A JOINT MODELING APPROACH TO STUDYING ENGLISH LANGUAGE PROFICIENCY DEVELOPMENT AND TIME-TO-RECLASSIFICATION

by

TYLER H. MATTA

A DISSERTATION

Presented to the Department of Educational Methodology, Policy and Leadership
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

December 2016

DISSERTATION APPROVAL PAGE

Student: Tyler H. Matta

Title: A Joint Modeling Approach to Studying English Language Proficiency Development and Time-to-Reclassification

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Educational Methodology, Policy and Leadership by:

Joseph Stevens	Chair
Ilana Umansky	Core Member
Yeow Meng Thum	Core Member
Sanjay Srivastava	Institutional Representative

and

Scott L. Pratt	Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded December 2016

DISSERTATION ABSTRACT

Tyler H. Matta

Doctor of Philosophy

Department of Educational Methodology, Policy and Leadership

December 2016

Title: A Joint Modeling Approach to Studying English Language Proficiency Development and Time-to-Reclassification

The development of academic English proficiency and the time it takes to reclassify to fluent English proficient status are key issues in monitoring achievement of English learners. Yet, little is known about academic English language development at the domain-level (listening, speaking, reading, and writing), or how English language development is associated with time-to-reclassification as an English proficient student. Although the substantive findings surrounding English proficiency and reclassification are of great import, the main focus of this dissertation was methodological: the exploration and testing of joint modeling methods for studying both issues. The first joint model studied was a multilevel, multivariate random effects model that estimated the student-specific and school-specific association between different domains of English language proficiency. The second model was a multilevel shared random effects model that estimated English proficiency development and time-to-reclassification simultaneously and treated the student-specific random effects as latent covariates in the time-to-reclassification model. These joint modeling approaches were illustrated using annual English language proficiency test scores and time-to-reclassification data from a large Arizona school district.

Results from the multivariate random effects model revealed correlations greater than .5 among the reading, writing and oral English proficiency random intercepts. The analysis of English proficiency development illustrated that some students had attained proficiency in particular domains at different times, and that some students had not attained proficiency in a particular domain even when their total English proficiency score met the state benchmark for proficiency. These more specific domain score analyses highlight important differences in language development that may have implications for instruction and policy. The shared random effects model resulted in predictions of time-to-reclassification that were 97% accurate compared to 80% accuracy from a conventional discrete-time hazard model. The time-to-reclassification analysis suggested that use of information about English language development is critical for making accurate predictions of the time a student will reclassify in this Arizona school district.

CURRICULUM VITAE

NAME OF AUTHOR: Tyler H. Matta

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, OR
Pratt Institute, Brooklyn, NY

DEGREES AWARDED:

Doctor of Philosophy, Educational Leadership, 2016, University of Oregon
Master of Science, Art and Design Education, 2007, Pratt Institute
Bachelor of Fine Arts, Fine Arts, 2007, Pratt Institute

AREAS OF SPECIAL INTEREST:

Latent Variable Modeling
Longitudinal Data Analysis
Missing Data

GRANTS, AWARDS AND HONORS:

Graduate Fellow Teaching Award, College of Education, University of Oregon, 2015
Travel Award, Society for Multivariate Experimental Psychology (SMEP), 2014
Scholarship for Education Research, Inter-university Consortium for Political and Social Research (ICPSR), 2014

ACKNOWLEDGEMENTS

My advisor, Joe Stevens, for his unconditional support throughout my doctoral studies. My dear friend, collaborator, and mentor, Yeow Meng Thum. John Wilson, for providing the data used for this study. All those who made this journey memorable, particularly, Meg Guerreiro, Michael Their, Jim Soland, Quinn Lathrop, and Dan McNeish. My family: Kaity, Deb, Pete, Jan, Paula, Debbie, and Peter. Finally, my best friend and wife, Bianca, to whom this work is dedicated.

For Bianca

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Table Page

CHAPTER I

INTRODUCTION

English learners (EL) are those students who come from homes where the active language is one other than English, and whose proficiency in English is not yet developed to the level where they can profit fully from instruction delivered solely in English (August & Hakuta, 1997). According to the National Center for Education Statistics (2015), the proportion of English learners enrolled in United States' (US) schools was higher in 2012-13 (9.2% or 4.4 million) than in 2002-03 (8.7% or 4.1 million), and with approximately 4.5 million English learners enrolled in PK-12 public schools nationwide in 2014-15, they are the fastest growing subgroup of students (Office of English Language Acquisition, 2015). This growing population of students faces the doubly difficult task of obtaining English proficiency while also meeting the same content standards as required for their fluent English speaking peers. Research indicates, however, that there are persistent gaps between English learners and their fluent English speaking peers on a number of key academic outcomes.

In 2015, there were achievement gaps on the National Assessment of Educational Progress (NAEP) between non-English learners and English learners in both mathematics and reading in Grades 4 and 8 of approximately one standard deviation (U.S. Department of Education, Institute of Education Sciences, & National Center for Education Statistics, 2015). This finding was not new, however. Research has shown persistent gaps between English learners and their fluent English peers generally (Callahan, Wilkinson, & Muller, 2010; Kim & Herman, 2009), and specifically for reading (Kieffer, 2008, 2011), and mathematics achievement (Roberts & Bryant, 2011). Additionally, research has shown significant

differences between English learners and their fluent English peers in their opportunity to learn (Abedi & Herman, 2010; Callahan, 2005). Furthermore, English learners have been found to be less likely to graduate from high school or persist into post-secondary education (Heilig, 2011; Kanno & Cromley, 2013; Kao & Thompson, 2003).

While there is general agreement that closing these gaps between English learners and their fluent English speaking peers is of critical importance, there is ongoing debate regarding the role that language and language policy plays in this effort (Hakuta, 2011). Current federal policy provides no guidance and states are left with the autonomy to serve English learners in any way they find to be effective as long as those programs are (a) based on sound educational theory, (b) implemented adequately, and (c) are periodically evaluated (Hakuta, 2011). Although states such as Arizona, Massachusetts, and California have passed laws that require English-only instructional programs, evidence suggests these policies have had no impact on academic outcomes for English learners (Koretz & Guo, 2012; Parrish, Perez, Merickel, & Linquanti, 2006). Furthermore, there is a sizable research literature that compares types of English-only instruction to types of dual language instruction (e.g., Cheung & Slavin, 2012; Valentino & Reardon, 2015). While language of instruction remains a key issue in English learner research, there remains little understanding of how English develops over time within these programs.

Academic English is characterized as the English language practices required of the four language *domains* (speaking, listening, reading, and writing) to successfully engage in academic content in school and display knowledge (Council of Chief State School Officers, 2012). When English learners reach academic

English proficiency, they are eligible for reclassification from EL status to what is generally referred to as *fluent English proficient*. Intertwined with the language of instruction debate is the issue of how long it should take for English learners to attain proficiency in academic English and reclassify to fluent English proficient. For example, in addition to four-hour intensive Structured English Immersion (SEI) instruction, Arizona's Title 15 (§15.752) requires that students be reclassified in one year. Early research has shown, however, that students take between three to five years to attain oral English proficiency and four to seven years to attain academic English fluency (Hakuta, Butler, & Witt, 2000).

While the act of reclassification is largely an administrative change, it results in students losing access to language support services and entering the mainstream classroom as well as altering accountability reporting of the EL subcategory of students. Thus, it is critical to ensure reclassification policies are aligned with the most current understanding of how academic English develops for ELs and accurately reflects the attainment of fluent English proficient status. There is scant longitudinal research, however, focused on describing the development of school-aged (K-12) English learners' academic English across the four language domains as measured by annual state English language proficiency assessments.

Understanding the factors that impact English learners' progress to reclassification is complex (Bailey & Carroll, 2015). There has been a small but growing literature focused on understanding this process, modeling the time it takes English learners to reclassify using discrete-time survival models (Slama, 2014; Thompson, 2015; Umansky & Reardon, 2014). However, this body of research does not account for the association between English language development and time to reclassification. Reclassification decisions are based largely on one's English

proficiency, which is a developmental outcome (Ramsey & O'Day, 2010). Thus, any factor that influences time to reclassification is likely to do so, in full or in part, through the development of English proficiency. The use of longitudinal English language proficiency measures as time-dependent covariates however, would likely be subject to issues of endogeneity.

