CRESST REPORT 764

Jinok Kim
Joan L. Herman

A THREE-STATE STUDY OF ENGLISH LEARNER PROGRESS

NOVEMBER, 2009

National Center for Research on Evaluation, Standards, and Student Testing

Graduate School of Education & Information Studies
UCLA | University of California, Los Angeles

A Three-State Study of English Learner Progress

CRESST Report 764

Jinok Kim & Joan L. Herman
CRESST/University of California, Los Angeles

November, 2009

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
GSE&IS Building, Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported under the national Research and Development Centers, PR/Award Number R305A050004, as administered by the U.S. Department of Education's Institute of Education Sciences (IES).

To cite from this report, please use the following as your APA reference:
Kim, J., & Herman, J. L. (2009). *A three-state study of English learner progress* (CRESST Report 764). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

A THREE-STATE STUDY OF ENGLISH LEARNER PROGRESS

Jinok Kim and Joan L. Herman
CRESST/University of California, Los Angeles

Abstract

In this three-state study, the authors estimate the magnitudes of achievement gaps between EL students and their non-EL peers, while avoiding typical caveats in cross sectional studies. The authors further compare the observed achievement gaps across three distinct dimensions (content areas, grades, and states) and report patterns of EL and non-EL achievement gaps within and across states. The study findings suggest that linguistic barriers and long-term EL designation may contribute to the observed achievement gaps. The findings further suggest that the differences in the stringency of state reclassification criteria may influence the reported size of the EL and non-EL achievement gaps between states.

Introduction

English learners (ELs) are a large and growing population in public schools across the country (e.g., see Shin, 2003; GAO, 2006). In California, for example, approximately one-fourth of all students and one-third at the elementary level are English learners (EdSource, 2007). The EL population faces complex challenges in needing simultaneously to acquire English language proficiency (ELP) and to achieve academic success in subject matter content. Ample research documents the extent of the challenge, evident in pervasive achievement gaps between EL and non-EL students. For example, on the basis of 2003–04 mathematics test scores across 48 states, EL students' math proficiency level averaged 20% lower than the overall population (GAO, 2006); for the 2005 National Assessment of Educational Progress (NAEP) in mathematics, 46% of Grade 4 and 71% of Grade 8 EL students scored below basic as compared to 18% and 30% of non-EL students at the two grades respectively (Perie, Grigg, & Dion, 2005).

Although such findings may provide a useful starting point to draw well-deserved attention to ELs' academic performance, such comparisons may pose serious limitations for evaluating EL progress within and across states. Among these challenges are issues related to: a) inherent composition of the EL population; b) confounding of EL and socioeconomic status (SES); and c) differences in state EL policies and practices.

A key challenge is the changing population of the EL group itself. As English learners improve, the most successful students are reclassified as English proficient, leaving the

remaining, less successful students in the EL group. The EL group continually receives new, mostly lower performing students. Thus, the same EL and non-EL comparison over time, even between short time periods, may be based on different samples. Students who performed well and exited EL status are not only missing from the EL group, but may automatically become part of the EL comparison group (i.e., non-ELs). Unlike comparisons between ethnic or gender groups, the EL group is inherently unstable, making accurate comparisons difficult.

Furthermore, existing comparisons may underestimate the fact that ELs tend to be from appreciably lower SES backgrounds than their non-EL peers. The over-representation of ELs in more disadvantageous conditions has been reported repeatedly (see, e.g., Artiles, Rueda, Salazar, & Higareda, 2005; Gándara, Rumberger, Maxwell-Jolly & Callahan, 2003; McCardle, Mele-McCarthy, Cutting, Leos, & D'Emilio, 2005) and is evident in our analysis of data from three states in the current study. EL and non-EL groups show striking differences in the proportions receiving free or reduced lunch (FRL). Analysis that do not account for socio-economic status may consequently produce EL achievement gap estimates that are confounded by socioeconomic status, which may mask barriers which are unique to EL status.

Lastly, simple nationwide comparisons may overlook substantial differences across states, which often have varying practices and policies identifying, assessing, reclassifying and instructing their EL students (Wolf, Kao et al., 2008). Such variations both impede cross-state comparisons and may produce notable differences in achievement gaps between EL and non-EL students.

Our three-state study attempts to provide more accurate estimates of inferences concerning the average achievement gaps between EL students and their non-EL peers in settings where student achievement data are available at single time points (i.e., in cross sectional studies). We attempt to reduce, to the extent possible, the influences of the above caveats through the following approaches:

First, we divide students into four subgroups instead of the usual two: a) *current ELs*; b) *recently reclassified students*, (i.e., reclassified as fluent in English within the last 2 years); c) *former EL* students who are reclassified more than 2 years earlier (thus, no longer monitored for NCLB purposes); and *non-EL* students. By using these more refined categories, we reduce some of the instability of the EL population. For example, by separating the former and usually more successful ELs from the non-EL comparison group,

we can better estimate differences in achievement between current ELs and other students, and also learn more about differences in achievement within the EL population.

Second, our analyses control for whether a student receives free and reduce lunch (FRL). This helps to reduce possible differences in academic achievement based on different socioeconomic backgrounds between ELs and non-ELs.

Third, in order to make within- and between-state comparisons possible between each state, we converted our results from analysis of state tests into standard deviation (*SD*) units. This conversion to *SD*s also allows us to make comparisons between content areas and grades, thus capturing achievement gap patterns among our four groups within and across states. In the process, we attempt to suggest sources underlying the observed EL and non-EL achievement gaps, as well as illustrate how the observed achievement gaps can also be a byproduct of different reclassification criteria. Such comparisons suggest important challenges in evaluating EL performance due to: a) the heterogeneity of the EL population; b) differences in states' reclassification criteria; and c) potential linguistic barriers.

Our research questions are as follows:

1. What are the expected differences in annual state assessments of current ELs, recently reclassified ELs (i.e., students who are reclassified during 2 recent academic years), and former ELs (i.e., students who are reclassified for more than 2 academic years) as compared to non-ELs in different grades, subject areas, and states?
2. Are there noticeable within-state patterns in the achievement gaps between ELs and non-ELs, such as patterns across subject areas or across grades?

3a. Are there achievement differences between states? If so, what might be probable factors?

3b. How can the stringency of the reclassification criteria be gauged across states in settings where states use different methods and testing instruments for reclassification?

Methods

Data and Sample

Data for this study come from three volunteer states participating in a larger 3-state research project. Data included state annual assessment scores in mathematics, reading, and, when available, science. Data also included demographic information for all students as well as EL-specific information for current and reclassified EL students.

Our study sample consisted of elementary and secondary school cohorts in each state. Table 1 shows the specific cohorts including proportions of ELs, reclassified ELs and non-ELs in each state.

Table 1

Frequencies and Percentages of English Learner (EL) Students by Grade in Three Participating States

State	Student Status	Grade 4		Grade 5		Grade 7		Grade 8	
		N	Percent	*N*	Percent	*N*	Percent	*N*	Percent
A	Non-EL			24380	73.34			25180	75.58
	RFEP[a]			3854	11.59			4304	12.92
	EL			5008	15.07			3833	11.50
	Total			33242	100.00			33317	100.00
B	Non-EL	70511	91.41			75404	94.10	83956	94.77
	RFEP[a]	2539	3.29			2160	2.70	2079	2.35
	EL	4086	5.30			2565	3.20	2554	2.88
	Total	77136	100.00			80129	100.00	88589	100.00
C	Non-EL	47467	85.19					52299	89.92
	RFEP[a]	2020	3.63					2355	4.05
	EL	6230	11.18					3510	6.03
	Total	55717	100.00					58164	100.00

Note. [a]Reclassified fluent English proficient.

