. Available: http://lmri.ucsb.edu/RESDISS/prop227effects.pdf.

Garcia, E. E., & Curry-Rodriguez, J. E. (2000). The education of Limited English Proficient students in California schools: An assessment of the influence of Proposition 227 in selected districts and schools. *Bilingual Research Journal, 24*, 15-36.

Genesee, F. (1984). On Cummins' theoretical framework. In C. Rivera (Ed.) *Language proficiency and academic achievement*. (pp. 20-27). Clevedon, UK: Multilingual Matters.

Glass, G. V, & Hopkins, K. D. (1996). *Statistical methods in education and psychology (3rd ed.).* Boston: Allyn & Bacon.

Gonzalez, D. (2000, November 8). Bilingual education gets rebuke from state voters. *Arizona Republic*, p. EX1.

Greene, J. P. (2001). *An evaluation of the Florida A-Plus Accountability and School Choice Program.* New York: The Manhattan Institute.

Groves, M. (2000, August 15). English skills still the key in **test scores;** Stanford 9: Results show vast divide in achievement between students who have language fluency and those who don't. *Los Angeles Times*, p. A3.

Gutiérrez, K. D., Asato, J., & Baquedano-Lopez, P. (2000). "English for the Children": The new literacy of the old world order, language policy, and educational reform. *Bilingual Research Journal, 24*, 87-112.

Hakuta, K. (2001). Silence from Oceanside and the future of bilingual education. Stanford University manuscript. Available at http://www.stanford.edu/~hakuta/SAT9/Silence%20from%20Oceanside.htm.

Harcourt Brace Educational Measurement. (1997a). *Sample questions for the Stanford Achievement Test, Ninth Edition*. San Antonio, TX: Author. [On-line]. Available: http://www.cde.ca.gov/statetests/star/stanford9.pdf

Harcourt Brace Educational Measurement. (1997b). *Spring norms book***.** San Antonio, TX: Author.

Harcourt Brace Educational Measurement. (1997c). *Technical data report***.** San Antonio, TX: Author.

Hayes, K., & Salazar, J. (2001). *Evaluation of the Structured English Immersion Program Final Report: Year I.* Language Acquisition and Literacy, Program Evaluation and Research Branch, Los Angeles Unified School District.

Juffs, A., & Harrington, M. (1995). Parsing effects in second language sentence processing: Subject and object asymmetries in wh-extraction. *Studies in Second Language Acquisition*, 17(4), 483-516.

Juffs, A., & Harrington, M. (1996). Garden path sentences and error data in second language sentence processing. *Language Learning*, 46(2), 283-326.

Kerr, J. (2000, August 15). English learners show gains. *Associated Press*.

Krashen, S. (1996). *Under attack: The case against bilingual education.* Culver City, CA: Language Education Associates.

Linn, R., Graue, E., & Sanders, N. (1990). Comparing state and district test results to national norms: The validity of claims that "everyone is above average". *Educational Measurement: Issues and Practice,* 5-14.

Macías, R. F., Nishikawa, S., & Venegas, J. (1998). *Summary report of the survey of the states' limited English proficient students and available educational programs and services, 1996-97.* Washington, DC: National Clearinghouse for Bilingual Education.

Maxwell-Jolly, J. (2000). Factors influencing implementation of mandated policy change: Proposition 227 in seven Northern California school districts. *Bilingual Research Journal, 24*, 37-56.

McNeil, L. (2000). *Contradictions of school reform: Educational costs of standardized testing*. New York: Routledge.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.) (pp. 13-103). New York: Macmillan Publishing Company.

Meyer, M., & Fienberg, S. (Eds.). (1992). *Assessing evaluation studies: The case of bilingual education strategies.* Washington, DC: National Academy Press.

Moran, C. (2000, July 23). School test gains may be illusory, critics say. *San Diego Union-Tribune*, p. A1.

Moran, C., & Spielvogel, J. (2000, August 15). Students share in test gains; however, the gap between advantaged, disadvantaged has widened, data show. *San Diego Union-Tribune*, p. B1.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28,* 338-354.

Myles, F. (1996). The acquisition of interrogatives by English learners of French: The role played by structural distance. *Jyvaskyla Cross Language Studies*, 17, 195-208.

New York Times News Service. (2000, August 20). Spanish-speaking pupils gain after cored move to English; students in California improve strikingly with end of bilingual education. *The Baltimore Sun*, P. 3A.