In research using cross-sectional data, *covariate endogeneity* for a fitted model exists when the covariance between the covariate and the residual does not equal zero (Angrist & Pischke, 2008). In the longitudinal context, covariate endogeneity becomes more complex and has been well studied in both the analysis of repeated measures data (Diggle, Zeger, Liang, & Heagerty, 2002) as well as the analysis of survival data (Kalbfleisch & Prentice, 2011). For survival analysis, a time-dependent variable is exogenous (referred to as external in the survival literature) if its process influences the rate of event occurrence over time, but its future path is not affected by the occurrence of the event (Kalbfleisch & Prentice, 2011). Such variables include defined covariates where the values of the covariate are established in the study design, and ancillary covariates where the stochastic process is outside of the individual under study. A time-dependent variable is endogenous (referred to as internal in the survival literature) if its future path is affected by the event occurrence (Kalbfleisch & Prentice, 2011). These variables are typically the output of a stochastic process associated with the participant, and therefore require their own statistical model (Kalbfleisch & Prentice, 2011). The joint modeling paradigm provides a useful approach for incorporating endogenous time-varying covariates into longitudinal analysis (Kalbfleisch & Prentice, 2011; Rizopoulos & Lesaffre, 2014).

This study proposed two statistical models for the analysis of both academic English language proficiency development and the time-to-reclassification. The intention of the first model was to describe how developments in each language domain correlated with the others. The second model was designed to describe the system of reclassification more accurately than conventional approaches. In what follows, I provide an overview of the recent research on English language proficiency development and reclassification followed by a review of the joint modeling literature. From there, I provide a theoretical rational that connects this joint modeling paradigm to questions pertaining to English learner reclassification. I then propose a series of research questions that motivated the use of these models.

English Proficiency Development

Hakuta et al. (2000) have been influential in the conversation regarding the time it takes for English learners to become proficient in English. Their findings, however, were based on cross-sectional data and did not describe the within-student developmental process for attaining academic English proficiency. The research within August and Shanahan's (2006) exceptionally thorough literature review of second-language literacy development was almost exclusively cross-sectional. In the context of setting progress and attainment benchmarks for English learners, Cook, Boals, Wilmes, and Santos (2008) described the academic English language development of 12,836 English learners using repeated measures over a three year period. This descriptive analysis also treated the English proficiency outcomes in a cross-sectional manner by reporting grade-level mean scores and did not attempt to model within-student developmental change or the variation in development between students. As has been well documented, the analysis of outcomes in a cross-sectional manner when interest is in describing intra-individual development

can confound cohort and age effects and/or conflate inter-individual and intra-individual variation (Diggle et al., 2002; Raudenbush, 2001; Thum, 1994). That is, cross-sectional analysis is valid only when all individuals grow in the same way, or when individual differences, even when present, are considered to be nuisance variation.

Longitudinal English proficiency outcomes allows for the fitting of statistical models that provide a smooth empirical approximation of the student-specific developmental process. There is an abundance of longitudinal research focused on the English development of early childhood populations (Conboy & Thal, 2006; Hammer, Lawrence, & Miccio, 2008; Kohnert & Conboy, 2010; Vagh, Pan, & Mancilla-martinez, 2009); whereas, fewer studies exist for school-aged children (Rojas & Iglesias, 2013; Slama, 2012; Uchikoshi, 2012). Of the longitudinal studies of English development for school-aged children, the average developmental process has been described using simple linear functions (Uchikoshi, 2012), piecewise functions (Rojas & Iglesias, 2013), and polynomial functions (Slama, 2012).

Much of the longitudinal research on English language development focuses on oral proficiency (Rojas & Iglesias, 2013; Uchikoshi, 2012). Uchikoshi (2012) studied the oral vocabulary development of Spanish and Cantonese English learners between Kindergarten and Grade 2. Limited by only three time points per student, oral vocabulary development was characterized as a simple linear trajectory with variation in initial status and rate of development (Uchikoshi, 2012). Rojas and Iglesias (2013) also analyzed the oral English vocabulary development of a 1,732 English learners during their first three years of schooling. With two measures per year, Rojas and Iglesias fit a piecewise model that described within-grade development as a positive linear process with stagnation over the summer. While

these studies contribute to the research on English vocabulary development of English learners, neither study considers the development of the literate domains (writing and reading).

To my knowledge, there has been only one published study that examined the English language proficiency growth trajectories of English learners using annual state English language proficiency assessments. Slama (2012) focused on the difference between US-born and foreign-born adolescent ELs in their development of academic English proficiency. Using the composite English proficiency score (i.e., weighted combination of reading, writing, listing and speaking sub-tests), this study described the average within-student trend as a curvilinear function where the rate of growth was relatively large early in the study and decelerated in late grades (Slama, 2012). However the study also found significant variation in initial status, linear growth, and the curvature, resulting in some students' English development accelerating over time (Slama, 2012). Slama did not describe the functional form of domain-specific academic English proficiency development.

Overall, there is scant longitudinal research focused on describing the development of school-aged English learners' academic English across the four language domains as measured by annual state English language proficiency assessments. Understanding the average developmental process for a group of students, and the individual differences in those developmental trajectories, can prove useful for English learner policy and practice. From a policy perspective, well estimated developmental trajectories can aid in the setting of realistic standards for annual English language development and understand the differential impact of English language programs. Additionally, such information can enable schools

to improve practices by accurately identifying individual students whose English language development is not keeping pace and may be in need of intervention.

Reclassification

Within the broader English learner literature is a body of research focused specifically on the time it takes for students to be reclassified as English proficient. Students are tracked from some original classification, often at the start of Kindergarten, until they are reclassified as fluent English proficient or are censored (i.e., they do not reclassify within the study window or are lost to follow-up). The time spent classified as an English learner —referred to as duration in the survival analysis literature —can then be analyzed using discrete time survival models. Covariates can also be added to the model in an attempt to understand the factors that predict reclassification and to understand how specific decision criteria may impact time to reclassification. The earliest example of this approach was used to investigate the impact of California's Proposition 227, also called the English Language in Public Schools Statute, on time to reclassification (Parrish et al., 2006).

More recently, a new wave of reclassification studies has emerged. Data used in these studies have been collected at the state or district level and have focused on different populations. Slama (2014) analyzed a single cohort of 5,345 English learners across the state of Massachusetts. Umansky and Reardon (2014) focused on Latino English learners from a large urban school district, using nine cohorts over 12 years (fall 2000 through spring 2012) for a total sample of 5,423 students. Finally, Thompson (2015) analyzed nine years of data that spanned 2001-02 through 2009-10 using eight cohorts for a total of 202,931 English learners from the Los Angeles Unified School District.

This body of research indicates that it can take anywhere from three to eight years for the average English learner to reclassify. Slama (2014) found the average English learner reclassified after three years (or in second grade) but 17% of the cohort had not been reclassified by the end of Grade 7. Umansky and Reardon (2014) found that 38% of students reclassified by the end of Grade 5, 75% reclassified by the end of Grade 11, and the median time to reclassification was Grade 7. Similarly, Thompson (2015) found the likelihood of reclassification increased during elementary school peaking after six years (5th grade), and then began to drop again. After nine years, students had a 74% chance of being reclassified (Thompson, 2015).

Student and program related factors were also found to influence the likelihood of reclassification. A non-Spanish speaking English learner was found more likely to reclassify than Spanish speaking English learners (Slama, 2014; Thompson, 2015). Slama (2014) also found free/reduced price lunch (FRL) status to be predictive of time to reclassification. In addition to FRL status, Thompson (2015) found that females were more likely to reclassify than males, and those with parents who had higher levels of education, or those who were never identified as special education students were also more likely to reclassify. Umansky and Reardon (2014) found that students in English immersion programs had higher rates of reclassification in early years but the students in two-language programs caught up over time. Thompson also showed that students who ever received bilingual education had a significant interaction with time, as did initial first language proficiency and initial second language proficiency.

Within this literature, secondary analyses have investigated which of the reclassification criteria proved to be barriers to reclassification. Rather than using

the actual event of reclassification, individual exit criteria are used to estimate the time a student would have reclassified if that sole criterion was used for reclassification. Of the seven criteria used to determine reclassification, Umansky and Reardon (2014) indicated that in early grades (K-5), English proficiency was a larger barrier than the English Language Arts (ELA) criterion, but this shifted after Grade 5. Thompson (2015) found that the barrier to reclassification was not any one specific criterion, but rather, a matter of meeting all criteria in the same year.

Joint Models for Longitudinal Data

The longitudinal analysis of multiple outcome measures has been, and remains, an active topic of research (Galecki, 1994; Reinsel, 1982; Rizopoulos & Lesaffre, 2014; Wu & Carroll, 1988). Early examples in the social sciences can be traced to the multivariate random effects models (e.g., Bock & Bargmann, 1966; Joreskog, 1970; Joreskog & Goldberger, 1975; Muthen, 1989). Multivariate hierarchal linear models (Raudenbush & Bryk, 2002; Thum, 1994, 1997) can be seen as extensions of a subset of models focused on multiple outcomes. Similarly, multiple outcomes analyzed with latent curve models (Blozis, 2004, 2007; Bollen & Curran, 2006) can perform longitudinal analyses where the functional form of time is unknown and is estimated from the data. A modern synthesis of the above developments can be found in generalized latent variable modeling (Muthén, 2002; Skrondal & Rabe-Hesketh, 2004).

