2016

Who is like Whom? Reclassification and Performance Patterns for Different Groupings of English Learners

Molly M. Faulkner-Bond

University of Massachusetts Amherst, faulkner.bond@gmail.com

Follow this and additional works at: http://scholarworks.umass.edu/dissertations_2

Part of the Bilingual, Multilingual, and Multicultural Education Commons, Educational Assessment, Evaluation, and Research Commons, Education Policy Commons, and the Survival Analysis Commons

Recommended Citation


http://scholarworks.umass.edu/dissertations_2/624

This Open Access Dissertation is brought to you for free and open access by the Dissertations and Theses at ScholarWorks@UMass Amherst. It has been accepted for inclusion in Doctoral Dissertations by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.
Who is like whom?
Reclassification and performance patterns for
different groupings of English learners

A Dissertation Presented

By

MOLLY M. FAULKNER-BOND

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2016

College of Education
Research, Educational Measurement, and Psychometrics
WHO IS LIKE WHOM?
RECLASSIFICATION AND PERFORMANCE PATTERNS FOR DIFFERENT GROUPINGS OF ENGLISH LEARNERS

A Dissertation Presented

By

MOLLY M. FAULKNER-BOND

Approved as to style and content by:

________________________________________
Stephen G. Sireci, Chair

________________________________________
Lisa Keller, Member

________________________________________
Aline Sayer, Member

________________________________
Joseph B. Berger, Senior Associate Dean
College of Education
DEDICATION

To Frank Bond, Sr.
(July 2, 1927 – April 22, 2013)

Despite being counseled out of school after 9th grade, my grandfather never stopped valuing education, and never tired of watching his children and grandchildren collect academic degrees and accolades. This would have been his first doctorate.
ACKNOWLEDGMENTS

I think it is safe to say that without the formidable Ellen Forte, none of this – my degree, my dissertation, my love for education research and policy – would have happened. It was Ellen who first introduced me to this field, and it was Ellen’s intellect, leadership, and dedication that inspired me to love this work as passionately as I do. It was Ellen, too, who not only encouraged me to go to graduate school, but essentially ensured as much by conspiring with Stephen Sireci to lure me to UMass Amherst under false pretenses so that I could meet the faculty and students there. Without that benevolent manipulation, who knows where I’d be in my career or studies – at best a few years behind where I am now, but quite possibly on a different track altogether. I owe Ellen a deep debt of gratitude for all she has done (and continues to do) to support and inspire me, both directly and indirectly. I will probably spend the rest of my life thanking her (gladly).

It’s worth noting that Ellen also connected me to the person who later helped me get my dissertation data – someone who has also become a highly-respected and beloved role model to me. Unfortunately, for confidentiality reasons, I cannot thank that person by name, nor can I thank any of several other individuals from that state who made time and space in their busy schedules to help me get this dataset. If any of them see these acknowledgments, though, I trust they will know who they are, and to them I say: THANK YOU for making the time. Without you, there would be no this. I would also like to extend my gratitude to my committee members – Steve Sireci, Lisa Keller, and Aline Sayer – for their supportive attitudes, helpful feedback, and gracious flexibility throughout this process.

Steve, in particular, deserves additional thanks. Ellen didn’t trick me into going to graduate school just anywhere – she specifically guided me towards Steve Sireci and the UMass program. Unsurprisingly, that turned out to be excellent advice. Steve has been a v
wonderful advisor to me, and I feel very lucky to have found him. We both care deeply about
doing research that is practical and policy-oriented, and he has challenged me to grow and
stretch myself as a thinker, a scholar, and a person. He has given me a great deal of
intellectual freedom, particularly for this dissertation, but also stayed available and present
for when I do need advice or guidance. He knows when to push and when to let up, and is
particularly good at recognizing those moments when what I really need to hear is, “Molly,
you should relax. Why don’t you take tonight off?” I am grateful for all Steve has done to
help me get to this point in my career.

Although Steve has been my primary advisor throughout my time at UMass, I have
enjoyed and appreciated the opportunities I’ve had to work with the other UMass faculty as
well. Ron Hambleton, Lisa Keller, Jennifer Randall, Craig Wells, and April Zenisky have all
been patient teachers, thoughtful mentors, and generally kind people to me since the day I
arrived on campus. I am particularly grateful to Lisa, Jennifer, and Craig, who together have
built my knowledge of statistics entirely from scratch – an especially relevant feat for this
dissertation, since I taught myself a whole new method to do this study. The mere idea of
such an act was inconceivable to me before I entered their classrooms.

There is a third organization whose support made this dissertation possible:
Educational Testing Service (ETS), and particularly Mikyung Kim Wolf in the English
Language Learning and Assessment (ELLA) group. Although Mikyung was not directly
involved in this particular study, she has generally helped me expand and refine my thinking
on EL research and policy in numerous ways. More specifically, she also made it possible for
me to work in the ELLA group as a graduate research fellow which, in addition to being a
rewarding experience, literally made it possible for me to support myself and live in the same
state as my husband while writing my dissertation. Because of the travel arrangements involved, it also allowed me to get to know Mary Pitoniak, who has been a generous and patient mentor and host. Mary sat through more defense presentation and job talk practice runs than anyone except my husband, providing thoughtful and helpful feedback every time. I am grateful to Mikyung, to Mary, and to ETS for their support.

And finally, there are those other folks – the ones who have forgiven me when I’ve missed their birthday parties and baby showers; who have opened their homes to me when I needed a place to work or sleep between semesters; who have nodded their heads supportively when I try to explain my field, my degree, my dissertation topic, even if they have no clue or interest in what I’m talking about. I am lucky to have such wonderful friends and family in DC, Maryland, Western Mass, Princeton, and beyond, who have cheered me along through this process, and waited patiently while I tell them, “just another six months or so!” for almost two years (or really, five years, if you count all of graduate school). The Faulkners, the Bonds, and the Griswolds (also the Elitzers and the Newton House family) have all made my graduate school experience more bearable through their love and support.

And, finally, no individual has done more for me than my favorite human of all, my husband Jeff. Thank you, Jeff, for being there to encourage, support, feed, and listen to me. Thank you for sitting through dozens of practice presentations. Thank you for enduring the bad days and celebrating the good ones, and for reminding me to sleep, bathe, and breathe throughout this dissertation process. Marrying you in the middle of all this was the best decision I’ve ever made. Soon, the tables will be turned and you’ll be the one dissertating. I only hope I do as good a job supporting you as you have for me.
ABSTRACT

WHO IS LIKE WHOM?
RECLASSIFICATION AND PERFORMANCE PATTERNS FOR DIFFERENT GROUPINGS OF ENGLISH LEARNERS

MAY 2016

MOLLY M. FAULKNER-BOND, B.A., HARVARD COLLEGE
Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Stephen G. Sireci

Approximately 10 percent of the US K-12 population consists of English learners (ELs), or students who are learning English in addition to academic content in areas like English language arts (ELA) and mathematics. In addition to meeting the same academic content and performance standards set for all students, it is also a goal for ELs to be reclassified – i.e., to master English so that they can shed the EL label and participate in academic settings where English is used without needing special support. Working with a longitudinal cohort of ~28,000 ELs in grades 3 through 8 from one state, this study uses discrete-time survival analysis to study the probability of being reclassified as a function of time, instructional covariates (e.g., type of language instruction), background covariates (e.g., the student’s home language), years of EL-related service, and district resources. The results suggest that, while probability of reclassification can vary considerably as a function of students’ instructional program, intrastate mobility, and grade retention, ultimately the best predictor of reclassification is the amount of time students have been receiving services. Policy recommendations are provided to support decision-making and resource-allocation, particularly for students most at risk of remaining ELs for a prolonged period of time. Future research ideas are also discussed.
# TABLE OF CONTENTS

ACKNOWLEDGMENTS ........................................................................................................... v  

ABSTRACT ............................................................................................................................. viii  

LIST OF TABLES ................................................................................................................... xi  

LIST OF FIGURES .................................................................................................................. xiii  

CHAPTER  

1. INTRODUCTION ................................................................................................................ 1  

1.1. Background ..................................................................................................................... 1  

1.1.1. The EL Lifecycle in US K-12 Settings ....................................................................... 4  

1.1.2. Different Ways of Grouping ELs ............................................................................. 8  

1.2. Statement of the Problem .............................................................................................. 10  

1.3. Significance of the Problem ......................................................................................... 11  

1.4. Purpose of the Study ..................................................................................................... 12  

2. LITERATURE REVIEW .................................................................................................... 14  

2.1. What is the Overall Purpose of the EL Classification System? ................................. 14  

2.2. How Should ELP Be Defined and Measured? ................................................................. 17  

2.3. How Should Reclassification Standards be Set? .......................................................... 20  

2.4. How Should Reclassification Be Studied? .................................................................... 24  

2.5. Previous Studies of Reclassification: Who Gets Reclassified, and When? ............... 27  

2.6. Common Findings Across Reclassification Studies ....................................................... 39  

2.7. Conclusions ................................................................................................................... 42  

3. METHODOLOGY .............................................................................................................. 46  

3.1. Data Characteristics and Variables ................................................................................ 49  

3.1.1. Student Sample .......................................................................................................... 50  

3.1.2. Reclassification Criteria ............................................................................................ 54  

3.1.3. Time Variables .......................................................................................................... 57  

3.1.4. Missing data ............................................................................................................... 62  

3.2. Data Analyses ................................................................................................................ 64  

3.2.1. Baseline discrete-time models .................................................................................. 65  

3.2.2. Duration Dependency and Time Specifications ......................................................... 66  

3.2.3. Factors Affecting Survival ......................................................................................... 69
3.2.4. Model Comparisons and Interpretation ........................................72

4. RESULTS ........................................................................................................79

4.1. Descriptive Survival Rates .........................................................................79

4.1.1. Observed Hazard and Survival by School Year ....................................79
4.1.2. Observed Hazard and Survival by EL Year ...........................................81
4.1.3. Characteristics of Censored Students .................................................82

4.2. Fitted Model: Duration Dependency and Time Specifications ....................87

4.2.1. Main Effect of Time ..............................................................................87

4.3. Fitted Model: Factors Affecting Survival ..................................................90

4.4. Fitted Model: Final Model and Interpretations ..........................................93

4.4.1. Time – Main Effects ...........................................................................95
4.4.2. Cohort, Sex, Immigrant Status, and Off-grade Status – Main Effects and Interactions ..........................................................97
4.4.3. Home Language Main Effects .............................................................98
4.4.4. District and Mobility Main Effects ......................................................99
4.4.5. Language Instructional Program – Main Effects and Interactions .......100
4.4.6. Time in the EL Subgroup – Main Effects and Interactions ..............103

4.5. Preliminary Summary of Results ...............................................................104

5. DISCUSSION ...................................................................................................124

5.1. Results by Research Question ....................................................................124

5.1.1. Reclassification probability over time, grade-level..........................125
5.1.2. Reclassification probability based on time in the EL subgroup ......127
5.1.3. Factors affecting probability of reclassification ...............................128

5.2. Policy Implications and Recommendations .............................................130

5.2.1. Considerations for Reclassification Policies .....................................130
5.2.2. Other Policy Considerations ............................................................133
5.2.3. Identifying and Supporting At-Risk Students ....................................136

5.3. Limitations and Further Research ............................................................136

APPENDIX: FOLLOW-UP STUDY .................................................................140

REFERENCES ..................................................................................................160
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1. Summary of prior survival analysis designs and findings</td>
<td>44</td>
</tr>
<tr>
<td>3.1. Summary of baseline year student characteristics, by cohort</td>
<td>75</td>
</tr>
<tr>
<td>3.2. Cross-tab of school years and grade-levels</td>
<td>76</td>
</tr>
<tr>
<td>3.3. Summary of missing data, by cohort</td>
<td>76</td>
</tr>
<tr>
<td>3.4. Covariates affecting conditional probability of reclassification</td>
<td>77</td>
</tr>
<tr>
<td>4.1. Life table by school year</td>
<td>109</td>
</tr>
<tr>
<td>4.2. Life table by EL year</td>
<td>110</td>
</tr>
<tr>
<td>4.3. Summary of censored student characteristics, based on timing</td>
<td>111</td>
</tr>
<tr>
<td>4.4. Summary of time specification models by school year (SY) and EL year (EL)</td>
<td>112</td>
</tr>
<tr>
<td>4.5. Model Fit and Comparisons for Different Predictor Specifications</td>
<td>113</td>
</tr>
<tr>
<td>4.6. Final point estimates for the effect of time on reclassification</td>
<td>114</td>
</tr>
<tr>
<td>4.7. Final point estimates for the effect of sex, immigrant, and off-grade status</td>
<td>114</td>
</tr>
<tr>
<td>4.8. Final point estimates for the effect of home language on reclassification</td>
<td>115</td>
</tr>
<tr>
<td>4.9. Final point estimates for the effect of district resources on reclassification</td>
<td>115</td>
</tr>
<tr>
<td>4.10. Final point estimates for the effect of language instruction program type</td>
<td>116</td>
</tr>
<tr>
<td>4.11. Final point estimates for the effect of time in the EL subgroup on reclassification</td>
<td>117</td>
</tr>
<tr>
<td>4.12. Incremental $R^2$ Increase Per Predictor, Controlling for Time Only</td>
<td>117</td>
</tr>
<tr>
<td>A.1. Distributional characteristics of proficiency-centered ELA and math scores by year, 2007 cohort</td>
<td>151</td>
</tr>
<tr>
<td>A.2. Distributional characteristics of proficiency-centered ELA and math scores by year, 2008 cohort</td>
<td>152</td>
</tr>
<tr>
<td>A.3. Distributional characteristics on proficiency-centered ELP assessments by year, 2007 cohort</td>
<td>153</td>
</tr>
<tr>
<td>A.4. Distributional characteristics on proficiency-centered ELP assessments by year, 2008 cohort</td>
<td>154</td>
</tr>
</tbody>
</table>
A.5. Accommodation use by type, content area, school year ........................................... 155
A.6. Summary of Missing Data, by Cohort ........................................................................ 156
A.7. Summary of Attrition, by Cohort ............................................................................... 157
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 Identification and reclassification lifecycle for an English learner</td>
<td>13</td>
</tr>
<tr>
<td>3.1. Number of students eligible for reclassification each year, by school year</td>
<td>78</td>
</tr>
<tr>
<td>3.2. Number of students eligible for reclassification each year, by EL year</td>
<td>78</td>
</tr>
<tr>
<td>4.1. Fitted probability of reclassification by school year for various time specifications</td>
<td>118</td>
</tr>
<tr>
<td>4.2. Cumulative failure rates based on sex, immigrant, and off-grade status</td>
<td>119</td>
</tr>
<tr>
<td>4.3. Cumulative failure rates based on home language</td>
<td>120</td>
</tr>
<tr>
<td>4.4. Cumulative failure rates based on district resource needs</td>
<td>121</td>
</tr>
<tr>
<td>4.5. Cumulative failure rates based on language instruction program type</td>
<td>122</td>
</tr>
<tr>
<td>4.6. Cumulative failure rates based on time in the EL subgroup</td>
<td>123</td>
</tr>
<tr>
<td>4.7. Fitted hazard rates by EL year for students in different school years (grades)</td>
<td>123</td>
</tr>
<tr>
<td>A.1. Average proficiency-centered ELA scores by school year, cohort</td>
<td>158</td>
</tr>
<tr>
<td>A.2. Average proficiency-centered math scores by school year, cohort</td>
<td>158</td>
</tr>
<tr>
<td>A.3. Average proficiency-centered ELP scale scores by school year, cohort</td>
<td>159</td>
</tr>
</tbody>
</table>
CHAPTER 1
INTRODUCTION

1.1. Background

Approximately 10 percent of the US K-12 population consists of students who are learning the English language in addition to learning academic content in areas like English language arts (ELA) and mathematics. The goal for these students, referred to in this paper as English learners (ELs), is to learn English as quickly as possible so they can participate in academic settings where English is used without needing special support. An additional goal is to ensure that these students all meet the same challenging academic content and performance standards expected of the general population. An implication of the first stated goal (of linguistic proficiency) is that, eventually, ELs will be folded into the general population for instruction and assessment. This expectation of reclassification, as the transition from EL to former EL is called, makes ELs unique relative to other special populations such as students with disabilities, because it implies that their special status is intended to be temporary.

The timing and criteria associated with reclassification can have direct and far-reaching impact on ELs’ educational opportunities and later achievement. On the positive side, reclassification can open the door to more challenging academic coursework (e.g., Advanced Placement courses) that may have been inaccessible to students as ELs due to scheduling conflicts or lack of teacher capacity. On the negative side, reclassification

1 Students in this subpopulation are referred to with many names including limited English proficient (LEP), English as a second language (ESL), English language learners (ELLs), emerging bilinguals (EBs), and English learners (ELs). I use the term English learner (EL) throughout this paper because of its person-first, non-deficit orientation, and because it is the terminology currently used by the US Department of Education.

2 This is not to say that a student’s having been an EL in the past does not remain relevant as a characteristic and possible covariate of that student’s achievement.
typically signifies the end of a student’s access to linguistic instruction, linguistic supports in the classroom, and accommodations on assessments; obviously, premature removal of any of these supports could have detrimental effects on achievement. States are essentially tasked with balancing these pros and cons so that ELs are retained in the subgroup for long enough that they will not suffer setbacks when their supports are removed, but not so long that they emerge from the subgroup unprepared to meet standards for college and career readiness.

Given these stakes, states and districts have a vested interest in designing their reclassification process so that it threads the needle between these two undesirable consequences. This requires two important considerations in particular: first, an appropriate cut score must be chosen for the level of ELP that ELs should obtain before transitioning. It should be high enough that their language no longer hinders their academic performance, but not so high as to be prohibitive. Second, a decision must be made about what aspects of a student’s linguistic or academic performance should (and should not) be considered as part of the reclassification decision. Many potential measures and scores are available to use (e.g., separate scores for the four language domains of reading, writing, listening, and speaking; composite scores combining some or all of these domains; scores from academic content assessments; grades or GPA, etc.), and each has its own set of pros and cons for inclusion. All of these considerations can and should be supported by data and research.

In practice, however, such data – and the research applied to them – can be difficult to parse, study, and interpret correctly. This is because ELs are a diverse group of students, both as individuals, and as a cohort over time. As individuals, ELs speak
different home languages, come from different backgrounds, and receive different types of linguistic instruction and supports during their time in the EL subgroup. Importantly, they also become ELs in the first place at different ages and with different levels of ELP. For all of these reasons, ELs spend different amounts of time in the EL subgroup; they also progress towards the goal of proficiency and reclassification at different rates, and they have different needs along the way. Despite this diversity, accountability reporting lumps all ELs together as one subgroup, differentiated only by grade-level. By ignoring other differences within the EL subgroup, current accountability models can mask the fact that the EL subgroup is a constantly shifting cohort, with students entering and exiting at different times, for different reasons, and having followed different journeys.

The study described in this paper will present some alternate approaches for grouping and tracking ELs over time. Using longitudinal data from a cohort of ELs in one state, it includes analyses designed to explore patterns in students’ progress, performance, and reclassification between grades 3 through 8 – critical years for accountability testing and reporting. Over this important period, I focus on students’ probability of being reclassified as a function of different time-based groupings; I also will investigate interactions such as the effect of student-level characteristics on the length or trajectory of the journey to reclassification, or how a student’s relative position within his or her journey may affect his progress or performance.

First, however, I provide more background and context about how ELs are educated and assessed in US schools. As these opening paragraphs suggest, this context is complex, and unique to this particular subgroup; it is, however, important for readers to understand this context for the analyses that follow to have meaning and value.
Accordingly, the next two sections will provide more detail regarding (1) how ELs are identified, served, and reclassified, (2) the many ways that ELs can be both differentiated from one another, and grouped together.

1.1.1. The EL Lifecycle in US K-12 Settings

Keeping in mind that there are many within- and across-state variations, the typical process by which ELs are identified, served, assessed, and reclassified is illustrated in Figure 1.1, and is discussed subsequently.

Students are initially **identified** as ELs via the school enrollment process. Most schools administer a home language survey (HLS) to new families to collect information about students’ language background. If a student’s HLS suggests that s/he uses or is exposed to non-English languages at home, he or she is referred to take a so-called “screener” assessment of ELP. The purpose of this screener is to determine whether the student’s language skills are likely to affect his or her ability to learn, to meet academic performance standards, or to “participate fully in society” when English is used. Students who demonstrate proficiency on this screener assessment are released back into the mainstream student population as initially fluent English proficient (IFEP) students, while students who score below their state’s proficiency threshold are labeled as ELs and referred for services.

The way ELs are **served** in the period between their identification and reclassification is a matter of district- or school-level discretion, though some states have passed laws mandating or prohibiting particular types of instructional programs (Parrish et al., 2006; Smith, Coggins, & Cardoso, 2008). There is also specific federal guidance prohibiting the segregation of ELs into dead-end or sub-par programming based on their EL-status. Broadly speaking, most language instruction programs can at least be
categorized as either English-only or bilingual, where a language other than English is also used for some instruction in the latter case (Genesee, 1999). Families also may refuse the offer of language services for their child, though districts still face a legal responsibility to ensure that all ELs, including those not receiving specialized services, have equal access to the same educational opportunities offered to all students. Under the two most recent reauthorizations of the Elementary and Secondary Education Act (ESEA), they are also still responsible for ensuring that these students meet the state’s performance standards in ELA, mathematics, and science.

However they are served, all ELs, as mentioned previously, are expected to transition out of the EL subgroup (i.e., be reclassified) eventually. Thus, starting in the year they are identified, all K-12 ELs are assessed annually in ELP, to track students’ progress towards proficiency and determine whether they are ready to be reclassified. Reclassification process models will be described in more detail in the literature review, but typically, students must meet a certain performance standard on the state ELP assessment to be considered for reclassification. Reclassified students are often referred to as reclassified fluent English proficient (RFEP).

It is important to note that the use of statewide summative ELP assessments was new in the era of No Child Left Behind (NCLB). Prior to NCLB, states typically did not have standardized or coordinated systems for defining ELP or deciding when ELs had achieved it; states or districts that did opt to use ELP assessments did so at their own discretion, and were not subject to federal regulation or monitoring. NCLB introduced, for the first time, mandated systems of statewide standards, assessments, and accountability for language proficiency.
Today, all states have articulated K-12 standards for ELP, which inform the assessment content and cut scores that guide the reclassification process\(^3\). The ELP standards are required to be “aligned” to the state’s academic content standards, in the sense that they represent the linguistic content and uses found in academic classrooms and on academic assessments. (Notably, ELP experts have advocated for using the term “correspondence” to describe this relationship, since it differs in nature from the more typical “alignment” relationship between standards and assessments of those standards.) Similarly, the cut scores and performance standards on the state ELP assessment are expected to be set so that a proficient performance signals mastery of the types of language a student would need to meet their state’s academic content and performance standards.

As the final step of the EL lifecycle, the law also requires that RFEP students be monitored for two years after they are reclassified, to ensure that the reclassification decision is serving the student’s best interests. Again, the content and nature of this monitoring process varies from school to school; in some settings, students may continue to receive minor instructional supports or accommodations, and may also continue to participate in the statewide ELP assessment. In others, monitoring may only consist of teachers checking students’ academic performance at the end of the year. Regardless, federal guidance allows states to include monitored RFEP students as part of the EL subgroup for academic accountability reporting purposes; many states follow this option, since RFEP students are likely to earn higher scores on academic assessments, thereby

\(^3\) Note that in the context of ELP, “content standards” refer to linguistic and communication skills and knowledge, just as content standards in academic areas such as mathematics refer to mathematical skills and knowledge.
helping to pull up the average performance of the EL subgroup. A student who makes it through his monitoring period without being referred back for more service will then exit the EL subgroup entirely, and cease to be included in its numbers for any further reporting.

Although the monitoring period marks the end of the federally mandated EL lifecycle, a student’s ever having been an EL may remain a characteristic of interest for policymakers, researchers, and educators alike. While it is important to track ELs’ achievement up to and including their point of reclassification, their performance after this point – even when they are no longer considered part of the special population – provides important evidence and feedback about the adequacy of their preparation and the reclassification standards and policies. Accordingly, a 2010 working group specifically highlighted the importance of tracking the academic achievement of ELs throughout their academic careers (not just during their time as ELs), as their post-reclassification achievement may be one of the most important indicators of language program quality, or equity within the school system (Working Group on ELL Policy, 2010).

One final critical point about the EL lifecycle is the complete absence of standardization across, and even within some states (see, e.g., Bailey & Kelly, 2013; Linquanti & Bailey, 2014; Linquanti & Cook, 2015; Mahoney & MacSwan, 2005; Ragan & Lesaux, 2006). Although the federal government mandates identification, service, assessment, reclassification, and monitoring, it does not go so far as to tell states how
they must achieve these requirements. As a result, every step, process, instrument, criterion, and timeline in the cycle is open-ended and subject to interpretation.

Particularly since NCLB marked the first time that states became involved as enforcers with respect to EL policies, many states opted to grant districts and schools considerable freedom and discretion to define and implement their own systems. In practice, this has meant that ELs who move across and even within states often will find themselves subject to different service systems with different criteria; as a result, such students might find themselves abruptly moved from one stage in the lifecycle to another, as a function only of geography or policy. This also means that ELs – at least those who move at some point in their K-12 lifetime – do not necessarily follow linear paths through the EL lifecycle, and may skip or repeat phases based on policy differences in their new schools.

1.1.2. Different Ways of Grouping ELs

All ELs follow some version of the life cycle described in the previous section, but it is important to also recognize the diversity of the students who become members of this group. While all ELs share the label of their subgroup – and, by extension, have been identified as needing help to reach an acceptable level of ELP – they also differ from one another in some important ways. First, as referenced in the introduction, ELs speak different home languages, have different household and demographic backgrounds, and receive different types of services to support their English development over time.

More importantly, however, are the ways in which ELs differ substantively, in terms of their ELP status and growth. It is not uncommon to see or think of terms like

---

4 Although NCLB does not include specifics about EL services, other federal documents such as non-regulatory guidance, memoranda, and federal court decisions, do provide some further specifics about minimum service requirements, illegal or inappropriate practices, and standards for program evaluation.
“beginner” or “advanced” to describe EL students, but it is important to disentangle three related concepts that these terms can reference: namely, a student’s (1) grade-level, (2) level of ELP, and (3) time since identification. In particular, one might expect to find differences in progress and performance among:

1. **Younger vs. older ELs**, who differ from one another in terms of their age and grade-level;

2. **Newer vs. long-term ELs**, who differ from one another in terms of how long they have spent in the EL subgroup; and

3. **Beginner vs. advanced ELs**, who differ from one another in terms of their level of English proficiency.

It is important to recognize that students who are similar in one of these traits may be very different on another – for instance, two third grade ELs might have very different levels of ELP, or two intermediate ELs may be in different grade-levels, or have spent very different amounts of time in the EL subgroup. Not all newly identified ELs are young, or are beginners, and not all beginners are young, or are newly identified.

These differences are relevant because, although all three of these dimensions matter for interpreting a student’s performance and needs, accountability reporting only takes one of them into account: grade-levels. Thus, as both individuals and a group, ELs’ performance in all areas – ELP, ELA, and mathematics – is defined, measured, and reported based on expectations for that student’s grade-level. Similarly, their progress over the past year or expected progress for the future do not take into account their current level of ELP or how long they have been an EL. This means, by extension, that when states or researchers use accountability data to study how long ELs take to reach
proficiency, or how their ELs are performing in academic content areas, or whether language affects content performance, they are usually studying ELs who have been grouped together based on their grade-levels, without accounting for these other ELP-relevant factors. One of the central premises of this study is that grouping ELs in this way, while expedient, may mask or distort important information about these students’ progress, performance, and needs.

1.2. Statement of the Problem

The present study sits at the intersection of the topics that have been summarized thus far: the EL lifecycle, reclassification, and different ways of grouping ELs to create relevant subgroups within the population. To date, while numerous studies have evaluated reclassification rates and covariates that affect them, none have systematically compared the outcomes or interpretations associated with different ways of grouping ELs. In other words, studies group students by either their time in the EL subgroup, or their grade-levels, or their ELP, but few have asked the question which of these groupings is more useful or informative for decision-making.

In addition, while a great deal of research has been conducted in the past few years about reclassification and longitudinal growth for ELs, most of the best-designed studies have used cohorts that start in kindergarten. Coincidentally, kindergarten is the one point at which a student’s grade and time in the subgroup are more or less conflated, meaning the beginning of a student’s EL status maps directly onto the beginning of their schooling. Such cohorts are excellent for developing a baseline sense of the amount of time it takes students to be reclassified, and the trajectories they follow to get there. Their generalizability is somewhat limited, however, because not all ELs begin in kindergarten.
For students who do not begin “at the beginning,” we might expect to see different developmental patterns; or, at least, we should ask whether the same patterns hold. For example, as I will discuss in the literature review, some studies have developed estimates of the average amount of time it takes students to achieve English language proficiency; others have identified certain schooling moments (e.g., the transition from elementary to middle school) as critical moments for transition. It seems worth asking: to what extent are such findings conditional on students’ having started in kindergarten, and to what extent are they generalizable to ELs who join the subgroup later? What milestones and patterns might we observe for students who become ELs in grade 2 or grade 3? Do they tend to follow parallel, but delayed, trajectories, or do we see entirely different patterns and milestones for these students?