Table 1 shows that, depending on the state, our samples included 5th- and 8th- grade students, 4th-, 7th-, and 8th-grade students, or 4th- and 8th-grade students. In the elementary school grades (i.e., 4th or 5th), current ELs comprise 15%, 5%, and 11% of the entire student population in corresponding grades, respectively in states A, B and C. Reclassified ELs comprise 12%, 3% and 4%, respectively in states A, B and C. In the secondary school grades (i.e., 7th or 8th) current ELs comprise 11%, 3%, and 6% of the entire student population and reclassified ELs comprise 13%, 2–3%, and 4% respectively in states A, B and C. The three states were located in the West and Southeast. Despite their geographic dispersion, the EL population in all three states was of similar ethnicity (more than 80% Hispanic in each state) and had low SES (more than 70% eligible for FRL in each state).

Table 2 presents percentages of FRL status by EL status in three states, showing that in all three states, the socioeconomic gaps are large. EL students who receive FRL were 67%–83%, whereas only 31%–39% of the non-EL students receive FRL. Our study sample from

three states agrees with other findings regarding disadvantageous backgrounds for the majority of EL students (Artilles et al., 2005; Gándara et al., 2003; McCardle et al., 2005).

Table 2

Frequencies and Percentages of Students Receiving Free or Reduced Lunch (FRL) by EL Status in Three Participating States.

State	Student Status	EO	EL		Total
			Former	Current	
A	Non-FRL	17183 (68.8)	1511 (35.1)	1261 (32.9)	19955 (60.3)
	FRL	7786 (31.2)	2793 (64.9)	2572 (67.1)	13151 (39.7)
B	Non-FRL	51350 (61.2)	554 (26.6)	446 (17.5)	52350 (59.1)
	Reduced-priced lunch	6947 (8.3)	235 (11.3)	237 (9.3)	7419 (8.4)
	Free lunch	25659 (30.6)	1290 (62.0)	1871 (73.3)	28820 (32.5)
C	Non-FRL	36170 (69.2)	768 (32.6)	596 (17.0)	37534 (64.5)
	Reduced-priced lunch	3504 (6.7)	249 (10.6)	252 (7.2)	4005 (6.9)
	Free lunch	12625 (24.1)	1338 (56.8)	2662 (75.8)	16625 (28.6)

Note. Percentages are enclosed in parentheses.

Measures

State assessment measures were those used for determining Adequate Yearly Progress as required by NCLB. Reading (or alternatively English Language Arts) and mathematics assessment data were available for students from Grades 3 to 8 (or in more grades in one state) and science assessments were available only at designated grade levels (e.g., Grades 5 and 8). We used the scale scores to preserve the psychometric properties of the tests and avoid the variations posed by achievement levels (see Linn, Kortez, Baker & Burstein, 1991).

Other measures included individual student scores on state English language proficiency (ELP) assessments, measuring EL skills in speaking, listening, reading, and writing. For all participating states, ELP scores were the primary, if not only criterion for reclassifying EL students. Each state used different ELP assessments as well as different ways to evaluate and reclassify EL students. In our analysis to gauge the stringency of EL reclassification criteria, we used the "overall" ELP scale scores in two states and the ELP reading scale scores in the third state. Our decisions were based on the primary roles of the chosen scores in evaluating and reclassifying ELs in their respective states.

Statistical Approach

This section presents the logic and statistical methods used in this study, focusing on: a) statistical models to estimate the expected achievement levels of and gaps among EL students, recently reclassified EL students, former EL students and non-EL students; b) logic in comparing the estimated achievement gaps across content areas, grades, and states; c) logic and statistical models to gauge the extent of stringency in states' reclassification criteria; and d) rationale for exploring the extent of stringency in reclassification criteria. We use two-level hierarchical models (HMs) in which students are nested in schools.

Because academic achievement of both EL and non-EL students in a school may be more similar than the academic achievement of students in other schools, (i.e., there may be significant intraclass correlations in outcomes), the use of HMs was strongly warranted (see, e.g., Goldstein, 2003; Raudenbush & Bryk, 2002). In nested settings with possible intraclass correlations, HMs support sound inferences by yielding accurate standard errors and associated parameters. HMs also provide estimates of the extent within-cluster parameters vary across clusters (i.e., schools) and the variability in outcome in cluster and student levels. Such partitioned variability can be further modeled as a function of predictors at each level.

Estimating the achievement gaps. We fit 2-level HMs with almost identical specifications repeatedly to different outcomes, since each of the different sets of subject areas, cohorts, and states involves different outcome scores, thereby resulting in separate analyses for different sets of subject area, cohort, and state combinations. Equation 1 (see Appendix A) shows the specification of two-level hierarchical model in which students are nested within schools.

As noted earlier, to learn more from group average comparisons given the inherent inconsistency of the EL group, we divided ELs to more refined categories: current ELs, recently reclassified ELs (reclassified within less than 2 previous years) and former ELs (reclassified for more than 2 years). The HM in Equation 1 captures the expected differences or gaps in achievement between the baseline group (non-ELs) and the other groups, holding free and reduced lunch status (FRL) constant. The HM in Equation 1 also controlled for FRL to reduce confounding by differences in socioeconomic backgrounds between EL students and non-EL students. EL and non-EL students in all cohorts in all three states revealed considerable disparity in their SES (see Table 2), which is important to consider given the well-known relationships between student academic achievement and FRL status (see, e.g., Tharp, 1997).

Comparing the achievement gaps within and across states. In estimating achievement gaps as shown in Equation 1 (Appendix A), the outcome measures differed by state, subject, and grade. Therefore, the estimated gaps from different scales were not comparable. We calculate the estimated differences in achievement (or achievement gaps) in terms of standard deviations (*SD*s) of outcomes in order to enable comparisons among states, subjects, and grades. Another benefit of using *SD*s is that *SD*s can provide a direct sense of the magnitudes of the estimated gaps or expected differences. In studies of treatment effects (e.g., Cohen, 1988), researchers often use rough approximations to gauge the magnitude of treatment effects. For example, under certain circumstances, 0.2 *SD*s is considered "small," 0.5 *SD*s is considered "medium," and over 0.8 *SD*s is considered "large." Although the results in this chapter are not effect sizes, these approximations serve as a reference to the magnitude of the achievement gap.

Gauging the extent of stringency in states' reclassification criteria. Although states use different criteria for reclassifying EL students, one primary criterion for states is student performance. One way of assessing the stringency of states' reclassification criteria is to gauge the levels of student English language proficiency required for reclassification in different states. This is challenging because states administer entirely different ELP tests, so that the required ELP levels are not directly comparable.

Our approach was to approximate the extent of stringency of the ELP-related reclassification criteria, based on a comparison between student ELP performance and content performance. If in states where EL students just meet the reclassification cut-off, tend to do better than the state assessment proficiency levels intended for all students; we may reasonably infer that the states' ELP-related criteria are rather stringent. On the other hand, in states where EL students just meet the reclassification cut-off, tend to do much worse than the state assessment proficiency levels, we may reasonably infer that the states' ELP-related criteria are relatively lenient.

The validity of the approach outlined above is contingent on strong and consistent relationships between performances in ELP assessment and state assessment. Specifically, we estimated the expected scores on content-area assessments given the ELP cut-offs at which students were reclassified. We then compared the scores to the content-area test cut-off scores at which students were considered meeting or above proficiency. To estimate the relationships between the ELP and state content-area assessments, we fitted the HMs repeatedly for each combination of subject area, cohort and state (See Appendix B for details).