There are two common frameworks under the joint modeling umbrella: multivariate random effects models (MVREM) and shared random effects models (SREM). MVREMs provide a robust framework for describing the dependence between multiple longitudinal outcomes (see Verbeke & Davidian, 2008; Verbeke,

Fieuws, Molenberghs, & Davidian, 2014). When interest is in the association between repeated measures outcomes and event times, SREMs has proven to be a useful modeling paradigm (see Proust-Lima, Séne, Taylor, & Jacqmin-Gadda, 2014; Tsiatis & Davidian, 2004). Both frameworks consist of separate submodels for each outcome and are linked together through a latent variable structure.

Multivariate random effects models (MVREM) account for the association between two or more repeated measures outcomes through subject-specific latent variables (e.g., random effects; MacCallum, Kim, Malarkey, & Kiecolt-Glaser, 1997; Verbeke et al., 2014). That is, for each outcome, $Y_1, \ldots, Y_m$, a mixed model is specified. Consider a random intercepts model for two outcomes, Y_1 and Y_2, for individual i at time j,

$$\begin{aligned} Y_{1ij} =& \beta_1 + \zeta_{1i} + \beta_2 t_{ij} + \epsilon_{1ij} \\ Y_{2ij} =& \beta_3 + \zeta_{2i} + \beta_4 t_{ij} + \epsilon_{2ij} \end{aligned}$$

But rather than assuming $\boldsymbol{\zeta}_i \equiv [\zeta_{1i}, \zeta_{2i}]$ follow independent zero-mean normal distributions —as would be assumed if the outcomes were analyzed separately —their dependence is estimated. Under the MVREM framework, the dependence between Y_1 and Y_2 is accounted for by the correlation of the latent variables, $\rho(\zeta_1, \zeta_2)$. Because each outcome is assumed interdependent conditional on the random effects, $\boldsymbol{\zeta}_i$, the submodels for Y_1 and Y_2 carry the same interpretation as their univariate expression.

With MVREM, one is required to specify the form of the multivariate random effects. While it is conventional to assume a multivariate normal distribution, Thum (1994, 1997) specified a multivariate-t distribution while Stoolmiller (1994) and Ferrer and McArdle (2003) both specified non-parametric random effects. Verbeke et al. (2014) note that the underlying idea of the MVREM

holds irrespective of the number of outcomes. However, as the number of outcomes grow, the dimensionality of the latent structure also grows, which can present computational issues (Verbeke et al., 2014). Fieuws and Verbeke (2006) show that high dimensional models with normal random effects can be estimated by averaging over all pairwise combinations of the outcomes.

The MVREM literature contains numerous examples of linear, generalized linear, non-linear, and combinations of such models for multivariate repeated outcomes. For the continuous case, Beckett, Tancredi, and Wilson (2004) estimated a multivariate growth model for four outcomes related to cognitive functioning in an aging population (ages 65 and older). They modeled growth in episodic memory, semantic memory, working memory, and perceptual speed, with a focus on the correlations between their respective rates of decline (linear slopes). Doss, Thum, Sevier, Atkins, and Christensen (2005) employed a latent variable regression of the change in a satisfaction outcome on the change in partner behavior outcome within a doubly-multivariate growth model in a comparison of two types of behavioral couple therapy. That is, two outcomes were tracked for the two members of each couple who are assigned to one of two types of behavioral couple therapy. In the context of generalized linear models, Ribaudo and Thompson (2002) extended the multivariate repeated model to deal with binary outcomes. They estimated the prevalence of reporting six quality of life outcomes between two treatment groups for subjects with previously untreated non-small-cell lung cancer. Finally, Fieuws, Verbeke, Maes, and Vanrenterghem (2008) provide an example of combined linear, generalized linear and non-linear mixed models to predict renal graft failure. For a review this literature, see Verbeke et al. (2014).

Like the multivariate random effects models, SREMs require the specification of submodels for each outcome, for example, a SREM for two longitudinal processes, Y_1 and and Y_2 may be specified as:

$$Y_{1ij} = \beta_1 + \zeta_{1i} + \beta_2 t_{ij} + \epsilon_{1ij}$$

$$Y_{2ij} = \beta_3 + \zeta_{1i}\lambda_1 + \beta_5 t_{ij} + \epsilon_{2ij}.$$

Unlike MVREMs, however, the dependence is specified through 'shared' random effects that link the two models, ζ_{1i}, where λ_1 is a scalar. While SREMs may be fit for any combination of outcomes, they are used primarily for the joint modeling of longitudinal data and event outcomes where the survival model is fit without a subject-specific random effect, or frailty as they're referred to in the survival literature[1]. Such a specification enables the researcher to test different hypotheses regarding the relationship between the longitudinal process and the time-to-event process. However, one drawback is that the shared random effects influence the correlation between repeated measures and the dependency between the repeated measure and the time to event (Verbeke et al., 2014). Tsiatis and Davidian (2004) provide a philosophical rational for the use of SREMs for repeated measures and event outcomes.

The SREM paradigm was developed in the statistics and biostatistics literatures. The first SREM was published by Wu and Carroll (1988) to deal with what Little (1995) termed latent variable dependent missingness. The use of SREMs to understand the association between a set of repeated measures and the time to some event was popularized by early HIV clinical trials (De Gruttola & Tu, 1994; Tsiatis, Degruttola, & Wulfsohn, 1995; Wulfsohn & Tsiatis, 1997). These

[1]The term frailty was introduced by Vaupel, Manton, and Stallard (1979) and is an unobserved random proportionality factor that modifies the hazard function of an individual.

studies aimed to understand how C4D T-lymphocytes were associated with onset of AIDS or death for those subjects with HIV. Henderson, Diggle, and Dobson (2000) utilized a shared random effects model to reanalyze the effect of drug therapy for schizophrenia patients while simultaneously accounting for attrition.

The SREM framework has been less commonly used in educational research. Muthen and Masyn (2005) fit a latent class growth model to students' aggressive behavior in Grades 1 and 2 that was used to predict the time to removal in Grades 3 through 7 using a discrete time survival process with a latent class frailty. Feldman and Rabe-Hesketh (2012) employed an SREM to understand if achievement trajectories were impacted by possibly non-random dropout in a large national dataset. Their discrete-time hazard submodel included separate parameters for the random intercept and random slope. Estimates from the SREM and a competing model fit to the data assuming the missing data mechanism was ignorable were consistent, suggesting the data were not sensitive to missing data assumptions (Feldman & Rabe-Hesketh, 2012). Finally, Thum and Matta (2015a, 2015b) employed a shared random effect model for longitudinal interim assessments between Grades 4 and 9, SAT and ACT scores between Grades 10 and 12, and a logistic regression model for the probability of taking a college test. The parameter estimates were then used to establish college readiness benchmarks for the interim assessment. While the SREM has had limited application in educational settings, it provides a framework for understanding relationships between multiple processes that cannot be revealed when separate models are used for each process.

Theoretical Framework

The development of English proficiency and reclassification from English learner to fluent English proficient are two related but separate phenomena.

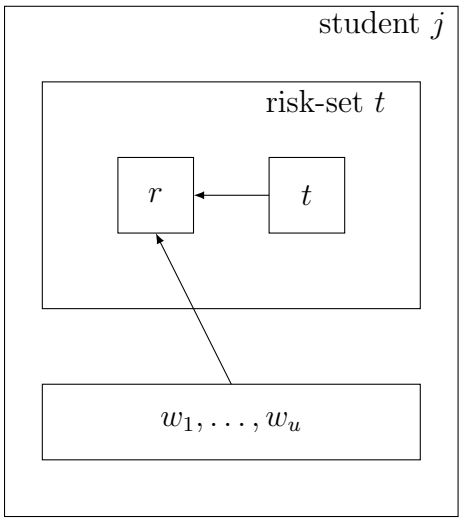
Academic English proficiency is a construct characterized by the English language practices required to successfully engage in academic content in school and display knowledge (Council of Chief State School Officers, 2012), what Schleppegrell (2001) described as "doing school" (p. 432). Teachers of English to Speakers of Other Languages (TESOL; 2006) operationalize academic language as "the language used to acquire a new or deeper understanding of content related to the core curriculum areas and communicate that understanding to others" (p. 18). While academic English is an evolving construct, and one not without criticism (Scarcella, 2003), the practical realization is that one cannot do school with ordinary language alone (Hakuta, 2011). Proficiency is assessed annually across the four language domains: speaking, listening, reading, and writing (Bailey & Huang, 2011). For federal reporting and decision purposes, the four domain measures are combined or assessed in total to create a single measure of academic English proficiency (Abedi, 2008; Cook et al., 2008).