1.3. Significance of the Problem

Reclassification is a critically important moment in an EL student’s career, as it directly affects his or her access to appropriate instruction and, by extension, his or her later academic achievement. As such, reclassification standards should be set carefully to ensure they are maximally beneficial to students, and evaluated regularly to ensure that students on both sides of the divide – meaning, those who are retained, and those who are reclassified – are served well by the outcomes of the reclassification criteria.

Evaluating reclassification criteria requires exploring two families of questions. First, which students meet the threshold, when do they reach it, and what factors affect their probability or timing of doing so? Secondly, for students who do reach the reclassification threshold, do we see evidence that they are well-prepared for the academic environments in which they will find themselves after exiting EL status? Both
sets of questions are critically important to EL success and reclassification validity; this paper will focus primarily on the former, however.

1.4. Purpose of the Study

The purpose of the current study is to explore reclassification patterns in a longitudinal panel of ELs from one state. It is guided by the following research questions:

1. What is the probability of reclassification over time for a cohort of ELs progressing from grade 3 to grade 8?
   a. When are third grade ELs most likely to be reclassified?
   b. Does the probability of reclassification vary by either grade-level, amount of time in the EL subgroup, or both?

2. Are students who have been in the EL subgroup for a long time (>5 years) more or less likely to be reclassified than more recently identified students?

3. What factors affect the probability of reclassification for students in the same grade-level, or who have been ELs for the same amount of time?
Figure 1. Identification and reclassification lifecycle for an English learner
CHAPTER 2

LITERATURE REVIEW

This chapter will build on the first by summarizing research on the major topics that have been discussed thus far, as well as on the quantitative methods that will be used in this study. I begin with a brief summary of current thinking about how reclassification should be defined and approached, followed by a summary of research on how different ways of defining ELP may affect who gets reclassified. Since reclassification is supposed to occur when language no longer exerts undue influence on students’ academic achievement, this is followed by a brief summary of research on how language and content performance relate for ELs generally, and subgroups of EL in particular, based on factors such as their grade-level and level of ELP.

Next, I discuss how research about reclassification has evolved over the past 15 years in terms of methods, data structuring, and interpretations. I argue that the field has experienced something of a learning curve since NCLB’s implementation in understanding how research of ELs’ performance can best be designed to produce valid and useful results. Having addressed this, I then use the final section to summarize literature about progress and performance leading up to reclassification, including factors affecting the likelihood of reclassification. Taken together, the topics reviewed in this chapter should set the stage for the methods I propose for this study, in chapter 3.

2.1. What is the Overall Purpose of the EL Classification System?

*Lau v. Nichols* was a 1974 Supreme Court case in which Chinese-American students in San Francisco sued their district based on its language policies. Their central complaint was that the district required students to be proficient in English to graduate, but did not provide support or instruction to help students develop English proficiency.
Delivering the Supreme Court’s opinion in the case, which the students won, Justice William O. Douglas wrote,

there is no equality of treatment merely by providing students with the same facilities, textbooks, teachers, and curriculum; for students who do not understand English are effectively foreclosed from any meaningful education.

(Lau v. Nichols, 1974, p. 566)

This court decision, which was decided based on Title VI of the Civil Rights Act of 1964, persists as the foundation for contemporary EL education policy. As Douglas’ wording suggests, the primary concern for ELs is essentially to prevent such educational “foreclosures,” and ensure that they are not excluded from educational opportunities on the basis of their linguistic proficiency. Because language has been established by courts as a proxy for national origin, excluding ELs based on their linguistic proficiency would amount to a violation of these students’ civil rights – Title VI specifically prohibits exclusion on the basis of “race, color, or national origin” (The Civil Rights Act of 1964. Pub. L. 88-352, 1964)

NCLB defines ELs as students whose difficulties in speaking, reading, writing, or understanding the English language may be sufficient to deny the individual the ability to meet the State’s proficient level of achievement on State assessments described in section 1111(b)(3); the ability to successfully achieve in classrooms where the language of instruction is English, or the opportunity to participate fully in society. §9101(25)(D)

Echoing the Lau language, the NCLB definition also makes an explicit link between language, inclusion, and opportunity to learn (or perform). Thus, as mentioned in the first chapter, the purpose of the EL identification and reclassification system as a whole is to identify students whose language skills threaten to exclude them from the curriculum, neutralize that threat through effective language instruction, and then remove the EL
label so that the student can participate in “regular” educational settings as a “normal” student.

This legal context is important for understanding reclassification, because it also provides the foundation for how to determine when a student should not be considered an English learner: specifically, when it can be shown that a student’s opportunities to learn or demonstrate academic proficiency are not compromised by their linguistic proficiency. The challenge, of course, is accurately judging when this criterion has been reached. As the varied performance of native speakers shows, even students with full linguistic access to instruction may yet fail to meet the academic standards set for them. Thus, one of the trickiest aspects of reclassification is figuring out how to isolate the role of language, specifically, in students’ performance.

In the next two sections, I discuss two specific aspects of this challenge. The first is defining linguistic proficiency, both conceptually, and in terms of how it should be scored and reported from a test. Because language is a multi-faceted construct involving skills in numerous semi-distinct domains (e.g., speaking, writing, comprehension, etc.), it can be measured, scored, and reported in multiple ways. The way domains are combined and weighted to represent proficiency, however, affects which students are deemed proficient and eligible for reclassification. The second aspect is how relationships between language and content should be considered in making reclassification decisions. It seems intuitive that content performance should be considered somehow, since a lack of access to content is part of the EL definition. Different ways of incorporating it, however, also affect which students are reclassified, as well as the reasons for which students may remain ELs over time.
2.2. How Should ELP Be Defined and Measured?

Based on the phrasing of NCLB, states report anywhere from two to nine different scores from their ELP assessments. Currently, the most common model (used in 42 states) is to report separate scores for the four domains of reading, writing, listening, and speaking, plus a composite comprehension score (a combination of listening and reading) and overall proficiency (a combination, sometimes weighted, of all four domains) (Faulkner-Bond, Shin, Wang, Zenisky, & Moyer, 2013). Although states may report all of these scores to students and teachers, they may use only some for the specific purpose of ELP standard-setting and reclassification. Thus, the first decision a state must make in designing a reclassification process is which ELP measures to consider for reclassification, and how to combine them.

One option is to combine all ELP subscores into one overall proficiency score, and use this score as the sole indicator of students’ ELP. The use of an overall score represents a compensatory approach, which is premised on the idea that relative weaknesses in certain language domains can be tolerated if they are complemented by strengths in other language domains. If this does not seem believable (conceptually, or in practice), states also may choose to make the overall proficiency score a weighted average of the four domains, with certain domains given more weight if they are shown or believed to be more important for academic success. In Texas, for example, 75 percent of a student’s overall ELP level is based on their performance on the state’s ELP reading test (Texas Education Agency & Pearson, 2012). One advantage of the compensatory approach is that the overall score is based on the most items and is thus the most reliable
score from a decision consistency standpoint. In other words, there is less of a chance that students might be classified differently if they were to retest.

The next step along this continuum is what Carroll and Bailey (2015) refer to as a combination decision rule, where students who meet an overall ELP standard also must score above certain minima in each domain to be reclassified – e.g., a student must have an overall ELP level of “proficient” and not score in the lowest performance level for any language domain. Or, finally, if it is believed or shown that students need to demonstrate mastery in all domains to be truly proficient, states may use a conjunctive decision rule, where students must earn proficient scores on all ELP subtests to be eligible for reclassification. A drawback to combination and conjunctive models is that they require making decisions based on subtests that may be relatively short (e.g., 15 to 20 items). Since test length directly affects reliability and standard errors of measurement, there is a higher chance that students might hit or miss a cut score due to random error, rather than actual linguistic proficiency (or a lack thereof). This fact can blur the distinction between students who are and aren’t reclassified – that is, students who are more or less indistinguishable in their performance may end up on either side of the cut score, simply due to measurement error.

Although standard setting panels may consider the practical impact of different proficiency models when making their final decisions, there are few published studies that have directly compared the impact of different proficiency definitions on student outcomes and performance. In a descriptive study of 875 ELs and 92 non-ELs in grade 5, Carroll and Bailey (2015) illustrated how decision rules can lead to sizable differences in the number and proportion of EL students who are recommended for reclassification. In
particular, they found that conjunctive rules that impose minimum performance requirements for all subtests (e.g., students must meet a cut score on all domain subtests) tend to identify the most students as non-proficient, whereas compensatory models that focus only on the overall proficiency score identify the fewest non-proficient students.

Some studies with California data have also evaluated which scores tend to be the limiting factor for students’ transitioning at different points in time. California uses a complex model, where students must meet three separate criteria: (1) an overall ELP score in the state’s performance level 4 (out of 5), (2) scores that are not below level 3 (out of 5) on each domain subtest, and (3) performance in at least level 2 (out of 4) for the state’s ELA assessment. In a longitudinal study of over 200,000 ELs from the Los Angeles Unified School District (LAUSD), Thompson (2012) found that the ELP reading subtest was the most likely barrier to reclassification for her students up through grade 5, after which the ELA content assessment became the limiting score. Robinson (2011) reached a similar conclusion in a study of 39,736 California ELs (all from one unnamed district), where he also observed that the ELP reading score held back the largest proportion of potentially eligible students in fourth, fifth and sixth grade (between 30 and 40 percent of students), but the ELA assessment limited between 40 and 50 percent of eligible students for grades 7 through 10. Umansky and Reardon (2014) also observed that the ELA content assessment became the limiting factor for reclassification starting in 6th grade in their study of 5,423 California ELs from one district.

Findings like these underscore the impact – intended or unintended – of using a conjunctive model compared to a compensatory one. For example, since neither Thompson or Robinson observed the overall score acting as a barrier for a large number
students, this suggests that, were California to use a compensatory model, as opposed to its conjunctive one, many more students would have been reclassified than under the current system. This is neither good nor bad per se, but does serve to illustrate how different ways of defining ELP can affect students’ time in the EL subgroup. These findings also point to the impact of incorporating content assessment scores into reclassification decisions, in addition to ELP scores; I discuss this idea further in section 2.3.

2.3. How Should Reclassification Standards be Set?

Setting reclassification standards requires both defining ELP and deciding what type of evidence is necessary to determine that it has been achieved. Needless to say, it is both complicated and high-stakes, as it affects which students make it through, and how reclassified students go on to perform as RFEP students.

As a baseline, Wolf and Farnsworth (2013) stress the importance of establishing the validity of the ELP assessment itself, as reclassification can only be as fair and as valid as the scores that form its primary – or perhaps only – criterion. In particular, they emphasize the importance of (1) appropriately articulating the ELP construct in terms of the language knowledge and skills necessary for academic contexts, and (2) properly aligning ELP standards with ELP assessments, including establishing correspondence between ELP standards and academic standards. They recommend that states collect evidence to support the reliability, construct validity, and consequences of the ELP assessment’s use for reclassification purposes. They also encourage states to consider impact data for different standards and models to see how they affect the number of students who transition, how long students spend as ELs, how students who are
reclassified go on to perform, etc. In a study of ELs from various grade levels in three different states, Kim and Herman (2009) explored these types of questions by seeing what the predicted content score was at the reclassification cut point, and comparing this predicted value to the average performance of non-ELs.

Beyond the ELP assessment itself, states must decide which scores to use, both from the ELP assessment, and possibly from other sources. In some states, ELP and reclassification are essentially synonymous, while in others, achieving ELP is a trigger to look for further evidence to make a reclassification decision. Ragan and Lesaux (2006) studied the identification and reclassification criteria used in the 10 states and 10 districts with the largest EL enrollments (at the time), and found that eight of the ten states and five of eight districts considered additional criteria in exit decisions such as academic test scores, grades, or teacher or committee recommendation. Generally, some states, such as California and Massachusetts, recommend that districts consult other measures for students who have met the state’s ELP standard (California Department of Education, 2015; Massachusetts Department of Elementary and Secondary Education & DePascale, 2012), whereas others, such as Oregon, Arizona, and the state used for this study, rely solely on ELP scores, as a rule.

The use of additional scores can obviously affect reclassification rates, as suggested with the examples from California referenced in section 2.2. Or, to use an example from a different state, Carroll and Bailey (2015) illustrate that a stepwise decision process, in which ELP scores serve as the first indicator for reclassification considerations, can also limit which students have the opportunity to transition. In particular, they found that a proportion of ELs in their sample of 875 demonstrated
proficient or advanced performance on all of their state’s academic content assessments, but fell short of meeting ELPA performance criteria. Such students – whose numbers varied from 2 to 34 individuals, depending on the ELPA decision rule used – would not have been considered for reclassification in their state since ELPA scores are considered first. The authors suggest, in response, that states should consider both language and content performance simultaneously – rather than sequentially – to ensure that students who may be eligible for reclassification are not overlooked.

Cook, Linquanti, Chinen, and Jung (2012) propose a different solution, which is to examine the relationships between ELP scores and content scores, rather than the scores themselves. They specifically recommend that:

Researchers can define “English language proficient” as the point at which EL students’ academic content achievement assessed using English becomes less related to their ELP. That is, there is a point at which EL students have sufficient English language skills to adequately function in English on content assessments; accordingly, there should be observable decreases in the relationship between the two assessments. At or beyond this point is where the ELP performance standard might be considered… (Cook et al., 2012, p. 8)

They propose three quantitative methods for empirically identifying this turning point using real data: decision consistency, logistic regression, and descriptive box plots. All three models require sorting or grouping ELs based on their level of ELP, and then plotting the language-content relationships for these different groups. For decision consistency, they plot the percentage of students scoring in the proficient range in both ELP and content for different levels of ELP, and recommend setting the cut score at or around the level of ELP that has the highest percentage of proficient-proficient agreement. For logistic regression, they predict the probability of scoring in the proficient range on each content assessment, conditional on ELP score. They recommend setting the
cut score at the point where the probability of meeting the ELA standard is at or above chance (0.5). For the descriptive box plots, they recommend creating box plots of content performance for each ELP performance level, and setting the cut score at or around the ELP level that is closest to being centered on the content performance cut score (i.e., half the students in the ELP level score above the content cut, and half score below).

Cook et al. (2012) illustrate all three of their methods using two-year data samples for grades 4, 7, or 10 from three different states ($n$-sizes ranged from 1,120 to 2,563 depending on the grade-level and state). Among other things, their illustration demonstrates both that the predicted point of divergence does occur in all samples, and, equally important, that its placement varies from state to state and grade-level to grade-level based on factors such as the state’s ELP standards, content standards, and the linguistic complexity of the test forms themselves. They also note that their design is premised on the fact that academic content cut scores are typically non-negotiable when ELP cut scores are set; thus, the resulting cut scores should be considered to maximize desirable outcomes conditional on academic content scores. Were both cuts being set concurrently, it is possible that both could be placed more optimally.

Having set standards for ELP and reclassification, the next step for any state or policymaker is to collect evidence for the validity of this standard. In this context, three particularly important validity considerations are (1) which students meet this standard, (2) how long it takes them to get there, on average, and (3) how they fare after they have met it and transitioned out of the EL subgroup. Due to the scope of the current study, only (1) and (2) will be discussed in depth, in the final section of this literature review.
First, however, I pause briefly to discuss some methodological considerations for how reclassification should be studied at all.

2.4. How Should Reclassification Be Studied?

In general, because of its mandates about standardized ELP assessment and disaggregated reporting, NCLB has played a major role in shaping research about ELs in the 21st century. In addition to codifying reclassification as a formalized event, NCLB’s mandates also forced states to pay attention to ELs’ progress and performance, and ensured that certain information about these students would be included in state-level datasets. In the fifteen years during which NCLB was law, research on these students proliferated (and continues to, despite a new ESEA reauthorization in December 2015). Over the same period, however, many experts came to recognize exactly how complex and difficult it can be to come up with appropriate designs and methods to get good, meaningful information about ELs’ performance and progress. Thus, the research of the past fifteen years also represents something of a learning curve for the field, as researchers have refined their thinking about how to validly structure their EL data, apply their models, and interpret their findings.

Choosing valid research designs to study reclassification can be surprisingly tricky, because of the way the EL subgroup works. Even before NCLB, Linquanti (2001) identified an important “redesignation dilemma,” which is that high achieving ELs are more likely to be reclassified, while struggling ELs are more likely to stay in the subgroup. This dynamic, while obvious, can distort the data, particularly since accountability data are primarily cross-sectional in nature, and do not track individual students over time. At the group level, this dynamic will make it consistently appear that
reclassified ELs do very well, which suggests that the EL programs and reclassification are functioning well and preparing students to succeed upon exiting. What it masks, however, is the fact that the same lower achieving ELs are remaining in the EL subgroup year after year without exiting. Thus, rather than each student making steady progress towards English proficiency, higher-achieving students are passing through and leaving these “long-term ELs” behind (Menken, Kleyn, & Chae, 2012; Olsen, 2010). This suggests, by contrast, that the programming is actually not working, precisely for the students who need the most help.

The primary implication of this dynamic is that typical accountability data, which groups ELs only by grade, is not enough on its own to give an accurate picture of ELs’ progress. A cross-sectional sample of third grade ELs, for example, might include students who have been ELs for 5 years, three years, and one year, and who have low, medium, and near-fluent levels of English proficiency. Substantively, it matters which of these students make progress or exit the subgroup. In particular, if students who have already received years of instruction are not likely to exit, this suggests that something is not working in either the instruction, the reclassification standard, or both.

A related secondary challenge in measuring ELs’ language development longitudinally is identifying the appropriate outcome measure for such studies. Since ELPA participation ceases within two years of being reclassified, and since reclassification may occur at any grade-level after any length of time in the subgroup, there is no common outcome datum for all ELs, other than reclassification itself. Given, again, the way EL status functions, this event could occur at any grade-level, and may come at the end of a trajectory of many or few years spent taking the ELP assessment and
participating in language instruction. Those differences matter, for design and interpretation.

It took a few years for researchers to appreciate these dynamics in their data. In time, however, states and researchers learned that the time an EL had spent in the subgroup needed to be accounted for in their designs, and the field has converged increasingly on a few appropriate methods for these purposes. First, descriptive analyses of state data are actually somewhat common, as they can effectively provide a portrait of who is transitioning, when, and after how long, even if they cannot answer questions about what factors predict reclassification. The remaining two sections include several studies that are based solely on descriptive analyses of state datasets (American Institutes for Research, 2013; Carroll & Bailey, 2015; Grissom, 2004; Massachusetts Department of Elementary and Secondary Education & DePascale, 2012).

Second, discrete-time survival analysis (also referred to as event-history modeling) has increasingly been used to study the time it takes for students to be reclassified. The primary advantage of survival analysis is its capacity to model censorship within a dataset – that is, to account for and model the fact that some subjects may not achieve the intended outcome within an observation period. In the context of reclassification, it can acknowledge that there is a difference between an EL who has been in the subgroup for 8 years and then is reclassified, and one who has been in the subgroup for 8 years and still has not been reclassified when the observation period of a study ends. Survival analysis also uses as its outcome the probability (referred to as the hazard) that a particular outcome will or will not happen for a given subject. Because the probability is the outcome, this solves the aforementioned problem of needing an
outcome variable to predict in these longitudinal studies. At least five studies to date have used survival analysis to study reclassification in the EL K-12 context, all of which will be discussed in the final literature review section (Cook et al., 2012; Parrish et al., 2006; Slama, 2014; Thompson, 2012; Umansky & Reardon, 2014).

2.5. Previous Studies of Reclassification: Who Gets Reclassified, and When?

Studies of reclassification typically focus on questions like how long it takes students to reach their state’s reclassification threshold, the proportion or likelihood of students transitioning in any given year, and factors that affect a student’s likelihood of being reclassified. Some studies have explored such questions descriptively, by simply looking at the proportion of students who are reclassified each year.

The Massachusetts Department of Elementary and Secondary Education (MDESE), for example, conducted a number of descriptive analyses of over 225,000 K-12 ELs who enrolled in the state’s public education system between 2002 and 2011 (Massachusetts Department of Elementary and Secondary Education & DePascale, 2012). They found that the proportion of students transitioning in each grade increased steadily from 10.1 percent in grade 1 up to a peak of 23 percent in grade 5, before decreasing again through middle and high school, save for one slight surge in grade 9. The highest transition rates occurred in grades 4 through 8. They also found that students spent 3.5 years as ELs, on average, with an interquartile range of 2 to 5 years. Contrary to most other studies discussed here – and possibly owing to regional differences in EL populations between Massachusetts compared to states like California – they found that Hispanic students (who are generally assumed to be Spanish speakers) spent less time as
ELs, on average, compared to speakers of other languages (2.3 years, compared to 3.9 years for non-Hispanic students).

The American Institutes for Research (AIR) conducted descriptive analyses of ELs from Oregon over a three-year observation period from 2009 to 2012 (American Institutes for Research, 2013). Their sample included 39,575 ELs who were in grades K-10 in the first data year and had complete data for the three-year period, either because they transitioned or because they remained ELs and had three years of test scores for analysis. AIR focused on students’ two- and three-year performance level patterns and calculated the percentage of students who followed various paths such as earning the same performance level in consecutive years, or exiting after earning a particular performance level in the previous year. They observed that between a third and a half of students at each performance level scored in the same level in consecutive years – for example, 46 percent of students scoring in Level 2 in 2009-10 scored in the same level again in 2010-11. This tendency was most pronounced for younger students in grades K-2. On the other hand, across all grades, 51 percent of students who scored in the second highest performance level (Level 4) in the first year went on to score in Level 5 in the next year, making them eligible for reclassification. These findings suggest that, in Oregon at least, students may take two years to progress through each performance level, at least until they reach Level 4.

In agreement with several other studies that are summarized in this section, AIR (2013) also found that late elementary school seemed to be a popular transition point for ELs. They observed that, compared to other grades and test forms, a higher percentage of students who took the grade 3-5 ELP test form were reclassified the next year, across all
performance levels. For example, although very few students who scored in Level 2 one year were reclassified the next year, this unusual pattern was more common for grade 3-5 students compared to other grades (9 percent of grade 3-5 students, compared to, for example, 5 percent of grade 6-8 students).

Grissom (2004) calculated the probability of reclassification over time for California ELs from grades 2 through 5 using three cohorts of data from 1998 through 2003 (pre-NCLB, notably). His cohorts ranged in size from approximately 58,000 ELs in the 1998 cohort to just under 79,000 in the 2000 cohort. In all three cohorts, he observed that less than a third of students had been reclassified by grade 5. Grissom did not have data on when the students in his cohorts were initially identified as ELs, so the most conservative reading of his findings is to consider grade 2 (his first year of data) as students’ first year as ELs. By this estimate, grade 5 represents their fourth year in the EL subgroup, at least; if students were identified earlier than grade 2 (which many probably were), grade 5 could represent five or six years of EL status. Notably, Grissom’s findings clash with several other studies described in this section, which found that fifth grade was often a peak reclassification year for ELs both within and beyond California. Consistent with some other studies, however, he found that slightly more girls transitioned each year than boys, that Spanish speakers were less likely to transition than speakers of other languages, and that lower-income students – identified based on their eligibility for free or reduced-price lunches (FRL) through the National School Lunch Program (NSLP) – were less likely to transition than non-eligible students.

While descriptive analyses of reclassification rates can provide a general sense of how likely reclassification is for certain groups of students, they are limited in their
generalizability and predictive value. As referenced in the preceding section, event history or survival analysis has become a go-to method for studies of this type. One of its strengths is the articulation of a set of continuous functions to describe the relationships among covariates and predictors for reclassification over time – these are known as the survival function and the hazard function (see section 3.2.1 for a more detailed discussion of these functions). These functions allow for stronger interpretations and comparisons, and can be useful to make predictions and evaluate possible outcomes. Several studies using this design are described next, and are also summarized in Table 2.1.

One of the earliest survival analysis studies of reclassification was by Parrish et al. (2006), in an evaluation study of a California law, Proposition 227 (Prop 227). Passed in 1998, Prop 227 is a controversial law that essentially mandated English-only instruction for ELs, and also set the expectation that ELs should be reclassified after one year. Among many analyses conducted to evaluate the law’s effects, these authors used survival analysis to determine how many students were meeting the law’s one-year expectation for reclassification. Although the methods are not spelled out in great detail, it appears that the authors used cross-sectional statewide California data for the 2002-03 school year (approximately 1.5 million ELs, though the exact number is not reported in the technical appendices), and grouped students according to the number of years they had been identified as ELs in US schools. Thus, the groups contained all ELs in grades K-12 who had been in the subgroup for one year, two years, three years, etc. The results of the survival analysis therefore represent the proportion of students who are reclassified after increasing numbers of years in the EL subgroup, regardless of their grade-level. For example, the Year 4 estimate would represent the proportion of students who were
reclassified after spending four years as ELs, regardless of what grade-level they were in on identification or reclassification.

They found that only 12 percent of students were reclassified after three years of instruction, and fewer than 40 percent had been reclassified after ten years of instruction. They acknowledge that their method of grouping students is somewhat imprecise, particularly since they could not differentiate between the number of years students had been ELs at all and the number of years they had been ELs in California, specifically – for the purpose of evaluating Prop 227, this would be an important distinction. Their estimates are lower than those found in other studies described here, but those differences could stem from the study’s cross-sectional design, or the fact that the data came from very early in NCLB’s passage. If nothing else, their findings certainly demonstrate that Prop 227’s expectation for reclassification after one year of instruction is unrealistic.

In the same report described above about studying relationships between ELP and content performance (see section 2.3), Cook et al. (2012) also conducted event history analysis using K-5 data from one state. These authors grouped their sample of 23,668 students by test form (K-2 or 3-5) and initial ELP (Level 1, 2, or 3) and then calculated reclassification rates over the next four years. Because of their grouping, this entailed calculating separate survival functions for six different groups: Students who took the K-2 form and scored in ELP Level 1 on their first administration; students who took the K-2 form and scored in ELP Level 2 on their first administration; students who took the 3-5 form and scored in ELP Level 1 on their first administration, and so on. Each group ranged in size from just under 10,000 ($n = 9,328$ for students in the K-2 group who
scored in ELP Level 3 on their first administration) to fewer than 200 ($n=168$ for students in the 3-5 group who scored in ELP Level 2 on their first administration).

The authors found markedly different, and telling, patterns of reclassification for their different groups. Not surprisingly, students who had higher levels of ELP on their first administration had very good odds of being reclassified within the observed period. Descriptively, over 50 percent of K-2 students with initial ELP at Level 3 had been reclassified by the next year, and 86 percent had been reclassified within 4 years. The predicted probabilities using survival analysis were comparable but slightly lower (e.g., a 52 percent chance predicted as opposed to a 57 percent chance observed). The 3-5 group was even more successful – over three-quarters (77 percent) were reclassified within one year, and nearly all (94 percent) had been reclassified within four years. The initial Level 1 students, by contrast, never exceeded a fifty percent chance of reclassification; their predicted probability of reclassification in fact never even exceeded 40 percent for either age group. On the one hand, because the data only cover a five-year period, it is possible that many of these students went on to be reclassified shortly beyond Cook et al.’s (2012) window of observation (e.g., in students’ Year 5). On the other hand, however, their findings paint a pessimistic picture of beginner ELs’ likelihood of reclassification.

Slama (2014), found slightly more optimistic results using data from Massachusetts ELs. She used survival analysis to study reclassification patterns in a sample of approximately 5,000 Massachusetts EL students using eight waves of data between kindergarten and grade 7. In contrast to the findings in Cook et al.’s (2012) state, and in agreement with the MDESE descriptive study described above (Massachusetts Department of Elementary and Secondary Education & DePascale, 2012), Slama (2014)
found that a majority of students in her dataset were reclassified by second grade (i.e., three years after being identified in kindergarten), and that fifth grade (six years after identification) was the year in which the largest proportion of students transitioned relative to the total EL population. (She noted that because most ELs had been reclassified earlier, the absolute number of ELs transitioning in Year 6 was smaller, even though the proportion was larger). She found that, generally, students’ probability of reclassification decreased over time, but followed a cubic function such that it accelerated going into grade 5 (the year the largest proportion of students transitioned) before dropping precipitously in middle school. She also found that low-income and Spanish-speaking ELs were significantly less likely to be reclassified than other ELs – the latter of which is a common finding (see also Grissom, 2004; Massachusetts Department of Elementary and Secondary Education & DePascale, 2012; Parrish et al., 2006; Thompson, 2012).

Thompson (2012) also found that fifth grade was a critical turning point for reclassification odds. She applied survival analysis to a sample of 202,931 ELs from the Los Angeles Unified School District (LAUSD); her sample comprised 8 cohorts of students with 10 waves of data each, covering grades K-9. She found that students needed six to seven years, on average, to reach a point where they had a 60 percent chance of being reclassified; she also found that about a quarter of her students (24 percent) were still ELs when her data ended, after grade 9. Thompson also explored the effects of a number of covariates including students’ sex, home language, initial level of ELP, and whether they had ever enrolled in bilingual education programs. As referenced in the preceding paragraph, she found, as Slama (2014) did, that Spanish speaking ELs were
significantly less likely to be reclassified than speakers of other languages. Indeed, Thompson found that speakers of Cantonese, Korean, or Filipino were nearly twice as likely to transition as Spanish speakers, though non-Spanish speakers represent a very small minority (6 percent for all non-Spanish languages combined) in the dataset and thus these estimates may not be representative. Thompson also observed that girls were more likely to be reclassified than boys throughout the K-9 span (about 16 percent more likely, on average), and, not surprisingly, students with IEPs were about 21 percent less likely to transition.