Relating the stringency of reclassification criteria and achievement gaps. In comparing observed EL and non-EL achievement gaps across states, we focused on one policy, EL reclassification criteria. Suppose that the only difference between two settings is the stringency of reclassification criteria, while other factors that may relate to achievement gaps being comparable. In such settings, more stringent reclassification criteria will lead to smaller achievement gaps, whereas more lenient reclassification criteria will be associated with larger achievement gaps. Fewer students are likely to be able to pass the more stringent reclassification criteria and those students are likely to be more academically accomplished than those who achieve only the lower criteria. We exemplify this idea in Figure 1.

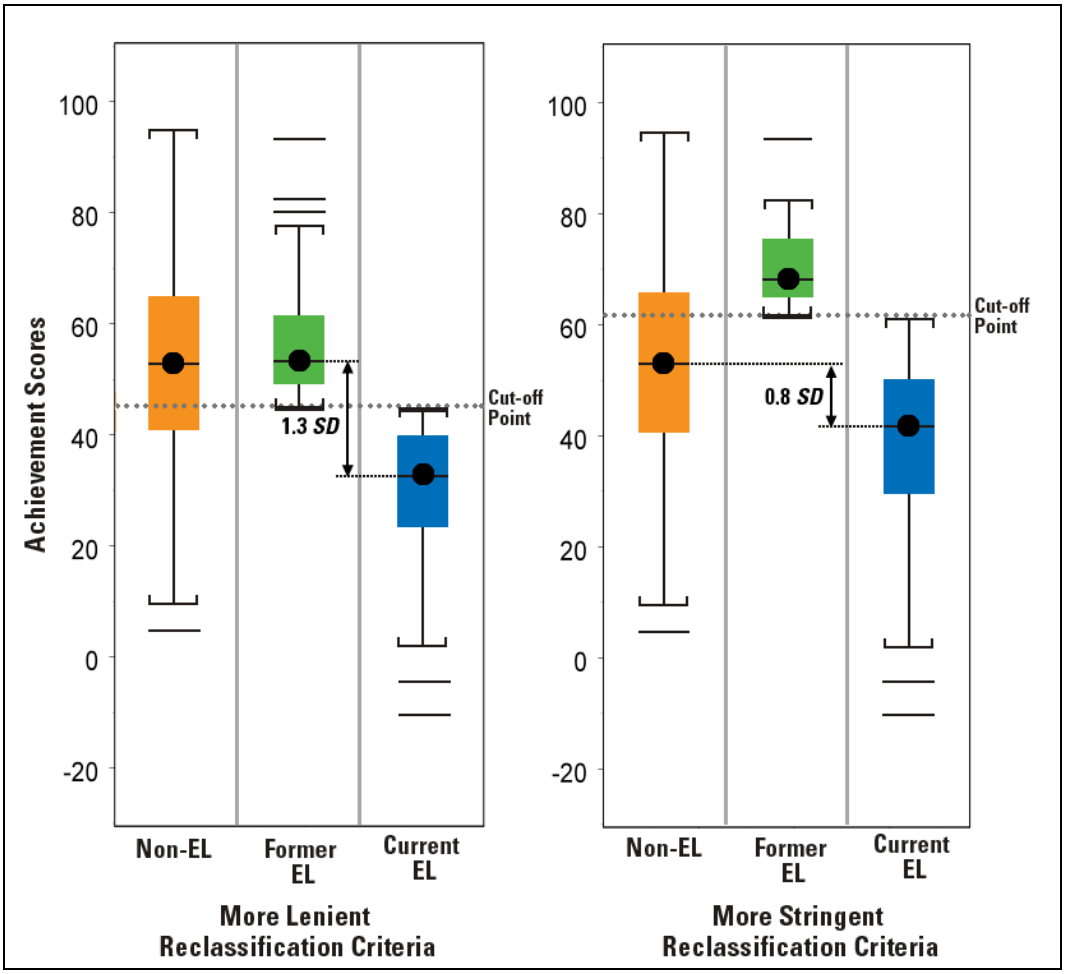

Figure 1. Achievement gaps between current EL students and non-EL students.

Student score distributions in Figure 1 are identical; however, the cut-off for EL students to exit is lower in the left panel, while higher in the right panel (See horizontal dotted lines). The two cut-off scores are set to be 1 *SD* apart of the generated data. As a result of the difference in cutoffs, one can see that with a lower cut-off point (i.e., more lenient

criteria) shown in the left panel, the average achievement gap between current ELs and non-ELs is greater (1.3 *SD*). Whereas, with a higher cut-off point shown in the right panel (i.e., more stringent criteria) the average achievement gap is smaller (0.8 *SD*). The lengths of the vertical arrows capture the average achievement gaps, displayed to facilitate visual comparison using the same score distributions for both EL and non-EL students.

Results

We present results by our research question, focusing on summaries and trends that emerged across all three states.[1] Focusing on trends helps illustrate various factors that may underlie the observed achievement gaps, or which may suggest challenges in studying EL progress from results at a single time point on state annual assessments.

Research Question 1. What are the expected differences in annual state assessments of current ELs, recently reclassified ELs (i.e., students who are reclassified during 2 recent academic years), and former ELs (i.e., students who are reclassified for more than 2 academic years) as compared to non-ELs in different subjects, grades, and states?

HM analysis was conducted to estimate average achievement gaps for scores on each of the 17 assessments, for each combination of the states, grades, and content areas. The specification of the HM is shown in Equation 1 in Appendix A. Table 3 show HM results under two selected settings, in 8th grade mathematics in State A (left panel) and B (right panel) respectively.

[1] Results from all analyses for each state, content area and grade level combination are presented in a report available in public (Kim & Herman, 2008, available at http://www.cse.ucla.edu/products/reports/R738.pdf). For sets of HM analyses estimating achievement gaps, see Chapter 3, Appendix CH3 of the cited report (Tables A1 to A7). For sets of HM analyses estimating the relationships between ELP and annual state assessments, see Chapter 3, Appendix CH3 (Tables B1 to B7).

Table 3

Results from Estimating EL and non-EL Achievement Gaps in 8th Grade Mathematics in States A and B.

	State A			State B		
Fixed effects	Coefficient	*(SE)*	*p* value	Coefficient	*(SE)*	*p* value
Intercept, γ00	296.31	(5.15)	<. 0001	360.11	(0.17)	<. 0001
ELL, γ10	-64.08	(2.64)	<. 0001	-2.82	(0.19)	<. 0001
ExitMonitor, γ20	-44.68	(3.74)	<. 0001	2.89	(0.53)	<. 0001
Exit, γ30	20.28	(2.58)	<. 0001	3.53	(0.23)	<. 0001
FRL, γ40	-38.45	(2.02)	<. 0001	-5.06	(0.11)	<. 0001
Random effects	Variance component	*(SE)*	*p* value	Variance component	*(SE)*	*p* value
Intercept, τ00	2294.41	(384.10)	<. 0001	18.31	(1.15)	<. 0001
ELL, τ11	192.98	(69.84)	0.0029	2.08	(0.80)	0.0048
ExitMonitor, τ22	307.58	(125.39)	0.0071	10.04	(4.45)	0.0121
Exit, τ33	151.04	(62.13)	0.0075	2.73	(1.06)	0.0052
FRL, τ44	148.41	(43.13)	0.0003	4.47	(0.38)	<. 0001
Residual	7492.15	(58.96)	<. 0001	62.89	(0.27)	<. 0001

Figure 2 displays the estimated differences in achievement in terms of *SD*s of outcomes, based on results from all 17 HM analyses. The three panels are the results from each of the three states. Each column in the figure represents results from one HM analysis. For example, the results for 8th grade math in states A and B in Table 3 are first converted in the *SD* differences of the corresponding outcomes; and next are shown respectively in the fifth column in the state A panel and the six column in the state B panel.