Pilkington, C. L., Piersel, W. C. & Ponterotto, J. G. (1988). Home language as a predictor of first-grade achievement for Anglo- and Mexican-American children. *Contemporary Educational Psychology, 13*, 1-14.

Public Schools Accountability Act. (1999). [On-line] Available: http://www.cde.ca.gov/psaa/

Ramirez, D., Pasta, D., Yuen, S., Billings, D., & Ramey, D. (1991). *Final report. Longitudinal study of structured English immersion strategy, early-exit and late-exit transitional bilingual education programs for language-miority children*. (Vols. 1 & 2).

San Mateo, CA: Aguirre International.

Rogosa, D. R. (1999). *How accurate are the STAR national percentile rank scores for individual students? An interpretive guide.* Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing [On-line]. Available: http://cresst96.cse.ucla.edu/CRESST/Reports/drrguide.html

Rosenthal, A., Milne, A., Ellman, F., Ginsburg, A., & Baker, K. (1983). A comparison of the effects of language background and socioeconomic status on achievement among elementary-school students. In K. Baker & E. de Kanter (Eds.). *Bilingual education: A reappraisal of federal policy*. Lexington, MA: Lexington Books.

Rossell, C., & Baker, K. (1996). The educational effectiveness of bilingual education. *Research in the Teaching of English, 30*, 7-74.

Sacks, P. (1999). *Standardized minds: The high price of America's testing culture and what we can do to change it.* Cambridge: Perseus Books.

Schirling, E., Contreras, F., & Ayala, C. (2000). Proposition 227: Tales from the schoolhouse. *Bilingual Research Journal, 24*, 127-140.

Simon, T. F., Fico, F., & Lacy, S. (1989). Covering conflict and controversy: Measuring balance, fairness, defamation. *Journalism Quarterly, 66,* 427-434.

Singer, E. & Endreny, P. (1993). *Reporting on risk: How the mass media portray accidents, diseases, and other hazards.* New York: Russell Sage Foundation.

Stabler, E. P. Jr. (1986). Possible contributing factors in test item difficulty: Research memorandum. Princeton, NJ: Educational Testing Service.

Stabler, E. P., Jr. (1994). The finite connectivity of linguistic structure. In C. Clifton, L. Frazier, & K. Rayner (Eds.), *Perspectives on sentence processing*. New York, NY: Erlbaum.

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

Steinberg, J. (2000b, August 20). English-only classes appear a success; after ending bilingual programs, California schools see boosts in test scores. *Milwaukee Journal Sentinel*, p. 3A.

Steinberg, J. (2000c, August 20). Rising test scores shock critics of English-only law. *The Plain Dealer,* p. 21A.

Tankard, J., & Ryan, M. (1974). News source perceptions of accuracy in science coverage. *Journalism Quarterly, 51*, 219-225, 334.

Thompson, M. S. (April, 2000). *Investigating dependency at the classroom and school levels in educational research: A multilevel approach.* Presented at the Annual Meeting of the American Educational Research Association, New Orleans, Louisiana.

Ulibarri, D. M., Spencer, M. L., & Rivas, G. A. (1981). Language proficiency and

academic achievement: A study of language proficiency tests and their relationship to school ratings as predictors of academic achievement. *NABE Journal, 5*, 47-80.

United States Department of Education. (1994). *Summary of the bilingual education state educational agency program survey of states' limited English proficient persons and available educational services (1992-1993): Final report.* Arlington, VA: Development Associates.

United States Department of Education. (1997). *1993–94 schools and staffing survey: A profile of policies and practices for limited English proficient students: Screening methods, program support, and teacher training*. Washington, D.C.: U.S. Department of Education. [On-line]. Available: http://nces.ed.gov/pubs/97472.pdf

Weiss, C., & Singer, E. (1987). *Reporting of social science in the national media.* New York: Russell Sage Foundation.

Willen, L., & Kowal, J. (2000, September 9). Study: Bilingual ed falling short. *Newsday*, p. A10.

Willig, A. (1985). A meta-analysis of selected studies on the effectiveness of bilingual education. *Review of Educational Research, 55,* 269-318.

Wright, C. R., & Michael, W. B. (1989). The relationship of average scores of seniors in 172 southern California high schools on a statewide standardized achievement test to percentages of families receiving federal aid and to percentages of limited-English proficient speakers. *Educational and Psychological Measurement, 49,* 937-944.