Although she incorporated it into her model in a different way, Thompson (2012) also corroborated Cook et al.’s (2012) finding that initial ELP made one of the most important differences in reclassification. She found that a difference of one standard deviation in initial ELP led to a 66 percent difference in the likelihood of being reclassified. Played out over the 10 year observation period, this translated to a probability difference of about 30 percent: students with the lowest initial ELP (Level 1) had about a 50 percent chance of being reclassified by 9th grade, while students with an initial ELP of Level 3 had an 80 percent chance. Interestingly, Thompson found that home language proficiency (HLP) also played a significant role – students whose initial HLP was lowest (Level 1) had about a 60 percent chance of being reclassified after 10 years, compared to about 80 percent for students with the highest level of initial HLP (Level 4). In this vein, Thompson also attempted to see whether students’ enrollment in a bilingual program affected their probability of reclassification; she found that students who were ever in bilingual programs had a very small – and ultimately non-significant – edge in the final years of her data. She acknowledged, however, that tracking the effects
of program type is difficult, as many bilingual programs are designed to last a certain number of years (e.g., K-5), such that students are not likely to transition out of such programs before they end, even if the students have met the necessary reclassification criteria.

Umansky and Reardon (2014) delved more deeply into the role of language instruction programs in their survival analysis study of ELs from an unnamed California district. They used a sample of 5,423 Latino ELs who all had between four and 12 years’ worth of data, depending on when they started kindergarten. Latino students in their district could enroll in one of four different language instruction program types: English immersion, transitional bilingual education, maintenance bilingual education, or dual immersion. The transitional and maintenance programs differ in that the former seeks to leverage home language skills without continuing to develop them per se, while the latter pursues proficiency in both languages as a goal for students. Dual immersion programs enroll native speakers of both English and Spanish and are designed to support full bilingualism (e.g., proficiency in English and in Spanish) for all participating students. Using English immersion as the baseline, the authors used survival analysis to evaluate whether program choice affected students’ probability of being reclassified over time. They also controlled for student-level covariates such as home language and initial ELP, as well as school-level covariates such as the proportion of EL students within a school.

They found that Latino ELs in English immersion were more likely to meet standards and be reclassified in earlier grades, but that students in the three types of dual-language programs caught up and eventually surpassed these students over time, in both performance and likelihood of reclassification. By high school, Latino ELs in any type of
dual-language program were more likely to be reclassified, and more likely to meet language or content performance standards than students in English immersion programs, with the most pronounced difference for dual-immersion students. The differences were small – they observed 7 percentage point differences in reclassification rates for two-way vs. English immersion students in 11th grade – but significant. These results were consistent with an earlier large-scale multistate study (Thomas & Collier, 2002), as well as a meta-analysis of 17 studies (Slavin & Cheung, 2005), which both found that students in English-only programs typically pull ahead in performance at first, but were eventually overtaken by students in bilingual programs.

Umansky and Reardon (2014) noted that their findings – and possibly others’ findings that preceded them – may partially be an artifact of how these different types of programs are designed: in the district they studied, the transitional program was designed to end after grade 3, and the maintenance and two-way programs go through at least fifth grade, with the option for students to continue taking some core content classes in Spanish through middle or high school, if they choose to do so. Although a program’s end does not automatically trigger reclassification, they speculated that,

in the dual immersion and maintenance bilingual programs teachers may have little incentive to reclassify EL students prior to the 5th grade given that students remain in the program through 5th grade regardless of their reclassification status.

(Umansky & Reardon, 2014, p. 25)

Overall, they found that students in the maintenance program fared “best” by many criteria: they were most likely to meet ELP achievement standards, meet ELA achievement standards, and be reclassified. Importantly, however, they also noted that the students enrolled in the different types of programs differed significantly on some
background characteristics, and that certain types of students were more likely to choose certain types of programs. Thus, the program effects cannot be interpreted causally, since students are not randomly assigned to the different types.

Conger (2010) also used survival analysis to model the effects of different language programs on reclassification. Her data came from New York City (NYC) and her panel included 10,182 students over the period of 1996 to 2002. Her dataset ended after grade 8, so her students had different amounts of data depending on their initial grade (1 through 6) in the baseline year, 1996. Her results both agree and disagree with Umansky and Reardon (2014). On the one hand, Conger observed higher and faster reclassification rates for students who received only ESL instruction, similar to Umansky and Reardon’s findings for the period up through high school. Thus, after six years of observation, roughly 80 percent of ESL students had been reclassified, compared to approximately 55 percent of bilingual students. After adding in student- and school-level covariates such as home language, IEP status, initial ELP, immigrant status, and FRL-eligibility, she still observed a difference in reclassification probability of roughly 33 percentage points between ESL and bilingual, which decreased only slightly over time.

On the other hand, Conger found no observable benefits to bilingual education, in contrast to Umansky and Reardon. In interpreting Conger’s results, it is important to remember that her data were cross-sectional, so that the students exiting in each observed year may be in grades 2 through 8, and may have been ELs for any length of time already prior to their exit. It is also important to note that, since her data end after grade 8, we do

---

5 Although Conger’s study takes place entirely before NCLB was implemented, she describes the EL education policy context in New York City at the time and it is essentially the same as the federal system established in NCLB.
not know whether she would have found the same reversal in achievement and reclassification that Umansky and Reardon (2014) observed starting in high school.

In addition, like Umansky and Reardon (2014), Conger also found patterns and differences among students who enrolled in different types of programs. She found that Spanish speakers were more likely to enroll in bilingual programs, but also that, across all language groups, bilingual program students were more likely to be FRL-eligible, to have lower initial ELP, and to transfer schools at least once than students who chose ESL programs. These traits were all associated with a lower probability of reclassification. She also observed that the vast majority of ESL students (87 percent) attended schools at some point where they could have enrolled in bilingual programs, underscoring the element of choice in language program enrollment. These selection differences again prevent any causal interpretations of program effects, since students are not randomly assigned to different types. As further corroboration of this fact, Conger also ran separate models to see whether the effects of bilingual education varied for speakers of different languages, and found no differences. Based on laws and policies in NYC at the time, she interprets her finding of no difference as evidence that negative effects associated with bilingual education are partially the result of non-random sorting into such programs of students more likely to struggle. 

---

*Conger summarized research about Spanish vs. Chinese bilingual education in NYC at the time of her study, which suggests that Spanish bilingual programs were more intense and involved than Chinese bilingual programs. Based on this difference, she hypothesized that, if bilingual education is harmful for students' development, the effects should be stronger for Spanish speakers, since their programs involve, essentially, “more” bilingual time/education than Chinese programs.*
2.6. Common Findings Across Reclassification Studies

It should be clear from the descriptions of these studies that several common findings have emerged consistently in studies of EL reclassification patterns. First, Spanish speaking ELs are less likely to be reclassified, or take longer to do so, than speakers of other non-English languages (Conger, 2010; Grissom, 2004; Parrish et al., 2006; Slama, 2014; Thompson, 2012). This finding, which emerged in New York City, Massachusetts, and California, is somewhat troubling, since the majority of ELs are Spanish speakers, both nationwide and within most states. One consideration for interpreting this finding is the possibility that Spanish-speaking ELs are more likely than speakers of other languages to enroll in bilingual language instruction programs. If so, Umansky and Reardon (2014) observation about bilingual program design affecting reclassification rates could partially explain this pattern (i.e., students in such programs are unlikely to be reclassified before the program comes to its natural end, regardless of their actual proficiency). In Conger’s (2010) study of New York City ELs, she did indeed find that Spanish speakers were more likely to enroll in bilingual programs. The finding also may stem from other traits that are correlated with Spanish as a home-language, rather than language itself. In some states and studies, Spanish speakers may be more likely than other students to come from high-poverty backgrounds or have parents with less education; both characteristics are associated with lower achievement in the general population.

---

7 To the extent that Spanish speakers are more likely than other ELs to come from low-SES backgrounds, this could also play a role, although most of the studies described here studied these terms in isolation, rather than as an interaction.
A second common finding is that the probability of a student being reclassified changes over time, with certain periods associated with peaks or dips in probability. Fifth grade, in particular, emerged somewhat consistently as a critical turning point for reclassification. Two California studies (Thompson, 2013; Umansky & Reardon, 2014) and two Massachusetts studies (Massachusetts Department of Elementary and Secondary Education & DePascale, 2012; Slama, 2014) observed a peak in reclassification at fifth grade, followed by a decrease in the probability or rate of reclassification in middle school. Although Grissom’s California study (2004) stops at grade 5, his findings also suggest a consistent increase in reclassification rates leading up to a peak in his final year of data.

A third finding is that, not surprisingly, students’ initial level of ELP plays a significant role in their reclassification probability (American Institutes for Research, 2013; Conger, 2010; Cook et al., 2012; Thompson, 2012). Within a single year, this finding is neither surprising nor problematic per se, as one would not necessarily expect a student who is a true beginner to master the entire English language in a single year (Prop 227 notwithstanding). What is more troubling is if students who begin as beginners remain unlikely to transition even after years of language instruction – this suggests that the instruction these students are receiving during their time as ELs is potentially not doing enough to help them learn the language and maintain their academic standing. It also suggests that, in studying reclassification, it is important to account for students’ ELP – either by conditioning or by grouping – as a check to make sure that these students, in particular, are not being “left behind” in the EL subgroup.
A fourth finding, related to the third, is that a considerable proportion of students remain ELs throughout their entire education careers. Slama (2014), Thompson (2012), and Umansky and Reardon (2014) all found that approximately a quarter of their students had not transitioned after 8, 10, and 12 years of observation, respectively. Parrish et al.’s (2006) estimate that only 40 percent of students had transitioned after 10 years in the subgroup is even more severe. There could be many possible causes for this stagnation that some students experience – poorly set reclassification cut scores; opportunity to learn issues in their academic courses; insufficient linguistic instruction, to name a few – but whatever the cause, it is problematic, and certainly cause for concern, that such a large portion of ELs can spend their entire K-12 careers in this subgroup without ever meeting the ELP achievement standards expected of them.

Finally, the one notable point of divergence across the studies summarized here was the average amount of time students took to reach proficiency. Here, we observed agreement within states, but diversity across states: the two Massachusetts studies found that students spent only about 3-4 years as ELs, on average, while the numerous California studies found estimates of five or more years to proficiency, at least. These differences likely stem from state-level differences in ELP standards or reclassification policies, and possibly also from how the authors chose to define and calculate the average time to reclassification. It is important to understand, however, that neither of these values is “good” or “better,” per se; as referenced previously in this section, reclassification is one metric of EL achievement that should be interpreted alongside other evidence – most notably, the post-reclassification achievement of students who have met this standard.
This final point highlights an important caveat to this study: namely, reclassification is not a positive goal in itself, until it can be shown that students who are reclassified go on to do “well” academically, however this may be defined. This is relevant when considering conclusions such as Conger’s appraisal that bilingual education is “harmful” for ELs, based on her finding that students in bilingual education take longer to be reclassified than students in ESL programs. While reclassification is an important goal that ELs should meet, it is only one indication of their achievement, or of a particular program’s benefits. A program that exits many students quickly should not necessarily be interpreted as “good” or “better,” until it is also shown that those exiting students fare well outside of the program. While the current study will not endeavor to address the post-reclassification performance of students (due to scope limitations, although such a study has been designed as a follow-up and is discussed further in the Appendix), it is important to recognize that this would be an important next step before drawing any strong policy conclusions about the validity or appropriateness of the state’s reclassification standards.

2.7. Conclusions

In this section I have attempted to summarize relevant research about (1) defining ELP, (2) setting ELP and reclassifications standards, (3) studying progress, performance, and reclassification, generally, and (4) determining which students are likely to meet reclassification standards and after how long. While much work has clearly been done in this area, there is still plenty of room for more. The current study adds to and builds on the research described here in certain ways. First, very few states have been studied in this area – readers may have noted that the majority of studies described above used data
from either California or Massachusetts. By using data from a different state, the current study will add new information to the conversation, which will contribute evidence either of cross-state differences or of similarities. Either will add to the field’s understanding of how reclassification may affect ELs’ growth and performance. Second, although some studies have attempted to account for multiple EL features at once (e.g., by grouping students in the same grade-span by initial ELP, or by using ELP-at-exit as a conditioning variable), none have systematically compared fit and variance explained for models that use different ways to account for ELs’ language proficiency and time in the subgroup. By fitting different models and making these comparisons, I may be able to provide insight that other researchers can use when applying similar designs to new data. Third, finally, most of the studies described above used data that began in kindergarten or first grade. Because my data begin later, both the methods and the findings used here may be useful for other researchers who have similarly messy or inconvenient datasets. It is to the methods I will use that I now turn in the next chapter.
Table 2.1. Summary of prior survival analysis designs and findings

<table>
<thead>
<tr>
<th>Authors</th>
<th>Year</th>
<th>Source</th>
<th>N</th>
<th>Span</th>
<th>Grouping</th>
<th>Conditioning</th>
<th>3 yrs</th>
<th>5 yrs</th>
<th>5+ yrs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parrish, Merickel, Perez, Linquanti et al.</td>
<td>2006</td>
<td>CA</td>
<td>~1.5M</td>
<td>K-12°</td>
<td># Years EL</td>
<td>None</td>
<td>12</td>
<td>&lt;40</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(10 yrs)</td>
</tr>
<tr>
<td>Conger</td>
<td>2010</td>
<td>NY District</td>
<td>10,182</td>
<td>1-8</td>
<td>Home language</td>
<td>Age/grade*</td>
<td>30-55</td>
<td>45-75</td>
<td>55-80</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(NYC)</td>
<td></td>
<td></td>
<td>ELP program type</td>
<td>Initial ELP*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Home language*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Low-income*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Immigrant*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ELP program type*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>IEP*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Transfer schools*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cook, Linquanti, Chinen &amp; Jung</td>
<td>2012</td>
<td>State</td>
<td>23,668</td>
<td>K-5</td>
<td>Test form</td>
<td>None</td>
<td>31-79</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(unnamed)</td>
<td></td>
<td></td>
<td>Initial ELP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Authors</td>
<td>Year</td>
<td>Source</td>
<td>$N$</td>
<td>Span</td>
<td>Grouping</td>
<td>Variables Used</td>
<td>% Reclassified after</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------</td>
<td>------</td>
<td>-------------------------</td>
<td>------</td>
<td>------</td>
<td>-------------</td>
<td>-------------------------------------------------------------------------------</td>
<td>----------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thompson</td>
<td>2012</td>
<td>CA District (LAUSD)</td>
<td>202,931</td>
<td>K-9</td>
<td>Grade</td>
<td>Sex* • Initial ELP* • Home language* • Low-income* • ELP program type • IEP* • HLP**</td>
<td>12 39 76 (10 yrs)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Umansky &amp; Reardon</td>
<td>2014</td>
<td>CA District (unnamed)</td>
<td>5,423</td>
<td>K-12</td>
<td>Grade</td>
<td>ELP program type*</td>
<td>38 ~75 (Gr5 12 yrs)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slama</td>
<td>2014</td>
<td>MA</td>
<td>5,354</td>
<td>K-7</td>
<td>Grade</td>
<td>Sex • Home language* • Low-income* • Immigrant</td>
<td>~50 --- 74 (8 yrs)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Parrish et al. used cross-sectional state-wide data and grouped students according to how long they had been classified as ELs.

*Indicates that this variable was found to have a significant effect on reclassification probability **HLP = Home language proficiency
CHAPTER 3

METHODOLOGY

As presented in the first chapter, the specific research questions I will address are:

1. What is the probability of reclassification over time for a cohort of ELs progressing from grade 3 to grade 8?
   a. When are third grade ELs most likely to be reclassified?
   b. Does the probability of reclassification vary by either grade-level, amount of time in the EL subgroup, or both?

2. Are students who have been in the EL subgroup for a long time (>5 years) more or less likely to be reclassified than more recently identified students?

3. What factors affect the probability of reclassification for students in the same grade-level, or who have been ELs for the same amount of time?

In this chapter I introduce the dataset used in this study – a longitudinal sample that follows a panel of students from grade 3 through grade 8 – and present my analysis plan for how to address my research questions.

First, however, I take a moment to introduce some specialized terms that are central and relatively unique to survival analysis. This method involves developing probability models based on the observed rates of survival and failure in a dataset. These terms are defined in the context of an event of interest, where survival refers to subjects who have not experienced the event, and failure refers to subjects who have experienced it. In the framework of this study, survival indicates that a student remains in the EL subgroup from one year to the next, whereas failure indicates that a student has been reclassified. Thus, counterintuitively, the survival rate represents the proportion of students who remain ELs each year, while the failure rate represents the proportion of
students who are reclassified. The ratio of these two events produces a **hazard rate**, which represents the likelihood, or risk, that an individual will “fail” (i.e., be reclassified).

Secondly, probability estimates are based on the **risk set** at each time point, which refers to the number of analytic units – here, students – that are eligible to experience the event in a given time period. In this study, the risk set each year comprises the number of former grade 3 EL students who could potentially be reclassified in that school year. Notably, the risk set excludes three important types of individuals, which I discuss in more detail below: those who have experienced the event already, those who have left the dataset for reasons other than experiencing the event, and those who become ELs later than grade 3. As I will discuss at length in this section, it is important to understand that, for my data, the risk set only includes students whose careers as ELs include grade 3. In other words, students who become ELs after grade 3, or students who cease to be ELs before grade 3, are not included in my dataset. By extension, therefore, they are not included in my risk set, and my findings thus cannot be generalized to such individuals.

Third, finally, one of the unique features of survival analysis is its ability to model **censorship**, a particular type of systematic missingness in data. If one is interested in the occurrence of a particular event – e.g., a person developing an illness – over a particular period of time, one can generally classify the population into three types of individuals: those who have already experienced the event, those who experience the event during an

---

As I discuss later, my risk sets do not represent all the students eligible to be reclassified each year, because my data do not include students who are identified as ELs later than grade 3. Thus, for example, it does not include students who are identified as ELs for the first time in grade 4, 5, 6, etc. Although there may not be high numbers of such students, it means that my risk sets are only a subset of all eligible ELs in grade levels higher than 3.
observed period, and those who do not experience the event during the same observed period. Each of these individuals contributes to the dataset and affects the model outcomes in different ways, and survival analysis is structured to account for this.

Individuals of the first type – those who have already experienced the events – are referred to as left-censored, and are not included in the survival analysis because they have no data to contribute. Although these individuals are not included in a study, it is important to be aware of their existence and characteristics, as their absence from the survival analysis sample may affect generalizability or interpretations of results. In this study, left-censored students are ELs who have already transitioned out of the subgroup prior to grade 3. As discussed in the literature review, some studies (e.g., Slama, 2014) found that many ELs who are identified in kindergarten do transition by grade 3, suggesting, again, that the sample here may be composed of lower achieving ELs who are more likely to become long-term ELs. This means, by extension, that the probability estimates from this study may be negatively biased and produce longer survival estimates than one would obtain were the left-censorship not present (i.e., were the data to begin in kindergarten, rather than grade 3).

Individuals who experience the focal event during the observed period are not censored. As I will show below, these individuals contribute information to the model for all time periods up to their experience the focal event, and then cease to contribute data thereafter. This is one of the more useful features for modeling reclassification since, as described above, students only take the state’s ELP exam when they are ELs. Individuals of the third type – those who do not experience the event during the observed period – are referred to as right-censored, because there is additional data about these individuals’
survival that is not contained in the dataset. In this study, this means that students will continue to take the state’s ELP assessment and contribute information about EL progress and performance beyond grade 8, where my dataset terminates. Right censorship is particularly useful because it allows me to differentiate among students who do and do not experience the focal event – reclassification – even if they have the same data history. For example, it allows me to differentiate between a student who has been an EL for six years and has not been reclassified by grade 8, and a student who is reclassified at some point between grades 3 and 8 after spending six years as an EL.

Having introduced these terms, I now discuss my student sample and provide descriptions and summary statistics of the different variables and measures I will use in my analyses. In the second half of the chapter, I provide detailed explanations of the survival analysis models and methods I will use to address my research questions.

3.1. Data Characteristics and Variables

Data for this study come from a longitudinal dataset of ELs from grades 3 through 8 in one state. The dataset contains two student cohorts of roughly equal size that enrolled in Grade 3 in the 2006-07 school year (13,913 students) and 2007-08 school year (13,598 students). It has a panel structure, meaning the same initial sample of students from the baseline year are tracked across the full time period (2006 through 2012, or 2007 through 2013, depending on the cohort); no additional students are added to the panel after the baseline year. The data are organized into a person-period format (Singer

---

9 As stated previously, the dataset only includes students who are (still) ELs as of the fall of their grade 3 year. Thus, former ELs who had already been reclassified prior to Grade 3 are not represented in this sample. Similarly, students who are identified as ELs after Grade 3 and become the classmates of students in this sample are also not present in the data. Based on these data features, the findings from this study are only generalizable to students who are ELs in third grade.
& Willett, 2003a), where each student has one row of data for each year that he or she remains an EL student. This means that individual students had anywhere from one to six waves of data, depending on when and whether they were either reclassified or censored for the first time.

3.1.1. Student Sample

In this section I summarize the background covariates of my student sample, including the definitions and derivations for my variables. The covariates for the baseline year (Grade 3) are summarized in Table 3.1 for each cohort, and are described in more detail in the sections that follow. For all covariates, it is important to note that the exact n-size contributing to parameter estimates will depend on whether the predictor is coded as time-invariant (the same for all school years) or time-varying (subject to change via school year). I discuss time-varying and time-invariant predictors in greater detail in section 3.2.3; here it is relevant to note only that the covariates are summarized in this section for the baseline year cohorts (as opposed to the full set of all observations), to give the reader a sense of the sample of unique students who contributed observations to the model estimates.

Immigrant Status

The dataset included an indicator variable for immigrant students, as opposed to those who were born in the United States. Roughly a quarter of each cohort were marked as immigrant students (24.5 percent of the 2007 cohort, and 23.9 percent of the 2008 cohort).

Home language

As is typical of the US EL population, a majority of students in both cohorts spoke Spanish as their home language (64.8 and 63.2 percent in the 2007 and 2008
cohorts, respectively). Overall, though, the students in this dataset spoke over 100 unique languages. For the purposes of this study, any language spoken by 150 students or more was dummy-coded for use as a predictor. Eleven “languages” met this criterion in at least one cohort, though it should be noted that one of the largest categories in both was “Undetermined,” meaning the student’s home language could not be determined by whomever was responsible for this task. It is also noteworthy that English was the second most common home language in both cohorts (6.0 and 5.6 percent, respectively), after Spanish. There could be many reasons for this finding, including the possibility that parents may attempt to report English as a home language in an effort to prevent their children from being screened and classified as English learners in the first place. It is also possible that some ELs come from households where they (the student) speak(s) English as a first language, but adult members of the household do not.

**LEP Program type**

The law in this state requires that all students identified as LEP receive services. Students are coded for whether they are enrolled in an English as a second language (ESL) program, a bilingual program, a two-way bilingual program (in which native speakers also participate and all students use both English and a second language), or any other LEP program that does not fall into one of the previous categories. The state also has a code for students who are eligible for LEP services but refuse them. This code did not appear anywhere in the dataset; however, 5.2 percent of the students in the 2007 cohort, and 3.2 percent of students in the 2008 cohort, were missing data for this variable. As I discuss below in my section on missing data (see section 3.1.4), I opted to keep these
students in for my survival analysis and use multiple imputation to accommodate their missing data.

**District resource need level**

School-level information is not available in this dataset, but it does include rough categorizations about needs and resources at the district level. The state categorizes districts according to a calculation of how well they can meet the needs of their students using local resources. This value is calculated as the ratio between the district’s estimated level of poverty (as measured by FRL eligible students and Census values of children in poverty) and a measure of the district wealth per pupil relative to the state average. Districts that are at or above the 70th percentile on this metric (meaning their poverty outstrips their local resources) are categorized as “high need” and further subdivided into rural or urban/suburban districts based on the density and size of their student populations. Districts with need ratios between the 20th and 70th percentiles are considered average needs, and districts with ratios below the 20th percentile are considered low needs. In addition, the state has special categories for (1) the single largest school district in the state (hereafter referred to as Central), (2) the next four largest cities (Large City Districts), and (3) charter schools. Thus, to summarize, districts are categorized according to the following set of labels: low needs, average needs, rural high needs, urban/suburban high needs, Central, large city, and charter school.

As one might expect, the majority of students in this dataset were enrolled in the Central district (64 percent in the full sample). After this, the next highest proportion of students were enrolled in sub/urban high needs districts (roughly 13 percent of students),
followed by average needs districts (9-10 percent). Very few students were enrolled in rural high needs districts or charter schools – fewer than 1 percent in either cohort.

**Intra-state mobility**

In 754 instances, students’ district codes changed in the dataset, suggesting that they had relocated to a district with a different resource needs profile. This was most likely to occur after the second observed school year (Grade 4 for most students), when 328 students changed district codes. Although these numbers are very small relative to the number of students observed in a given year (328 represents less than 2 percent of the students observed in school year 2), information about district moves was incorporated as a predictor because previous research in New York City has found that students who change schools are less likely to be reclassified than comparable peers (Conger, 2010; Menken et al., 2012). To increase power and help stabilize estimates of this effect, it was incorporated in a time-invariant manner, meaning it signals only that a student has ever changed to a different type of district (as opposed to indexing the change to a particular school year). Nonetheless, this variable undoubtedly provides underestimates of student mobility, since it fails to include either school-level moves within the same district, or district-level moves to a new district with the same level of need.

**Time as EL as of Fall Grade 3**

The dataset included information about how long students had already been part of the EL subgroup at the beginning of the baseline year (Grade 3). In both cohorts, roughly half the Grade 3 students had been ELs for four or more years already, suggesting that many had either been identified as ELs in Pre-K, or had already been
retained for a year prior to Grade 3. Unfortunately, it is not possible to determine from the data which of these possibilities is more likely or more common.

After that, the cohorts differed slightly – in the 2007 cohort, roughly equal proportions of students had been ELs for one, two, or three years (12 to 15 percent each), and 4.5 percent of students had been identified as ELs in third grade. In the 2008 cohort, meanwhile, just over a quarter of the students (26.7 percent) had been ELs for three years already, suggesting they had been identified in kindergarten. Smaller but similar proportions of ELs in this cohort had been in the subgroup for one or two years already (11.6 and 10.3 percent, respectively), and 7.1 percent had been identified in the current school year, Grade 3. These values are summarized in Table 3.1, and an “LEP to grade 3” variable was included in survival analysis models as a time-invariant predictor, to control for the amount of time that students had already spent as ELs prior to the observation period.

3.1.2. Reclassification Criteria

For the purposes of this study, reclassification was coded as a dichotomous variable that took on a value of 1 for a student’s final year as an EL, and 0 for all prior years. This variable is the outcome in the survival analysis models, which are in essence predicting the probability that a student will meet the criteria necessary to be reclassified starting in the following school year.

According to technical documentation, reclassification decisions in this state are to be made based solely on students’ ELP scores (see, e.g, [State 2011 ELPA Technical Report], p.4). Up through the 2012-13 school year (the final year of data for the 2008 cohort, only), the state used an ELPA that produced two scale scores: oral language, and written language. The test comprised four subtests in reading, writing, listening, and
speaking, and the two scale scores were created by pairing these raw scores – listening and speaking for oral language, and reading and writing for written language – and scaling the two combinations separately. In their individual score reports, students received scaled scores and proficiency levels for oral and written language, as well as one overall proficiency level.

The four performance levels for all three reporting categories were Beginning, Intermediate, Advanced, and Proficient. The cut score for proficiency and reclassification was between Advanced and Proficient, meaning students had to score in the highest performance level (Proficient, Level 4) to be reclassified. Similarly, to be reclassified, or generally to advance from one overall proficiency level to the next, students needed to clear the cut score on both scales. In sum then, only students who scored in the highest performance levels on both subscores were eligible for reclassification, according to the state’s criteria.

A student’s reclassification is formally signaled in the dataset when the student ceases to be marked with the state’s special LEP code. According to the state’s data manual,

Students who test out of LEP by reaching proficiency on the [ELPA] are still entitled to accommodations and some types of services for two years; however, once the students have tested out of LEP, they must NOT be recorded as LEP with [the state’s LEP] code. (p. 188-9)

Thus – assuming that districts adhere to the state’s policies about reclassification – beginning in the next school year after a student meets the reclassification cut score, the student is no longer coded as an EL in the state records.

For the purposes of this study, it was instructive to examine the extent to which the state’s use of this LEP flag code reflected the state’s documented reclassification
policies – in other words, did the test scores of students who were reclassified meet the performance standards specified for this transition? To explore how well schools and districts adhered to the state’s reclassification policies, I created contingency tables that crossed students’ performance levels with their EL status in the following school year. Overall, out of 20,457 instances where students’ ELP performance made them eligible for reclassification, 93 percent (19,029 cases) were reclassified in the following school year, as evidenced by the disappearance of the state’s limited English proficient code from the student’s file in the following school year. These 19,029 students also represented 96 percent of all students who were reclassified during the observation period. Thus, the vast majority of reclassification decisions followed the state’s prescriptions.