In each column, the horizontal line of 0 indicates the performance level of the baseline group who are non-ELs, and the estimated differences in *SD* scales are indicated by circles, triangles, and squares respectively for the current ELs, former ELs, and recently reclassified ELs. The further the estimates are located from the horizontal line of 0, the greater the differences in achievement between the respective group and the baseline group (i.e., non-ELs).

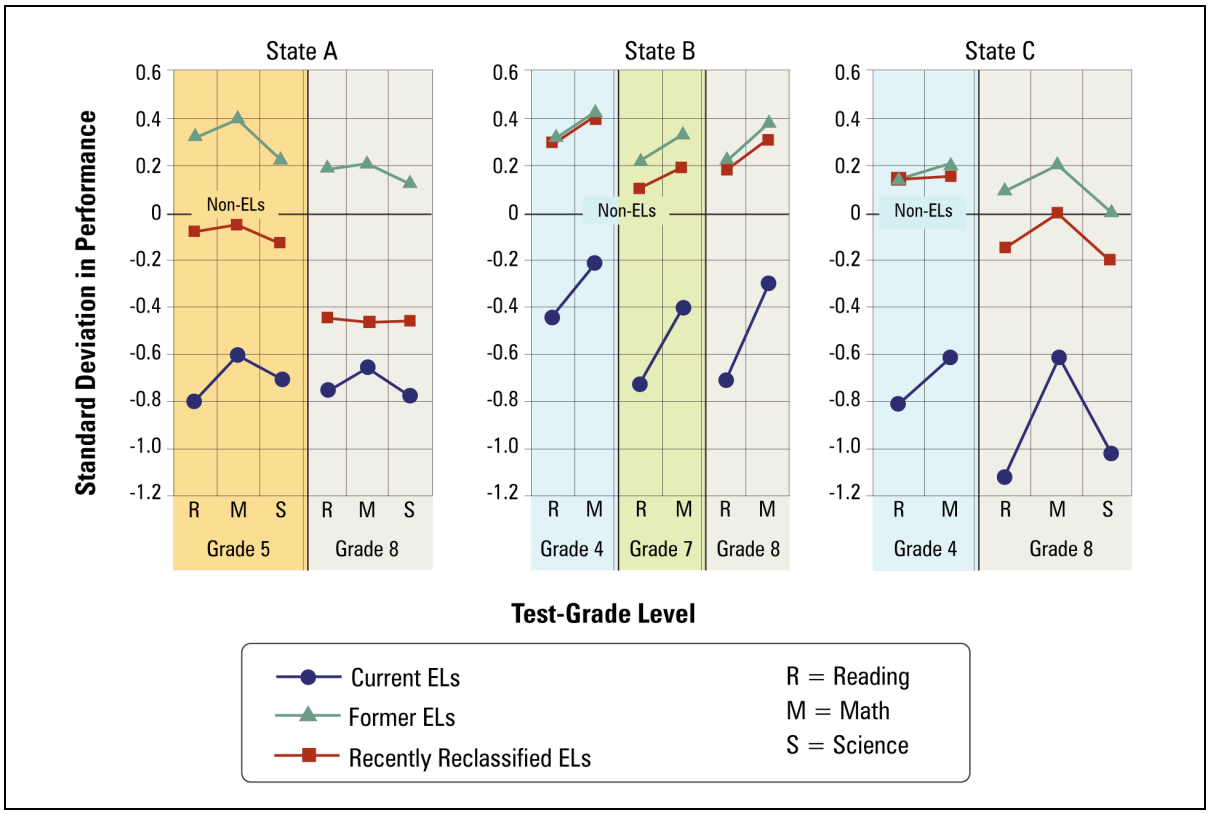

Figure 2. Estimated EL achievement gaps for three states.

We found statistically significant achievement gaps between current EL students and their non-EL peers in all three states. As shown by circles in Figure 2, the gaps range from fairly modest magnitudes of about 0.2 to 0.3 *SD*s, to large magnitudes greater than 1 *SD*, depending on the subject, grade, and state combination. The magnitudes of average achievement gaps ranged from small to medium in mathematics (0.2 to 0.6 *SD*s), whereas, in reading or science, they ranged from medium to large sizes (0.4 to 1.1 *SD*s). As indicated by triangles, in all states, former EL students who were reclassified at least 2 years earlier performed, on average, significantly better than or as well as non-EL students after controlling for student FRL status. Former EL students performed better by varying magnitudes (from 0.1 *SD*s to 0.4 *SD*s), depending on the subject, grade, and state combinations.

Looking specifically at recently reclassified students (see squares in Figure 2), we found mixed results within and between states. As represented by squares, in State A, recently reclassified students tended to perform lower overall than non-ELs, while in State B, recently reclassified students tended to perform higher than non-ELs. Students in State C had mixed results.

Research Question 2. Are there noticeable within-state patterns in the achievement gaps between ELs and non-ELs, such as patterns across subject areas or across grades?

Comparisons across subjects. Within states but across subject areas, current ELs, recently reclassified ELs, and former ELs all tended to perform worse in reading and science than in math, when compared to non-ELs. Figure 2 shows that most of the lines connecting from math to other subjects dropped for all states and grades, with faster drops representing greater magnitudes of differences in gaps between subjects. Although this pattern is consistent in all EL groups (i.e., current ELs, recently reclassified ELs, and former ELs), the pattern across subjects is more pronounced among current EL students, especially in all grades in State B and also in upper grades in State C. In more prominent cases, the differences in average gaps between math and other subjects can be equal to or greater than 0.4 *SD*s (e.g., the average gaps of 0.3 *SD*s in math; and 0.7 *SD*s in reading among 8th grade current ELs in State B).

Comparisons across grade levels. As for grade levels with states, current ELs, recently reclassified ELs and former ELs all tended to perform worse in upper grades than in lower grades when compared to non-ELs. This pattern was more prominent among recently reclassified students in States A and C, among current EL students in State B, and among current EL students in reading in State C. In more prominent cases, the differences in average gaps between upper and lower grades ranged from 0.2 *SD*s to 0.4 *SD*s (e.g., the average reading gaps of 0.8 *SD*s for lower grade current ELs; and 1.1 *SD*s for upper grade current ELs in State B).

Research Question 3a. Are there achievement differences between states? If so, what might be probable factors?

The magnitudes of the expected differences in achievement between EL and non-EL students varied across states, as shown in the three panels of Figure 2. Comparison between States A and B shows that: a) the average gaps between current ELs and non-ELs were appreciably greater in State A than in State B (with an exception of reading in upper grades[2]);

[2] In reading in the upper grade (8th grade), unlike all other combinations, the achievement gap between current ELs and non-ELs of State A is rather similar to that of State B (instead of the gap of State A being greater than that of State B). Further examination of all states helped us find that this is due to an anomalous designated level of meeting proficiency in upper-grade reading in State B. This paper compares how well ELs just meeting reclassification criteria would perform to the state-designated level of meeting proficiency in the annual state assessment for all students. While in the other subject, grade, and state combinations, non-ELs tended to perform similarly to the level of meeting proficiency. However, in upper grade reading in State B, non-ELs tended to perform better by far than the designated level of meeting proficiency, which results in the greater gap in outcome than expected. Given that the level of meeting proficiency in state annual assessment is beyond the scope of this paper, we just consider this as an exceptional result.

and that b) recently reclassified ELs tended to perform worse relative to non-ELs in State A, but tended to perform better than non-ELs in State B. State C shows mixed results within the state.