About the Authors

Marilyn S. Thompson
Marilyn Thompson is Assistant Professor of Measurement, Statistics, and Methodological Studies in the Division of Psychology in Education at Arizona State University. Her research interests include methodological techniques for large data set analysis, applications of structural equation modeling and multilevel modeling, and policy pertaining to the assessment of English learners. Dr. Thompson is Director of the EDCARE laboratory at ASU, providing measurement, statistics, and evaluation services to the educational community. She previously taught physics and chemistry and was science department chair in an urban high school, giving her a first-hand perspective on the role of classroom assessment. She may be reached at m.thompson@asu.edu.

Kristen E. DiCerbo
Kristen DiCerbo is a Ph.D. student in the School Psychology program in the Division of Psychology in Education, College of Education at Arizona State University. Her research interests include psychoeducational assessment, especially the assessment of minority children, and internalizing problems in children. She is currently a psychology intern in a district serving a large population of English learners. Ms. DiCerbo was formerly Editor of *Current Issues in Education*. She may be reached at kristen.dicerbo@asu.edu.

Kate Mahoney
Kate Mahoney is a Ph.D. student in Curriculum and Instruction with a concentration in bilingual education at Arizona State University. Her teaching experience includes teaching

mathematics with English language learners. Her research interests include psychometrics and validity issues concerned with testing English language learners.

Jeff MacSwan
Jeff MacSwan is an Assistant Professor of Language and Literacy in the College of Education at Arizona State University. He has published articles in *Hispanic Journal of Behavioral Sciences, Bilingual Review, Bilingual Research Journal, Bilingualism: Language and Cognition*, and *Southwest Journal of Linguistics*. He is the author of *A Minimalist Approach to Intrasentential Code Switching* (Garland, 1999), and editor of the forthcoming volume *Grammatical Theory and Bilingual Codeswitching* (MIT Press). He is chair of the organizing committee of the Fourth International Symposium on Bilingualism, to be held at Arizona State University in April, 2003. He may be reached at macswan@asu.edu.

Appendices

Appendix A
Weighted Mean Scaled Scores

			1998		1999			2000		
Subject	Grade		LEP	All	LEP	EP	All	LEP	EP	All
Reading										
	2	M	542.93	570.14	550.11	578.98	575.35	555.91	584.00	580.64
		SD	15.13	23.35	15.23	18.55	23.23	15.40	18.25	23.17
		N	2388	4843	2557	2557	3319	2741	2741	3356
	3	M	567.19	599.53	573.53	606.94	604.06	578.52	611.32	608.33
		SD	13.51	25.97	13.00	19.15	25.17	12.50	18.37	24.57
		N	2445	4874	2608	2608	4936	2788	2788	4979
	4	M	593.31	626.58	597.35	630.15	629.20	601.07	634.35	632.77
		SD	11.91	25.10	12.08	18.19	24.52	11.84	17.47	24.03
		N	2295	4837	2422	2422	4894	2606	2606	4958
	5	M	610.03	643.35	612.81	646.04	644.92	615.14	647.72	646.84
		SD	9.96	22.42	10.05	15.66	21.79	10.00	15.05	21.36
		N	2147	4798	2293	2293	4868	2436	2436	4924
	6	M	624.30	655.79	627.47	660.78	657.78	628.96	662.54	659.37
		SD	9.04	19.55	9.12	14.70	18.99	9.49	14.64	18.98
		N	1449	3395	1522	1522	3319	1646	1646	3356
	7	M	633.17	670.85	635.71	679.00	672.17	638.05	680.54	674.06
		SD	10.34	20.31	9.75	15.49	19.61	10.85	15.72	19.74

		N	885	1776	941	941	1769	997	997	1797
	8	M	649.52	684.52	651.42	692.03	686.13	653.21	693.55	687.63
		SD	9.09	18.32	8.37	13.81	17.75	9.04	14.19	17.79
		N	860	1803	919	919	1819	982	982	1851
	9	M	650.43	683.79	651.81	690.86	684.25	652.97	691.93	685.67
		SD	7.81	116.91	7.52	13.96	16.83	7.31	14.10	16.73
		N	602	1288	641	641	1288	693	693	1321
	10	M	654.50	689.40	656.25	697.05	689.85	656.93	697.56	691.05
		SD	8.65	16.74	7.81	13.84	16.50	7.88	14.29	16.70
		N	602	1440	614	614	1435	704	704	1479
	11	M	662.13	697.18	663.57	703.39	696.76	664.35	703.50	697.81
		SD	9.35	16.35	8.26	13.26	15.83	8.29	13.91	16.07
		N	571	1395	598	598	1393	661	661	1446