An analysis of the deviations from the rule suggested that it was more common for students who met the reclassification criteria to be retained as ELs than for students who fell short to be reclassified. Among all students who met the reclassification criteria, 1,428 (approximately 7 percent) were not reclassified. Meanwhile, among all students who were reclassified, 723 (approximately 4 percent) fell short of the prescribed criteria. In the latter case, one fifth (154 cases; 21 percent) were students who had met the criterion on one test and were in the second highest performance level on the other, and roughly one tenth (10.6 percent; 77 cases) were students who scored in the second highest category on both assessments. The remaining 492 cases were mostly students who were missing data altogether. In general, although it is possible that coding errors may be at play in some of the observed instances of deviation from the reclassification rule, for the purposes of this study I assumed that any deviations from the rule were
genuine situations (i.e., not coding errors) where school- or district-level judgment was
allowed to override state policy.

As a final point, it is important to note that the state did switch to a different ELP
assessment with different reclassification rules starting in the 2012-13 school year. This
new assessment not only had new cut scores, but also used an entirely new decision rule
for reclassification – a combination model where students had to meet an overall
performance level and certain minimum scores on each of the four subtests to be
reclassified. Because of the timing of this change, no reclassification decisions in my
observation period were made on the basis of this new test (those decisions would be
reflected in student records starting in the 2013-14 school year). Scores from this
assessment are available in the final year of the 2008 cohort’s data, however, and are
relevant for considerations about students who are still ELs at the end of the observation
period (see section 4.1.3 for this discussion). This shift is also important, because it
clearly limits the generalizability of my conclusions to years beyond those observed here.
Since the entire reclassification shifted, it remains to be seen whether the relationships
observed here changed in subsequent school years.

3.1.3. Time Variables

To use discrete-time survival analysis, it is necessary to “define” time –
specifically, to determine how it is counted and where it begins. In general, because
reclassification decisions are made once per school year in this state, it made sense for the
units of time to be school years. (As a note, the fact that reclassification can only occur at
certain pre-determined time points is also why time is considered a discrete, rather than a
continuous, variable). Thus, the remaining question is how to count years. Following
Slama’s (2014) example, I derived two different approaches for modeling this, both of which have been shown to affect students’ likelihood of reclassification in past studies.

The first approach uses grade-levels to count time. In this framework, the model produces estimates of the odds that a student will be reclassified in each grade between 3 and 8, given that they have not already been reclassified. This way of structuring time to group students is typical for accountability reporting, and serves as something of a baseline for analysis – i.e., the findings and interpretations one could expect to find if the data are structured and reported as NCLB demands.

If grades are used to count time, it is important account for the fact that some students are retained or promoted over the observation period. The extent of off-grade occurrence is summarized in Table 3.2, which shows the proportion of students in each school year who were enrolled in each grade-level. The Total column on the far right represents the total number of students observed in each calendar school year – these are the risk sets used for the survival analysis. The bolded values on the diagonal are the expected grade progression for a student who neither skips nor is retained for any grades. Values above the diagonal represent higher than expected grade-levels (i.e., students who skipped a grade), while values below the diagonal represent lower than expected grade-levels (i.e., students who repeated a grade). Percentages are based on the school year risk sets – thus, for example, in the second school year, 96.9 percent of observed students were in the expected grade-level (Grade 4), while 3 percent (580 students) were in a lower-than-expected grade level, and 0.1 percent (18 students) were in a higher-than-expected grade-level. In reading the table, it is important to remember that the off-grade numbers are essentially cumulative – for example, the 671 students who are in Grade 4 in
school year 3 includes all or most of the 580 students who repeated Grade 3 the previous year and are now progressing again, plus some additional students who are repeating Grade 4.

Grade retention means that the risk set for each grade-level may include some students twice – for example, a student who repeats Grade 3 will contribute two cases to the Grade 3 risk set. Technically this would be appropriate, since the student was eligible for reclassification after Grade 3 on two different occasions; statistically, however, the student’s two cases would not be independent of one another, which is an assumption required for logistic regression. Given the non-ignorable prevalence of repeated grades, I opted to use school years, rather than grade levels, to group students and code time. Under this approach, years are counted based on calendar years, and numbered as 1 through 6, since the actual calendar years associated with each year of observation differs across the two cohorts. This approach means that each student appears only once in each risk set, thus preserving the necessary conditions for logistic regression.

On the flip side, however, this decision also means that each risk set includes some students who are not in the expected grade-level. To account for this, a dummy code was created to mark students who were off-grade in each school year. Conceptually, it would be preferable to have separate dummy codes for students who were above versus below grade-level; practically, however, the extremely small incidence of students who were above grade-level (never more than 20 students in a given year) made it unlikely that they could produce meaningful estimates as a subgroup. Thus, all off-grade students are coded as one group; in practice, these estimates will need to be interpreted essentially as representing below-grade students.
As an alternative to the grade-driven structure, I also counted time by the number of years students spent in the state’s EL subgroup between grades K-8. Using this variable for survival analysis essentially restructures the data to be cross-sectional, so that grades are collapsed and students are grouped instead by how long they have been ELs. Students who have been ELs for 4 years thus become a risk set for estimating the odds of being reclassified after this amount of service, regardless of whether they are in Grade 3 or Grade 7 when it occurs. This model would produce findings similar in structure to those by Parrish et al. (2006) and Conger (2010), who both used cross-sectional data in their survival analyses (see section 2.5).

To calculate this time variable, which I called EL Year (\(EL_{Yr}\)), I used the “LEP to grade 3” variable (described above) essentially as a start value: if the student had been an EL for three years already prior to grade 3, then grade 3 represented the student’s fourth EL Year (i.e., \(EL_{Yr} = 4\)). For this variable, it is important to understand that although it may be known that a student has been an EL prior to grade 3, they will still only contribute to model estimates for the six years that are observed in the current dataset\(^{10}\). Thus, in the example just given, this student will only contribute to

---

\(^{10}\) Although it would be technically possible to reverse engineer data rows for students prior to grade 3, there are several reasons why this would likely not help, and might actively bias the analytic results. First, although the number of years as an EL may be known, there is no way of knowing what those years comprised in terms of service, grade-level retention, or school or district transfers. It is unknowable, for example, whether a students who has been an EL for four years prior to grade 3 started in pre-K as an EL, or started in kindergarten and repeated a grade prior to grade 3. Secondly, while it is possible to impute values for the students contained in this dataset, it is not possible to construct values or data for the many thousands of students who are not observed in this dataset – namely, all students who were reclassified out of EL status prior to grade 3. Because this dataset begins in grade 3 and includes only ELs, it excludes, by definition, all students who already experienced the event of reclassification prior to this particular period of observation. Thus, not only would there be no variance among the observed students in their reclassification prior to grade 3 (because none were reclassified), it would be impossible to compute non-biased estimates of the probability of reclassification in these earlier years, because survival analysis requires that the sample include all individuals who could have experienced the event of interest.
estimates of the probability of being reclassified starting from $EL_{Yr} = 4$. The reasons for this are discussed at length in a footnote; briefly, the primary reason for this is that it would not be possible to produce unbiased estimates using these students’ unobserved data.

By extension, this fact means that estimates of the probability of students being reclassified after any amount of service will only be based on students who received some amount of that service from grade 3 onwards. The effects of students’ time as ELs prior to grade 3 will be modeled separately using the “time as EL” variable described in section 3.1.1. Similarly, because the data do not include students who are identified as ELs after grade 3 (e.g., newcomer students who are identified in grades 4, 5, 6, etc.), estimates of the probability of students being reclassified in any particular grade-level will only apply to students who have been ELs since at least grade 3. They will not apply to “all grade 4 ELs,” “all grade 6 ELs,” etc.

To make this idea more concrete, Figure 3.1 and Figure 3.2 summarize the number of students “at risk” of reclassification for each observed time value using the two different timing variables. In Figure 3.1, the 19,306 students eligible for reclassification in grade 4 do not represent all the ELs in grade 4 in the state, but only the grade 4 ELs who were also ELs in grade 3. Similarly, in Figure 3.2, the 1,599 students

The examples given here use grade-level as the time variable, but they would also apply to the EL Year variable – for example, there are likely many students who spent one, two, and three years as ELs but were reclassified prior to the beginning of the period observed in this dataset. Thus, imputing data for the current sample back to EL Year =1 for all students would not produce useful results because they would exclude the many thousands of students who were reclassified in this same amount of time prior to the beginning of these data.

Ultimately, these concerns affect the generalizability of the results found here, which I emphasize throughout this dissertation. In short, it is imperative to remember that any findings here are applicable only to students who are ELs in 3rd grade; they cannot be generalized to younger EL students, or to ELs who are identified as ELs after grade 3.
who are in their first EL year represent only the students whose first year as an EL occurred in grade 3. The 4,495 students in $EL_{Yr} = 2$ are either students in Grade 4 who were first identified in grade 3, or students for whom grade 3 was their second EL year. As mentioned in the beginning of this chapter, this configuration serves to limit the generalizability of the results from this study, regardless of which time variable is used. Specifically, it entails that the results here are only generalizable to students who are (still) ELs at the beginning of grade 3.

In terms of modeling, these two timing frameworks contribute to probability estimates through a series of dummy variables for each observed unit of time (e.g., Grade 3, Grade 4, etc., and $EL_{Yr} = 3, EL_{Yr} = 4$, etc). Students who are ELs during any given period receive a dummy code of 1 for that period, while students who were not observed – due to censorship, reclassification, or any other reason – receive a value of 0. As I will discuss in section 3.2, this means that students will contribute data to model estimates only for years that they are observed as ELs.

3.1.4. Missing data

All students in the sample had complete data for their home language, district needs categorization, sex, immigrant status, and cohort. A very small number of individuals ($n = 19$) were missing information about how long they had been English learners prior to Grade 3. Of these 19 students, 15 were reclassified after the first year of observation, while the other four survived as ELs for anywhere from two to six more years. Because these individuals accounted for less than 1 percent of students and less than 0.001 percent of observations, their missingness did not pose a threat to final estimates in terms of bias or stability.
A more problematic level of missingness was observed with students’ program data. On this level, 1,145 individuals (4.2 percent of the sample) lacked data in the baseline year, and 5,183 observations (6.6 percent) lacked program data across the six-year period. Because the state uses dedicated codes for both “other program type” and refusal of service, missing program data would suggest a failure or error in using these codes; there is, however, no way to be certain of this. In the interest of retaining these students in the sample, I used multiple imputation to create 10 datasets with imputed program values, using the other predictors in this model as predictors for the imputations as well. The results presented for the fitted logit hazard model in section 4.4 all represent the pooled estimates across all 10 imputed datasets.

As discussed previously, it is also important to ensure that a survival analysis risk set includes all subjects who could have experienced the event of interest, otherwise the absence of certain individuals or groups could bias the results. In this sense, left- and right-censorship can also be thought of as a type of missingness that affect model estimates. In terms of left-censorship, it has already been noted that this dataset categorically excludes students who have experienced reclassification prior to Grade 3. Given that higher achieving ELs tend to be reclassified more quickly, this censorship could mean that my estimates are negatively biased, meaning they will predict longer time to reclassification and lower probabilities of meeting the threshold. I will discuss the characteristics of right-censored students – those who leave the dataset without being reclassified – in my results section (see section 4.1), alongside descriptive survival and failure rates.
In addition to these two types of censorship, it is also important to clarify that this dataset also excluded ELs with disabilities as a rule. This included both students who were identified as having cognitive disabilities in grade 3, and any students who went on to be diagnosed as having disabilities at any point over the six year period. The motivation for this decision was that the effects of learning and cognitive disabilities on these students’ language and content performance could complicate their performance interpretations, and also jeopardize the validity of any models or interpretations in which students with and without disabilities are grouped together. As a result, the findings from this study are not generalizable to ELs with disabilities. Because I never had these students’ records, I cannot determine precisely how many were excluded by this criterion. A review of the state’s NCLB report cards, however, suggests that approximately 5 percent of the students participating in the state’s grade 2-4 ELP assessment in the year 2008 were identified as having disabilities. Thus, it seems probable that at least 1,400 students (700 per cohort) may have been excluded based on this rule.

3.2. Data Analyses

I use survival analysis as the primary method to explore reclassification between grades 3 and 8 for students who are ELs at the beginning of grade 3. Survival analysis, also referred to as event history analysis, is a longitudinal method designed to estimate the likelihood that an event will occur over time. The outcome in these models is a dichotomous variable that takes on a value of 1 in the time period when the individual experiences an event, and 0 in all other time periods. In predicting the probability of the event’s occurrence in a given period, survival analysis models can take different forms depending on whether the time data are continuous, meaning the target event could
happen at any time within an observed period, or discrete, meaning the event can only occur at certain specified times (e.g., once per month, or once within an established cycle). The data for this study are discrete because reclassification decisions are made once per school year.

3.2.1. Baseline discrete-time models

Survival analysis models produce two types of functions to represent survival and failure over time. The function of primary interest is the **hazard function** \( h(t_j) \), which models the conditional probability that an individual \( i \) will experience the focal event at a particular time \( j \), given that the event has not already occurred for that individual:

\[
h(t_j) = \Pr[T_i = j | T_i \geq j]. \tag{3.2.1}
\]

The hazard function is complemented by the **survival function** \( S(t_j) \), which expresses the probability that an individual will survive or persist past the time period in question:

\[
S(t_j) = \Pr(T_i \geq j) \tag{3.2.2}
\]

In both equations, \( T_i \) represents the time of the event’s occurrence, and \( j \) represents the current time period; the two functions represent, in essence, the probability of failing ‘now’ (however defined) versus failing in the future.

The sum of the probability of hazard and survival will always be 1 in a given time period, which means that the survival function can be re-expressed as \( S(t_j) = 1 - h(t_j) \). Accordingly, the odds of failing (i.e., experiencing the outcome event) can be expressed as the ratio of these probabilities:

\[
\frac{h(t_j)}{1 - h(t_j)} \tag{3.2.3}
\]
This expression then lends itself to the logit transformation, which offers several benefits that allow for easier model comparisons and visualizations relative to other ways of modeling event probability. Thus, a baseline hazard model for an individual’s probability of being reclassified at some point during the observed period is displayed in Equation 3.2.4.

\[
\text{logit } h(t_i) = [a_1 \text{Time}_{i1} + a_2 \text{Time}_{i2} + \cdots + a_J \text{Time}_{ij}] + \epsilon_{ij} 
\] (3.2.4)

In this equation, each term represents the probability that the \( i \)th individual will be reclassified at the \( j \)th time point as a function of regression coefficients \( (a_j) \) and time indicators \( (\text{Time}_j) \). In this study, separate models will be built and compared using school years (which correspond to grades 3 through 8 for most students) and time as an EL \( (EL\_Yr) \) to count time.

The outcome of this model is the logit hazard rate of reclassification, which represents the relationship between time and the risk of event occurrence. (Substantive predictors can also be added to this model, as I discuss below). The interpretations of these effects are not always straightforward on the logit scale, and are thus often transformed into odds ratios (which represent the odds of reclassification for one group relative to another), or probabilities (which represent the probability of being reclassified for different groups). As I will discuss in Chapter 4, results from my final fitted model will be presented in terms of odds ratios and a cumulative failure (reclassification) rate over time based on the model’s predictions.

### 3.2.2. Duration Dependency and Time Specifications

In Equation 3.2.4, the inclusion of separate regression coefficients for each time point is based on the assumption that the effect of time on event probability may differ
from one time point to the next. Singer and Willett (2003b) refer to this type of specification as a completely general model, based on the fact that it does not follow any polynomial specification and is thus free to take whatever shape the data dictate. This general model will always fit the data better than any alternative, but there may be advantages to adopting a polynomial model or some other spline function (e.g., a linear function, quadratic function, logarithmic, etc.) if it can represent the data equally as well (Box-Steffensmeier & Jones, 2004a; Singer & Willett, 2003b). This is because polynomial functions generally have known forms and characteristics, and may also allow for better prediction of new data. Singer and Willett (2003b) specifically recommend exploring polynomial specifications when hazard rates are expected to be near zero in some time periods (p. 409), which is the case in the final year of data (grade 8 for most students) observed here. In effect, because students’ Grade 9 status is unknown, all Grade 8 students must be treated as censored, even though presumably some went on to be reclassified. A fully general model will not be able to produce an estimate for this, but a polynomial model might.

To compare different specifications for time, Singer and Willett (2003b) recommend a sequential process where a completely general model is fit, followed by sequentially more complex polynomial models. The baseline polynomial model is a constant one, where the probability of event occurrence is unaffected by time at all (i.e., \( \logit h(t_o) = a_0 \text{ONE} \), where \( \text{ONE} \) is specified as 1 for all rows in the dataset and \( a_0 \) thus becomes an intercept, which by definition is constant). As a note, this model is something of a null hypothesis with respect to the use of longitudinal analysis; if a model with no
time variable fits the data well, let alone best, it suggests that time is unnecessary or inappropriate to understanding the probability of event occurrence.

To this baseline constant model, the effects of time can be modeled across time periods as a linear function (i.e., \( \logit h(t_{ij}) = a_0 \text{ONE} + a_1 \text{Time} - c \)), where the effect of one school year on the probability of reclassification is estimated as an additive effect over and above the intercept, and is the same for all observed periods), a quadratic function (i.e., \( \logit h(t_{ij}) = a_0 \text{ONE} + a_1 \text{Time} - c + a_2 \text{Time}^2 \)), where the effect of one school year increases or decreases towards a maximum or minimum value over time before reversing, and this acceleration is modeled as an additive effect over and above the constant linear effect), a cubic function (i.e., the trajectory and acceleration change directions twice), and so on.

Singer and Willett recommend fitting to at least a 4th order polynomial (i.e., quartic) model, which represents the likely upper bound of utility for a polynomial specification compared to the fully general one. The fit of each polynomial model should be compared both to the fully general model and, when applicable, to a simpler polynomial model with one fewer terms, using likelihood ratio tests and degrees of freedom based on the difference in parameters across the two models being compared. Thus, for example, for my data using grades to mark time, a cubic model should be compared both to a quadratic model using \( df = 1 \) and to a fully general model where \( df = 3 \), since the fully general model has one time predictor for each of the six observed school years. The ideal model should be the simplest model that fits the data (1) as well as the completely general model (i.e., the likelihood ratio test comparing it to the general model is non-significant), and (2) significantly better than models with simpler
specifications of time (i.e., the likelihood ratio test to the preceding simpler model is significant). Of course, the final choice of model may also be informed by other considerations such as utility, substantive theory, and past research.

3.2.3. Factors Affecting Survival

As with any model, it is possible to add covariates to a hazard function, so that the conditional probability of the event’s occurrence is modeled as a function both of time, and of one or more time-varying or invariant covariates. Thus, once an appropriate time variable and specification have been determined for the baseline hazard function, I will add in covariates to complete the survival analysis model. The general form for the survival function with covariates is displayed in Equation 3.2.5

\[
\text{logit}(h_{ij}) = [a_1 Time_{1ij} + a_2 Time_{2ij} + \cdots + a_j Time_{Kij} ] + B_X X_i + B_W W_{ij} + \epsilon_{ij} \quad (3.2.5)
\]

In this equation, \( X_i \) is a vector of time invariant characteristics, and \( W_{ij} \) is a vector of time-varying characteristics, for student \( i \). Each vector has a corresponding vector of regression coefficients (e.g., \( B_X \), which consists of \( \beta_{x_1}, \beta_{x_2}, \ldots, \beta_{x_k} \)).

As discussed in the literature review, previous survival analysis studies have explored the effects of program type, SES, initial ELP, sex, and the proportion of ELs within a school, on the probability of reclassification (see Table 2.1 for a more complete summary of covariates that have been tested in other studies). Because my data come from a different state than the ones studied in these prior publications, I will also explore the effects of variables such as these on students’ probability of reclassifying. Specifically, I will use the covariates summarized in section 3.1 (most of which are listed in Table 3.1): home language, LEP program type, district resource needs, sex, and immigrant status. For those variables that are categorical in nature (i.e., home language,
LEP program type, and district resource needs), I will determine whether they are best modeled categorically (i.e., with a series of binary indicators for each category), or dichotomously, where students are either in the most common condition (i.e., Spanish home language, Central district, or ESL program) or an “other” group. I will also include binary indicators for students’ original cohort (2007 versus 2008), whether they have ever moved to a district with a different level of need, and whether they were off-grade in a given school year. The model using school years to count time will also incorporate students’ time as an EL prior to grade 3, while the EL year model will incorporate students’ grade-levels.

Singer and Willett (2003b) specify three assumptions about covariate effects that can be explored and, if necessary, relaxed, to improve model fit and utility. The first is whether certain predictors can take on different values in different time periods. Time-varying predictors of this sort essentially imply that a student’s odds of reclassification will change if, at any point during the observation period, his or her predictor status changes. For example, if a student is in a bilingual program and then switches to an ESL program subsequently, this change would be associated with a change in odds of reclassification. An associated implication is that time-invariant coding may lend a variable slightly more power as an estimate, since each student will automatically be assigned the same code for all of their observations, which could lead to slightly higher numbers of observations than if the variable is left as time-varying.

Three variables used here functioned in this way: program type, district needs, and off-grade status. Given the ample power of this sample, these variables were maintained as time-varying. Two of the three – program type and off-grade status – were
further explored via interaction terms, to see whether group differences varied for different grade levels (e.g., whether the difference in odds for an ESL program compared to a bilingual program is different for students in fourth grade as opposed to eighth grade). These interactions are discussed in more detail in the discussion of the proportionality assumption, below.

The second assumption to explore is that of linear additivity – that is, whether any interaction effects are present among two or more substantive predictors. (Interactions with time are addressed via the proportionality assumption, discussed below). One interactions of this type will be tested – program type and home language – based on Conger’s (2010) findings that students enrolling in bilingual programs tend to be a self-selecting group, but that the effects of such programs do not appear to differ for speakers of different languages.

A third and final check about covariate effects is the proportionality assumption, which addresses whether certain effects may vary across time (e.g., the effect of being female is different for third graders than it is for eighth graders). Two research-based possibilities warrant exploration here. The first is the aforementioned possibility that EL year and grade-level may interact, which would suggest that whether or not four years of EL status is “enough” to get a student reclassified (all other things equal) may vary depending on the student’s grade-level. Given the shifting language demands across grade-levels, such an effect seems plausible. Another interaction stems from Umansky and Reardon’s (2014) finding that students in bilingual programs were less likely to be reclassified in early grades, but later became more likely to exit (possibly as a result of bilingual program design). To explore these possibilities, interactions between program
type and school year (Program * School year) and EL year and school year
(EL_Yr*School year) will be tested for significance. Table 3.4 summarizes the covariates
that will be incorporated in various ways during this piece of the model-building process.

### 3.2.4. Model Comparisons and Interpretation

For explorations of duration dependency and covariates, models will first be
compared for overall fit to determine what type of specification provides the most
efficient explanation of variance in odds of reclassification. Sequentially, I will first
identify a best-fitting model for time specifications, then add covariates to this model and
determine the best configuration for those. To compare models throughout this process,
two indices will be used.

For nested models (e.g., different polynomial models using the same time
variable, or covariate models with and without certain interaction terms), fit can be
compared using a likelihood ratio test (LRT) where the difference in deviance (log
likelihood) values for the two models is compared to a $\chi^2$ statistic with degrees of
freedom equal to the difference in the number of parameters between the two models
(Box-Steffensmeier & Jones, 2004b; Slama, 2014).

For non-nested models (e.g., a model that uses school year to measure time,
compared to a model that uses EL years to measure time), LRTs are not possible. I will
compare these models using the sample adjusted Bayes information criterion (SABIC)
which is given by

$$SABIC = -2LL + \ln \left( \frac{N + 2}{24} \right) q$$

(3.2.6)

where $LL$ is the model’s log-likelihood, $q$ is the number of parameters in the model, and
$N$ is the sample size. While there is no significance test for this metric and it does not
address overall fit of the model to the data, smaller values indicate better fit relative to alternative models with higher SABIC values. In a review of simulation literature, Pastor and Gagné (2013) found that the SABIC performed best relative to other information criteria such as the Akaike information criterion (AIC).

In addition to fit, models can also be compared by an effect size of how much variance they explain in students’ odds of reclassification. For logistic regression, this quantity can be expressed using the Nagelkerke $R^2$, which is calculated with the expression

$$R^2_N = \frac{1 - \exp\left[-\frac{2}{n}(LL_{\text{predictors}} - LL_{\text{baseline}})\right]}{1 - \exp\left[-\frac{2}{n}(LL_{\text{baseline}})\right]}.$$  

(3.2.7)

In the equation, $n$ is the number of observations, $LL_{\text{predictors}}$ is the log-likelihood (deviance) of a model including whatever time and substantive predictors are included, and $LL_{\text{baseline}}$ is the baseline constant model with only an intercept. Although it is ultimately a pseudo- $R^2$ statistic, the Nagelkerke $R^2$ is scaled so that it is bounded by 0 and 1, and can thus be interpreted using typical rules of thumb such as Cohen’s (1992), which posit that values between 0.3 and 0.5 can be interpreted as moderate effect sizes, and values above or below this threshold are large or small, respectively.