The substantial differences in magnitudes and directions of the expected differences in achievement across states occur especially in two groups, current ELs and recently reclassified ELs. This supports our hypothesis that differences in state reclassification criteria contribute to the between-state achievement results. As illustrated earlier (see Figure 1 and the discussions around it), because current ELs are the students who did not reach reclassification criteria, with recently reclassified ELs being the students who just passed the criteria, the stringency of the state reclassification criteria can directly affect the compositions and characteristics of the EL groups. This is especially true for current ELs and recently reclassified ELs.

Research Question 3b. How can the stringency of the reclassification criteria be gauged across states in settings where states use different methods and testing instruments for reclassification?

We estimated the stringency of the reclassification criteria on the basis of relationships between ELP and state assessments (see the Methods section for more details). The validity of this approach depends on a number of issues including: how important the role of ELP assessment is in reclassifying ELs; and how stable the relationships are between ELP and state assessments.

To estimate the relationships between ELP levels or scores and annual state assessments and to examine the stability of relationships for students with different characteristics and settings, we used 2-level HMs (See Equation 2, Appendix B). The results showed that for all grades and subjects, ELP scores or levels were strongly and positively associated with performance in content-area assessments across all three states.

Two other results are of particular note. First, after ELP levels or scores are controlled, student FRL status tends not to contribute to substantively meaningful difference in achievement despite statistical significance (Note that due to very large sample sizes, any small differences may turn out significant). Second, the relationships between ELP and content-area assessments do not vary across schools. In other words, the positive and strong relationships remain fairly consistent under appreciably different settings. Table 4 shows the results under two selected settings, 8th grade mathematics in State A (left panel) and State B (right panel). Figures 3 and 4 display the ELP-to-content area assessment relationships based on the results in Table 4.

Table 4

Results from Estimating ELP-to-Content Area Achievement Relationships in 8th Grade Mathematics in State A and State B

	State A			State B		
Fixed effects	Coefficient	*(SE)*	*p* value	Coefficient	*(SE)*	*p* value
Intercept	163.92	(4.27)	<.0001	353.57	(0.30)	<.0001
ELP	26.61	(1.38)	<.0001	0.06	(0.00)	<.0001
ELP^2	7.98	(0.77)	<.0001	0.00	(0.00)	<.0001
FRL	-6.63	(2.58)	0.01	-0.87	(0.31)	0.0057
ELP^2 x FRL				0.00	(0.00)	0.0179
cohort4	35.04	(4.32)	<.0001			
cohort5	47.1	(3.85)	<.0001			
cohort6	30.82	(4.20)	<.0001			
Random effects	Variance component	*(SE)*	*p* value	Variance component	*(SE)*	*p* value
Intercept	439.91	(127.61)	0.0003	3.99	(0.68)	<.0001
Residual	5011.02	(108.69)	<.0001	36.71	(0.95)	<.0001

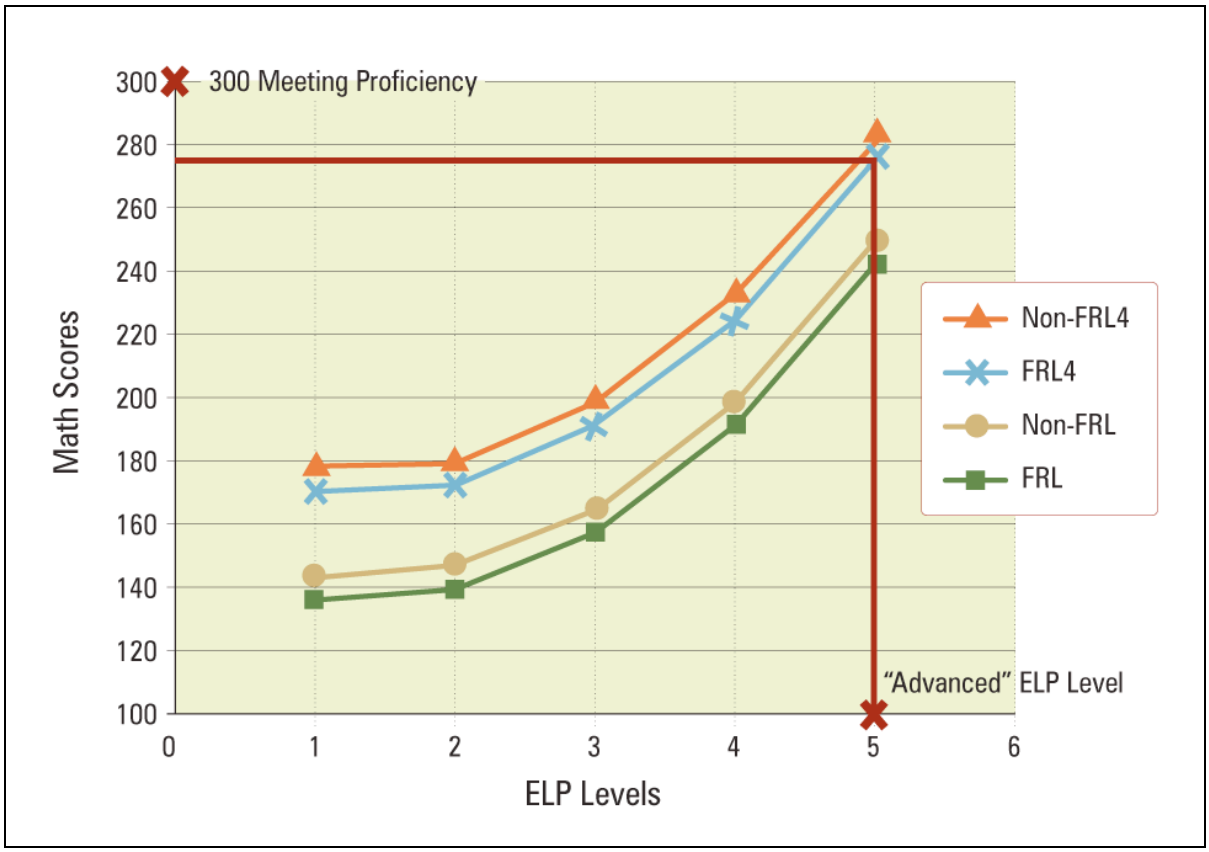

Figure 3. The relationships between ELP and math state assessments for 8th graders in State A.