			1998			1999			2000	
Subject	Grade		LEP	All	LEP	EP	All	LEP	EP	All
Math										
	2	M	547.65	564.58	555.79	572.58	571.91	562.12	579.37	579.14
		SD	15.75	20.96	15.97	19.20	21.07	16.47	19.05	21.38
		N	2540	4875	2555	2557	4914	2737	2741	4969
	3	M	571.11	590.43	579.49	598.43	598.36	587.14	606.81	606.51
		SD	15.58	22.36	15.30	19.45	21.95	15.25	19.07	21.66
		N	2493	4883	2602	2608	4938	2783	2788	4980
	4	M	592.14	614.08	597.92	618.28	618.95	603.46	625.57	625.71
		SD	13.03	21.56	13.42	18.09	21.24	13.43	17.77	21.23
		N	2385	4846	2419	2420	4902	2604	2606	4959
	5	M	614.86	638.55	619.22	641.50	642.36	623.74	646.64	647.75
		SD	11.84	21.09	11.99	16.85	20.73	12.15	16.63	20.92
		N	2209	4808	2286	2292	4871	2431	2434	4927
	6	M	629.16	656.14	633.79	661.83	660.34	637.24	665.71	664.57
		SD	12.45	21.67	12.99	18.08	21.65	13.62	18.30	22.03
		N	1486	3400	1520	1522	3322	1644	1646	3361
	7	M	643.68	667.56	646.47	674.23	670.16	648.62	676.83	673.15
		SD	10.90	19.16	10.61	17.09	18.85	11.44	17.61	19.61
		N	886	1777	941	941	1772	996	997	1793

	8	M	652.85	676.41	655.18	682.99	679.36	656.96	685.46	681.71
		SD	11.70	18.74	10.83	16.65	18.80	11.43	17.09	19.10
		N	868	1803	912	917	1810	979	981	1847
	9	M	667.00	688.24	668.78	698.79	689.71	670.22	696.90	692.25
		SD	11.20	16.98	10.28	14.97	16.76	9.91	15.42	17.08
		N	617	1285	638	641	1295	690	693	1326
	10	M	677.47	694.70	679.93	701.79	696.78	680.26	702.34	698.09
		SD	12.17	15.66	11.17	14.53	15.76	10.81	15.02	15.98
		N	609	1438	609	614	1433	696	704	1478
	11	M	680.76	699.80	684.06	707.38	702.07	684.69	708.36	703.99
		SD	13.66	17.52	13.05	16.18	17.57	12.84	16.89	18.03
		N	575	1397	594	598	1392	658	661	1439
			1998			1999			2000	
Subject	Grade		LEP	All	LEP	EP	All	LEP	EP	All
Language										
	2	M	558.26	580.56	563.58	588.20	585.12	568.34	592.11	589.54
		SD	13.09	20.79	13.37	17.11	20.75	13.78	17.15	20.93
		N	2483	4868	2546	2556	4910	2730	2740	4969
	3	M	571.63	596.13	578.35	603.84	602.17	584.12	609.41	607.56
		SD	13.33	22.57	13.49	18.14	22.27	13.64	17.63	21.92
		N	2442	4873	2586	2601	4933	2771	2786	4977
	4	M	595.30	620.61	598.68	623.71	622.82	602.67	628.02	626.78
		SD	11.92	20.89	12.25	16.60	20.62	11.93	15.95	20.11
		N	2367	4839	2407	2413	4895	2596	2604	4960
	5	M	607.68	634.34	610.14	636.88	636.29	613.07	639.29	639.03
		SD	10.76	20.49	10.96	15.88	20.29	11.02	15.59	20.15
		N	2190	4805	2278	2291	4871	2424	2433	1927
	6	M	617.70	643.43	620.59	648.11	645.75	622.65	650.80	648.31
		SD	9.64	17.68	10.11	14.36	17.75	10.55	14.56	17.96
		N	1472	3394	1513	1521	3314	1634	1646	3354
	7	M	626.52	655.69	628.79	663.36	657.72	631.26	665.77	660.41
		SD	9.74	17.65	9.34	14.56	17.57	10.18	14.77	17.78
		N	883	1768	938	941	1768	991	996	1790
	8	M	632.29	661.89	633.81	669.37	664.12	635.63	671.49	666.25