In the final model, point estimates for various predictors’ effects on the odds of reclassification will also be interpreted. Because the final model may include 50 or more separate predictors, I will use the Holm’s sequential procedure (Holm, 1979) to control for familywise error. In this method, the various regression predictors are ranked by their
$p$-values and numbered accordingly (e.g., $\beta_1, \beta_2, \ldots, \beta_m$); their $p$-values are then compared to decreasingly stringent alpha levels, using the rule

$$P_n \leq \frac{\alpha}{m + 1 - h} \quad (3.2.8)$$

Where $h$ is the number’s rank (e.g., 23rd, out of 50 total parameters), and the overall $\alpha = 0.05$. 
Table 3.1. Summary of baseline year student characteristics, by cohort

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Full Sample</th>
<th>2007</th>
<th>2008</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Students in baseline year (Grade 3)</td>
<td>27,511</td>
<td>13,913</td>
<td>13,598</td>
</tr>
<tr>
<td><strong>Student-level covariate percentages</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>49.4</td>
<td>49.1</td>
<td>49.6</td>
</tr>
<tr>
<td>Immigrant</td>
<td>24.2</td>
<td>24.5</td>
<td>23.9</td>
</tr>
<tr>
<td><strong>Home language</strong>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanish</td>
<td>64.0</td>
<td>64.8</td>
<td>63.2</td>
</tr>
<tr>
<td>English</td>
<td>5.8</td>
<td>6.0</td>
<td>5.6</td>
</tr>
<tr>
<td>Undetermined</td>
<td>5.4</td>
<td>5.2</td>
<td>5.6</td>
</tr>
<tr>
<td>Arabic</td>
<td>2.5</td>
<td>2.7</td>
<td>2.4</td>
</tr>
<tr>
<td>Bengali</td>
<td>2.5</td>
<td>2.2</td>
<td>2.7</td>
</tr>
<tr>
<td>Chinese</td>
<td>2.2</td>
<td>2.0</td>
<td>2.4</td>
</tr>
<tr>
<td>Urdu</td>
<td>1.8</td>
<td>1.8</td>
<td>1.7</td>
</tr>
<tr>
<td>Russian</td>
<td>1.7</td>
<td>1.7</td>
<td>1.8</td>
</tr>
<tr>
<td>Haitian Creole</td>
<td>1.4</td>
<td>1.4</td>
<td>1.4</td>
</tr>
<tr>
<td>Albanian</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>Other</td>
<td>11.8</td>
<td>11.4</td>
<td>12.3</td>
</tr>
<tr>
<td><strong>Program Type (as of fall Grade 3)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ESL</td>
<td>72.6</td>
<td>71.5</td>
<td>73.7</td>
</tr>
<tr>
<td>Bilingual</td>
<td>16.2</td>
<td>16.7</td>
<td>15.7</td>
</tr>
<tr>
<td>Two-Way</td>
<td>5.0</td>
<td>4.7</td>
<td>5.3</td>
</tr>
<tr>
<td>Multiple LEP Programs</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>Other LEP</td>
<td>1.7</td>
<td>1.6</td>
<td>1.9</td>
</tr>
<tr>
<td>Refused services</td>
<td>n = 10</td>
<td>n = 0</td>
<td>n = 10</td>
</tr>
<tr>
<td>Missing</td>
<td>4.2</td>
<td>5.2</td>
<td>3.1</td>
</tr>
<tr>
<td><strong>District Type (as of fall Grade 3)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Central</td>
<td>66.4</td>
<td>67.5</td>
<td>65.2</td>
</tr>
<tr>
<td>Sub/urban High Needs</td>
<td>12.9</td>
<td>12.6</td>
<td>13.3</td>
</tr>
<tr>
<td>Average Needs</td>
<td>9.5</td>
<td>9.1</td>
<td>9.9</td>
</tr>
<tr>
<td>Large City</td>
<td>5.9</td>
<td>5.5</td>
<td>6.2</td>
</tr>
<tr>
<td>Low Needs</td>
<td>4.3</td>
<td>4.4</td>
<td>4.3</td>
</tr>
<tr>
<td>Rural High Needs</td>
<td>0.7</td>
<td>0.6</td>
<td>0.8</td>
</tr>
<tr>
<td>Charter</td>
<td>0.4</td>
<td>0.4</td>
<td>0.4</td>
</tr>
<tr>
<td>Moved to district w/different need category</td>
<td>2.7</td>
<td>3.1</td>
<td>2.3</td>
</tr>
<tr>
<td><strong>Previous Time as EL (as of fall Grade 3)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5+ years (2 years Pre-K, or retained)</td>
<td>5.9</td>
<td>6.7</td>
<td>5.2</td>
</tr>
<tr>
<td>4 years (ID’ed Pre-K)</td>
<td>43.3</td>
<td>47.5</td>
<td>39.0</td>
</tr>
<tr>
<td>3 years (ID’ed K)</td>
<td>21.0</td>
<td>15.5</td>
<td>26.7</td>
</tr>
<tr>
<td>2 years (ID’ed Gr1)</td>
<td>11.9</td>
<td>12.2</td>
<td>11.6</td>
</tr>
<tr>
<td>1 year (ID’ed Gr2)</td>
<td>11.9</td>
<td>13.5</td>
<td>10.3</td>
</tr>
<tr>
<td>0 years (ID’ed Gr3)</td>
<td>5.8</td>
<td>4.5</td>
<td>7.2</td>
</tr>
<tr>
<td>Average years (#)</td>
<td>3.0 ± 1.3</td>
<td>3.0 ± 1.4</td>
<td>3.0 ± 1.3</td>
</tr>
<tr>
<td>Missing</td>
<td>n = 19</td>
<td>n = 13</td>
<td>n = 6</td>
</tr>
</tbody>
</table>

*Only languages spoken by 150 students or more were included in the analysis.
Table 3.2. Cross-tab of school years and grade-levels

<table>
<thead>
<tr>
<th>School Year*</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>27,511</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>27,511</td>
</tr>
<tr>
<td></td>
<td>(100.0)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(100.0)</td>
</tr>
<tr>
<td>2</td>
<td>580</td>
<td>18,707</td>
<td>18</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>19,306</td>
</tr>
<tr>
<td></td>
<td>(3.0)</td>
<td>(96.9)</td>
<td>(0.1)</td>
<td>(0.0)</td>
<td></td>
<td></td>
<td></td>
<td>(100.0)</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>671</td>
<td>12,490</td>
<td>20</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>13,196</td>
</tr>
<tr>
<td></td>
<td>(0.1)</td>
<td>(5.1)</td>
<td>(94.6)</td>
<td>(0.2)</td>
<td>(0.0)</td>
<td></td>
<td></td>
<td>(100.0)</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>25</td>
<td>611</td>
<td>7,682</td>
<td>20</td>
<td>1</td>
<td>0</td>
<td>8,339</td>
</tr>
<tr>
<td></td>
<td>(0.3)</td>
<td>(7.3)</td>
<td>(92.1)</td>
<td>(0.2)</td>
<td>(0.0)</td>
<td></td>
<td></td>
<td>(100.0)</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td>35</td>
<td>771</td>
<td>4,760</td>
<td>14</td>
<td>0</td>
<td>5,581</td>
</tr>
<tr>
<td></td>
<td>(0.0)</td>
<td>(0.6)</td>
<td>(13.8)</td>
<td>(85.3)</td>
<td>(0.3)</td>
<td></td>
<td></td>
<td>(100.0)</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>78</td>
<td>825</td>
<td>3,149</td>
<td>9</td>
<td>4,065</td>
</tr>
<tr>
<td></td>
<td>(0.1)</td>
<td>(1.9)</td>
<td>(20.3)</td>
<td>(77.5)</td>
<td>(0.2)</td>
<td></td>
<td></td>
<td>(100.0)</td>
</tr>
<tr>
<td>Total</td>
<td>28,104</td>
<td>19,404</td>
<td>13,158</td>
<td>8,552</td>
<td>5,607</td>
<td>3,164</td>
<td>9</td>
<td>77,998</td>
</tr>
<tr>
<td></td>
<td>(100)</td>
<td>(100)</td>
<td>(100)</td>
<td>(100)</td>
<td>(100)</td>
<td>(100)</td>
<td></td>
<td>(100)</td>
</tr>
</tbody>
</table>

*Because the two cohorts are observed over different but overlapping calendar year spans, school years have been renumbered as 1 through 6 for both cohorts. For the 2007 cohort, school year 1 = the 2006-07 school year; for the 2008 cohort, school year 1 = the 2007-08 school year.

NOTE. The table reads, “Out of 19,306 students observed in the second school year, 96.9 percent (18,707 students) were in Grade 4, while 3 percent (580 students) were repeating Grade 3, and 0.1 percent (19 students) had been promoted to Grade 5 or 6.”

Table 3.3. Summary of missing data, by cohort

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Total Sample</th>
<th>2007</th>
<th>2008</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Students</td>
<td>27,511</td>
<td>13,913</td>
<td>13,598</td>
</tr>
<tr>
<td>Missing Data Element</td>
<td>n</td>
<td>%</td>
<td>n</td>
</tr>
<tr>
<td>LEP program information</td>
<td>1145</td>
<td>4.2</td>
<td>726</td>
</tr>
<tr>
<td>Pre-Grade 3 Time as EL</td>
<td>19</td>
<td>0.001</td>
<td>13</td>
</tr>
<tr>
<td>Variable</td>
<td>Type</td>
<td>In/Varying</td>
<td>Possible Values</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>---------------</td>
<td>------------</td>
<td>----------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Sex</td>
<td>Binary</td>
<td>Invariant</td>
<td>Male*, Female</td>
</tr>
<tr>
<td>Immigrant</td>
<td>Binary</td>
<td>Invariant</td>
<td>Not immigrant*, Immigrant</td>
</tr>
<tr>
<td>Cohort</td>
<td>Binary</td>
<td>Invariant</td>
<td>2007*, 2008</td>
</tr>
<tr>
<td>Off-grade</td>
<td>Binary</td>
<td>Varying</td>
<td>On-grade*, Off-grade</td>
</tr>
<tr>
<td>Home language</td>
<td>Categorical</td>
<td>Invariant</td>
<td>Spanish*, Albanian, Arabic, Bengali, Chinese, English, Haitian Creole, Other, Russian, Undetermined, Urdu, (Not Spanish)</td>
</tr>
<tr>
<td>LEP program type</td>
<td>Categorical</td>
<td>Varying</td>
<td>ESL*, Bilingual, Two-way bilingual, Multiple, Other, Refused, Missing, (Not ESL)</td>
</tr>
<tr>
<td>District resource needs</td>
<td>Categorical</td>
<td>Varying</td>
<td>Central*, Average Needs, Charter, Large City, Low Needs, Rural High Needs, Sub/urban High Needs, (Not Central)</td>
</tr>
<tr>
<td>Change in district type</td>
<td>Binary</td>
<td>Invariant</td>
<td>Never changed*, Ever changed</td>
</tr>
<tr>
<td>Pre-Gr3 Time as EL</td>
<td>Discrete</td>
<td>Invariant</td>
<td>0-4+</td>
</tr>
<tr>
<td>EL Year * School year</td>
<td>Interaction</td>
<td>Varying</td>
<td>3+</td>
</tr>
<tr>
<td>Program * School year</td>
<td>Interaction</td>
<td>Varying</td>
<td>0 – 6</td>
</tr>
<tr>
<td>Off-grade * School year</td>
<td>Interaction</td>
<td>Varying</td>
<td>0 – 6</td>
</tr>
<tr>
<td>Program * Home language</td>
<td>Interaction</td>
<td>Varying or Invariant</td>
<td>0 – 1</td>
</tr>
</tbody>
</table>

* Indicates baseline (0) condition. Values in parentheses are alternate codings for variables that may be categorical or binary, or may be time varying or invariant.
NOTE. Grade-levels in parentheses are the expected grade-level for each school year, assuming students have not been retained or promoted. Actual risk sets contain students in off-grades as well; see Table 3.2.

Figure 3.1. Number of students eligible for reclassification each year, by school year.

Figure 3.2. Number of students eligible for reclassification each year, by EL year.
CHAPTER 4
RESULTS

This chapter presents the results of the model-building process described in the second half of Chapter 3. The first section presents descriptive survival rates based on the observed data; the next two sections summarize the model-building process for a fitted logit hazard model using time (section 4.2) and other substantive covariates (section 4.3) as predictors. In the final sections (4.4 and 4.5), point estimates from the final fitted model are presented, interpreted, and briefly discussed.

4.1. Descriptive Survival Rates

Descriptive information about sample survival rates are commonly presented using life tables, as in Table 4.1 and Table 4.2. These tables display the number and proportion of students from the original risk set for each time interval, as well as the number and proportion who are reclassified or censored in each time period. Because one focus for this study was on how best to group EL students for analysis, descriptive results are presented for students grouped both by school year (essentially grade-level; see section 3.1.3 to review how this variable was specified) and by EL year (i.e., how long they have been in the EL subgroup, regardless of grade-level).

4.1.1. Observed Hazard and Survival by School Year

Table 4.1 summarizes rates of survival, event occurrence (i.e., reclassification), and censorship for ELs grouped by school year. Column A shows the risk set for each school year (i.e., the number of unique students who were eligible for reclassification in...
that school year\textsuperscript{11}, while Columns B and C show the number of students who were reclassified each year, or who left the dataset without being reclassified (i.e., were censored). Columns D through G show the hazard, censorship, and cumulative survival rates, as well as the proportion of students who were off grade-level for each school year. (The latter is not a typical inclusion in a survival table, but is included here as an additional piece of information to support interpretations of the data).

Table 4.1 shows several important landmarks about grade 3 ELs’ reclassification patterns by grade. First, it shows that the largest number of students ($n = 6,432$) experienced reclassification after the first observed school year (Grade 3 for all students), while the largest proportion of students (roughly a third) experienced reclassification after the third school year (when 95 percent of students were in Grade 5). Second, it also shows that just over half the initial cohort has been reclassified by the beginning of the third school year, as evidenced by the number of students in the risk set at the beginning of that year ($n = 13,196$). Thus, among students who are ELs in third grade, roughly half have either been reclassified or left the state’s public school system within two school years.

Third, the final row of the table shows that 15 percent of the original cohort (4,065 students) were right-censored – meaning they were still ELs – by the end of the observation period. Although reclassification decisions for these students would be recorded in the next school year’s data and are thus not observed, 1,019 of them earned ELP scores that would qualify them for reclassification, suggesting a hazard rate of 0.25.

\textsuperscript{11} Note that, as referenced previously, this does not represent the total number of students who are eligible for reclassification in grades higher than 3, since the dataset does not include new ELs who enroll or identified in grades 4 through 8.
The remaining 3,046 students would presumably have continued as ELs into high school, representing 11 percent of the initial risk set of 27,511 students. Although this number may seem low, it is troublingly high considering how long these students have already been ELs: by definition, they have been ELs for at least six years (i.e., since Grade 3), but the median time as EL for these students was actually 9 years, with an average of 8.5 years.

As a final note, the Table also shows that censorship was most likely to occur after grade 3, when 1,774 students were censored. (Recall that censorship after Grade 8 is due to the observation period ending, rather than students’ disappearing from the dataset without being reclassified). Although this number may sound high, Column E shows it represented roughly 6 percent of the risk set overall. The characteristics of censored students, which are relevant for understanding the observed and fitted results from this study, are presented and discussed in section 4.1.3, after the descriptive results by EL year.

4.1.2. Observed Hazard and Survival by EL Year

Table 4.2 shows rates of survival, reclassification, and censorship for ELs grouped by EL year. Based on how these data are structured, the display and interpretations differ slightly from those in Table 4.1. This Table breaks down the risk set in terms of students who began the observation period in each EL year, as well as those who reached each EL year during the observation period. It shows that, when the sample is grouped by how long they have been ELs, the largest number of students (n = 5,204) was reclassified after 5 years of service, while the largest proportion of students (30 percent) were reclassified after 7 years of service. Censorship rates appear to be relatively higher within this structure (e.g., 10 percent of students or more are censored after half of
the observed years), but it is important to remember that the risk sets formed by this time variable are relatively small and thus will produce larger proportions even with smaller numbers of students.

Due to the fact that “new” students join the risk set each year, it is not possible to calculate a cumulative survival rate based on an “original” student cohort. Thus, in contrast to Column F in Table 4.1, Column H in Table 4.2 represents the proportion of students who survive from the risk set for a given year. Similarly, it is not strictly possible to calculate a median lifetime for students here, though the risk sets suggest that more than half of the total student sample of 27,511 students have either been reclassified or censored at some point between EL year 6 \((n = 14,408)\) and EL year 7 \((n = 9,561)\).

### 4.1.3. Characteristics of Censored Students

A total of 7,759 students were censored during the observation period, meaning they disappeared from the dataset without being reclassified. Within this group, 4,065 students were censored due to the end of the observation period, meaning they were still ELs at the end of Grade 8, where the data end (see Table 4.1). I refer to these students as “time-censored,” to indicate that some, if not all, were censored for this reason. Based on the structure and nature of the dataset (see section 3.1.1), the remaining 3,694 right-censored students presumably were censored because they either (1) left the state, (2) enrolled in private school, or (3) dropped out of school altogether. For the purposes of understanding how this right-censorship might affect the remaining risk sets and findings based thereon, I grouped these students according to when they were censored (after Grade 8 or earlier) and examined their characteristics using chi-squared tests to compare them to one another, and to students who were reclassified at any point during the observation period.
These characteristics are summarized in Table 4.3, which shows that significant differences emerged frequently, not only between censored and reclassified students, but also between students censored before and after Grade 8. Students censored before Grade 8 were significantly more likely to be immigrants (27 percent were, compared to 10 percent of time-censored students, and 17 percent of reclassified students), and significantly less likely to be Spanish speakers, particularly compared to time-censored students (roughly 60 percent, compared to roughly 80 percent of time-censored students). Their immigrant status is also reflected in the fact that they tended to have spent less time in the EL subgroup prior to Grade 3, relative to the other two groups. Interestingly, these students also were significantly more likely to be enrolled in bilingual education programs compared to reclassified or time-censored students. They were also significantly less likely to have ever moved districts. One possible explanation for these characteristics among right-censored students is that their families are leaving the state to relocate after an initial settlement period in the United States.

In contrast to students censored before Grade 8, time-censored students were significantly more likely to be off-grade level, at 22.5 percent (916 students). Notably, they were also significantly more likely to be boys (by at least 4 percentage points) and to have moved districts at some point (at least twice as likely). While students censored before Grade 8 were significantly more likely to be immigrants, students censored after Grade 8 were overwhelmingly likely to be US natives (roughly 90 percent were), even compared to those who were reclassified at some point. These students were also significantly more likely to be Spanish speakers and, interestingly, were significantly more likely to be enrolled in ESL programs, and significantly less likely to be enrolled in
bilingual programs, relative to students who were reclassified or censored earlier. This fact is a relevant consideration in the context of previous research that has observed reversals in reclassification patterns for bilingual students in high school (Umansky & Reardon, 2014). Since relatively few remaining ELs are enrolled in bilingual programs at the end of middle school, this suggests that no such reversal would be likely to emerge from these data, were the observation period to continue.

The presence or absence of ELP scores and performance levels in students’ final year of data serves as a potential indicator of whether the student left during or after the school year. While the presence of scores does not guarantee that the student stayed through the end of the school year, the absence of scores suggests that the student did not make it far enough through the school year to take the ELP assessment before leaving. By this measure, students who left the district before being reclassified stood out: just over 9 percent of these students (346 individuals) lacked scores, compared to 2 – 3 percent of students who were reclassified, or were still ELs at the end of the observation period. This finding suggests that students who leave the state without being reclassified may be leaving mid-school-year.

For time-censored students who did have ELP scores, information about their performance can provide further information about why these students may be missing out on reclassification. As referenced earlier, the state switched to a new ELP assessment and a different reclassification rule in the 2012-13 school year. This means that the two cohorts were subject to different reclassification standards in their final year of observation; they will be discussed separately, as a result.
For the 1,935 time-censored students from the 2007 cohort, 408 students (21 percent) earned scores that would qualify them for reclassification in Grade 9, while the remaining 79 percent fell short. Among the 1,527 students who fell short, nearly half (748 students) earned scores placing them in the highest performance level for oral language, and the second highest level for written language. This was by far the most common performance pattern among students in this year and cohort. An additional 311 students scored in the highest level for oral language, but scored in the second lowest level (Level 2) for written language, and another 213 students scored in Levels 2 and 3, respectively, in written and oral language – in other words, written performance still lagged oral performance. Taken together, this suggests that 70 percent of students who fell short of reclassification did so due to their written language performance, and over 80 percent did better in oral language than in written language. By contrast, only 48 students (1.2 percent) from this cohort scored better in written language than they did in oral language.

Students in the 2008 cohort were subject to a combination decision rule (Carroll & Bailey, 2015), wherein their overall ELP scale score (a scaled combination of reading, writing, listening, and speaking scores) had to place them in the highest performance level (Level 4), and their raw scores on each subtest had to clear a certain minimum value. In essence, this new rule required students to clear 5 separate cut scores, as opposed to two for the older ELPA. Among 2,026 time-censored students in this cohort, 611 (30.2 percent) cleared the necessary thresholds to qualify for reclassification – notably, a higher proportion than had met the performance standards on the older ELPA. Among the remaining 1,415 students in the 2008 cohort who fell short, the vast majority
(1,144 students; 80.1 percent) earned scores placing them in the second highest overall performance level (Level 3). The most common performance pattern among these students, observed among 321 individuals (28.1 percent), was also to fall short of reclassification solely based on their writing scores. Interestingly, the second most common condition for this cohort, observed for 200 individuals (17.4 percent of Level 3 students) was to fall short based solely on listening. The third most common performance pattern (142 individuals; 12.4 percent) was to fall short based on both writing and listening.

Taken together, this information suggests that, among students who are still ELs by the end of Grade 8, roughly a quarter (approximately 1,109 students) would have qualified for reclassification starting in Grade 9, while just over a third (1,380 students) fell short of reclassification based solely on their written language skills. Interestingly, the single year of data from the new state ELP assessment suggests that the updated listening test or performance standard may be more challenging than its predecessor, and thus play a larger role in reclassification decisions moving forward; more data are needed to establish whether this is a pattern, however.

More generally, the information presented here about students who were censored during the observation period offers a few insights to the other results that follow. In general, while the findings here do corroborate prior research about which groups are less likely to be reclassified (Conger, 2010; Grissom, 2004; Slama, 2014; Thompson, 2012; Umansky & Reardon, 2014), the systematic censorship of such students may bias the study results in non-negligible ways. For example, past research has found that boys, Spanish speakers, and students in bilingual programs are all less likely to be reclassified.
In these data, however, students who were censored before Grade 8 were significantly more likely to be boys, and to be enrolled in bilingual programs. This could mean that a lower likelihood of reclassification associated with either of these conditions – if observed – stems both from such students’ performance and from the fact that these students are more likely to leave the state before they are reclassified. In other words, it may be less the case that bilingual programs do not prepare students for the state’s reclassification standard, and more likely that students in bilingual programs are more likely to leave the state before meeting said standard.

Relatedly, the students in these conditions who remain may not be appropriately representative of their populations. For example, given that immigrant students are significantly more likely to leave the state before Grade 8, the immigrant students who remain in the dataset throughout the observation period may differ in some important ways from those who are censored. This, too, could lead to biased estimates of reclassification among immigrants as a larger group. This fact is an important one to keep in mind for the remainder of this section; its statistical, practical, and political implications will also be considered more thoroughly in section 4.4.

4.2. Fitted Model: Duration Dependency and Time Specifications

The remaining sections of this chapter present the results of the process described in section 3.2, to build a fitted logit hazard model of survival and reclassification.

4.2.1. Main Effect of Time

Different specifications for time were fit to the data and compared as described in section 3.2.2, using both school years and EL years to group students and count time. The results of this process are summarized in Table 4.4, which shows summary fit information for the fully general, constant, and various polynomial models using each
time variable. As a baseline, these results confirm that time does affect the probability of reclassification, as evidenced by the fact that the two constant models (which posit that time has no effect) fit the data worse than any other option in which time was used as a predictor. Beyond this baseline, the table also shows that school years functioned better than EL years as a counting mechanism, as evidenced by the lower deviance statistics and SABIC values associated with these models, as well as slightly higher $R^2$ values. Even the fully general EL year model – the best possible option for using this variable – fared worse than all but two of the school year models (the constant and the linear models) in terms of fit and variance explained. Based on this result, school years were used as the timing variable for subsequent model building with other predictors (see section 4.3).

The best way to specify the relationship between reclassification and school years required some judgment. As suggested previously, the fact that all students were censored by definition in the final school year (i.e., none were known to be reclassified in the following school year) created convergence problems for the fully general model. Although it returned plausible values for overall fit statistics, the fully general model failed to meet a log-likelihood convergence criterion of 0.001, even after 500 iterations, and the estimated regression coefficient for the final school year was implausibly large ($\alpha_6 = -101.20$). As a result, a fully general model using only the first five school years was used for comparisons to polynomial models. Not surprisingly, the reduced fully general model fit the data poorly; as the table shows, all likelihood ratio tests favored a polynomial option in comparison.

Figure 4.1 shows the fitted hazard rate by school year for the various polynomial specifications. In the fully general model, shown in solid black, the probability of
recategorization ascends to a peak in the third school year (Grade 5 for most students), and then descends for the remaining years. The constant and linear models clearly do a poor job approximating this shape, while the other polynomial options approximate it more closely. Statistically, the addition of each new polynomial term conferred a significant improvement in fit, such that the quartic (4th degree) model emerged as the preferred model according to the LRTs. The Figure suggests that this model approximates the general model closely at the extremes, though diverges more in the middle school years. Most notably, it peaks a year later than the general model, in the fourth school year (Grade 6 for most students).

As noted previously, a quartic model is generally considered overly complex for use or interpretation, and is typically not preferable if a simpler model may be tenable. This led to a consideration of the cubic model as an alternative. As Table 4.4 shows, the difference in variance explained ($R^2$) between the quartic school year model and the simpler cubic model is very small ($\Delta R^2 = 0.002$), suggesting that sacrificing the fourth degree polynomial term does not drastically reduce the amount of variance explained. Figure 4.1, meanwhile, shows that the cubic model mirrors the general model closely in the first three years, particularly in contrast to the simpler quadratic model, which produces consistently higher estimates over this period compared to the general model. Moreover, although it technically peaks in the fourth school year as well, its third year probability is essentially the same as the general model ($\alpha_{3\_Cubic} = \alpha_{3\_General} = 0.33$), and only slightly less than the peak probability in school year 4 ($\alpha_{4\_Cubic} = 0.334$). It is also worth noting that Slama (2014) also chose a cubic baseline hazard function in her study of Massachusetts ELs using a similar model building process to the one described here.
Based on these factors, the cubic model is retained as the best fitting time specification when students are grouped by grade-level; this is the model that is carried forward to test the effects of covariates on the likelihood of reclassification. The final point estimates for the effects of time will be presented in section 4.4, once all predictors have been added to the model.

4.3. Fitted Model: Factors Affecting Survival

As described in section 3.2.3, covariates were added in a structured sequence to compare how certain predictors could best be included. As a first step, I determined whether the program, home language, and district needs categories functioned better as a single binary indicating the most common condition versus an “other” condition (e.g., Spanish speaking versus other, for home language), or as a series of binary indicators grouping students into each alternate condition relative to a common baseline (e.g., Bengali home language versus Spanish, Haitian Creole versus Spanish, etc.). I refer to this second condition as a categorical specification, although it should be noted that all indicators are binary.

The results of this covariate specification process are summarized in Table 4.5, which shows the fit statistics and effect sizes for models where one, two, or all three variables are included as either binary or categorical. All models shown use a cubic specification of time measured by school years (see section 4.2), and four demographic variables that are inherently binary: sex, immigrant status, cohort, and whether the student is off-grade in a given year. (These variables are all time invariant, except for the off-grade variable, which can vary by year). As will be elaborated in the next section, these variables were coded so that the baseline condition is a male student from the 2007
cohort who is on grade-level and not an immigrant. It is also important to note that for this step onwards, the multiple imputation datasets were used, to control for the fact that 6.6 percent of the observations (across all years) lacked program data (see section 3.1.4; to reiterate, because the data included variables to indicate a refusal of LEP service for eligible students, missing data were interpreted as truly missing, as opposed to students’ opting out of language instruction). Thus, all model comparisons were run across the original data and all 10 imputed datasets; fit, variance explained, and point estimates are all pooled across all datasets throughout the remainder of this report.

Overall, the Table shows that the all categorical model, which uses separate indicator variables for each of 10 home languages, five program types, and six district types (see Table 3.1 and Table 3.4), fit the data best. This all-categorical model explained the most variance \( R^2 = 0.359 \), fit significantly better than all simpler models, and had the lowest deviance and SABIC values, despite having 29 estimated parameters. It is interesting to note, however, that none of the three covariates explained very much additional variance (no more than 0.004, for the district type), and the type of instructional program students enrolled in seemed to contribute the least information relative to district and home language \( \Delta R^2 = 0.001 \). Despite these small effect sizes, most indicators for each covariate were significant as predictors (note that point estimates for the final model will be presented and interpreted section 4.4); in addition, an interest

\[ \text{In these categorical models, the most common condition is excluded as an indicator, thus becoming the baseline condition. Each binary indicator represents the average group difference between this baseline and an alternate. Thus, for example, in the language model, there is no indicator for Spanish, and each of the 10 indicator variables represents the average difference in odds between Spanish speakers and speakers of another language.} \]
in exploring research-based interactions with program type, as I discuss next, led to the decision to retain the fully categorical model to proceed in the model-building process.

As discussed in section 3.2.3, four interaction terms were explored: (1) program type by home language, (2) program type by school year, (3) off-grade status by school year, and (4) EL year by school year. Because program type and home language were modeled categorically, a full interaction between home language and program type would require creating separate interaction terms for each program*language combination (e.g., Spanish*ESL, Spanish*bilingual, Bengali*ESL, etc.). As this would have added over 50 additional terms and also produced many empty cells (e.g., if there are no Bengali speakers enrolled in two-way bilingual programs), a simpler approach was used to model the interaction: all program types were interacted with a binary language specification, separating Spanish speakers from speakers of all other languages.

The rationale for this approach was based primarily on the number of program types relative to the number of home languages. Because there were fewer program types (baseline plus five, as opposed to baseline plus nine for home language), collapsing across home languages created more power for analysis (i.e., more observations per cell) and also avoided the problem of empty cells for home languages where no students enrolled in a particular program type. In addition, although the point estimates are not reported here, it was noted that the direction of difference from the baseline home language condition was the same for all non-Spanish languages. That is, all non-Spanish languages whose reclassification odds differed from the baseline condition were more likely to be reclassified, never less\(^\text{13}\). If different conditions had differed from the

\(^{13}\) Strictly speaking, there was one exception to this – Haitian Creole – but its negative difference was non-significant.
baseline in different directions, combining them would have been problematic; however, since alternate conditions differed only in magnitude, and not direction, combining them seemed justified, for the sake of parsimony.

Thus, three additional models were fit and compared to the model with categorical program, language, and district predictors. One incorporated five interaction terms between program type and home language (one for each program type other than the baseline condition of ESL), one incorporated five interaction terms between program type and school year, and one incorporated both types of interactions using ten additional terms. These models are also summarized in Table 4.5, which shows that the interaction between home language school year contributed very little to model fit and effect size ($\Delta R^2 = 0.001$), while the program type and home language interaction was more useful ($\Delta R^2 = 0.007$). Again, for research purposes, and because power is not at issue for this dataset, both interactions were retained for the final model.

To create the final model, students’ time in the EL subgroup prior to grade 3 was added as a predictor, as well as a time-invariant indicator for whether they had ever changed districts. In addition, two further interaction terms were added: EL year by school year, and off-grade status by school year. The results of this final model, including overall fit, effect size, and point estimates for all predictors and effects, are presented and interpreted next, in section 4.4.

4.4. Fitted Model: Final Model and Interpretations

In this section, point estimates for the effects of various predictors on the odds of reclassification in the final fitted logit hazard model are displayed and interpreted. The effects of various predictors are presented and discussed separately in a series of
subsections; the next chapter will address combinations of these predictors with regard to this study’s specific research questions. As a reminder, for all sections that follow, the predictors are coded so that the baseline student is a male, Spanish speaking, native US-born student from the 2007 cohort who is on-grade, enrolled in an ESL program in the Central district, and has just been identified as an EL for the first time in Grade 3. It is also important to recall that all values that follow are estimated values from the fitted model, rather than observed values from the sample data.