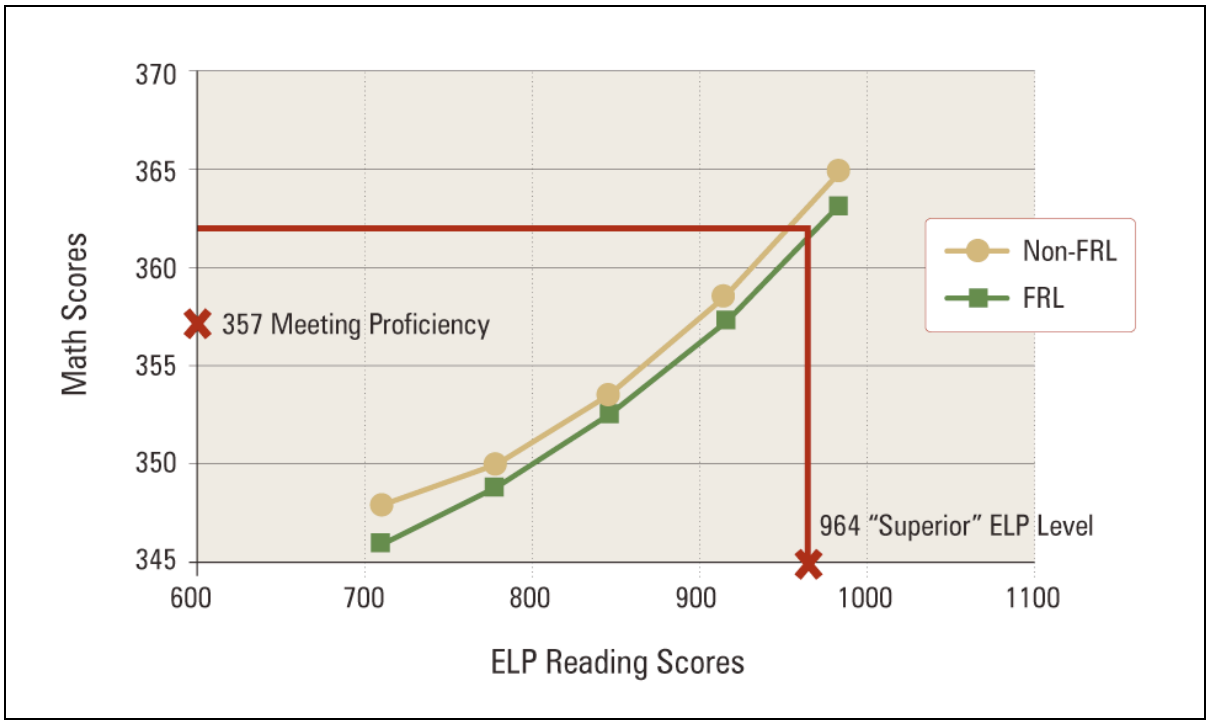

Figure 4. The relationships between ELP and math state assessments for 8th graders in State B.

Based on the estimated relationships, we attempt to gauge the stringency of the states' reclassification criteria. Since State C did not have uniform reclassification criteria, we used the ELP cut-off levels (State A) or cut-off scores (State B) and estimated how EL students who just meet the English proficiency cut-offs would tend to perform in state annual assessments. In Figures 3 and 4, the vertical lines depict the ELP cut-off at which students are reclassified (x-axis), while horizontal lines indicate the estimated scores in state exams (y-axis) given the ELP level or cut-off (x-axis); and so these lines assist us in seeing how well an EL student tends to perform right when he or she meets the reclassification criteria. The cut scores set by the state in state annual assessments to categorize all students as meeting the standard or being at or above proficient are marked as asterisks in the y-axis for visual comparisons.

The results show that, unlike the relationships between ELP and content-area assessments, findings about reclassification criteria fluctuated across states, as we hypothesized. The figures illustrate the results with regard to the stringency of reclassification criteria. In State A, students who just get reclassified tend to perform worse than the math proficiency level in state assessment. On the other hand, State B appears to have more stringent criteria given that students who just get reclassified already tend to perform higher than the state-designated math proficiency level intended for all students. These contrary patterns in States A and B were consistent across other subject areas and grades. Based on the primary role of the ELP test in determining reclassification and the strong and consistent relationships, such results may indicate that State A has relatively lenient criteria for reclassification, while State B has relatively stringent criteria for reclassification.

Discussion

EL to non-EL Achievement Differences in Various Subjects, Grades, and States

Consistent findings from all three states reveal that significant gaps exist between current EL and non-EL students, while former ELs (i.e., students who are reclassified for more than 2 years and no longer monitored) generally perform better than non-EL students, regardless of content areas and grade levels. These trends are consistent and evident across all three states. An important caveat is that these cross sectional study results should *not* be understood to mean that the achievement gap between EL and non-EL students tends to narrow or close after more than a few years following reclassification. Rather, it should be seen to underscore the heterogeneity of the EL population. Some ELs improve their ELP and academic performance and exit EL status. They also may close achievement gaps with their

non-EL peers but others do not. In fact there may be a substantial portion of EL students who never exit EL status nor catch up with their non-EL peers, the so-called long-term ELs. In California, for example, more than 50% to 60% of ELs do not exit the EL status after 10 years of schooling (Grissom, 2004; Mitchell, Destino, & Karam, 1997; Parrish, Perez, Merickel, & Linquanti, 2006).

Within-State Achievement Patterns

Comparisons across content areas deal with the same set of students although the analyses are conducted separately for each content area.[3] By using the same set of students, we gain unique control over many variables that may relate to student achievement such as student characteristics, school characteristics, and practice and policy of schools, districts, and states, because the same set of students will share all such characteristics or experiences. Thus, our comparisons focus on differences in difficulties that ELs encounter in different content areas.

Therefore, the results showing smaller achievement gaps between ELs and non-ELs in mathematics than in reading or science may indicate that linguistic barriers are one of the primary underlying sources of achievement gaps. That is, differences in achievement gaps across subject areas may be due to the varying extent of linguistic difficulty that ELs encounter in various subjects. ELs may have relatively less difficulty in mathematics than in other subjects, because instruction and assessment in mathematics may depend less on English proficiency than in other subjects.

Such an inference is supported by several findings. First, the achievement pattern is more prominent among current ELs who presumably have more linguistic difficulties than recently reclassified ELs or former ELs. Second, a linguistic analysis of the items comprising the annual state assessments used in this study found greater complexity (especially in vocabulary) in reading and science than in math (Wolf, Chang et al., 2008). Similarly, other studies found supporting evidence. For example, Abedi and Lord (2001) found that, minor changes in the wording of test items to reduce linguistic complexity can raise ELL student performance.

Although the pattern is more prominent for current ELs, it is notable that lower performance in reading and science rather than math is also evident for recently reclassified ELs and former ELs compared to non-ELs. This is despite the vast differences in performance levels for current ELs versus recently reclassified and former ELs. Our findings

[3] The exceptions would be a few students whose scores are differentially missing across content areas (e.g., students who took ELA test, but did not take math test, or vice versa).

may imply that, although the linguistic barrier is most significant among current ELs, it appears to be a shared difficulty for reclassified students.

Differences across grade levels reveal greater achievement gaps between ELs and non-ELs in upper grades than in lower grades. These differences, may further confirm that linguistic barriers are a primary cause for achievement gaps. As EL students move to upper grades, they face increasing linguistic complexity in curricular materials, instruction, and assessments; potentially increasing the challenge of being able to transition to mainstream classes without sufficient language skills.

In addition to linguistic barriers, greater average gaps in upper grades than in lower grades may be due to further sources that are related to the inconsistency of the EL population. Note that comparisons across grades include a different set of students (i.e., students in different grades) unlike comparisons across subject areas. Students who improve their English language proficiency and academic achievement and are reclassified in the 4th grade, will likely differ substantially in various characteristics from students who are reclassified in the 8th grade. Therefore, the composition of "current ELs," who were not able to meet the reclassification criteria, may differ appreciably between 4th and 8th grades, despite the identical group identification.

These potential differences in the composition of the EL population are worth noting when we compare achievement gaps of ELs across grades. Greater performance gaps of EL students in higher grades may have been overestimated, not only reflecting their actual performance that may fall farther behind over grades, but also because the current EL group in higher grades consists of higher proportion of long-term ELs.

Although our extant state data did not allow us to empirically examine the performance of long-term ELs specifically in the current study, researchers from previous studies have raised concerns about potential adverse consequences of such long term designation, such as less access to classes required for high school graduation and admission to post-secondary education (see, for example, Parrish et al., 2006, Callahan, 2005; Harklau, 2002); potentially negative affective consequences of EL status during adolescence (Gándara, Gutierrez, & O'Hara, 2001; Maxwell-Jolly, Gándara, & Méndez Benavídez, 2007); dropping out of high school (Silver, Saunders, & Zarate, 2008; Watt & Roessingh, 1994).