Reclassification, on the other hand, is an administrative change in status from English learner to fluent English proficient. Federal regulations require English language proficiency assessments to be part of the criteria used in reclassification decisions (National Research Council, 2011). States are granted the autonomy, however, to include other criteria including (but not limited to) academic content standards, teacher recommendations, or information from parent conferences (Ramsey & O'Day, 2010). Furthermore, states are free to define whether the reclassification decision rules are conjunctive, compensatory, complementary, or mixed (Bailey & Carroll, 2015). For example, Arizona makes reclassification decisions based on the total English language proficiency score only, whereas Californias decision rules use a students combined English language

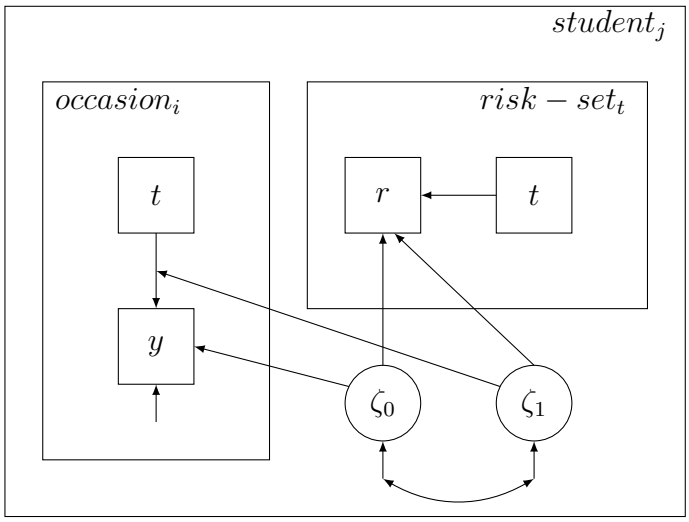
proficiency score, the four domain-specific scores, and their annual ELA score. This suggests that within some states, reclassification is a highly mechanistic process, based largely on a students initial status and rate of development for each decision criterion. In other states, however, the reclassification process may be less direct and also depend on clinical judgments and evaluations. In what follows, I propose three theoretical models that illustrate the possible relationships between the development of English proficiency and reclassification for a state the uses a total English language proficiency score as the only basis for reclassification (e.g., Arizona).

Prior research on reclassification has used student-specific covariates as predictors with no explicit connection to English language proficiency, as illustrated in Figure 1a. Variables inside the box labeled "student j" have an j subscript and vary between students while variables inside the box labeled "risk-set t" vary by both students and risk-sets and have both j and t subscripts. Here, r is a dummy variable taking the value 1 if student j is reclassified in risk-set t and 0 otherwise. Student-level covariates, generically referred to as w, are time invariant. The arrow from time t to r represents the baseline survival model, or the regression of whether or not reclassification occurred for student j at time t. The arrows from the student-level covariates to r represent the regression of those covariates on reclassification under the proportional hazards assumption.

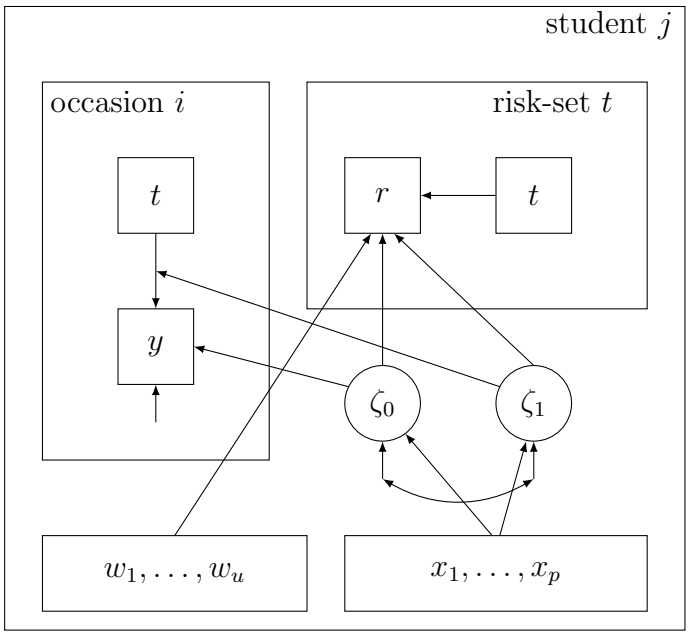
Figure 1b illustrates a theoretical model where reclassification is predicted by the attributes of a student's English proficiency development, y. The figure adds a box labeled "occasion i" to represent those variables that vary by measurement occasion and student, and have both i and j subscripts. The arrow from t to y represents a regression of the observed English proficiency score for student



(a) Survival model with covariates



(b) Shared random effects model



(c) Shared random effects model with covariates

Figure 1. Directed graphs for time-to-reclassification

j on time i. The latent variable ζ_0 represents the random intercept while ζ_1, which interacts with time t, represents the random slope (Skrondal & Rabe-Hesketh, 2004). This aspect of the directed graph represents a mixed effect growth model where the residual variance is constrained to be equal across occasions. The student-specific deviation in initial status and growth from the population mean, ζ_0 and ζ_1, are then used as latent covariates to predict reclassification at time t. Figure 1c combines Figure 1a and 1b by using student covariates (w and x) to condition English proficiency development and time to reclassification simultaneously. This theoretical model could be expanded to include a multivariate growth processes with multiple true scores if a state based reclassification on more than one criterion.

This dissertation draws on a dataset from a district in Arizona to demonstrate the utility of joint models for English learner research. In particular, the first purpose of this research was to describe the relationship between the development of overall academic English proficiency and each language domain for students in the school district. The second goal was to develop a model that accurately predicts time to reclassification for these students. Specifically, the research questions were:

1. What is the functional form of overall academic English proficiency and for each language domain between Grades 3 and 7?

2. How do developments in each language domain correlate with each other?

3. Does a shared random effects model yield more accurate predictions of time to reclassification than a conventional discrete-time survival model?

For Research Question 1, scores from the overall English language proficiency assessment and from each language domain assessment will be modeled across Grades 3 to 7 using a mixed effect modeling framework. For Research Question 2, each of the univariate models from Research Question 1 will be combined to form a multivariate random effects model. Research Question 3 will evaluate the predictive accuracy of three models, a discrete-time survival model for time to reclassification and two shared random effects models for the overall English language proficiency assessment scores and time to reclassification.

CHAPTER II

METHODOLOGY

Sample

This study used extant data from one large urban school district in Arizona. The data were from a single cohort of English learners tracked longitudinally from third grade in academic year 2007-08 through seventh grade in 2011-12[1]. During this time, Arizona implemented an English-only instruction law resulting in homogeneity of language programs across schools (Gándara & Orfield, 2010). For each student, annual English language proficiency assessment scores were collected as well as indicators of classification status. Due to the nature of the English learner classification, the data were highly unbalanced where some students may contribute data for only one time point while others contributed observations for the entire time span.

The Grade 3 English learner cohort consisted of 277 students, or 20.77% of all third grade students in the district. All 277 English learners were enrolled in self-contained ELD classes. By Grade 7, there were 20 English learners remaining. Of the 20 English learners, 19 were in self-contained English language development classes and 1 was in a classroom with Non-EL students.

There were 19 schools serving elementary Grades 3 - 5 and six schools serving middle school Grades 6 and 7. Elementary schools served between 37 and 121 students in Grade 3 and middle schools served between 47 and 297 students in Grade 7. English learners were nested within 18 of the 19 elementary schools and made up between 1.39% and 42.15% of students in their school. In seventh grade,

[1]Grade 8 data was excluded from the analysis because Arizona implemented a new English proficiency assessment in the 2012-13 academic year and there was limited information regarding construct and scale alignment between the two assessments.

English learners attended five of the six middle schools and made up between 11.02% and 26.46% of students in their school.

Variables

Outcomes. The outcome variables for this study included students' scores on the Arizona English Language Learner Assessment (AZELLA) as well as the binary indicator of student reclassification from English learner to fluent English proficient. The AZELLA is Arizona's annual English language proficiency assessment and consists of four sub-tests: reading, writing, listening, and speaking. Each sub-test is a vertically scaled, IRT based assessment that consists of multiple choice, constructed response, short response, and extended response items (Harcourt, 2007). The listening and speaking tests are combined to produce an oral language proficiency score. In addition, items from all four domains are combined to produce an overall English proficiency score, referred to as the total score. The total score is not a composite scores, but an independently constructed and vertically scaled IRT-based score that has its own conditional standard errors with no direct reference to the domain scores (Harcourt, 2007). The vertical scale of the sub-tests and the total score provide a foundation for analysis of the measures across grades.