		SD	9.07	17.89	8.74	14.86	18.05	9.28	15.21	18.22
		N	866	1799	913	918	1802	978	982	1843
	9	M	642.68	668.38	643.96	676.22	669.97	644.84	678.01	672.07
		SD	9.29	15.47	8.98	13.19	15.67	9.07	13.76	16.08
		N	608	1285	633	640	1280	688	693	1318
	10	M	639.13	669.03	640.67	678.03	670.48	641.48	679.24	672.59
		SD	9.64	17.34	8.64	14.64	17.43	8.95	15.26	17.67
		N	597	1431	606	613	1425	693	703	1463
	11	M	650.25	678.16	652.09	686.12	679.64	652.70	686.90	681.33
		SD	9.42	15.69	8.83	13.37	15.99	8.91	14.46	16.44
		N	565	1386	588	597	1383	658	661	1438

			1998			1999			2000	
Subject	Grade		LEP	All	LEP	EP	All	LEP	EP	All
Spelling										
	2	M	533.17	558.19	542.89	568.99	565.13	550.35	575.46	572.04
		SD	19.86	23.42	19.94	19.41	23.26	20.75	19.27	23.49
		N	2502	4872	2553	2553	4912	2739	2739	4970
	3	M	567.11	589.48	574.74	598.86	595.70	582.39	604.80	601.92
		SD	17.49	20.30	16.97	16.15	19.96	16.67	15.92	19.38
		N	2487	4883	2603	2603	4936	2785	2785	1981
	4	M	583.53	612.64	588.39	617.46	615.69	593.67	623.30	620.95
		SD	14.27	23.21	14.58	17.86	22.60	14.21	17.27	22.14
		N	2384	4849	2420	2420	4900	2604	2604	4961
	5	M	603.56	629.88	606.59	634.13	632.42	610.18	636.86	635.52
		SD	11.46	19.07	11.63	14.77	18.90	11.73	14.16	18.61
		N	2200	4810	2286	2286	4874	2430	2430	4929
	6	M	611.54	642.94	615.15	649.85	645.85	618.38	653.33	649.06
		SD	11.76	20.18	12.39	16.11	20.07	12.68	16.17	20.33
		N	1486	3400	1521	1521	3323	1645	1645	3361
	7	M	623.16	657.22	625.20	665.34	658.31	627.47	667.90	660.98
		SD	11.08	18.26	10.91	14.55	17.92	12.25	14.86	18.40
		N	887	1775	938	938	1771	995	995	1798
	8	M	641.29	668.60	642.20	675.44	670.07	643.64	677.14	671.75
		SD	8.00	14.65	7.87	12.03	14.68	8.27	12.32	14.82

		N	870	1807	916	916	1816	978	978	1848
			1998		1999			2000		
Subject	Grade		LEP	All	LEP	EP	All	LEP	EP	All
Science										
	9	M	651.24	670.30	652.79	675.05	671.21	653.62	675.96	672.39
		SD	6.60	12.30	6.05	10.35	12.00	5.95	10.47	11.98
		N	616	1281	639	640	1290	689	693	1325
	10	M	656.39	676.94	657.96	682.40	678.00	657.81	682.15	678.20
		SD	7.20	13.22	6.04	11.54	13.07	6.03	11.85	13.15
		N	606	1428	609	614	1434	691	704	1471
	11	M	659.77	681.99	661.60	687.48	683.23	662.08	687.60	683.98
		SD	7.41	13.66	6.79	12.15	12.00	6.47	12.50	13.80
		N	571	1388	593	598	1387	658	661	1436
Social Studies										
	9	M	632.42	649.34	633.57	652.97	649.61	634.83	653.88	650.91
		SD	5.09	11.28	4.54	9.69	10.89	4.01	9.54	10.65
		N	614	1279	638	641	1283	689	693	1323
	10	M	633.48	652.44	634.03	656.65	652.61	634.04	656.44	652.87
		SD	5.75	11.72	4.93	10.24	11.51	4.71	10.40	11.46
		N	604	1425	608	614	1428	693	704	1466
	11	M	645.77	665.91	647.20	671.00	666.65	647.86	670.94	667.45
		SD	6.56	12.49	5.81	10.39	12.00	5.70	10.74	12.02
		N	573	1384	590	598	1379	657	661	1434