Each subsection presents a table showing the odds ratios associated with a group of predictors (e.g., different types of language instruction programs), followed by a graph mapping out the cumulative failure rate for different subgroups relative to those predictors (e.g., the cumulative failure rate of students in bilingual instruction programs, relative to the baseline ESL condition). The cumulative failure rate is an inverse of the survival rate, and provides an estimate of how many students from the original cohort are predicted to experience the event by the end of the observation period. When interpreting the graphs, it is also important to remember that the proportionality assumption in survival analysis (see section 3.2.3) implies that constant effects over time are multiplicative, as opposed to additive (as they are in ordinary least squares (OLS) regression). Thus, main effects will produce curves that are increasingly divergent, while interaction effects may produce curves that are parallel – i.e., patterns that are opposite to those observed with interactions and main effects in OLS.

The tables include confidence intervals for the odds ratios, as well as $n$-sizes for the total number of observations across all years; these are the values that produce the actual estimates in the survival analysis model (as opposed to the number of unique
students, which would be lower). Because the \((\exp(\beta))\) values are odds ratios, effects are symmetric around the value of 1, which represents equivalent odds between the two groups in question. For binary indicators, odds ratios represent the difference in odds between the baseline group and an alternate subgroup (e.g., a Chinese-speaking baseline student, compared to a Spanish-speaking baseline student). For continuous indicators, odds ratios represent the change in odds associated with a unit change in the predictor (e.g., the change in odds associated with having spent one more year as an EL prior to grade 3). Odds ratios greater than 1 can be interpreted as percentages – for example, \(\exp(\beta) = 1.5\) indicates that the non-baseline condition is 50 percent more likely to be reclassified, and \(\exp(\beta) = 2\) indicates that this condition is twice as likely as the baseline. Odds ratios less than 1 have been converted to a similar metric using the conversion \(-1/\exp(\beta)\), which allows them to be expressed as percentage or times less likely (i.e., \(\exp(\beta) = -1.5\) indicates that the non-baseline condition is 50 percent less likely to be reclassified, etc.). Only results that remained significant after controlling for familywise error (see section 3.2.4) are highlighted throughout this section.

**4.4.1. Time – Main Effects**

As discussed in section 4.2, time was specified using a cubic polynomial form. This entails that the odds of reclassification vary at each time point, as do the direction and rate of change in odds per school year. The point estimates for the effects of time are presented in Table 4.6; because the cubic term is negative, this indicates that the odds descend to an initial trough before climbing to a later peak. Using the quadratic equation
below, it is possible to calculate the exact values of the low and high points in the
function\(^{14}\):

\[
1 + \frac{-\alpha_2 \pm \sqrt{\alpha_2^2 - 3\alpha_1\alpha_3}}{3\alpha_3}
\]  

(4.4.1)

(where \(\alpha_n\) represent the various regression coefficients for the linear, quadratic, and cubic
time terms). The results of this formula reveal, consistent with the results in section 4.2,
that the function reaches a peak in the fourth school year (Grade 6 for most students),
after an initial trough in the first school year. These values are reflected in the hazard
rates across time, which begin at \(h(SY_1) = 0.14\) (meaning the model predicts of a
reclassification rate of 14 percent after this year) and ascend to \(h(SY_4) = 0.37\) (a 37
percent reclassification rate) before decreasing again. It is worth recalling that, when time
was included as the sole predictor, predicted hazard rates in Grades 5 and 6 were roughly
equivalent \((h(SY_3) = h(SY_4) = 0.33)\), and close to the actual Grade 5 peak in the observed
data. In the final fitted model presented here, the Grade 6 peak is more pronounced
\((h(SY_3) = 0.30)\), and serves as a reminder of the compromises necessary to fit a
polynomial model to the data, as opposed to using the fully general model that would
more closely mirror the observed values.

As a final note with respect to time, it is also important to note that the fitted
model also controlled for students’ time in the EL subgroup by including the variable for
the amount of time students had already been ELs at the beginning of grade 3, and an

---

\(^{14}\) As a note, only the exponentiated regression coefficients are presented here and throughout this section,
even though the non-exponentiated coefficient estimates are used in this formula. These coefficients can be
provided to the interested reader upon request.
interaction term between school year and EL year. These effects are discussed later in this section.

4.4.2. Cohort, Sex, Immigrant Status, and Off-grade Status – Main Effects and Interactions

Table 4.7 shows the final point estimates and confidence intervals for the sex, immigrant status, and off-grade status variables, including the interaction effect between off-grade status and school year. The baseline condition for these variables was a male, native-born student in the 2007 cohort who is on grade-level each year, and was initially identified as an EL in Grade 3. As Figure 4.2 shows, the model predicts that 77 percent of the students in this baseline condition would be reclassified by the final year.

Although there were no significant differences between the two cohorts, other demographic indicators did reveal significant differences. Female students were 21 percent more likely than male students to be reclassified in a given year \((\exp(\beta)_{\text{female}} = 1.213)\), and immigrant students were 16 percent less likely to be reclassified than native born students \((\exp(\beta)_{\text{immigrant}} = -1.156)\). For the latter finding, it is important to remember that immigrant students were disproportionately likely to be censored before the end of the observation period (see section 4.1); this pattern almost certainly means that estimated reclassification rates for immigrant students are negatively biased (i.e., biased downwards) and do not reflect the true probability of reclassification for the full population of immigrant students.

The effects of being retained were the most pronounced: students who were off-grade were more than three times less likely to be reclassified than peers who were on grade-level \((\exp(\beta)_{\text{off-grade}} = -3.509)\). Because the interaction between school year and
grade-level was significant and positive, however, the effects of being off-grade became less problematic with each additional school year. Most likely, this suggests that students who were retained ultimately fell in step with comparable peers in their new grade-level cohort and did not continue to suffer further for having been retained. Put differently, this suggests that being retained tends to have a one-time effect, rather than a lasting one over time. Nonetheless, the effects were non-trivial – by the end of the observation period, only 62 percent of retained students were predicted to be reclassified; 15 percentage points fewer than the baseline condition.

4.4.3. Home Language Main Effects

Table 4.8 shows the numeric results for the effects of home language on reclassification. As referenced in the earlier discussion of model specification (see section 4.3), all significant differences with respect to home language were positive, meaning that Spanish speakers were uniformly less likely to be reclassified than comparable peers who spoke other languages. The two exceptions were speakers of Arabic and Haitian Creole, neither of whom differed significantly from Spanish speakers in their likelihood of reclassification. Speakers of other languages were anywhere from 21 percent more likely (for students reporting English as their home language) to roughly twice as likely (for speakers of Russian) to be reclassified, relative to Spanish speakers. This finding is consistent with other research that has been referenced in this study (Conger, 2010; Slama, 2014; Thompson, 2012), though it should also be interpreted cautiously, since those same studies have found non-random relationships between home language and other factors that affect performance, such as parental education and language program type. In other words, as will be discussed in the final chapter of this dissertation, home
language should almost certainly be interpreted as a correlate to performance and reclassification, rather than a cause.

As Figure 4.3 shows, the difference in likelihood between Spanish and the highest performing language (Russian) was approximately 16 percentage points. Thus, by the end of the sixth year, Russian speakers had a 93 percent chance of being reclassified, while Spanish speakers, the least likely, had only 77 percent chance.

### 4.4.4. District and Mobility Main Effects

The effects of district resources and needs on reclassification are summarized in Table 4.9 and Figure 4.4. Unsurprisingly, the results generally show that students enrolled in districts with more local resources and fewer needs are more likely to be reclassified than comparable students enrolled in districts that need more support from the state. By the end of the observation period, ELs in low- and average-needs districts had predicted reclassification rates of 92 and 90 percent, respectively, compared to a rate of 77 percent for the baseline Central district, where a majority of students in the sample were enrolled. Students enrolled in charter schools had similar long-term outcomes to students in average needs districts, and rural high needs districts were no more or less likely to be reclassified than students in the baseline Central district; in both cases, however, small sample sizes may affect the validity of these findings.

Perhaps the most notable results with respect to district resources are that students enrolled in the three largest cities other than Central are significantly less likely to be reclassified (by 37 percent) while students enrolled in other urban or suburban high-needs districts are significantly more likely to be reclassified, by a somewhat similar margin (27 percent). Given that Central is both a large city and likely a high-needs district, this split is interesting. On the one hand, it could suggest that the other three Large City districts
might benefit from increased state support; on the other hand, since the high needs
category lumps urban and suburban districts together, the discrepancy could stem from
the effects of suburban settings, even if they lack wealth, which effects would
presumably be absent for the three Large City districts. In either case, again, this finding
is correlational – the higher reclassification rates likely reflect more than simply the
amount of money provided by the state – but may nonetheless be useful for consideration
at the level of policy and decision-making, as I discuss in the final chapter.

Also noteworthy is the fact that mobility had a strong negative effect on
reclassification. Students who moved districts at least once at any point during the
observation period were more than two times less likely to be reclassified than students
who ever moved \( \exp(\beta_{\text{moved}}) = -2.653 \). This finding, too, has been observed in past
research (e.g., Menken et al., 2012), but is notable, nonetheless, since it suggests that any
kind of movement – even movement to a district with better resources – can be
detrimental to ELs’ learning. As I discuss further in the final chapter, this effect could
stem from instructional disruptions that might occur if the student’s new district has a
different instructional environment in terms of language program curricula or types.

4.4.5. Language Instructional Program – Main Effects and Interactions

As described earlier (see sections 3.2.3 and 4.3), the effects of language
instructional programs on reclassification were modeled using one main effect for
program type, and two interaction effects between program type, school year, and home
language. This specification meant that estimates of the effect of program type on
reclassification (1) varied by year, and (2) differed for speakers of Spanish versus
speakers of (all) other languages. These results are presented in Table 4.10; in reading the
Table, it is important to keep in mind that the interaction effects represent group
differences with respect to the program main effects for speakers of each language. Thus,
a significant interaction suggests that the effect of a particular program type is different
for speakers of non-Spanish languages than it is for speakers of Spanish. The exact nature
of the difference depends on the values of the two parameters; their sum will determine if
the difference is one of direction, magnitude, or both.

Table 4.10 shows that, for all students, only bilingual and two-way bilingual
programs differed significantly from the baseline ESL condition. Their effects were
strongest for Spanish speakers, who were nearly two times less likely to be reclassified if
they enrolled in bilingual as opposed to ESL, and roughly 30 percent less likely if they
enrolled in two-way bilingual programs. The positive interaction effects with home
language entail that these two program-types have less of an effect for non-Spanish
speakers. Summing and exponentiating the regression coefficients reveals that bilingual
programs still reduce the probability of reclassification for non-Spanish speakers, but to a
lesser extent than it does for Spanish speakers

\[
(\exp(\beta_{\text{bilingual}} + \beta_{\text{bilingual-NoSpan}}) = \exp(-0.246) = 0.782 = -1.279). \text{ Roughly speaking, non-Spanish speakers in bilingual programs are 28 percent less likely to be reclassified, while Spanish speakers are more than 90 percent less likely.}
\]

The difference for two-way programs is strong enough to produce an opposing
effect for speakers of different languages – while Spanish speakers in two-way programs
are roughly 30 percent less likely to be reclassified compared to Spanish speakers in ESL,
non-Spanish speakers in two-way programs are actually more likely to be reclassified, by
the same margin (30 percent) \((\exp(\beta_{\text{2way}} + \beta_{\text{2way-NoSpan}}) = \exp(0.263) = 1.301)\). In the latter
case, it is relevant to note the very small sample size for non-Spanish speaking students in two-way programs \(n = 221\); given this fact, this finding should be interpreted with caution, and ideally corroborated with additional data before it is given too much weight.

Both types of bilingual programs also had significant, positive interactions with time (for speakers of all languages), suggesting that the difference in reclassification probability between each program and ESL decreased over time (by about 10 percent per year). A significant positive interaction also emerged for “other” language programs that could not be categorized as ESL, bilingual, or two-way bilingual. (Without knowing this state’s definition for ESL instruction, it is difficult to know what this category indicates, although it could potentially include students who only receive pull-out one-on-one language instruction). Whatever “other” indicates, the model suggests that this does not differ significantly from ESL at first, but it begins to significantly increase students’ odds of reclassification over time.

The collective impact of all these effects is illustrated more concretely in Figure 4.5, which shows the cumulative reclassification probability for Spanish and non-Spanish speakers in ESL, bilingual, two-way, and other LEP programs over time. Despite the positive interaction effect, Spanish speakers in bilingual programs remain the least likely to be reclassified by the end of the observation period, at only 65 percent. Non-Spanish speakers in bilingual programs were equally as likely to be reclassified as Spanish speakers in ESL programs (78 and 77 percent reclassified by year 6, respectively), while non-Spanish speakers in ESL programs had a reclassification rate of 87 percent. Non-Spanish speakers in two-way programs were the most likely to be reclassified, at 90 percent. I discuss potential interpretations of these findings in more detail in the final
chapter; generally speaking, however, it is relevant to keep in mind that prior research discussed earlier (Conger, 2010; Umansky & Reardon, 2014) has found evidence of (a) selection bias in terms of which students tend to enroll in bilingual programs, and (b) a long-term reversal in reclassification likelihood for bilingual programs relative to ESL once students enter high school.

4.4.6. Time in the EL Subgroup – Main Effects and Interactions

Finally, Table 4.11 and Figure 4.6 summarize the effects of time in the EL subgroup on reclassification likelihood. The Table shows that each additional year of service prior to Grade 3 significantly increased a students’ probability of reclassification by approximately 15 percent. There was also a small but significant negative interaction between school year and EL year, meaning that the relative advantage of having received more service decreased over time. After six years, the difference in predicted reclassification between ELs who had started in grade 3 versus those who started in Pre-K or earlier was only 11 percentage points (77 percent reclassified versus 88 percent reclassified). These numbers are nonetheless troubling, considering that students who had been ELs for 4 years prior to Grade 3 would be finishing their tenth year in the EL subgroup by the end of the observation period. The fact that 12 percent of the initial cohort of such students are still unlikely to be reclassified after this amount of time is problematic, and will be discussed in more detail in the final chapter.

A specific research question for this study pertained to whether students who had been English learners for a specific amount of time (five or more years) were any more or less likely to be reclassified in any given year. Although the answer to this question can be derived from Figure 4.6 (based on the fact that the curves do not cross), a more straightforward way of answering this question is to display the fitted hazard rate in each
school year for students who have been ELs for one year, two years, three years, etc., as in Figure 4.7. These hazard rates, which represent the proportion of students who are predicted to be reclassified after this school year, are based only on the students in each grade in $EL_{Yr}=1$, $EL_{Yr}=2$, etc. (Due to the structure of the data and the definition of the $EL_{Yr}$ variable, there are not data for all combinations of school year and EL year; see sections 3.1.3 and 4.1). The main take-away from this Figure, again, is that the curves do not cross. Were they to do so, this would suggest that, after a certain point, students who have been ELs for relatively longer become less likely to be reclassified relative to other students. Since the lines do not cross, this means that, in any school year, the students who have been ELs for the longest have the highest hazard rate, and are thus most likely to be reclassified (at least according to the fitted model).

4.5. Preliminary Summary of Results

Taken together, the results presented in this section suggest that ELs vary considerably in how likely they are to be reclassified over time. Time itself affects this likelihood but, as this section has shown, many other factors such as a student’s home language, district resources, and time in the EL subgroup also affect a student’s immediate and long-term likelihood of reclassification. A relevant question is the extent to which these substantive factors vary in their relative impact on students’ reclassification probability. In other words, which substantive factors make the biggest difference in a student’s probability of reclassification? This question was explored in two ways.

First, the cumulative reclassification rates displayed above for various subgroups provide some indication of which factors lead to relatively higher cumulative rates over
time for students who are otherwise the same (i.e., controlling for all other predictors).

For instance, the different district types led to a considerable range of reclassification probabilities (controlling for other variables) relative to the baseline Central district. As Figure 4.4 showed, 92 percent of students in low-needs districts were predicted to be reclassified by the end of the six year observation period, whereas only 67 percent of similar students in the state’s large cities (other than Central) were predicted to be reclassified over the same period. These differences of 15 and 10 percentage points relative to the baseline rate suggest that, after controlling for other variables, the type of district in which a student is enrolled – which likely is a proxy for socioeconomic status – makes a considerable difference in that student’s probability of reclassification over time.

By this same logic, the impact of moving from Central to a different type of district appears to have the single biggest impact on reclassification, lowering the probability by over 30 percentage points, from 77 to 46. After mobility, being retained in grade was the other largest effect observed in this model: As shown in Figure 4.2, only 62 percent of retained students were predicted to be reclassified by the end of the observation period; 15 percentage points fewer than comparable students who were not retained.

An advantage to the approach of comparing cumulative reclassification rates is that they are based on estimates in which all other variables in the model are controlled for. Thus, to reiterate, the difference in cumulative reclassification rates for any two subgroups represent differences for students who are the same on all other model predictors (e.g., the sex effect shows the relative probability for a student who meets all other baseline conditions, but is female). A downside to this method of comparison is that correlations among the various predictors may inflate or deflate the regression
coefficients associated with each predictor, making it difficult to parse out unique contributions of each.

To address this potential downside, a second comparison method was employed. In this framework, each predictor or set of predictors was added separately to the baseline cubic time model, and the incremental change in the Nagelkerke $R^2$ was recorded. The predictor or set of predictors that produces the largest increase in $R^2$ explains the most additional variance in reclassification rates after controlling for time (only). The results of this exploration are displayed in Table 4.12. As the first row of the table shows, time on its own accounted for one-third of the variance observed in the sample ($\Delta R^2 = 0.333$). The various predictors increased the variance explained by anywhere from one-tenth of 1 percent to just over 1 percent.

In reading the table, it is relevant to keep in mind how many additional coefficients are associated with each predictor, since, as a rule, increasing the number of predictors will increase the amount of variance of explained by at least some amount. Thus, although program type and home language clearly explain the most additional variance ($\Delta R^2_{\text{program}} = 0.011; R^2_{\text{language}} = 0.009$), their contributions might still be considered somewhat inefficient, since their effects are specified using 15 and 10 regression coefficients, respectively. (On the other hand, it is also worth noting that not all home language and program coefficients were significant. For program type, for example, only seven of the 15 coefficients were significant (see Table 4.10), suggesting that these coefficients likely do most of the work explaining the additional 1.1 percentage points of variance gained.)
In this sense, from an efficiency standpoint, the effect of moving districts is, again, the single most potent, as it explains an additional 0.3 percentage points of variance using a single coefficient. Particularly on top of the main effects of district type on its own ($\Delta R^2_{\text{district}} = 0.008$), this suggests that a student’s mobility and district profile play a considerable role in his or her likelihood of being reclassified. After mobility, a student’s time as an EL prior to grade 3 also emerges as having a considerable effect – together, this variable’s main effect and interaction with time account for an additional 0.4 percentage points of variance. After this, the third most impactful effect was that of being retained ($\Delta R^2 = 0.002$, for 2 coefficients).

As a final note on these effect sizes, it is worth remembering that all the substantive predictors in combination only explained an additional 4 percentage points of variance beyond time alone. This suggests, on the one hand, that a great deal of reclassification variance remains unaccounted for in this model. On the other hand, it provides some perspective for interpreting the relatively small increments contributed by each substantive predictor. For example, the 1.1 percentage points explained by program type represent over a quarter of the four percentage points explained by all substantive predictors together.

Taken together, these two methods of considering impact and effect size suggest that district mobility and needs, time in the EL subgroup, and being retained in grade may be three effects with relatively higher impact on a student’s probability of being reclassified. Home language and program type also play an important role, though their complex specification in this model suggests that they may tell us relatively less about individual students and be more useful for identifying at-risk groups that might benefit.
from intervention. These types of policy considerations and possible actions will be addressed more deeply in chapter 5.
### Table 4.1. Life table by school year

<table>
<thead>
<tr>
<th>School year*</th>
<th>Number of ELs who were</th>
<th>Proportion of eligible ELs who were</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Enrolled at year’s beginning (Risk set) (Observed)</td>
<td>Reclassified at year’s end (Hazard rate) (Observed)</td>
</tr>
<tr>
<td>1 (G3)</td>
<td>27,511 (Observed)</td>
<td>6,432 (Observed)</td>
</tr>
<tr>
<td>2 (G4)</td>
<td>19,306 (Observed)</td>
<td>5,237 (Observed)</td>
</tr>
<tr>
<td>3 (G5)</td>
<td>13,196 (Observed)</td>
<td>4,304 (Observed)</td>
</tr>
<tr>
<td>4 (G6)</td>
<td>8,339 (Observed)</td>
<td>2,474 (Observed)</td>
</tr>
<tr>
<td>5 (G7)</td>
<td>5,581 (Observed)</td>
<td>1,305 (Observed)</td>
</tr>
<tr>
<td>6 (G8)</td>
<td>4,065 (Observed)</td>
<td>(1,019) (Observed)</td>
</tr>
</tbody>
</table>

Note. “Cum. Survival rate” = cumulative survival rate, meaning the cumulative proportion of students who remain from the initial cohort of 27,511 students. Per-year survival rates can be calculated via (A – B – C) / A for each row. For the final school year, Columns B through F are estimates based on students’ observed ELP scores. Actual reclassification results are not available for these students, as they would be observed in the following school year.

*Because the two cohorts are observed over different but overlapping calendar year spans, school years have been renumbered as 1 through 6 for both cohorts. For the 2007 cohort, school year 1 = the 2006-07 school year; for the 2008 cohort, school year 1 = the 2007-08 school year. All students are enrolled in Grade 3 in the first school year in which they are observed.
### Table 4.2. Life table by EL year

<table>
<thead>
<tr>
<th>EL year</th>
<th>Number of ELs who were</th>
<th>Proportion of eligible ELs who were</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>In this EL year in Grade 3</td>
<td>Survivors from a previous year</td>
</tr>
<tr>
<td></td>
<td>(Observed)</td>
<td>(C – D – E)</td>
</tr>
<tr>
<td>1</td>
<td>1,599</td>
<td>---</td>
</tr>
<tr>
<td>2</td>
<td>3,278</td>
<td>1,217</td>
</tr>
<tr>
<td>3</td>
<td>3,274</td>
<td>3,340</td>
</tr>
<tr>
<td>4</td>
<td>5,790</td>
<td>4,568</td>
</tr>
<tr>
<td>5*</td>
<td>11,907</td>
<td>7,008</td>
</tr>
<tr>
<td>6</td>
<td>1,547</td>
<td>12,861</td>
</tr>
<tr>
<td>7</td>
<td>79</td>
<td>9,482</td>
</tr>
<tr>
<td>8</td>
<td>11</td>
<td>5,799</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>3,679</td>
</tr>
<tr>
<td>10+</td>
<td>---</td>
<td>2,111</td>
</tr>
</tbody>
</table>

*Cases where the reported initial EL year is >5 likely contain errors, although it is plausible that some portion of the observations may be genuine cases where students have repeatedly been retained. For model building purposes, reported values were used as is.
Table 4.3. Summary of censored student characteristics, based on timing

<table>
<thead>
<tr>
<th>Student percentages</th>
<th>Censored</th>
<th>Not censored</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre Grade 8</td>
<td>Post Grade 8</td>
</tr>
<tr>
<td>Female*</td>
<td>47.4</td>
<td>43.8</td>
</tr>
<tr>
<td>Immigrant **</td>
<td>27.4</td>
<td>10.6</td>
</tr>
<tr>
<td>Spanish speaker*</td>
<td>59.8</td>
<td>79.6</td>
</tr>
<tr>
<td>Off-grade level**</td>
<td>4.0</td>
<td>22.5</td>
</tr>
<tr>
<td>Enrolled in ESL*</td>
<td>72.2</td>
<td>84.6</td>
</tr>
<tr>
<td>Enrolled in bilingual*</td>
<td>13.6</td>
<td>5.9</td>
</tr>
<tr>
<td>Enrolled in Central District*</td>
<td>62.6</td>
<td>70.7</td>
</tr>
<tr>
<td>Ever moved districts*</td>
<td>1.2</td>
<td>5.6</td>
</tr>
</tbody>
</table>

Previous Time as EL (as of fall Grade 3)

<table>
<thead>
<tr>
<th>Time</th>
<th>Censored</th>
<th>Not censored</th>
</tr>
</thead>
<tbody>
<tr>
<td>5+ years (2 years Pre-K, or retained)</td>
<td>5.2</td>
<td>8.9</td>
</tr>
<tr>
<td>4 years (ID’ed Pre-K)</td>
<td>30.1</td>
<td>40.1</td>
</tr>
<tr>
<td>3 years (ID’ed K)</td>
<td>19.9</td>
<td>18.3</td>
</tr>
<tr>
<td>2 years (ID’ed Gr1)</td>
<td>14.5</td>
<td>10.0</td>
</tr>
<tr>
<td>1 year (ID’ed Gr2)</td>
<td>20.8</td>
<td>14.2</td>
</tr>
<tr>
<td>0 years (ID’ed Gr3)</td>
<td>9.6</td>
<td>8.5</td>
</tr>
</tbody>
</table>

Lacked ELP scores*         | 9.4      | 2.8          | 1.8         |

*Indicates that all three groups differ significant from one another (p <0.01).
**Indicates that all three groups differ significant from one another (p <0.001).
Table 4.4. Summary of time specification models by school year (SY) and EL year (EL)

<table>
<thead>
<tr>
<th>Format</th>
<th>Parameters</th>
<th>Deviance</th>
<th>LRT to Prev. ((df))</th>
<th>LRT to Gen. ((df))</th>
<th>SABIC</th>
<th>(R^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SY Constant</td>
<td>1</td>
<td>88272.49</td>
<td>---</td>
<td>2730.01*</td>
<td>88279.25</td>
<td>0.300</td>
</tr>
<tr>
<td>SY Linear</td>
<td>2</td>
<td>88119.62</td>
<td>152.87* ((1))</td>
<td>2882.88*</td>
<td>88133.14</td>
<td>0.302</td>
</tr>
<tr>
<td>SY Quadratic</td>
<td>3</td>
<td>86294.32</td>
<td>1825.30* ((1))</td>
<td>4708.18*</td>
<td>86314.61</td>
<td>0.326</td>
</tr>
<tr>
<td>SY Cubic</td>
<td>4</td>
<td>85731.99</td>
<td>562.33* ((1))</td>
<td>5270.50*</td>
<td>85759.04</td>
<td>0.333</td>
</tr>
<tr>
<td>SY Quartic</td>
<td>5</td>
<td>85563.02</td>
<td>168.97* ((1))</td>
<td>5439.48*</td>
<td>85596.83</td>
<td>0.335</td>
</tr>
<tr>
<td>SY General(\text{e})</td>
<td>6</td>
<td>85367.21</td>
<td>195.81* ((1))</td>
<td>---</td>
<td>85407.79</td>
<td>0.337</td>
</tr>
<tr>
<td>SY General (no SY6)</td>
<td>5</td>
<td>91002.49</td>
<td>-5439.47 ((0))</td>
<td>---</td>
<td>91036.31</td>
<td>0.263</td>
</tr>
<tr>
<td>EL Constant</td>
<td>1</td>
<td>88211.94</td>
<td>---</td>
<td>1563.35*</td>
<td>88218.70</td>
<td>0.300</td>
</tr>
<tr>
<td>EL Linear</td>
<td>2</td>
<td>88209.77</td>
<td>2.18 ((1))</td>
<td>1561.18*</td>
<td>88223.29</td>
<td>0.300</td>
</tr>
<tr>
<td>EL Quadratic</td>
<td>3</td>
<td>86916.12</td>
<td>1293.65* ((1))</td>
<td>267.53*</td>
<td>86936.41</td>
<td>0.317</td>
</tr>
<tr>
<td>EL Cubic</td>
<td>4</td>
<td>86820.18</td>
<td>95.94* ((1))</td>
<td>171.59*</td>
<td>86847.23</td>
<td>0.318</td>
</tr>
<tr>
<td>EL Quartic</td>
<td>5</td>
<td>86707.31</td>
<td>112.87* ((1))</td>
<td>58.72* ((5))</td>
<td>86741.12</td>
<td>0.320</td>
</tr>
<tr>
<td>EL General</td>
<td>10</td>
<td>86648.59</td>
<td>58.72* ((5))</td>
<td>---</td>
<td>86716.22</td>
<td>0.321</td>
</tr>
</tbody>
</table>

\(p < 0.001 \)  
\(\text{The general model by school year failed to converge after 500 iterations, likely due to the fact that all students are censored in the final year because their EL status in the next year is not known. The deviance statistics for this model are reported here because they seem plausible, but this model was not considered for final retention due to the convergence error.}\)
Table 4.5. Model Fit and Comparisons for Different Predictor Specifications