The binary reclassification indicator was coded 0 for each grade a student was classified as an English learner and was coded 1 for the first grade in which a student was reclassified as English language proficient. After reclassification, the student was no longer tracked. For these data, a student was reclassified when they earned an AZELLA total score in the proficient category. The AZELLA technical manual indicates that the total score of the 3 to 5 grade band test is composed of 31% speaking items, 19% listening items, 23% reading items and 27% writing items

Table 1.
*Grade of reclassification for the sample (*N = *277)*

	Reclassified		Censored	
Grade	*n*	(%)	*n*	(%)
3	6	(2.17)	2	(0.72)
4	130	(46.93)	31	(11.19)
5	41	(14.80)	11	(3.97)
6	40	(14.44)	6	(2.17)
7	4	(1.44)	6	(2.17)

*N*ote. Grade of reclassification refers to the last year a student was classified as an English learner.

compared to 30% speaking, 18% listening, 25% reading and 28% writing items for the 6 to 8 grade band test (Harcourt, 2007).

This study used AZELLA reading, writing, oral language, and total scores. The AZELLA scale was rescaled by dividing observed scores by 10 for computational purposes. Figure 2 plots the median, first and third quartile, and minimum and maximum rescaled score at each grade for each of the AZELLA domain scores and the total score. However, it is important to realize that English learners reclassify due to their performance on the AZELLA, so each subsequent grade contained fewer students, and those students remaining do not score as high as those students who reclassified. Table 1 provides the frequencies and percentages of students who reclassify at the end of each grade. Censored students are either those students who do not reclassify by the end of seventh grade or who were lost to follow-up due to begin retained a grade or leaving the district. The largest number of students reclassified in Grade 4 (130), and 56 students were censored by the end of Grade 7.

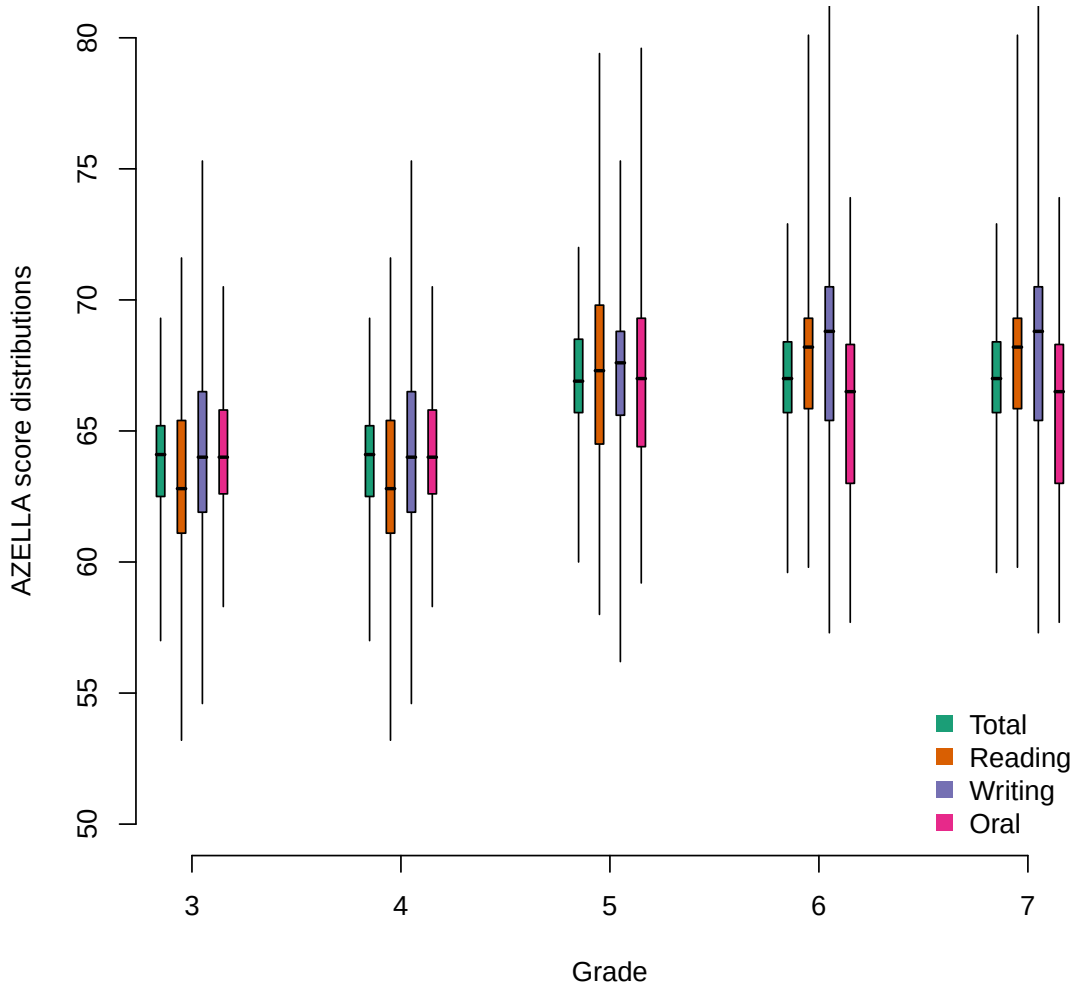


Figure 2. Distributions for AZELLA total and sub-test scores by grade

Covariates. In addition to the outcome measures, information pertaining to home language, race/ethnicity, free/reduced priced lunch status, special education status, and sex was collected (see Table 2). Among the sample of English learners, most students reported Spanish as their home language (93.5%). The next most common language was generically described as Other, Non-Indian (1.81%) and the other languages spoken by students in the cohort included Malay (0.36%), Navajo (0.36%), Somali (0.72%), Thai (0.36%), Urdu (0.36%), Vietnamese (0.72%), and Yaqui (0.36%). Within the cohort, most English learners identified as Hispanic (88.45%). While all Hispanic English learners reported their home language to be Spanish, 88.45% of students who reported speaking a home language of Spanish also identified as Hispanic. Because Spanish is the dominant language in the district, Table 2 pools home language into Spanish and non-Spanish. However, preliminary analyses were unable to detect any differences by home language or by race/ethnicity for this sample and therefore will not be discussed any further.

Free/reduced price lunch status (FRL) was used as a proxy measure for socioeconomic status. However, eligibility for free/reduced price lunch can vary from year to year based on family income levels. For example, student 54 in Figure 3 did not receive FRL in Grade 3, did in Grades 4 and 5, did not in Grade 6, but did so again in Grade 7. Across the sample, 90.25% of students were eligible for FRL in Grade 3 and 84.48% were eligible every year between Grades 3 and 7. Likewise, 9.75% of students were not eligible to receive FRL in Grade 3 but only 3.97% of students remained ineligible for FRL until Grade 7. A sensitivity analysis indicated little difference in mean overall AZELLA scores depending on FRL status. Thus, FRL was coded as ever being eligible for FRL as 1 and never

being eligible for FRL as 0. Furthermore, preliminary analysis was unable to detect any differences between these groups and will not be discussed further.

Like free/reduced price lunch status, classification as a student with a disability (SWD) is a time-varying process. While the SWD status for most English learners was invariant over time, some students in the sample were identified with a disability one year, and not the next. Across the sample, 83.75% of students were never diagnosed with a disability and 8.30% were classified as a SWD every year between Grades 3 and 7. The remaining 7.95% transitioned over time. For example, student 150 in Figure 3 was classified as a SWD in Grades 3, 4, and 7, but not in Grades 5 and 6. Because the group of students with varying SWD status was small, sensitivity analysis suggested little difference in mean overall AZELLA scores depending on the definition of FRL. Thus, SWD was coded those English learners who were ever classified as a SWD as 1 and those English learners who were never classified as a SWD as 0. Finally, the sample was 48.38% female and 51.62% male. For the analysis, a code of 0 indicated female English learners while a code of 1 indicated male English learners.