Appendix B
Weighted Mean Within-grade Gains

Subject	Grade		1998-1999		1999-2000			1998-2000	
			LEP	ALL	LEP	EP	ALL	LEP	ALL
Reading									
	2	M	7.84	5.88	4.20	4.40	5.39	12.90	11.21
		SD	9.82	8.89	10.54	9.67	8.31	10.88	9.66
	3	M	8.30	5.02	3.12	4.72	4.57	11.58	9.60

		SD	10.87	8.36	10.98	9.41	7.80	10.24	8.71
	4	M	6.47	2.69	2.16	3.58	3.79	7.60	6.48
		SD	10.27	8.02	10.07	8.69	7.48	9.65	8.44
	5	M	3.81	1.65	1.27	1.88	1.83	4.54	3.44
		SD	7.68	7.09	7.92	7.75	6.90	8.88	7.56
	6	M	3.56	2.12	1.96	1.97	1.51	4.18	3.68
		SD	8.01	5.93	6.84	6.52	5.52	7.35	6.23
	7	M	3.64	1.37	3.94	2.46	1.96	4.26	3.30
		SD	7.44	5.10	8.65	5.61	4.77	8.39	5.45
	8	M	3.55	1.62	2.35	2.37	1.52	3.46	3.15
		SD	6.97	4.59	7.85	5.94	4.57	8.08	5.25
	9	M	1.07	0.41	1.29	1.32	1.35	2.61	1.81
		SD	6.07	4.11	5.62	4.23	3.99	6.50	4.20
	10	M	1.30	0.44	0.83	1.14	1.13	2.44	1.59
		SD	6.67	4.44	6.60	4.68	4.42	7.99	4.56
	11	M	1.52	-0.36	0.91	0.98	0.92	2.31	0.60
		SD	7.28	5.15	7.38	5.30	5.26	7.65	5.23

Subject	Grade		1998-1999		1999-2000			1998-2000	
			LEP	ALL	LEP	EP	ALL	LEP	ALL
Math									
	2	M	10.08	7.72	6.91	7.63	7.29	14.37	15.04
		SD	12.82	10.02	14.14	12.01	10.18	13.47	11.66
	3	M	11.82	8.17	7.35	9.60	8.38	16.43	16.56
		SD	13.23	9.87	12.92	10.59	9.38	13.08	11.03
	4	M	7.84	4.88	4.95	7.33	6.91	11.30	11.83
		SD	10.33	8.71	11.39	9.51	8.35	11.08	9.69
	5	M	4.90	3.78	4.49	5.68	5.34	8.44	9.09
		SD	9.10	8.31	10.03	8.84	8.18	10.37	9.39
	6	M	4.66	4.30	3.67	4.89	4.11	7.55	8.49
		SD	10.06	7.54	9.66	8.61	7.51	9.51	8.51
	7	M	3.80	2.60	2.04	3.13	3.08	4.61	5.69
		SD	7.73	5.25	7.72	6.30	5.37	7.01	5.97
	8	M	4.09	2.97	1.92	2.73	2.40	4.19	5.38
		SD	7.16	5.04	7.88	6.85	5.27	7.03	5.93

9	M	1.71	1.46	1.68	2.53	2.52	3.52	3.97
	SD	5.75	4.30	4.96	4.60	4.23	6.05	4.54
10	M	2.36	2.14	0.68	1.24	1.25	3.16	3.40
	SD	5.82	4.01	5.28	4.48	3.99	6.50	4.33
11	M	3.42	2.35	0.88	1.92	1.83	4.30	4.22
	SD	6.64	4.68	6.30	5.23	4.82	6.69	5.14

Subject	Grade		1998-1999		1999-2000			1998-2000	
			LEP	ALL	LEP	EP	ALL	LEP	ALL
Language									
	2	M	5.69	5.85	3.82	4.07	4.55	9.95	9.46
		SD	9.25	8.24	11.01	9.05	8.07	10.34	9.26
	3	M	8.34	6.42	3.54	6.02	5.62	12.88	12.00
		SD	10.92	8.32	10.58	9.77	7.96	10.59	9.02
	4	M	5.56	2.24	1.83	3.43	4.15	7.30	6.35
		SD	9.62	7.56	10.09	8.41	7.43	9.93	8.20
	5	M	3.25	1.99	2.02	2.13	2.66	4.85	4.59
		SD	8.57	7.66	8.71	8.25	7.56	9.74	8.17
	6	M	3.08	2.37	1.99	3.19	2.49	4.50	4.93
		SD	8.78	6.62	7.80	6.77	6.31	7.92	6.96
	7	M	3.48	2.04	3.11	2.75	2.76	4.24	4.80
		SD	7.50	5.15	8.82	5.84	5.02	7.66	5.58
	8	M	2.92	2.24	2.56	2.49	2.18	3.31	4.41
		SD	5.97	4.96	8.55	7.37	5.10	7.55	5.63
	9	M	1.01	1.54	1.11	2.23	2.10	2.41	3.67
		SD	5.85	4.21	5.43	4.54	4.24	6.72	4.52
	10	M	1.24	1.46	1.24	2.06	2.05	2.53	3.53
		SD	6.08	4.89	6.08	5.36	4.95	7.38	5.06
	11	M	2.00	1.56	0.61	1.68	1.58	2.62	3.21
		SD	6.30	4.67	6.59	5.22	5.03	7.01	5.09