Between-State Achievement Differences

Our results well corroborate our hypothesis about the stringency of reclassification criteria as a state-level factor that underlies between-state differences in the observed achievement gaps. In State A with seemingly more lenient criteria, recently reclassified ELs

tend to perform significantly lower than non-ELs, while in State B with more stringent criteria, recently reclassified ELs already perform higher than non-ELs (see Figure 2). At the same time, these results with regard to the stringency of reclassification help explain the reason why the gaps between current ELs and non-ELs are larger in State A and in State B. Relative to State A, State B includes more ELP and more high-performing students in the current EL group since the reclassification criteria is very stringent, which results in smaller average gaps between current EL and non-EL students relative to State A.

State C does not have statewide criteria for reclassification, which may explain the diverging achievement patterns of recently reclassified students within the state. Recently reclassified students in the 4th grade, on average, perform better than non-ELs, while the same group in the 8th grade performs appreciably worse than non-ELs (see Figure 2).

Using the same logic as above, now in the opposite direction (i.e., from the results from achievement gaps in Figure 1 to the extent of stringency of the reclassification criteria), it may be that State C districts or schools tend to use relatively more stringent criteria for a lower grade, while they tend to use relatively more lenient criteria for the upper grades. This pattern of achievement of reclassified students (i.e., recently reclassified students performing appreciably better in the lower grade than in the upper grade, as compared to non-ELs) is also true for State A. Although State A employs statewide criteria, the stringency may differ for lower and upper grades, making it increasingly lenient for upper grades.

This finding suggests another challenge in evaluating EL progress in state annual assessments. The differences in EL performance across states, or across schools or districts with different reclassification criteria, may reflect substantively important differences, such as varying demographics (e.g., some areas that induce more of high achieving newly arrived EL students, or other areas that are concentrated with long–term ELs), or differences in English Language Proficiency (ELP) programs or in teacher capacities. The differences may also be a byproduct of differences in reclassification criteria.

The results concerning reclassification criteria as a challenge in evaluating EL progress also underscore an important policy question with regard to what may constitute optimal reclassification criteria. Reclassification is a key milestone for ELs, the point at which students are expected to fully function in mainstream classrooms, without any further ELP instructional services or assessment accommodations (Linquanti, 2001). Despite the importance, states and local schools show variation in their reclassification criteria and procedures both within states (see, e.g., Abedi, 2008; Jepsen & de Alth, 2005; Linquanti,

2001; Parrish et al., 2006) and between states (see, e.g., statewide practice review by Wolf, Kao et al., 2008).

Summary and Conclusion

This study highlights the difficulty of evaluating EL achievement progress in cross sectional studies. Both the approach and the findings of this paper strongly suggest that we should make cautious interpretations of EL and non-EL performance gaps. In addition to the English language barrier, which is the admitted and shared source for all ELs, other factors brought up in this study (i.e., long-term EL designation, and reclassification criteria) are closely related to the inherent instability of the EL population.

It is also important to be aware that estimated achievement gaps may be confounded by differences in cut-off scores before assuming that such differences reveal important substantive differences, such as the quality of instructional programs and methods; implementation of instructional programs; teacher capacities; and school, district, and state EL policies.

Our findings about the relationship between the stringency of criteria and post-reclassification EL performance do not necessarily mean that more stringent criteria are better. Moreover, our findings are silent on other important issues in formulating EL policy, for example, the costs and benefits of those who remain in EL status. As noted earlier, EL students suffer a number of negative consequences, including increased probability of dropping out of school and of lower access to college and future success.

Conducting longitudinal studies (see, e.g., Singer & Willet, 2003) in which we track individual students over time can be one of the fundamental solutions for the above illustrated challenges. Longitudinal studies that examine the growth trajectories of long-term ELs over years in terms of ELP and academic achievement may have important implications for policies and practices, as for example, early identification of long-term ELs. Also, longitudinal studies that track individual EL students over time from identification to reclassification, and to post-reclassification years, in terms of both English and academic proficiency, may provide important recommendations for reclassification policy; for example, by implying which type of reclassification may be the best interest for ELs in terms of their post-reclassification achievement.

References

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*(3), 219–234.

Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometrics issues. *Educational Assessment, 8*(3), 231–257.

Abedi, J. (2008). Classification system for English language learners: Issues and recommendations. *Educational Measurement Issues and Practice, 27*(3), 17–31.

Artiles, A. J., Rueda, R., Salazar, J. J., & Higareda, I. (2005). Within-group diversity in minority disproportionate representation: English language learners in urban school districts. *Exceptional Children, 71*(3), 283–300.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

EdSource (September, 2007). *Summary report: Similar English learner students, different results*. Palo Alto, CA: Author.

Gándara, P., Gutierrez, D., & O'Hara, S. (2001). Planning for the future in rural and urban high schools. *Journal of Education for Students Placed At Risk*, *6*,(1–2), 73–93.

Gándara, P., Rumberger, R., Maxwell-Jolly, J., & Callahan, R. (2003). English learners in California schools: Unequal resources, unequal outcomes. *Educational Policy Analysis Archives, 11*(36).

Government Accountability Office (2006). *No Child Left Behind Act: Assistance from education could help states better measure progress of students with limited English proficiency.* Washington, DC: United States Government Accountability Office.

Goldstein, H. (2003). *Multilevel Statistical Models,* (3rd ed.). London: Edward Arnold.

Grissom, J. B. (2004). Reclassification of English language learners. *Education Policy Analysis Archives, 12*(36), Retrieved August 2008 from http://epaa.asu.edu/epaa/v12n36

Harklau, L. (2002). ESL versus mainstream classes: Contrasting L2 learning environments. In V. Zamel & R. Spack. (Eds.), *Enriching ESOL pedagogy: Readings and activities for engagement, reflection, and inquiry,* (pp. 127–158). Mahwah, NJ: Lawrence Erlbaum Associates.

Jepsen, C., & De Alth, S. (2005). *English learners in California schools.* San Francisco, CA: Public Policy Institute of California.

Kim, J., & Herman, J. (2008). Investigating ELL assessment and accommodation practices using state data. In *Providing validity evidence to improve the assessment of English language learners* (CRESST Tech. Rep. No. 738, pp. 81–138). Los Angeles: University of California, National Center for Research on Evaluation, Standards and Student Testing (CRESST).

Linquanti, R. (2001). *The redesignation dilemma: Challenges and choices in fostering meaningful accountability for English learners* (Report No. 2001-1). Oakland, CA: University of California, Linguistic Minority Research Institute.

Linn, R. L., Koretz, D. M., Baker, E. L., & Burstein, L. (1991). *The validity and credibility of the achievement levels for the 1990 national assessment of educational progress in mathematics* (CRESST Tech. Rep. No. 330). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

McCardle, P., Mele-McCarthy, J., Cutting, L., Leos, K., & D'Emilio, T. (2005). Learning disabilities in English language learners: Identifying the issues. *Learning Disabilities Research & Practice, 20*(1), 1–5.

Maxwell-Jolly, J., Gándara, P., & Méndez Benavídez, L (2007). *Promoting academic literacy among secondary English language leaners: A synthesis of research and practice*. Oakland, CA: University of California, Linguistic Minority Research Institute Education Policy Center.

Mitchell, D. E., Destino, T., & Karam, R. (1997). *Evaluation of English language development programs in the Santa Ana Unified School District: A report on data system reliability and statistical modeling of program impacts.* Riverside, CA: University of California, Riverside School of Education: California Educational Research Cooperative.

Parrish, T., Perez, M., Merickel, A., & Linquanti, R. (2006). *Effects of the implementation of Proposition 227 on the education of English learners, K-12: Findings from a five year evaluation* (Final Report). Palo Alto, CA and San Francisco: American Institutes for Research and WestEd.

Perie, M., Grigg, W., & Dion, G. (2005). *The Nation's report card: Mathematics 2005* (NCES 2006-453). Washington, DC: U. S. Department of Education.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods,* (2nd ed.). Newbury Park, CA: Sage.