Missing data

As noted above, the data were highly unbalanced as students who met the reclassification criteria were no longer tested. Much of this imbalance should not be considered missing data as English proficiency scores were never intended to be collected from those students who reclassify. Covariates such as SWD or FRL were missing only when a student's entire record for that year was missing. Of the 277 English learners, 50 (18.05%) were lost to follow-up. Table 1 indicates two English learners were lost after Grade 3, 31 were lost after Grade 4, 11 were lost after Grade 5, and six were lost after Grade 6. Within each grade, Little's missing

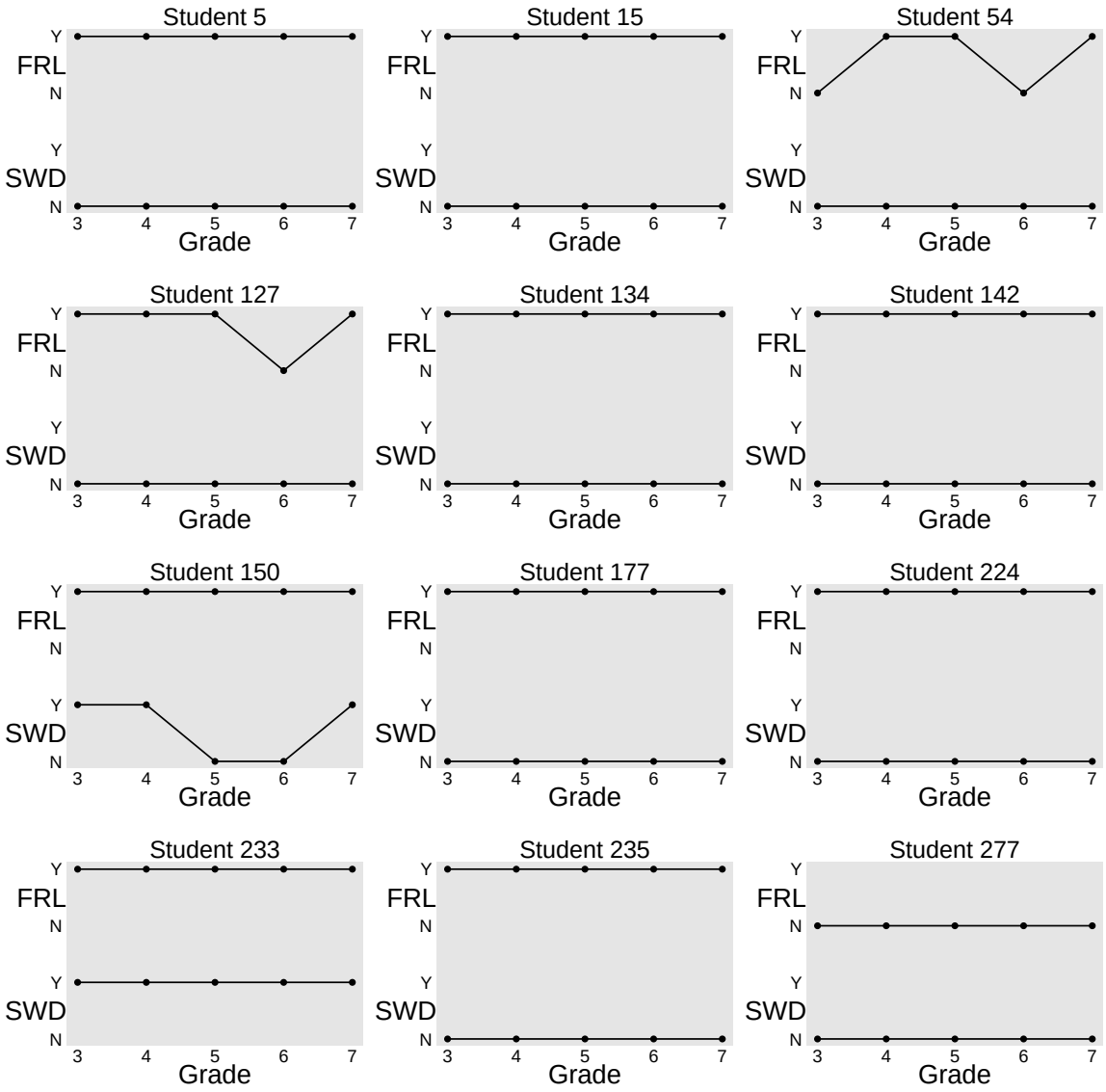


Figure 3. A random sample of students illustrating the variation in free/reduced price lunch status and special education status between Grades 3 and 7

Table 2.
*Demographic characteristics of the sample (*N = *277)*

Characteristic	*n*	%
Free or reduced price lunch		
No	11	3.97
Yes	266	96.03
Sex		
Female	134	48.38
Male	143	51.62
Student with disabilities		
No	232	83.75
Yes	45	16.25
Race / Ethnicity		
American Indian or Alaska Native	15	5.42
Asian	9	3.25
Black or African American	5	1.81
Hispanic	245	88.45
White	3	1.08
Home language		
Spanish	259	93.50
Other	18	6.50

completely at random (MCAR) test (Little, 1988) was not statistically significant ($p > .05$) indicating that within each grade, missing data were a random sample of the complete data. However, this does not guarantee that missing data across time were MCAR.

Missing data were handled using likelihood-based estimation assuming an ignorable missing data mechanism (Little, 1995; Rubin, 1976). For all growth models, each student contributed all available data. For all survival models, students who were lost to follow-up and thus, never experienced the event of

reclassification, were censored assuming they were non-informative. It is important to note, however, that there is no empirical procedure to determine if a missing mechanism (or censoring mechanism) is ignorable. Molenberghs, Beunckens, Sotto, and Kenward (2008) proved that any NMAR model fit to observed data can be reproduced exactly by an MAR counterpart.

Analysis

All models were fit within the probabilistic modeling language Stan (Gelman, Lee, & Guo, 2015) using Hamiltonian Monte Carlo estimation (Hoffman & Gelman, 2014). For these data, there were $k = 1, \ldots, 18$ schools, $j = 1, \ldots, n_k$ students in school k, and $i = 1, \ldots, n_{jk}$ observations for subject j in school k. For each student j in school k, consider the vectors of repeated measures outcomes, $\boldsymbol{y}_{Ojk}$, $\boldsymbol{y}_{Rjk}$, $\boldsymbol{y}_{Wjk}$, for the oral, reading, and writing AZELLA sub-tests, respectively, and $\boldsymbol{y}_{Tjk}$ being the AZELLA total score. Furthermore, consider the vector of reclassification indicators, $\boldsymbol{r}_{jk}$, where $r_{tjk} = 1$ for the time t a student reclassified to fluent English proficient student, and 0 until reclassification occurred.

For all analyses, model fit was evaluated using approximate leave-one-out cross validation (LOO) using Pareto smoothed importance sampling (Gelman, Hwang, & Vehtari, 2014; Vehtari, Gelman, & Gabry, 2016). LOO computes the expected log pointwise predictive density for a new dataset, *elpd*, and may be multiplied by -2 to be placed on the deviance scale. The deviance scaled elpd is referred to as the LOOIC and is a fully Bayesian information criterion that is viewed as an improvement over the deviance information criterion (DIC) and is more robust than the widely applicable or Watanabe-Akaike information criterion (WAIC) in the finite case with weak priors or influential observations. Model comparison may be conducted by taking the difference of the LOOCI, the

ΔLOOIC, and evaluate the difference in ratio to its standard error. Vehtari et al. (2016) suspect that this method of model comparison provides a better sense of uncertainty than conventional approaches that evaluate the difference of deviances in comparison to a χ^2 distribution. However, Vehtari et al. warn that such a comparison should not be used to select a single model from a set of potential models. To that end, posterior probability distributions were also evaluated based on their 95% highest posterior density (95% *HPD*) interval.