<table>
<thead>
<tr>
<th>#</th>
<th>Predictors</th>
<th>Parameters</th>
<th>Deviance</th>
<th>LRT compared to</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>All Binary (df)</td>
<td>All Categorical (df)</td>
<td>SABIC</td>
</tr>
<tr>
<td>1</td>
<td>All Binary</td>
<td>11</td>
<td>84313.85</td>
<td>---</td>
<td>651.84 (18)</td>
<td>84316.23</td>
</tr>
<tr>
<td></td>
<td>One Categorical</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Program only</td>
<td>15</td>
<td>84216.44</td>
<td>97.42 (4)</td>
<td>554.43 (14)</td>
<td>84218.82</td>
</tr>
<tr>
<td>3</td>
<td>District only</td>
<td>16</td>
<td>83947.15</td>
<td>366.70 (5)</td>
<td>285.14 (13)</td>
<td>83949.55</td>
</tr>
<tr>
<td>4</td>
<td>HL only</td>
<td>20</td>
<td>84122.18</td>
<td>191.67 (9)</td>
<td>460.17 (9)</td>
<td>84124.57</td>
</tr>
<tr>
<td></td>
<td>One Binary</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Program only</td>
<td>25</td>
<td>83572.03</td>
<td>561.82 (14)</td>
<td>90.02 (4)</td>
<td>83754.46</td>
</tr>
<tr>
<td>6</td>
<td>District only</td>
<td>24</td>
<td>84021.57</td>
<td>292.28 (15)</td>
<td>359.56 (5)</td>
<td>84023.97</td>
</tr>
<tr>
<td>7</td>
<td>HL only</td>
<td>20</td>
<td>83859.20</td>
<td>454.65 (9)</td>
<td>197.19 (9)</td>
<td>83861.62</td>
</tr>
<tr>
<td>8</td>
<td>All Categorical</td>
<td>29</td>
<td>83662.01</td>
<td>651.84 (18)</td>
<td>---</td>
<td>83664.44</td>
</tr>
<tr>
<td></td>
<td>Adding Interactions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Model 8 + HL*Program</td>
<td>34</td>
<td>83604.71</td>
<td>709.14 (23)</td>
<td>-57.30 (5)</td>
<td>83607.14</td>
</tr>
<tr>
<td>10</td>
<td>Model 8 + SY*Program</td>
<td>34</td>
<td>83052.43</td>
<td>1261.42 (23)</td>
<td>-609.58 (5)</td>
<td>83054.90</td>
</tr>
<tr>
<td>11</td>
<td>Model 8 + both interactions</td>
<td>39</td>
<td>82986.89</td>
<td>1326.96 (28)</td>
<td>-675.12 (10)</td>
<td>82989.37</td>
</tr>
</tbody>
</table>

Note: All models include a cubic specification for time, plus binary indicators for sex, immigrant status, cohort, and off-grade status (8 parameters in total). To see the factor levels for the categorical specifications, please refer to Table 3.4. All results are based on the pooled results of a multiple-imputation dataset with 10 imputations for missing program data.
Table 4.6. Final point estimates for the effect of time on reclassification

<table>
<thead>
<tr>
<th>Predictor</th>
<th>exp(β)</th>
<th>95% CI for exp(β)</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-6.024</td>
<td>[-6.536, -5.556]</td>
<td>p &lt; 0.001</td>
</tr>
<tr>
<td>School year</td>
<td>1.068</td>
<td>[0.979, 1.164]</td>
<td>p = 0.138</td>
</tr>
<tr>
<td>School year&lt;sup&gt;2&lt;/sup&gt;</td>
<td>1.461</td>
<td>[1.391, 1.534]</td>
<td>p &lt; 0.001</td>
</tr>
<tr>
<td>School year&lt;sup&gt;3&lt;/sup&gt;</td>
<td>-1.091</td>
<td>[-1.099, -1.082]</td>
<td>p &lt; 0.001</td>
</tr>
</tbody>
</table>

Table 4.7. Final point estimates for the effect of sex, immigrant, and off-grade status

<table>
<thead>
<tr>
<th>Predictor</th>
<th>exp(β)</th>
<th>95% CI for exp(β)</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female (n = 37,462)</td>
<td>1.213</td>
<td>[1.173, 1.254]</td>
<td>p &lt; 0.001</td>
</tr>
<tr>
<td>Immigrant (n = 15,140)</td>
<td>-1.156</td>
<td>[-1.217, -1.099]</td>
<td>p &lt; 0.001</td>
</tr>
<tr>
<td>2008 Cohort (n = 38,794)</td>
<td>-1.037</td>
<td>[-1.073, -1.003]</td>
<td>p = 0.030</td>
</tr>
<tr>
<td>Off-Grade (n = 3,699)</td>
<td>-3.509</td>
<td>[-4.587, -2.681]</td>
<td>p &lt; 0.001</td>
</tr>
<tr>
<td>Off-grade*School Year</td>
<td>1.348</td>
<td>[1.232, 1.475]</td>
<td>p &lt; 0.001</td>
</tr>
</tbody>
</table>
### Table 4.8. Final point estimates for the effect of home language on reclassification

<table>
<thead>
<tr>
<th>Predictor</th>
<th>exp(β)</th>
<th>95% CI for exp(β)</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Home language (baseline: Spanish, n = 54,121)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Albanian (n = 648)</td>
<td>1.628</td>
<td>[1.375, 1.929]</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Arabic (n = 1,781)</td>
<td>1.093</td>
<td>[0.974, 1.226]</td>
<td>0.133</td>
</tr>
<tr>
<td>Bengali (n = 1,593)</td>
<td>1.944</td>
<td>[1.745, 2.165]</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Chinese (n = 1,663)</td>
<td>1.644</td>
<td>[1.471, 1.837]</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>English (n = 3,383)</td>
<td>1.211</td>
<td>[1.103, 1.330]</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Haitian Creole (n = 1,179)</td>
<td>-1.138</td>
<td>[-1.311, -0.986]</td>
<td>0.076</td>
</tr>
<tr>
<td>Russian (n = 1,012)</td>
<td>2.127</td>
<td>[1.864, 2.428]</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Undetermined (n = 3,417)</td>
<td>1.782</td>
<td>[1.650, 1.925]</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Urdu (n = 1,173)</td>
<td>1.450</td>
<td>[1.275, 1.649]</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Other (n = 8,028)</td>
<td>1.345</td>
<td>[1.269, 1.425]</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

### Table 4.9. Final point estimates for the effect of district resources on reclassification

<table>
<thead>
<tr>
<th>Predictor</th>
<th>exp(β)</th>
<th>95% CI for exp(β)</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>District Type (baseline: Central, n = 52,624)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Charter (n = 712)</td>
<td>1.764</td>
<td>[1.432, 2.173]</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Large City (n = 5,270)</td>
<td>-1.372</td>
<td>[-1.488, -1.266]</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Low Needs (n = 2,416)</td>
<td>1.974</td>
<td>[1.802, 2.162]</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Rural High Needs (n = 509)</td>
<td>1.087</td>
<td>[0.886, 1.334]</td>
<td>0.423</td>
</tr>
<tr>
<td>Sub/Urban High Needs (n = 10,096)</td>
<td>1.276</td>
<td>[1.207, 1.349]</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Average Needs (n = 6,364)</td>
<td>1.698</td>
<td>[1.596, 1.807]</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Ever moved districts</td>
<td>-2.653</td>
<td>[-2.985, -2.353]</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>
Table 4.10. Final point estimates for the effect of language instruction program type

<table>
<thead>
<tr>
<th>Predictor</th>
<th>exp(β)</th>
<th>95% CI for exp(β)</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Program main effects, Spanish speakers (baseline: ESL, n = 39,048)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bilingual, Spanish speakers (n = 9,967)</td>
<td>-1.927</td>
<td>[-2.101, -1.764]</td>
<td>p &lt;0.001</td>
</tr>
<tr>
<td>2-way Bilingual, Spanish (n = 3,469)</td>
<td>-1.307</td>
<td>[-1.479, -1.153]</td>
<td>p &lt;0.001</td>
</tr>
<tr>
<td>Multiple Programs, Spanish (n = 290)</td>
<td>1.366</td>
<td>[0.843, 2.211]</td>
<td>p= 0.204</td>
</tr>
<tr>
<td>Other LEP, Spanish (n = 717)</td>
<td>-1.366</td>
<td>[-1.835, -1.016]</td>
<td>p= 0.039</td>
</tr>
<tr>
<td>Refused Service, Spanish (n = 629)</td>
<td>-1.110</td>
<td>[-1.988, 0.620]</td>
<td>p= 0.724</td>
</tr>
<tr>
<td><em><em>Program main effects, non-Spanish speakers</em> (baseline: ESL, n = 21,633)</em>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bilingual, non-Spanish speakers (n = 791)</td>
<td>1.506</td>
<td>[1.190, 1.906]</td>
<td>p= 0.001</td>
</tr>
<tr>
<td>2-way Bilingual, non-Spanish (n = 221)</td>
<td>1.700</td>
<td>[1.220, 2.368]</td>
<td>p= 0.002</td>
</tr>
<tr>
<td>Multiple Programs, non-Spanish (n = 39)</td>
<td>1.298</td>
<td>[0.547, 3.078]</td>
<td>p= 0.553</td>
</tr>
<tr>
<td>Other LEP, non-Spanish (n = 917)</td>
<td>-1.425</td>
<td>[-1.887, -1.075]</td>
<td>p= 0.014</td>
</tr>
<tr>
<td>Refused Service, non-Spanish (n = 277)</td>
<td>1.423</td>
<td>[0.930, 2.177]</td>
<td>p= 0.104</td>
</tr>
<tr>
<td><strong>School Year*Program Interactions</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SY*Bilingual</td>
<td>1.117</td>
<td>[1.063, 1.173]</td>
<td>p &lt;0.001</td>
</tr>
<tr>
<td>SY*2-way Bilingual</td>
<td>1.111</td>
<td>[1.036, 1.192]</td>
<td>p= 0.003</td>
</tr>
<tr>
<td>SY*Multiple Programs</td>
<td>-1.059</td>
<td>[-1.294, -0.867]</td>
<td>p= 0.575</td>
</tr>
<tr>
<td>SY*Other LEP</td>
<td>1.195</td>
<td>[1.085, 1.315]</td>
<td>p &lt;0.001</td>
</tr>
<tr>
<td>SY*Refused Service</td>
<td>-1.088</td>
<td>[-1.330, -0.889]</td>
<td>p= 0.410</td>
</tr>
</tbody>
</table>
Table 4.11. Final point estimates for the effect of time in the EL subgroup on reclassification

<table>
<thead>
<tr>
<th>Predictor</th>
<th>exp(β)</th>
<th>95% CI for exp(β)</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-Grade 3 Time as EL</td>
<td>1.155</td>
<td>[1.132, 1.179]</td>
<td>p &lt;0.001</td>
</tr>
<tr>
<td>EL Year*School Year</td>
<td>-1.035</td>
<td>[-1.046, -1.026]</td>
<td>p &lt;0.001</td>
</tr>
</tbody>
</table>

Table 4.12. Incremental R² Increase Per Predictor, Controlling for Time Only

<table>
<thead>
<tr>
<th>Predictor(s)</th>
<th># of predictors</th>
<th>Nagelkerke $R^2$</th>
<th>Total $\Delta R^2$ (relative to time only)</th>
<th>$\Delta R^2$ per predictor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time (baseline)</td>
<td>4</td>
<td>0.333</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Program Type</td>
<td>15</td>
<td>0.344</td>
<td>0.011</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Home language</td>
<td>10</td>
<td>0.342</td>
<td>0.009</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>District Type</td>
<td>6</td>
<td>0.341</td>
<td>0.008</td>
<td>0.001</td>
</tr>
<tr>
<td>Time as EL</td>
<td>2</td>
<td>0.337</td>
<td>0.004</td>
<td>0.002</td>
</tr>
<tr>
<td>District Moves</td>
<td>1</td>
<td>0.336</td>
<td>0.003</td>
<td>0.003</td>
</tr>
<tr>
<td>Off-Grade</td>
<td>2</td>
<td>0.335</td>
<td>0.002</td>
<td>0.001</td>
</tr>
<tr>
<td>Sex</td>
<td>1</td>
<td>0.334</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>Immigrant</td>
<td>1</td>
<td>0.334</td>
<td>0.001</td>
<td>0.001</td>
</tr>
</tbody>
</table>
Figure 4.1. Fitted probability of reclassification by school year for various time specifications
Note. Data labels in the figure represent the cumulative failure (reclassification) rate at the end of the observation period. The values can be interpreted as percentages; for example, 0.77 indicates that 77 percent of the students in the baseline condition would be reclassified by the end of the observation period.

Figure 4.2. Cumulative failure rates based on sex, immigrant, and off-grade status
Note. Data labels in the figure represent the cumulative failure (reclassification) rate at the end of the observation period. The values can be interpreted as percentages; for example, 0.77 indicates that 77 percent of the students in the baseline condition would be reclassified by the end of the observation period.

Figure 4.3. Cumulative failure rates based on home language
Note. Data labels in the figure represent the cumulative failure (reclassification) rate at the end of the observation period. The values can be interpreted as percentages; for example, 0.77 indicates that 77 percent of the students in the baseline condition would be reclassified by the end of the observation period.

Figure 4.4. Cumulative failure rates based on district resource needs
Note. Data labels in the figure represent the cumulative failure (reclassification) rate at the end of the observation period. The values can be interpreted as percentages; for example, 0.77 indicates that 77 percent of the students in the baseline condition would be reclassified by the end of the observation period.

Figure 4.5. Cumulative failure rates based on language instruction program type
Note. Data labels in the figure represent the cumulative failure (reclassification) rate at the end of the observation period. The values can be interpreted as percentages; for example, 0.77 indicates that 77 percent of the students in the baseline condition would be reclassified by the end of the observation period.

**Figure 4.6.** Cumulative failure rates based on time in the EL subgroup

**Figure 4.7.** Fitted hazard rates by EL year for students in different school years (grades)
CHAPTER 5  
DISCUSSION

This study was motivated by an interest in the process and policies by which English learners meet their state’s language proficiency standards to exit the EL subgroup and be reclassified as former ELs. Using a large, longitudinal sample of students from one state, I used discrete-time survival analysis to build a fitted logit hazard model to predict the likelihood that different subgroups of ELs would meet the reclassification standard over time. The model included several types of substantive predictors, to explore what factors other than time might decrease or improve students’ odds of reclassification. In this final chapter I discuss the results of this survival analysis first with respect to the specific research questions that guided this study, and subsequently with respect to more general policy and research implications.

5.1. Results by Research Question

As stated in the first chapter of this dissertation, the present study was guided by three specific research questions. They were:

1. What is the probability of reclassification over time for a cohort of ELs progressing from grade 3 to grade 8?
   a. When are third grade ELs most likely to be reclassified?
   b. Does the probability of reclassification vary by either grade-level, amount of time in the EL subgroup, or both?

2. Are students who have been in the EL subgroup for a long time (>5 years) more or less likely to be reclassified than more recently identified students?

3. What factors affect the probability of reclassification for students in the same grade-level, or who have been ELs for the same amount of time?
This discussion chapter begins with focused responses to each of these questions before branching out to more general recommendations and observations about reclassification patterns in this state.

**5.1.1. Reclassification probability over time, grade-level**

Generally speaking, the probability of reclassification did change over time, whether counted by school years or EL years. By both metrics, the probability increased to a peak, before decreasing thereafter. By school year, students were most likely to be reclassified after their third observed school year, which was Grade 5 for 95 percent of students. Thirty-three percent of eligible ELs were reclassified based on their performance in this school year, and the probability of reclassification decreased thereafter, meaning students were less likely to be reclassified in middle school, particularly towards the end. Grouping students by EL year, meaning the amount of time they spent in the EL subgroup receiving services, yielded similar results: the largest proportion of eligible students (30 percent) were reclassified after their seventh year as an EL, which corresponded to Grade 5 for a majority of students (58 percent).

Taken together, these findings suggest that, in line with previous studies in California and Massachusetts (Slama, 2014; Thompson, 2012), Grade 5 appears to be a critical window for reclassification. A large proportion of students make the transition in this school year, and students who do not manage to meet their state’s reclassification standard by this horizon face decreased odds of being reclassified thereafter. Having noted this, it is also relevant to note that large numbers of ELs were reclassified earlier than Grade 5; indeed, more students were reclassified after both Grades 3 and 4 relative to Grade 5 (see Table 4.1 for exact values; the differences across years were roughly 1,000 students per), such that more than half of the original sample (roughly 13,500
students) had already been reclassified by the beginning of Grade 5. The proportions in earlier years are smaller, however, because the total number of eligible students is larger. On the flip side, the proportion of students reclassified after Grade 6 was similar to the Grade 5 proportion, at 30 percent. Thus, while Grade 5 appears to be an important turning point for reclassification, it is also important to recognize that large numbers of ELs are reclassified throughout elementary school, and a similar proportion of students were reclassified after Grade 6.

In considering this finding, it is also important to remember that the observation period for this study begins in Grade 3. Since the sample excludes students who have already been reclassified prior to Grade 3, it is likely that the median lifetime and peak reclassification years are positively biased (i.e., they are higher than they likely would have been had the observation period begun in kindergarten and included all eligible ELs from that starting point). As referenced in section 2.5, Slama (2014) observed a median lifetime of just over 3 years in her study of Massachusetts ELs, which began in kindergarten. (It is important to note that California studies did not find this pattern, due to policy: since the ELA test is used in reclassification decisions in that state, the absolute earliest an EL can be reclassified is Grade 3, regardless of when he or she is first identified). Although it cannot be assumed that students in this state would follow a similar pattern, it is plausible, and even likely, that large numbers of students experienced reclassification prior to the beginning of the current observation period. Had they been included in this sample, the median lifetime and peak reclassification year might be lower (although, notably, Slama (2014) also found Grade 5 to be the year with the highest
probability of reclassification). This caveat also sets the stage for some policy considerations about reclassification, which I discuss in section 5.2.

5.1.2. Reclassification probability based on time in the EL subgroup

In addition to the findings reported in the preceding section about peak reclassification points based on time in the EL subgroup, this study also included a question specifically about whether students who have been ELs for relatively longer (5 years or more) are more or less likely to be reclassified than ELs who have been students for less time. This research-driven question was motivated by an interest in long-term ELs, and potentially separating the confluence of EL years and school years in previous studies that have used cohorts starting in kindergarten (Slama, 2014; Thompson, 2012). In other words, would the fifth or sixth year of service emerge as an important reclassification peak even for students who begin their journeys as ELs later than kindergarten?

It has already been noted in the preceding subsection that students were most likely to be reclassified after their seventh year of being an EL (which corresponded to Grade 5 for the majority of students in this sample). This suggests that, to a point at least, students who have been ELs for five years or longer are actually more likely to be reclassified than students who have been ELs for less time. Further evidence for this can be viewed in Figure 4.7, which shows that, in every school year, students who have been ELs for longest are always the most likely to be reclassified, according to the fitted hazard model. Thus, although all ELs experience a decrease in their probability of being reclassified in middle school, students who have been ELs relatively longer still maintain a slight edge in their hazard rates compared to students who have been ELs for relatively
less time. The amount of time students spend in the EL subgroup does not appear to become a disadvantage at any point during the observation period used for this study.

5.1.3. Factors affecting probability of reclassification

While the first two research questions focus primarily on the effects of time on the probability of reclassification, the final research question allows for an opportunity to consider the effects of predictors other than time on students’ probability of meeting the state’s ELP standard. As a starting point, it is important to acknowledge that time alone accounted for one third of the observed variance in students’ reclassification rates, and explained the most variance by far relative to other predictors included in the fitted hazard model. All other substantive predictors combined – home language, district type, program type, sex, immigrant status, time as an EL prior to grade 3, off-grade status, and cross-district mobility – explained only an additional 4 percentage points of variance beyond the baseline set by time alone. Ultimately, this may be good news, as it suggests that demographic or instructional differences among students tell us far less about their likelihood of meeting the state’s ELP standard compared to how long they have been enrolled in school learning English.

Nonetheless, group differences did emerge. In fact, the effects on reclassification probability of all substantive predictors (see Table 3.4 to review the full list) were significant, with the exceptions of: cohort (i.e., whether students were in Grade 3 in the 2006-07 school year versus the 2007-08 school year; see Table 4.7), two specific home languages (Arabic and Haitian Creole; see Table 4.8), one district type (rural high needs; see Table 4.9), and three program “types” (multiple programs, “other” programs, and refusal of service; see Table 4.10). Among the significant effects, three emerged as particularly influential based on their cumulative effects on reclassification over time, and
on the amount of additional variance they explained. These were: district mobility, retention in grade, and district resource levels relative to the baseline of the Central district (see the final subsection of Chapter 4 for a more in-depth discussion on this). Controlling for other factors, the cumulative effects of mobility led to a gap of over 30 percentage points over time, while retention and district profile led to gaps of approximately 15 percentage points. As discussed in section 4.1, students who had moved or been retained were also significantly more likely to still be ELs at the end of the observation period. Combinations of program type and home language also led to a considerable spread in reclassification rates over time, with gaps of 12 – 13 percentage points in each direction depending on students’ home language and program type.

Since the baseline cumulative reclassification rate was 77 percent over the six-year observation period, differences of 12 and 15 percentage points are considerable in their implications. Consider, for example, that a negative gap of this size leads to less than two-thirds of students being reclassified after six years (62 percent, in the case of retained students, or 65 percent for Spanish speakers in bilingual programs) whereas a positive gap leads to nearly all students being reclassified over the same period (92 percent, in students in low-needs districts, or 90 percent for non-Spanish speakers in two-way programs). From a policy- and decision-making perspective, such differences are important, and would likely be of a sufficient scale to warrant interventions or policy changes, if they are feasible. Such interventions and changes will be considered in section 5.2, which comes next.
5.2. Policy Implications and Recommendations

The findings from this study have several implications for potential policy actions and changes, both within this state and with respect to reclassification more generally. This policy discussion is divided into three subsections: the first focuses on this state’s reclassification policies and standards, and the effects these have on the findings observed here. To the extent that such findings are positive or negative, changes to the reclassification standards and policies could mitigate the outcomes and lead to more desirable observations in the future. The second subsection considers other state- and district-level policies that could be considered to improve students’ chances of reclassification (or to decrease their chances of remaining ELs after many years of service). The third subsection summarizes traits or patterns that could be used to identify students potentially at risk of not being reclassified.

5.2.1. Considerations for Reclassification Policies

As discussed in section 3.1.2, reclassification decisions in this state during this observation period were made on the sole basis of the ELP assessment. Students must earn scores that place them in the highest of four performance levels on each of two score scales, representing oral language (a combination of speaking and listening) and written language (a combination of reading and writing). This two-score conjunctive model is simpler than the models used in both California and Massachusetts, where reclassification decisions depend on four domain subtest scores, as well as an overall composite score, and scores on the state’s ELA content assessment (Linquanti & Cook, 2015). (As a note, California and Massachusetts do use the ELA score differently).

As noted in section 2.3 (specifically Carroll and Bailey’s (2015) study), this suggests that students should have a relatively easier time being reclassified in this state;
and this, indeed, is born out in the results observed here. Specifically, whereas studies in Massachusetts and California found that approximately a quarter of all ELs remained in the subgroup after eight (Slama, 2014), ten (Thompson, 2012), and twelve (Umansky & Reardon, 2014) years of service (the latter based solely on Latino ELs), this study found that 85 percent were reclassified by the end of the observation period (see descriptive survival rates in section 4.1), which represented an average of 9 years of service. In the fitted models, meanwhile, 77 percent of students were predicted to be reclassified after only six years of service, and 88 percent were predicted to be reclassified after 10 years of service (see Figure 4.6). In summary, then, the proportion of students who remain ELs after 10 years of service is roughly half as large in this state as it is in others that have been studied.

Whether or not this is a good thing depends, importantly, on how reclassified students actually fare once they have stopped receiving services. An analysis of students’ academic content performance was beyond the scope of this dissertation; however, a follow-up study is planned in which relationships between language and content scores are explored over time (see Appendix). It will be of specific interest to determine whether these relationships are lower for students who meet the state’s reclassification standard relative to those fall short of the standard15. Other types of future research will also be discussed in the final section of this chapter (see section 5.3); for the purpose of the current section, the primary point is that reclassification rates do appear to be relatively better – specifically, higher – in this state compared to others that have been studied.

---

15 As discussed in sections 2.3 and 2.4, this is a more appropriate metric than students’ content scores alone, since the goal of reclassification is not to ensure that students meet the academic content standards. Rather, it is only to ensure that language no longer prevents them from meeting said standards.
While this finding likely has multiple causes – e.g., differences in the student population, the types and quality of language instruction programs offered (Conger, 2010; Umansky & Reardon, 2014), and the state’s academic content performance standards (Cook et al., 2012) – the reclassification standard and model is undoubtedly among these. If the state had concerns about reclassification rates being too high or too low, one way to address these would be to adjust the standard to be either simpler or more complex. In fact, as referenced in section 3.1.2, and as I discuss further in the final section 5.3, the state has changed to a different and more complex reclassification standard in the years subsequent to those observed here. It is thus likely that a follow up study using data produced by this new rule would produce different results.

It is also worth noting here that, as discussed in the summary of time-censored students (see section 4.1), the majority of students who remained ELs at the end of the observation period studied here were falling short based on their written language scores. Again, to the extent that the written language standard accurately reflects the reading and writing skills that ELs need to engage in core academic content without support, this finding is not necessarily bad. Indeed, allowing students to reclassify based on lower scores might set them up for subsequent failure, as they would lose access to supports that might otherwise benefit them while they are still developing the written language skills they need for success. If, however, research revealed that students who remain ELs past middle school are more likely to drop out or fall short of expectations, the state might consider allowing schools and districts more discretion in their reclassification decisions in Grade 8 and beyond. Some of the designs proposed for future research in
section 5.3 could provide evidence to determine whether such a consideration is warranted.

5.2.2. Other Policy Considerations

As discussed in sections 4.5 and 5.1, intrastate mobility and retention in grade were two factors that accounted for some of the biggest differences in reclassification rates among otherwise similar students. This information may be useful from a policy perspective, as there are concrete steps this or any state could consider taking to try to mitigate the otherwise negative effects of these factors. First, the state might consider maximizing and facilitating the sharing of student information across districts, to try to ensure that students who move within the state can bring as much information as possible with them to their new district. Information about their past ELP instruction and performance is likely to be particularly useful in this regard; relatedly, the extent to which the state can encourage districts to use the same or similar curricula and courses of study for EL language programs may also help minimize disruptions for students who move. Taken together, the idea behind these two suggestions is to try to minimize situations in which an EL moves to a new district and finds himself placed into an entirely new or different learning environment with respect to what he is taught, how he is taught, and what he is expected to already know. On the retention front, since the effects of retention seemed to diminish over time, the state might just wish to consider monitoring retained ELs to ensure that they do, in fact, fall into step with their new cohort after being retained, rather than continue to fall behind.

Another malleable factor that affected reclassification considerably was the type of language instruction a student enrolled in, particularly if that student was a Spanish speaker (as most ELs are). On this front, a notable finding was that students in two-way
bilingual programs outperformed comparable students in EL-only bilingual programs by a considerable margin, regardless of whether they spoke Spanish or another home language (see Figure 4.5). They also performed comparably to similar students enrolled in ESL programs (who, generally speaking, tended to have higher reclassification rates). By contrast, Spanish speakers enrolled in EL-only bilingual programs were less likely to be reclassified than any other home language or program subgroup.

This is an important and potentially useful piece of information for the state, particularly since parents typically have a say in what types of programs their children enroll in. Reclassification aside, there are many reasons why parents may feel strongly that they want their students enrolled in bilingual programs. And indeed, even if students in such programs take longer to reclassify, the use of their home language may have other benefits that counterbalance the negative effects on time to reclassification. Several studies have found that students who enroll in bilingual programs tend to perform better than students in ESL programs across the full K-12 career (Collier & Thomas, 1989; Rolstad, Mahoney, & Glass, 2008; Slavin & Cheung, 2005; Thomas & Collier, 2002; Umansky & Reardon, 2014). From a more pragmatic standpoint, many states – including the one studied here – have so-called “parent trigger” laws where a quorum of parents can demand that their district offer some type of bilingual education for their children. In short, it is neither practical nor, potentially, desirable to recommend that states not offer bilingual education as an option.

In light of this reality, the findings here suggest that states should potentially prioritize two-way bilingual programs as the bilingual model of choice, to the extent practicable. In other words, given that states are essentially required to have some kind of
bilingual offering, it may be prudent to focus primarily or exclusively on two-way programs to satisfy this requirement. Two-way programs – also known as dual immersion programs – enroll both ELs and native English speakers; all students are taught in both languages, and the goal is for all students to achieve bilingualism and biliteracy by the program’s end (most programs are designed to last from either K-5 or K-8) (Genesee, 1999; Howard, Christian, & Genesee, 2004; US Department of Education; Office of Planning, Evaluation and Policy Development; Policy and Program Studies Service, 2012; Valentino & Reardon, 2015). As this description suggests, there may be several factors beyond the instruction itself that lead two-way programs to serve ELs well. For example, buy-in from native English-speaking families will mean there are more parents invested in the program’s quality and rigor – not to mention more parents who may feel empowered or entitled to intervene or speak up if this is not the case. Similarly, two-way programs also may help to promote a more positive school culture with respect to the value of other languages and cultures. Two-way programs, in other words, compel a greater proportion of the school’s community to care about language development and instruction. While there are many reasons why two-way programs may not always be feasible to implement – for instance, they require buy-in from non-EL students and parents; they also can be challenging for students who transfer into the program after kindergarten; they also require teachers who are trained and qualified to teach both content and language – the findings from this study suggest that such challenges may be worth surmounting, if possible.
5.2.3. Identifying and Supporting At-Risk Students

Finally, not all factors affecting reclassification can be directly addressed through policy. While sex, home language, and immigrant status may affect a student’s probability of reclassification, these are not factors that districts or schools can address directly through intervention. At best, however, they may serve as useful markers for identifying students who are at greater risk of not being reclassified. To this end, the information about which students were still ELs at the end of this observation period – as well as why they were still ELs – can also be useful for decision-making. As discussed in section 4.1, ELs who were male, Spanish speaking, and native born, were significantly more likely to still be ELs at the end of the six year observation period studied here. They were, moreover, overwhelmingly likely to be falling short based on their written language performance. These facts suggest that, if the state is interested in increasing its reclassification rates, these are students who would likely benefit from additional monitoring and support, particularly related to reading and writing skills, and particularly if they are still ELs after Grade 6.