Subject	Grade		1998-1999		1999-2000			1998-2000	
			LEP	ALL	LEP	EP	ALL	LEP	ALL
Spelling									
	2	M	9.49	7.46	7.70	5.89	6.96	16.91	14.36
		SD	12.38	9.80	11.84	25.10	9.46	13.83	10.96
	3	M	7.80	6.61	8.06	5.55	6.41	15.76	13.01
		SD	10.99	8.59	10.66	20.02	8.50	12.16	9.46
	4	M	4.44	3.12	5.68	4.91	5.48	10.14	8.60
		SD	9.68	8.45	9.91	20.82	8.17	10.43	9.14
	5	M	2.80	2.59	3.25	2.12	2.99	6.12	5.56
		SD	9.04	7.36	8.60	21.36	7.21	9.80	7.91
	6	M	3.05	2.98	2.94	2.44	3.19	6.14	6.26
		SD	8.75	7.17	8.71	24.24	6.99	9.50	7.75

	7	M	1.65	1.06	1.98	2.42	2.78	3.68	3.86
		SD	7.25	4.89	7.67	14.15	4.92	8.21	5.42
	8	M	0.83	1.44	1.23	0.97	1.70	2.20	3.14
		SD	5.91	4.49	5.99	22.11	4.57	6.99	5.25
Science									
	9	M	1.38	0.87	1.06	0.89	1.15	2.56	2.03
		SD	5.21	3.60	4.59	12.65	3.28	5.53	3.69
	10	M	1.38	1.10	0.01	-0.35	0.14	1.55	1.24
		SD	5.34	4.03	4.60	17.65	3.76	5.99	4.14
	11	M	1.86	1.33	0.46	0.20	0.65	2.34	2.02
		SD	5.25	4.42	5.10	16.08	4.39	5.65	4.56
Social Studies									
	9	M	1.08	0.22	1.46	0.93	1.26	2.58	1.49
		SD	4.53	3.10	3.77	12.45	2.97	4.36	3.20
	10	M	0.35	0.14	0.09	-0.19	0.23	0.61	0.37
		SD	4.56	3.43	4.05	15.78	3.22	5.09	3.43
	11	M	1.58	0.77	0.78	0.39	0.71	2.16	1.52
		SD	5.46	4.21	5.30	14.81	3.95	5.57	4.26

Appendix C
Weighted Mean Cohort Gains

Subject	Grade		1998-1999		1999-2000			Grades	1998-2000	
			LEP	ALL	LEP	EP	ALL		LEP	ALL
Reading										
	2 to 3	M	30.59	34.00	28.7	34.21	32.66	2 to 4	58.33	62.19
		SD	9.51	8.48	9.83	9.36	8.61		11.01	9.20
	3 to 4	M	30.27	29.54	27.87	27.76	28.35	3 to 5	47.99	47.05
		SD	8.89	7.28	8.97	7.85	7.11		10.06	8.93
	4 to 5	M	19.51	18.31	17.91	16.75	17.53	4 to 6	38.75	33.53
		SD	8.43	7.06	8.35	7.49	6.79		10.94	9.77
	5 to 6	M	19.75	15.72	18.88	14.91	15.92			
		SD	7.41	7.24	8.23	7.27	7.00			
Math										
	2 to 3	M	31.61	33.74	31.76	35.30	34.37	2 to 4	55.96	60.70
		SD	12.23	11.03	12.61	12.02	11.20		13.21	11.64
	3 to 4	M	26.81	28.30	24.38	27.49	27.07	3 to 5	52.83	57.00
		SD	11.02	9.65	11.12	10.74	9.76		12.85	11.17
	4 to 5	M	27.13	28.26	26.07	28.91	28.73	4 to 6	50.46	53.54
		SD	9.43	8.16	9.53	9.03	8.08		12.65	10.95
	5 to 6	M	22.79	24.58	22.60	25.67	25.43			