Research question 1. The first research question was addressed by fitting a series of univariate growth models for each of the four AZELLA outcomes using a systematic search for the best functional form in each case. The goal of this question was to understand the functional form of each outcome which guided model specification in the subsequent analyses. With measures nested in students and students nested in schools, I adopted the notation of (Skrondal & Rabe-Hesketh, 2004) to express the unit-level models generally as

$$y_{ijk} = \mathbf{x}'_{ijk}\boldsymbol{\beta} + \mathbf{z}^{(2)\prime}_{ijk}\boldsymbol{\zeta}^{(2)}_{jk} + \mathbf{z}^{(3)\prime}_{ijk}\boldsymbol{\zeta}^{(3)}_{k} + \epsilon_{ijk} \tag{2.1}$$

Where $\boldsymbol{\beta}$ was the $p \times 1$ vector of fixed effects corresponding to the known $p \times 1$ vector of explanatory variables, $\mathbf{x}_{ijk}$; $\mathbf{z}^{(2)}_{ijk}$ was the known $q^{(2)} \times 1$ vector of explanatory variables with student-specific random effects $\boldsymbol{\zeta}^{(2)}_{jk}$; and $\mathbf{z}^{(3)}_{ijk}$ was the known $q^{(3)} \times 1$ vector of explanatory variables with school-specific random effects $\boldsymbol{\zeta}^{(3)}_{k}$. I assumed $\boldsymbol{\zeta}^{(2)}_{jk} \sim \mathcal{N}(\mathbf{0}, \mathrm{T}^{(2)})$ and $\boldsymbol{\zeta}^{(3)}_{k} \sim \mathcal{N}(\mathbf{0}, \mathrm{T}^{(3)})$ where $\mathrm{T}^{(2)}$ and $\mathrm{T}^{(3)}$ are $q^{(2)} \times q^{(2)}$ and $q^{(3)} \times q^{(3)}$ unstructured variance-covariance matrices, respectively, and it was assumed that $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma)$. For those models that excluded the school-level random effects, Equation 2.1 reduced to

$$y_{ijk} = \mathbf{x}'_{ijk}\boldsymbol{\beta} + \mathbf{z}^{(2)\prime}_{ijk}\boldsymbol{\zeta}^{(2)}_{jk} + \epsilon_{ijk} \tag{2.2}$$

Because so many students in the sample were reclassified between third and fifth grade, the number of potential models was limited. Both linear and quadratic functions for the fixed effects were assessed while variance components were specified for the intercept and liner growth component. Unless otherwise noted, a random variable, x, was $\mathcal{U}(a,\, b)$ where $a \leq x \leq b$. If x was a standard deviation, $0 < x$ and $x < \infty$. For computational efficiency, Equation 2.1 was specified such that the Cholesky factorized random effects were given independent $\mathcal{N}(0, 1)$ priors while $\mathrm{T}^{(2)}$ and $\mathrm{T}^{(3)}$, re-specified as correlation matrices, were given $\mathcal{LKJ}(1.5)$ priors. While an $\mathcal{LKJ}(1)$ prior results in a uniform density over all correlation matrices of a given order (Lewandowski, Kurowicka, & Joe, 2009), an $\mathcal{LKJ}(1.5)$ was found to drastically reduce the autocorrelation in the chains. The complete set of models specified for Research Question 1 and their formulation is provided in Appendix A.

Research question 2. With univariate models specified, Research Question 2 required the joint analysis of the sub-test outcomes for the three measures in a multivariate random effects modeling framework,

$$y_{Rijk} = \mathbf{x}'_{Rijk}\boldsymbol{\beta}_R + \mathbf{z}^{(2)\prime}_{Rijk}\boldsymbol{\zeta}^{(2)}_{Rjk} + \mathbf{z}^{(3)\prime}_{Rijk}\boldsymbol{\zeta}^{(3)}_{Rk} + \boldsymbol{\epsilon}_{Rijk} \tag{2.3}$$

$$y_{Wijk} = \mathbf{x}'_{Wijk}\boldsymbol{\beta}_W + \mathbf{z}^{(2)\prime}_{Wijk}\boldsymbol{\zeta}^{(2)}_{Wjk} + \mathbf{z}^{(3)\prime}_{Wijk}\boldsymbol{\zeta}^{(3)}_{Wk} + \boldsymbol{\epsilon}_{Wijk} \tag{2.4}$$

$$y_{Oijk} = \mathbf{x}'_{Oijk}\boldsymbol{\beta}_O + \mathbf{z}^{(2)\prime}_{Oijk}\boldsymbol{\zeta}^{(2)}_{Ojk} + \mathbf{z}^{(3)\prime}_{Oijk}\boldsymbol{\zeta}^{(3)}_{Ok} + \boldsymbol{\epsilon}_{Oijk}. \tag{2.5}$$

Focus was on the estimated student-level covariance matrix for the random effects which estimated the student-specific dependency among the three outcomes. I defined $\boldsymbol{\zeta}^{(2)}_{jk} \equiv \left[\boldsymbol{\zeta}^{(2)\prime}_{Rjk}, \boldsymbol{\zeta}^{(2)\prime}_{Wjk}, \boldsymbol{\zeta}^{(2)\prime}_{Ojk}\right]'$, as a $q \times 1$ vector of student-specific random effects where q was the total number of student-specific random effects for all three submodels. Then I assumed $\boldsymbol{\zeta}^{(2)}_{jk}$ to have a zero mean and variance-covariance

matrix

$$\mathrm{T}^{(2)} = \begin{bmatrix} \tau_{11}^{(2)} & \tau_{12}^{(2)} & \cdots & \tau_{1q}^{(2)} \\ \tau_{21}^{(2)} & \tau_{22}^{(2)} & \cdots & \tau_{2q}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ \tau_{q1}^{(2)} & \tau_{q2}^{(2)} & \cdots & \tau_{qq}^{(2)} \end{bmatrix}$$

where $\sqrt{\tau_{pp}^{(2)}}$ was the standard deviation of $\zeta_{jkp}^{(2)}$ and $\rho_{pg} = \tau_{pg}^{(2)} / \sqrt{\tau_{pp}^{(2)} \tau_{gg}^{(2)}}$ was the correlation between $\zeta_{jkp}^{(2)}$ and $\zeta_{jkg}^{(2)}$ for $p \neq g$. Priors for Equation 2.3, including $\mathrm{T}^{(2)}$, were the same as those specified for Equation 2.1. The complete set of model specified for Research Question 2 and their formulation is provided in Appendix A.

Research question 3. Research Question 3 focused on the predictive accuracy of time to reclassification for a conventional survival model and a shared random effects model. English learners are eligible for reclassification at discrete times, $t = 1, 2, \ldots, T$. Consider the $T_{jk} \times 1$ vector of reclassification indicators $\boldsymbol{r}_{jk}$, where $r_{tjk} = 1$ when $t = T_{jk}$ indicating student j in school k was reclassified to fluent English proficient, and was 0 until reclassification.

The first model fit was a discrete-time survival model of the form:

$$\mathrm{logit}\{\mathbb{P}(r_{tjk} = 1 | \mathbf{w}_{tjk},\, \nu_k)\} = \mathrm{logit}(h_{tjk}) = \mathbf{w}_{tjk}' \boldsymbol{\alpha} + \nu_k \tag{2.6}$$

where h_{tij} was the hazard estimates at for student j in school k at time t, $\mathbf{w}_{tjk}$ was a $u \times 1$ vector of fixed effects, including indicators for each discrete time $1, \ldots, T$, $\boldsymbol{\alpha}$ was a $u \times 1$ vector of corresponding fixed effects, and ν_k was the school-level random intercept.

The second model fit was a shared random effects model which jointly estimated a growth submodel for the AZELLA total score as expressed in Equation 2.1 and a discrete-time survival submodel. The hazard submodel used the student-specific random effects estimated by the growth model as latent

Table 3.
Classification table of cell counts or proportions

	True Outcome	
Predicted Outcome	$r_{tjk} = 1$	$r_{tjk} = 0$
$\hat{r}_{tjk} = 1$	True Positive (TP)	False Positive (FP)
$\hat{r}_{tjk} = 0$	False Negative (FN)	True Negative (TN)

covariates to predict time to reclassification. The shared random effects model can be expressed as a system of equations:

$$y_{Tijk} = \mathbf{x}'_{ijk}\boldsymbol{\beta} + \mathbf{z}^{(2)\prime}_{ijk}\boldsymbol{\zeta}^{(2)}_{jk} + \mathbf{z}^{(3)\prime}_{ijk}\boldsymbol{\zeta}^{(3)}_{k} + \epsilon_{ijk} \tag{2.7a}$$

$$\text{logit}(h_{tjk}) = \mathbf{w}'_{tjk}\boldsymbol{\alpha} + \boldsymbol{\zeta}^{(2)\prime}_{jk}\boldsymbol{\lambda} + \nu_k \tag{2.7b}$$

where $\boldsymbol{\zeta}^{(2)}_{jk}$ in Equation 2.7b was a $q^{(2)} \times 1$ matrix of student-specific random effects estimated by the overall English proficiency growth model specified in Equation 2.7a, and $\boldsymbol{\lambda}$ was a $q^{(2)} \times 1$ vector of fixed effects corresponding to the latent covariates.

Priors for Model 2.7a were the same as those specified for Model 2.1. Due to the complexity of the shared-parameter model however, priors for Model 2.7b, in particular $\boldsymbol{\alpha}$ and $\boldsymbol{\lambda}$, were specified to constrain the support of the parameters (Gelman, Jakulin, Pittau, & Su, 2008). Specifically, $\boldsymbol{\alpha}$ and $\boldsymbol{\lambda}$ were specified with $\mathcal{N}(\mu, \sigma)$ where $-3 \leq \mu \leq 3$, and $6 \leq \sigma \leq 10$, which, on the logit scale, provided ample space for the data to dominate the posterior estimates. The complete set of models specified for Research Question 3 and their formulation is provided in Appendix A.