Shin, H. (2003). *Language use and English-speaking ability*. Washington, DC: U.S. Census. Retrieved June 2008 at http://www.census.gov/prod/2003pubs/c2kbr-29.pdf

Silver, D., Saunders, M., & Zarate, E. (2008). *What factors predict high school graduation in the Los Angeles Unified School District?* Santa Barbara, CA: University of California, California Drop Out Project.

Singer, J. D., & Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Methods for Studying Change and Event Occurrence.* New York: Oxford University Press.

Tharp, R. G. (1997). *From at-risk to excellence: Research, theory, and principles for practice* (Research Rep. No. 1). Washington, DC and Santa Cruz, CA: Center for Research on Education, Diversity & Excellence.

Watt, D., & Roessingh, H. (1994). ESL Dropout: the myth of educational equity. *Alberta Journal of Educational Research*, *40,* 283–296.

Wolf, M. K., Chang, S. M., Jung, H., Farnsworth, T., Bachman, P. L., Noller, J., & Shin, H., (2008). An Investigation of language demands in States' content-area and English Language Proficiency Tests. In *Providing validity evidence to improve the assessment of English language learners.* (CRESST Tech. Rep. No. 738, pp. 9–54). Los Angeles: University of California, National Center for Research on Evaluation, Standards and Student Testing (CRESST).

Wolf, M. K., Kao, J., Griffin, N., Herman, J. L., Bachman, P. L., Chang, S. M., & Farnsworth, T. (2008). *Issues in Assessing English Language Learners: English Language Proficiency Measures and Accommodation Uses—Practice Review.* (CRESST Tech. Rep. No. 732). Los Angeles: University of California, National Center for Research on Evaluation, Standards and Student Testing (CRESST).

Appendix A:
Specification of HMs Estimating the Average Achievement Gaps in State Annual Assessments

The student-level or Level-1 model is specified as

$$Y_{ij} = \beta_{0j} + \beta_{1j}(EL)_{ij} + \beta_{2j}(ExitMonitor)_{ij} + \beta_{3j}(Exit)_{ij} + \beta_{4j}(FRL)_{ij}, \quad r_{ij} \sim N(0, \sigma^2),$$

while the school-level or Level-2 model being

$$\begin{aligned}
\beta_{0j} &= \gamma_{00} + u_{0j} \quad u_{0j} \sim N(0, \tau_{00}) \\
\beta_{1j} &= \gamma_{10} + u_{1j} \quad u_{1j} \sim N(0, \tau_{11}) \\
\beta_{2j} &= \gamma_{20} + u_{2j} \quad u_{2j} \sim N(0, \tau_{22}) \\
\beta_{3j} &= \gamma_{30} + u_{3j} \quad u_{3j} \sim N(0, \tau_{33}) \\
\beta_{4j} &= \gamma_{40} + u_{4j} \quad u_{4j} \sim N(0, \tau_{44}).
\end{aligned} \tag{1}$$

In Equation 1, the outcome Y_{ij} is the achievement score in reading, math, or science in state assessment of student i in school j. All predictors are binary indicators: *EL* is coded as 1 when a student was an EL and as 0 otherwise; *ExitMonitor* is coded as 1 when a student was recently reclassified and still monitored; *Exit* is coded as 1 when a student was reclassified more than 2 years ago and no longer monitored; and *FRL* is coded as 1 if a student was eligible for or receives FRL.

The parameters at Level 1 represent expected levels or differences in the outcome (i.e., state assessment score in a specific subject area, cohort in a state) within school j. The intercept β_{0j} captures the expected achievement of non-EL students who did not receive FRL in school j. The key parameters are β_{1j}, β_{2j}, and β_{3j}, which represents how well current ELs, reclassified ELs, and former ELs do in state assessments as compared to non-EL students. β_{1j} captures the expected difference or gap between EL students and non-EL students in the outcome in school j controlling for whether or not the student was receiving FRL; β_{2j} captures the expected difference between recently reclassified students and non-EL students in school j; and β_{3j} captures the expected difference between reclassified students and non-EL students in school j. Lastly, β_{4j} estimates the expected decrement in achievement associated with students who were receiving FRL in school j.

At Level 2, these within-school parameters are posed to vary across schools. The extent to which each within-school parameter varies across schools was captured by the associated variance components, τ_{00} to τ_{44}.

Appendix B:
Specification of HMs Estimating Relationships between ELP and Content Area Assessments

The student-level or Level-1 model is

$$Y_{ij} = \beta_{0j} + \beta_{1j}(ELP)_{ij} + \beta_{2j}(ELP)^2_{ij} + \beta_{3j}(FRL)_{ij} +$$

$$\beta_{4j}(FRL)_{ij}(ELP)_{ij} + \beta_{5j}(FRL)_{ij}(ELP)^2_{ij} +$$

$$[\beta_{6j}(Cohort4)_{ij} + \beta_{7j}(Cohort5)_{ij} + \beta_{8j}(Cohort6)_{ij}] + r_{ij}\ ,\ r_{ij} \sim N(0, \sigma^2)\ ,$$

while the school-level or Level-2 model being

$$\beta_{0j} = \gamma_{00} + u_{0j} \qquad u_{0j} \sim N(0, \tau_{00})\ ,$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20} \qquad (2)$$

$$\beta_{3j} = \gamma_{30}$$

$$\beta_{4j} = \gamma_{40}$$

$$\beta_{5j} = \gamma_{50}\ .$$

In Equation 2, Y_{ij} is a score in content-area assessments in math, science, or reading of student i in school j; ELP_{ij} is a score or a level in ELP assessment for student i in school j, ELP^2_{ij} was a quadratic term of ELP_{ij}; and FRL_{ij} indicates whether student i in school j received FRL. Cohort variables, were included only in State A due to the availability of the variables in the existing data. The variables, $Cohort4_{ij,}$ $Cohort5_{ij,}$ and $Cohort6_{ij,}$ indicate students who were identified as ELs in the 2003–04, 2004–05, and 2005–06 academic years, respectively, with students who were identified in the 2002–03 year being the base category.[4]

Given the specification of the model, the key parameters of interest are β_{1j} and β_{2j}, which represent the relationship between the ELP levels or scores and the state assessment scores. The quadratic term of ELP capture the extent of curvature in the relationships.

[4] Because data were available with regard to the relationships between ELP and content-area assessments, we conducted additional analysis on one state concerning when EL students are identified. The base cohort, a group of students who were designated as EL students in the 2002–03 school years, comprised the majority of the EL population. The research results show that later cohorts who were designated as EL students in the 2003–06 academic school years performed better in the content-area assessments given the same level of ELP assessment than the base cohort who were designated as EL students in the 2002–2003 academic year, controlling for FRL status. We used only the results from the later cohorts to be consistent with the year of the annual state assessment scores, which are the outcomes of the analysis.

In estimating the ELP-content assessment relationships, we do not only control for the FRL status, but also estimate interactions between FRL status and ELP scores, with both linear and quadratic terms, of which the coefficients were β_{4j} and β_{5j}. The interaction terms were posed based on the hypothesis that the relationships between ELP and content-area assessment may depend on student FRL status. Note that the ELP scores are centered around their grand means. By virtue of the grand-mean centering, the intercept β_{0j} represents the expected scores in state test scores for a student who had a mean value of ELP scores (i.e., average level of ELP among ELs) and did not receive FRL. In a state in which ELP levels instead of ELP scale scores are used due to the data availability (State A), the ELP variable is centered around the medium level (i.e., 3 in a 5-point scale). The intercept β_{0j} represents the expected scores in state test scores for a student who had a medium level of ELP and did not receive FRL.