		SD	9.36	8.96	9.64	9.32	8.80			
Spelling										
	2 to 3	M	41.78	37.55	40.24	35.15	36.66	2 to 4	61.12	62.30
		SD	11.00	9.19	10.68	9.88	9.17		12.79	9.55
	3 to 4	M	21.57	26.03	19.51	26.15	24.82	3 to 5	43.41	45.79
		SD	10.08	8.04	10.13	8.91	8.21		11.73	8.38
	4 to 5	M	23.27	19.82	22.04	18.34	19.77	4 to 6	37.95	37.41
		SD	8.70	7.91	9.09	8.46	7.72		11.58	9.49
	5 to 6	M	13.88	17.32	13.88	18.61	18.04			
		SD	8.89	7.47	9.28	7.96	7.54			
Language										
	2 to 3	M	20.06	21.70	21.02	22.48	22.29	2 to 4	44.52	45.78
		SD	9.40	8.53	9.64	9.64	8.52		10.63	9.00
	3 to 4	M	27.13	26.50	24.71	23.62	24.30	3 to 5	41.50	42.56
		SD	9.21	7.92	9.35	8.61	7.77		10.98	9.03
	4 to 5	M	14.83	15.69	14.54	15.84	16.14	4 to 6	31.03	29.07
		SD	8.73	6.82	8.62	7.56	6.74		10.67	8.85
	5 to 6	M	15.37	12.84	15.01	12.89	13.72			
		SD	8.08	7.70	8.56	8.11	7.58			

Appendix D
Unweighted Mean Scaled Scores

			1998		1999			2000		
Subject	Grade		LEP	All	LEP	EP	All	LEP	EP	All
Reading										
	2	M	547.69	572.91	554.73	570.92	578.67	560.34	578.08	583.94
		SD	17.95	23.06	17.70	19.18	22.82	17.47	19.08	22.65
		N	2388	4843	2557	2557	3319	2741	2741	3356
	3	M	571.02	602.21	577.13	596.33	607.21	581.95	605.26	611.58
		SD	15.33	25.72	15.24	19.47	24.87	14.44	18.95	24.2
		N	2445	4874	2608	2608	4936	2788	2788	4979
	4	M	596.17	628.96	600.18	616.57	631.78	604.38	624.12	635.56
		SD	13.98	24.82	13.71	18.02	24.31	13.74	17.46	23.72
		N	2295	4837	2422	2422	4894	2606	2606	4958
Math										
	2	M	550.59	566.41	558.44	576.57	573.95	565.5	582.29	581.36

		SD	18.48	21.18	18.75	18.52	21.26	18.98	18.1	21.31
		N	2540	4875	2555	2557	4914	2737	2741	4969
	3	M	574.02	591.96	582.39	604.47	600.25	590.49	609.41	608.48
		SD	17.78	22.41	17.83	19.09	22.09	17.56	18.15	21.48
		N	2493	4883	2602	2608	4938	2783	2788	4980
	4	M	594.48	615.47	600.48	627.87	620.52	606.75	632.41	627.5
		SD	15.32	21.45	15.29	17.9	21.18	15.43	17.15	20.98
		N	2385	4846	2419	2420	4902	2604	2606	4959
Language										
	2	M	561.73	582.3	566.96	586.13	587.17	571.59	590.57	591.59
		SD	15.57	20.55	15.57	17.03	20.31	15.7	17.1	20.41
		N	2483	4868	2546	2556	4910	2730	2740	4969
	3	M	574.98	597.75	581.86	601.61	604.16	587.39	607.58	609.49
		SD	15.5	22.33	15.78	18.01	21.93	15.71	17.47	21.53
		N	2442	4873	2586	2601	4933	2771	2786	4977
	4	M	597.9	622	601.72	621.92	624.21	605.9	626.58	628.25
		SD	13.9	20.48	13.97	16.46	20.22	13.68	15.68	19.61
		N	2367	4839	2407	2413	4895	2596	2604	4960

The World Wide Web address for the *Education Policy Analysis Archives* is **epaa.asu.edu**

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. The Commentary Editor is Casey D. Cobb: casey.cobb@unh.edu .

EPAA Editorial Board

EPAA Spanish Language Editorial Board