5.3. Limitations and Further Research

I have made an effort to acknowledge limitations of this sample and study throughout my presentation of the data, methods, results, and discussion. To reiterate the primary shortcomings of this study, they are that the sample excludes EL students who (1) are reclassified prior to Grade 3; (2) are identified later than Grade 3; (3) have individualized education plans (IEP) or other known disabilities; (4) are outside the single state being studied here. The specific ways in which these facts may bias the results found here are discussed in particular in sections 3.1, 4.1, 5.1, and 5.2. In general, these limitations entail that the findings here are specific to this state and sample, and cannot
necessarily be generalized to other states, or to other types of EL students within this particular state.

Beyond this oft-repeated condition, there are a few other limitations worth noting. First, the labeling of program types from the state database (e.g., bilingual, ESL, two-way bilingual) may suggest more standardization or continuity than is actually reflected in classroom practice and implementation. There are numerous types of bilingual programs (Genesee, 1999; US Department of Education; Office of Planning, Evaluation and Policy Development; Policy and Program Studies Service, 2012; Valentino & Reardon, 2015) that vary in their exact design and goals, not to mention variations from school to school in the quality of instruction or implementation for programs intended to have “the same” design. As a result, it must be acknowledged that the program type labels used here are something of a blunt instrument and may still mask considerable variety in on-the-ground services for students.

Also related to program type is the fact that multiple studies (Conger, 2010; Umansky & Reardon, 2014; Valentino & Reardon, 2015) have found evidence for selection bias in which types of students enroll in which programs. In other words, since parents have some say in their children’s program type, the students in bilingual programs are a self-selecting bunch who may differ from students in other programs in important ways. This suggests, by extension, that differing reclassification rates across different program types may stem from more than just the program type itself. Based on this point and the preceding one, conclusions about the relative impact or effectiveness of specific program type should be interpreted with caution.
It is also important to note, as mentioned in section 3.1.2, that the state studied here subsequently shifted to a new reclassification model. Starting in the 2012-13 school year, the state adopted a new model with a more complex decision rule involving five total scores, representing the four domains and an overall linguistic proficiency composite score. Based on this increased complexity, it is likely that observed reclassification rates dropped in the state following the implementation of this new rule. Even if they have not, it follows by extension that the patterns observed here are not generalizable past the end of the 2011-12 school year.

This point underscores the relevance and need, however, for ongoing research about reclassification patterns in this state. As a starting point, a repeat survival analysis study using data under the new reclassification model will provide information about how the shift in criteria has affected the probability of reclassification over time, both in general, and for similar subgroups of students. From there, several designs would be particularly well-suited to flesh out such findings – using both the old and new reclassification models – by evaluating the validity of the respective reclassification standards. First, as mentioned in section 5.2, follow-up studies about the relationships between language and content performance would provide information about whether the role of language is sufficiently diminished at the chosen reclassification cut points. The methods proposed by Cook et al. (2012; see section 2.3) could be used for this purpose to (1) see how language and content relate for students with a variety of ELP levels, and (2) compare the relationships and patterns observed using the old and new ELP assessments and standards.
Second, it would be informative to study the academic content performance of R-FEP students once they have been reclassified, to see whether it supports the conclusion that they have the linguistic skills they need to meet academic content standards without additional support. Regression discontinuity is a useful method for these types of studies, as it manages selection bias by comparing the trajectories of students with comparable ELP performance (i.e., within a standard error of measurement) whose scores place them just above and below the selected cut score. If significant differences in performance or growth are observed subsequent to the students’ divergence, the condition (EL vs. reclassified) that produces better outcomes is presumably the one that serves the students better, and the state might consider adjusting its reclassification standards or policies accordingly to maximize the number of students who will be placed in this condition.

While there are many promising avenues for future research on this topic, the current study provides important and, hopefully, useful foundational information about reclassification patterns in a state that has not, to date, been studied or reported on. The findings reported should enhance our understanding of how reclassification works, both within and across states – including, in particular, which factors seem to affect students regardless of their state’s policies, and which seem to depend on context and policy.
APPENDIX: FOLLOW-UP STUDY

RELATIONSHIPS BETWEEN LANGUAGE AND CONTENT ASSESSMENT PERFORMANCE FOR ELS IN GRADES 3-8

As discussed in the final chapter of this dissertation, an important and useful follow-up study involves exploring the relationships between language and content performance for ELs in this sample, particularly those who qualify for reclassification based on their ELP scores. These analyses ultimately went beyond the scope of the dissertation, but the design for this follow-up study has been planned and will be pursued as an independent analysis. This appendix provides a brief overview of the study’s design.

A.1. Research Questions

This study has one general research question, which is explored by three specific subquestions:

1. How do ELs’ language assessment scores relate to their content assessment scores, and how do these relationships change over time?
   a. To what extent do relationships between language and academic content vary over time as a function of ELP, as measured by proximity to reclassification?
   b. To what extent do relationships between language and academic content vary over time as a function of grade-level?
   c. To what extent do relationships between language and academic content vary over time as a function of time in the EL subgroup?

These questions will be explored using a longitudinal multilevel model for change, which will estimate relationships between language and content scores for different groups of students over time.
A.2. Data Characteristics

This follow-up study will use the same dataset used for the dissertation, though it will make use of additional variables that are not included in the survival analysis used for the main study. The characteristics of these additional variables are described here.

**Content Assessment Scores**

Students’ content achievement is measured in this study by the state’s large-scale summative assessments in ELA and mathematics. The dataset includes students’ content assessment scores and the corresponding performance level for as many years as they remained in the state, including any post-reclassification years that fall within the observed time period. For most students, this meant there were six years of content assessment scores, although some students had fewer (see section 3.1.4 for a discussion of missing data and attrition). Across all years, scores from these assessments are used to sort students into one of four performance categories. The labels for these performance categories changed several times over the observed time period, but they all followed a consistent rule that students scoring in or above the third performance category are considered proficient in either content area.

Up until the 2008-09 school year, the state used the same cut score of 650 for both content areas and all grade-levels. In the 2009-10 school year, they adjusted their cut scores and also adopted different cuts for different grade levels. They continued with this model for the next three years, and then shifted to entirely new content assessments in the 2012-13 school year (the final year for the 2008 cohort), which were aligned to the Common Core State Standards (CCSS) and reported on an entirely new score scale. To

---

16 As a clarifying note, the dataset does not include item-level data for students’ assessment performance – only scaled scores and performance levels.
work around these variations in cut scores and scales across years and grade levels for both ELA and mathematics, all scale scores have been recentered around their own proficiency cut score, by year and content area. By subtracting the cut score from every student’s scale score for each test, the cut score becomes 0, and students are then sorted and scaled based on their proximity to this benchmark. Students who earn scale scores above the cut will have positive scale score values that represent the positive difference between their score and the cut. Similarly, students below the cut score will have negative scale score values representing the negative distance between their score and the cut.

Figure A.1 and Figure A.2 display the average performance in ELA and mathematics, respectively, for both cohorts using this recentered scale. The graph clearly reflects the two policy changes described above, as drops in average performance are visible in the 2009-10 school year and the 2012-13 school year (particularly the latter).17 Apart from the 2012-13 school year, the figures show that the students in the sample were actually meeting performance standards in math, on average, and were falling just shy, on average, in ELA. In interpreting this graph, it is important to remember that the composition of ELs and former ELs is shifting each year, as more and more students are reclassified. These averages are summarized in greater detail in Table A.1 and Table A.2. Here again, the mean scores illustrate the shifts in achievement standards in 2009-10 and 2012-13; we also see that the score distributions are generally leptokurtic, but approximately symmetric (skewness is between -0.5 and 0.5). Two notable exceptions are the ELA score distributions in the 2008-09 and 2009-10 school years, which are very

17 Although it is not possible to tell from my data whether non-ELs experienced a similar drop in performance, a review of the state’s accountability data on www.eddataexpress.ed.gov suggests that the proportion of students meeting standards in this state dropped by nearly half from 2011-12 to 2012-13 (roughly 50-70 percent proficient in 2011-12, compared to 30-40 percent proficient in 2012-13).
skewed (positively, in three of four instances). It is not immediately clear why this should be the case, though I will keep it in mind as I conduct analyses and interpret my findings.

**ELP Assessment Scores and Reclassification**

Students’ ELP is measured in this study by scores from the state’s large-scale ELP assessment. Two different English language proficiency assessments were used in the state over the seven years of data covered by the two cohorts. Both ELP assessments comprised four subtests in reading, writing, listening, and speaking, but they produced different types of scores and were incorporated into reclassification decisions in different ways. In the earliest version of the assessment, which I will call ELPA_1, two scale scores were reported: oral language, and written language. These scale scores were created by combining the raw scores on two subtests – listening and speaking for oral language, and reading and writing for written language – and scaling these combinations separately. To be reclassified on this assessment, students needed to meet or exceed cut scores on both scales. Similarly, to advance from one proficiency level to the next, students needed to clear the cut score on both scales.

In the 2012-13 school year, the state shifted to a new ELP assessment, which I refer to as ELPA_2. On this assessment, raw scores from all four subtests are combined, and the combined raw score is scaled to produce one scale score representing overall proficiency. To be reclassified with this test, students needed to meet the overall cut score and have raw scores exceeding certain minima on each of the four subtests (a combination model, by Carroll & Bailey’s (2015) definitions). Students who meet the cut score but do not clear the minimum raw score requirements are relegated to the second highest performance level, and are not eligible for reclassification. For both
ELPA_1 and ELPA_2, students are placed in one of the same four performance categories: Beginning, Intermediate, Advanced, and Proficient. The cut score for proficiency and reclassification is between Advanced and Proficient, meaning students must score in the highest performance level to be reclassified.

As with the content scores, I recentered all ELPA scores (for ELPA_1 and ELPA_2) on the proficiency cut score each year, so that the cut score became 0 on the score scale. On this transformed scale, positive scores indicate positive distances beyond the cut score, and negative scores indicate negative distances below the cut score. It is also important to note that, in contrast to content scores, students have different amounts of ELPA data depending on when they were reclassified. Also, due to the shift to ELPA_2 in the 2012-13 school year, the 2008 cohort has only one score in their final year. Average scores on this proficiency-centered scale are displayed in Figure A.3. What is clearest in these figures is that the written language cut score is likely to be the limiting factor in reclassification decisions in this state; save one exception each, students in all years and both cohorts are meeting or exceeding the oral language standard, on average, but falling short on the written language test.

In addition, in contrast to the content assessments where patterns were clearer based on school year, patterns emerge on the language assessment according to grade-level. The figures show that both cohorts follow very similar patterns for grades 4 through 7; in 2012-13, when the 2008 cohort was assessed with the ELPA_2 assessment, the average total score is almost exactly the average of the 2007 cohort’s mean oral and written scores on the ELPA_1.
The score distributions for the centered ELP scale scores are summarized in Table A.3 and Table A.4. An additional observation here is that the score variance appears to decrease across grade-levels in both cohorts. This is likely due to the fact that higher achieving ELs are being reclassified after each passing school year, rendering the remaining subgroup slightly more homogenous (and lower achieving) with each passing year. These distributions are also more symmetric than the content score distributions, with no notable skewness except, perhaps for the 8th grade data for the 2007 cohort.

As a final note about these scores, it is important to appreciate that the transition from ELPA_1 to ELPA_2 affected both the number of scores students receive, and the scale on which these scores are reported. For both ELPA_1 and ELPA_2, state guidance specifies that students’ ELP scores are meant to be the sole criterion for determining their reclassification status. Students who meet the reclassification criteria in their year are reclassified, and students who do not are not. Given the research findings about different types of reclassification systems that are discussed above (see sections 2.2 and 2.3), the shift from a conjunctive system with two scores for ELPA_1 to a combination system with five scores for ELPA_2 undoubtedly would have impacted which students qualify for reclassification in each model.

**Accommodations**

For each assessment and school year, the dataset included information about whether students received any accommodations on their assessments. Available accommodations included provision or administration of: extended time, a separate testing location, a bilingual dictionary, translated test forms, oral test translation, a third reading or listening section, and the option to provide written responses in a student’s
native language. For this study, specific accommodations were grouped into the two more general categories of “direct linguistic accommodation,” which included bilingual dictionary, translated test forms, oral translation, and written responses in native language, and “indirect linguistic accommodation,” which included extra time, separate location, and additional reading/listening section.

It is important to remember that for ELs generally, both types of accommodations are expressly designed to reduce or remove construct-irrelevant language barriers during testing, so they can demonstrate their content skills and knowledge more validly. Thus, for example, while non-ELs might receive extra time due to a learning disability like dyslexia, or a separate testing location due to attention deficit hyperactivity disorder (ADHD), an EL who has no learning or cognitive disabilities might receive these accommodations simply because it will take him longer to read and respond to the test questions (particularly if he is looking words up in a bilingual dictionary). Also noteworthy is the fact that, in this state, RFEP students who are still in their two-year monitoring period are eligible to receive (or continue receiving) accommodations on academic content assessments.

Since students with IEPs were systematically removed during the dataset’s construction, it was expected that relatively few students would receive accommodations on the ELP assessment. This is because the ELP assessment targets students’ language skills only; as such, direct linguistic accommodations are not provided, and students who need indirect accommodations would presumably need and receive them due to cognitive or learning disabilities (rather than to reduce language-related performance issues, as on a
content assessment). As expected, the data revealed that no students received accommodations on the ELP assessment at any point.

On content assessments, by contrast, it was expected that many students would receive accommodations. Accommodation use on the content assessments is summarized in Table A.5. Interestingly, the table shows that accommodation use was essentially non-existent for both cohorts until the 2011-12 school year, which was generally grade 8 for the 2007 cohort, and grade 7 for the 2008 cohort. Starting in this school year, hundreds of current ELs and RFEPs within their monitoring period began receiving direct and indirect accommodations on both assessments. (For the latter cohort, accommodation use was also high in their 8th grade year, 2012-13).

As referenced previously, 2011-12 was the year before the implementation of the new CCSS-aligned assessments, which began in the 2012-13 school year. It is possible that, in anticipation of these new and more challenging content assessments, schools made greater use of accommodations than they had in the past. It is also possible that the state demanded better record-keeping when the new assessments were administered – in other words, it is possible that the null-findings for accommodation use prior to 2011-12 are due to coding and record-keeping errors, rather than non-use. Unfortunately, it is not possible to determine the source of these patterns from these data; it would be particularly useful to see whether accommodation use also increased for non-ELs starting in 2011-12. For the purposes of this study, I created dummy variables for whether students received direct or indirect accommodations on their ELA or math assessments in these two school years.
Missing data, attrition, and limitations

Information about missing data and attrition is summarized in Table A.6 and Table A.7. An important type of missingness in these data, which has a partial explanation in federal policy, is the absence of ELA scale scores for roughly 12 percent of students. Current federal guidance allows states to grant students who have been enrolled in US schools for less than 12 months a one-time exemption on the state’s ELA assessment (only). As the data in Table A.6 show, the overwhelming majority of missing ELA scores were in the first two years of data and in ELA only, suggesting that this policy is likely the explanation for this missingness (note that some students who are identified late in the year in grade 3 might still be within the window for exemption in grade 4). In later years, and on the math assessment, the rate of missingness is 0.6 percent or less, meaning that even if there is a relationship between missingness and outcomes, the number of students missing data are unlikely to affect the overall outcomes. Of course, the missingness due to ELA exemption is non-random, and is very likely related to my outcomes of interest, since these students are unlikely to be reclassified, and cannot contribute to estimations of covariance between language and content scores.

A.3. Data Analyses

The research questions for this follow-up study focus on the relationship between ELP and content performance over time for students who are still ELs. The baseline question is whether we observe differences in the ELP-content relationship across students based on how close they are to reclassification. As referenced previously, this line of inquiry is based on the research-based framework proposed by Cook et al. (2012), wherein these two types of assessment scores should become less related over time as
students approach linguistic proficiency. Because this state used two ELPA scores for reclassification decisions for most of the period observed here, a slightly different approach is necessary than the bivariate models proposed by Cook et al. (2012). Thus, I will attempt to a multivariate approach to understanding how language and content relate as students approach proficiency. The specific hypothesis I investigate is whether the covariance among ELP and content scores decreases over time as students get closer to reclassification. If the ELP cut score is well set, students should have covariances in their final year of EL status that are negligible, or at least that are lower than previous years when they were not ready to be reclassified.

To compare the covariance among scores across time, I will build a multilevel longitudinal model and then explore different structures for the model’s error covariance matrix, to see (a) what type of error covariance matrix best describes the data, and (b) what this matrix ultimately tells us about the relationships between language and content as students approach reclassification. This analysis will take advantage of the \( YrToRcls \) variable, which groups students according to how close they are to reclassification. Because of the way reclassification decisions are made in this state, grouping students this way will also differentiate students based on their ELP – only students in their final year as ELs will be in the highest level of ELP, by definition. The question then becomes, for students in any grade-level who are close to reclassification, do we see weaker relationships between language and content scores than we observe between their scores at times that are more distal to their reclassification?

As a reminder, all scores contributing to the calculations for these tests will have been recentered around their cut scores, as described above. Although this transformation
will not affect the correlations or covariances themselves, any final interpretations about score relationships will be made relative to this version of the scale, rather than their original numeric values. Thus, for example, the actual “score” underlying the center of the standardized distribution in each group will be the average distance from the proficiency cut score. For final interpretations, I may further standardize these scores using the standard error of measurement (SEM) at the cut score, or the standard deviation of the sample.
Table A.1. Distributional characteristics of proficiency-centered ELA and math scores by year, 2007 cohort

<table>
<thead>
<tr>
<th>2007 Cohort (nGr3 = 13,913)</th>
<th>Mathematics</th>
<th>ELA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Grade 3 (SY06-07)</td>
<td>25.2</td>
<td>31.0</td>
</tr>
<tr>
<td>Grade 4 (SY07-08)</td>
<td>23.7</td>
<td>30.4</td>
</tr>
<tr>
<td>Grade 5 (SY08-09)</td>
<td>30.2</td>
<td>28.3</td>
</tr>
<tr>
<td>Grade 6 (SY09-10)*</td>
<td>1.3</td>
<td>27.1</td>
</tr>
<tr>
<td>Grade 7 (SY10-11)</td>
<td>3.0</td>
<td>24.6</td>
</tr>
<tr>
<td>Grade 8 (SY11-12)</td>
<td>2.3</td>
<td>25.7</td>
</tr>
</tbody>
</table>

NOTE. SY06-07 = the 2006-07 school year. The center of the scale for each score is the proficiency cut score for that assessment and year. Thus, positive values indicate performance beyond the cut score, and negative values indicate performance that fell short. *In the 2009-10 school year, the state reset its cut scores. The new cut scores were on a similar scale to the preceding years, but varied by grade-level and year.
Table A.2 Distributional characteristics of proficiency-centered ELA and math scores by year, 2008 cohort

<table>
<thead>
<tr>
<th>2008 Cohort (n&lt;sub&gt;G/3&lt;/sub&gt; = 13,598)</th>
<th>Mathematics</th>
<th>ELA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Grade 3 (SY07-08)</td>
<td>28.8</td>
<td>27.1</td>
</tr>
<tr>
<td>Grade 4 (SY08-09)</td>
<td>31.9</td>
<td>32.2</td>
</tr>
<tr>
<td>Grade 5 (SY09-10)*</td>
<td>5.0</td>
<td>28.0</td>
</tr>
<tr>
<td>Grade 6 (SY10-11)</td>
<td>2.6</td>
<td>26.2</td>
</tr>
<tr>
<td>Grade 7 (SY11-12)</td>
<td>4.3</td>
<td>27.7</td>
</tr>
<tr>
<td>Grade 8 (SY12-13)**</td>
<td>-26.1</td>
<td>32.1</td>
</tr>
</tbody>
</table>

NOTE. SY07-08 = the 2007-08 school year. The center of the scale for each score is the proficiency cut score for that assessment and year. Thus, positive values indicate performance beyond the cut score, and negative values indicate performance that fell short.

*In the 2009-10 school year, the state reset its cut scores. The new cut scores were on a similar scale to the preceding years, but varied by grade-level and year.

**In the 2012-13 school year, the state adopted new content assessments aligned to the Common Core State Standards (CCSS). These assessments were reported on an entirely different score scale and had new and reflected achievement relative to a very different content domain definition.
Table A.3. Distributional characteristics on proficiency-centered ELP assessments by year, 2007 cohort

<table>
<thead>
<tr>
<th>2007 Cohort (n_{Gr3} = 13,913)</th>
<th>Oral Language</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Written Language (or Overall)</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>Range</td>
<td>Kurtosis</td>
<td>Skew</td>
<td>M</td>
<td>SD</td>
<td>Range</td>
<td>Kurtosis</td>
</tr>
<tr>
<td>Grade 3 (SY06-07)</td>
<td>-2.6</td>
<td>49.1</td>
<td>-305 – 118</td>
<td>1.7</td>
<td>-0.5</td>
<td>-21.6</td>
<td>46.6</td>
<td>-250 – 180</td>
<td>1.2</td>
</tr>
<tr>
<td>Grade 4 (SY07-08)</td>
<td>15.3</td>
<td>43.5</td>
<td>-184 – 123</td>
<td>0.6</td>
<td>0.2</td>
<td>-12.4</td>
<td>43.4</td>
<td>-194 – 125</td>
<td>1.2</td>
</tr>
<tr>
<td>Grade 5 (SY08-09)</td>
<td>11.7</td>
<td>33.4</td>
<td>-142 – 128</td>
<td>1.7</td>
<td>0.5</td>
<td>-8.1</td>
<td>34.5</td>
<td>-161 – 139</td>
<td>1.2</td>
</tr>
<tr>
<td>Grade 6 (SY09-10)</td>
<td>7.8</td>
<td>33.4</td>
<td>-144 – 131</td>
<td>1.8</td>
<td>0.4</td>
<td>-10.6</td>
<td>32.0</td>
<td>-126 – 139</td>
<td>1.4</td>
</tr>
<tr>
<td>Grade 7 (SY10-11)</td>
<td>19.3</td>
<td>33.1</td>
<td>-110 – 142</td>
<td>1.9</td>
<td>0.4</td>
<td>-17.1</td>
<td>32.8</td>
<td>-159 – 133</td>
<td>1.5</td>
</tr>
<tr>
<td>Grade 8 (SY11-12)</td>
<td>25.8</td>
<td>36.4</td>
<td>-110 – 129</td>
<td>1.2</td>
<td>0.7</td>
<td>-17.7</td>
<td>29.7</td>
<td>-140 – 124</td>
<td>1.4</td>
</tr>
</tbody>
</table>

NOTE. SY06-07 = the 2006-07 school year. The center of the scale for each score is the proficiency cut score for that assessment and year. Thus, positive values indicate performance beyond the cut score, and negative values indicate performance that fell short.

*In the 2012-13 school year, the state adopted a new ELP assessments that produced only one overall scaled score. The range of this score scale was considerably smaller than the previously used ELPA (i.e., roughly 225 points, compared to roughly 400 points previously).
Table A.4 Distributional characteristics on proficiency-centered ELP assessments by year, 2008 cohort

<table>
<thead>
<tr>
<th>2008 Cohort (n_{Gr3} = 13,598)</th>
<th>Oral Language</th>
<th>Written Language (or Overall)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Grade 3 (SY07-08)</td>
<td>8.1</td>
<td>48.7</td>
</tr>
<tr>
<td>Grade 4 (SY08-09)</td>
<td>24.6</td>
<td>44.9</td>
</tr>
<tr>
<td>Grade 5 (SY09-10)</td>
<td>13.6</td>
<td>33.6</td>
</tr>
<tr>
<td>Grade 6 (SY10-11)</td>
<td>8.3</td>
<td>32.8</td>
</tr>
<tr>
<td>Grade 7 (SY11-12)</td>
<td>20.6</td>
<td>33.1</td>
</tr>
<tr>
<td>Grade 8 (SY12-13)*</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>

NOTE. SY07-08 = the 2007-08 school year. The center of the scale for each score is the proficiency cut score for that assessment and year. Thus, positive values indicate performance beyond the cut score, and negative values indicate performance that fell short.

*In the 2012-13 school year, the state adopted a new ELP assessments that produced only one overall scaled score. The range of this score scale was considerably smaller than the previously used ELPA (i.e., roughly 225 points, compared to roughly 400 points previously).
Table A.5. Accommodation use by type, content area, school year

<table>
<thead>
<tr>
<th>Content Area</th>
<th>English Language Arts (ELA)</th>
<th>Mathematics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accommodation Type</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Direct</td>
<td>Indirect</td>
</tr>
<tr>
<td>Number of Students</td>
<td>Total  EL  RFEP</td>
<td>Total  EL  RFEP</td>
</tr>
<tr>
<td>2011-12 School Year</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2008 Cohort (Grade 7)*</td>
<td>1439   656   783</td>
<td>2541  1215   1326</td>
</tr>
<tr>
<td>2007 Cohort (Grade 8)</td>
<td>715    367   348</td>
<td>1198  677   521</td>
</tr>
<tr>
<td>2012-13 School Year</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2008 Cohort (Grade 8)</td>
<td>141   95     46</td>
<td>1195  672   523</td>
</tr>
</tbody>
</table>

NOTE. RFEP = reclassified fluent English proficient. According to test administrator manuals, RFEP students who have been reclassified within the past two school years may receive language-related accommodations on state content assessments.

* For all rows, reported numbers reflect all students in the cohort, including those who are off-level due to previous retention.
Table A.6. Summary of Missing Data, by Cohort

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Total Sample</th>
<th>2007</th>
<th>2008</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Sample</td>
<td>27,511</td>
<td>13,913</td>
<td>13,598</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data Element</th>
<th>n</th>
<th>%</th>
<th>n</th>
<th>%</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>LEP program information</td>
<td>1154</td>
<td>4.2</td>
<td>726</td>
<td>5.2</td>
<td>428</td>
<td>3.1</td>
</tr>
<tr>
<td>Pre-Grade 3 Time as EL</td>
<td>35</td>
<td>0.001</td>
<td>13</td>
<td>&lt;0.001</td>
<td>6</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

ELA scale score

- Overall: 4381, 2.9; 2420, 3.1; 1961, 2.6
- Data Year = 0 (Grade 3): 3365, 12.2; 1818, 13.1; 1547, 11.4
- Data Year = 1 (Grade 4): 686, 2.6; 398, 3.0; 288, 2.2
- Data Year = 2 (Grade 5): 107, 0.4; 82, 0.6; 25, 0.2
- Data Year = 3 (Grade 6): 72, 0.3; 39, 0.3; 33, 0.3
- Data Year = 4 (Grade 7): 72, 0.3; 40, 0.3; 32, 0.3
- Data Year = 5 (Grade 8): 79, 0.3; 43, 0.4; 36, 0.3

Math scale score

- Overall: 94, 0.1; 39, 0.1; 55, 0.1
- Data Year = 0 (Grade 3): 34, 0.1; 10, 0.1; 24, 0.2
- Data Year = 1 (Grade 4): 9, <0.1; 6, <0.1; 3, <0.1
- Data Year = 2 (Grade 5): 10, <0.1; 5, <0.1; 5, <0.1
- Data Year = 3 (Grade 6): 10, <0.1; 4, <0.1; 6, <0.1
- Data Year = 4 (Grade 7): 16, 0.1; 6, <0.1; 10, <0.1
- Data Year = 5 (Grade 8): 15, 0.1; 8, 0.1; 7, 0.1
Table A.7. Summary of Attrition, by Cohort

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Total Sample</th>
<th>2007</th>
<th>2008</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of Students</td>
<td>27,511</td>
<td>13,913</td>
</tr>
<tr>
<td>Number of data years</td>
<td>n</td>
<td>%</td>
<td>n</td>
</tr>
<tr>
<td>6 (complete)</td>
<td>22,154</td>
<td>80.5</td>
<td>11,119</td>
</tr>
<tr>
<td>5</td>
<td>1,341</td>
<td>4.9</td>
<td>710</td>
</tr>
<tr>
<td>4</td>
<td>916</td>
<td>3.3</td>
<td>462</td>
</tr>
<tr>
<td>3</td>
<td>951</td>
<td>3.5</td>
<td>496</td>
</tr>
<tr>
<td>2</td>
<td>925</td>
<td>3.4</td>
<td>479</td>
</tr>
<tr>
<td>1</td>
<td>1,219</td>
<td>4.4</td>
<td>647</td>
</tr>
</tbody>
</table>
Figure A.1. Average proficiency-centered ELA scores by school year, cohort

Figure A.2. Average proficiency-centered math scores by school year, cohort
Figure A.3. Average proficiency-centered ELP scale scores by school year, cohort
REFERENCES