The predictive accuracy of the models can be understood through the cross-classification of $(r_{tjk}, \hat{r}_{tjk})$. Here, $\hat{r}_{tjk} = 1$ when $h_{tjk} \geq \pi$ where π is some probability threshold between 0 and 1. This cross-tabulation results in four

conditions: (a) true positives, (b) false positives, (c) true negatives, and (d) false negatives as seen in Table 3. False positives, or type I errors, are the counts or proportions of those subjects who have not reclassified at time t but were predicted to reclassify based on the model prediction, ($r_{tjk} = 0$, $\hat{r}_{tjk} = 1$). False negatives, or type II errors, are the counts or proportions of those students who do reclassify at time t but were not predicted to do so by the model, ($r_{tjk} = 1$, $\hat{r}_{tjk} = 0$). The true positive rate (TPR) for the model, defined as $\mathbb{P}(\hat{r}_{tjk} = 1 | r_{tjk} = 1)$, gives the probability that the model correctly predicted reclassification at time t. The true negative rate (TNR) for the model, defined as $\mathbb{P}(\hat{r}_{tjk} = 0 | r_{tjk} = 0)$, gives the probability that the model correctly predicted those students who were not ready to reclassify at time t. The overall classification accuracy of the model is then given by the probability that the model made an accurate prediction,

$$\mathbb{P}(\hat{r}_{tjk} = 1 | r_{tjk} = 1)\mathbb{P}(r_{tjk} = 1) + \mathbb{P}(\hat{r}_{tjk} = 0 | r_{tjk} = 0)\mathbb{P}(r_{tjk} = 0)\,.$$

The receiver operating characteristic (ROC) curve was generated by computing the TPR and TNR for $0 \leq \pi \leq 1$ and provides complete information on the set of all possible TPR/TNR rates. The area under the ROC curve was used to compare the predictive power of the hazard models.

CHAPTER III

RESULTS

In what follows, the notation for all models fit to answer the three research questions were suppressed and are presented in Appendix A. All models were estimated using four chains run for 2500 iterations using the first 1250 iterations for warm-up in each chain, resulting in 5000 samples from the posterior distribution. Evidence of convergence included $0.99 < \hat{R} < 1.01$ as well as confirming evidence from visual inspection of trace plots. Appendix B provides the convergence evidence for the five total score models and readers interested in reviewing additional convergence evidence are encouraged to contact the author.

English Language Proficiency Development

A series of mixed effect models were fit to each of the AZELLA tests to understand how AZELLA English proficiency scores developed across Grades 3 to 7 in an Arizona school district. The same model building protocol was followed for the AZELLA total score and each domain score. The first model specified linear development across grades and allowed intercepts and slopes to vary between students. The second model added a quadratic term to the fixed part of the model to determine if development was non-linear across grades. After the curvature of development was determined, the third model allowed the intercepts and slopes to vary between students and schools. The fourth model added student covariates to the model. Finally, the fifth model altered the specification of the priors used in the fourth model to assess the sensitivity of posteriors to the choice of model priors.

Total English proficiency. As noted above, five models were fit to identify the best model that described the functional form for overall English

proficiency development as measured by the AZELLA total score. Results for AZELLA total score growth models are shown in Table 4.

Linear growth model. The first model included fixed and student-specific random effects for the intercept, Grade 3 AZELLA total score, and linear growth in AZELLA total score across grades. Based on the estimates from the linear growth model, the average third grade English learner had a score of 63.72 on the AZELLA total English proficiency test, $\beta_1 = 63.72$, 95% $HPD = [63.48, 63.97]$, and had an increase of 3 points per year, $\beta_2 = 3.00$, 95% $HPD = [2.78, 3.22]$. Third grade test scores were found to vary between students by a standard deviation of 1.5 points, $\sqrt{\tau_{11}^{(2)}} = 1.54$, 95% $HPD = [1.31, 1.77]$, while annual growth was found to vary between students by a standard deviation of 0.62 points, $\sqrt{\tau_{22}^{(2)}} = 0.62$, 95% $HPD = [0.44, 0.83]$. The correlation between initial status and linear growth on the AZELLA total English proficiency test was $\tau_{21}^{(2)}/\sqrt{\tau_{11}^{(2)}\tau_{22}^{(2)}} = .75$, 95% $HPD = [.43, .97]$. Finally, the estimated within-student residual $\sqrt{\sigma}$ was 1.48, resulting in an estimated proportion of total variance between-students of .41.

The linear growth model provides a benchmark for comparing the model fit of subsequent models. The $\text{elpd}_{\text{psis-loo}}$ was -1261.22 with a standard error of 7.40, which on the deviance scale translated to a LOOIC of 2522.44 with standard error 14.80. The estimated effective number of parameters, p_{loo}, was 42.06 with a standard error of 2.02.

Quadratic growth model. The second model added a quadratic term to describe potential non-linearity in average AZELLA total score growth. In the quadratic growth model, the average student was estimated with a score of 63.81 in Grade 3 and had an initial linear growth of 2.44 points and an initial

quadratic acceleration of 0.32 points , $\beta_3 = 0.32$, 95% $HPD = [0.17, 0.46]$. This model included a quadratic term in the fixed part of the model only as there was insufficient within-student information to estimate variability in the curvature.

The quadratic model was a better fitting model than the linear model based on two criteria. First, given the data, the 95% *HPD* did not contain zero, rather, there was a 95% chance that the parameter for the quadratic term was between 0.17 and 0.46. Second, the LOOIC for the quadratic growth model was less than the LOOIC for linear growth model, $\Delta\text{LOOIC} = -30.79, SE = 4.2$, suggesting that the quadratic growth model would produce better out-of-sample predictions than the linear growth model.

School-level model. The third model extended the quadratic growth model by further partitioning the initial status and linear growth variance into between-student and between-school components. With the school-level model, the estimates for the fixed effects were unchanged but their interpretation evolved. The average student in the average school was estimated to have an AZELLA total score of 63.88 in Grade 3 and an initial linear change of 2.37 points and an initial quadratic acceleration of 0.36 points.

In addition to intercept and linear change varying between students, third grade test scores varied between schools by a standard deviation of 0.39 points, $\sqrt{\tau_{11}^{(3)}} = 0.39$, 95% $HPD = [0.03, 0.88]$, while linear growth varied between schools by a standard deviation of 0.54 points, $\sqrt{\tau_{22}^{(3)}} = 0.54$, 95% $HPD = [0.15, 1.01]$. The correlation between school-specific intercepts and slopes was estimated to be 0, $\tau_{21}^{(3)} / \sqrt{\tau_{11}^{(3)} \tau_{22}^{(3)}} = -.02$, 95% $HPD = [-0.77, 0.79]$. The estimated proportion of total variance between-schools was .10.

Determining if the school-level model fit the data better than the quadratic growth model was less straight forward. First, the 95% *HPD* for the between-school intercept and slope variance components did not contain zero, which favored their inclusion in the model. However, the 95% *HPD* for the intercepts-slopes correlation did contain 0. Furthermore, the ΔLOOIC of 41.0, $SE = 4.0$, suggested that the quadratic growth model would make better out-of-sample predictions than the school-level model. For these data with a limited number of schools, one could argue for the more parsimonious quadratic growth model (Bates, Kliegl, Vasishth, & Baayen, 2015). However, by including the school-level variance components, the student-specific random effects, which are the primary research interest, were arguably better estimated as they did not contain school-level variance. Additionally, when using a Bayesian estimation approach, there is no harm done by fitting a model with a full variance-covariance matrix for both students and schools.

School-level model with covariates. The fourth model extended the fixed part of the school-level model by adding two time-invariant dichotomous covariates, ever identified with a disability (SWDe) and sex, to the model. SWDe assigned the value of 0 to those English learners who were never identified with a disability up to Grade 7, and 1 to those English learners who had ever been identified with a disability up to Grade 7. Sex assigned the value of 0 to female English learners and 1 to male English learners.

Controlling for all other covariates, English learners who were ever identified with a disability were expected to score 2.37 points lower in third grade than those English learners who were never identified with a disability, $\beta_4 = -2.37$, 95% $HPD = [-2.93, -1.80]$. An interaction between SWDe and grade was also estimated, indicating that English learners who were ever identified

with a disability were expected to grow linearly by .45 points less than those English learners who were never identified with a disability, controlling for all other covariates, $\beta_5 = -0.45$, 95% $HPD = [-0.85, -0.05]$. Finally, male English learners were estimated to score 1.23 points lower in third grade than their female peers, controlling for other covariates, $\beta_6 = -1.23$, 95% $HPD = [-1.68, -0.78]$. Evaluation of the 95% HPD and the change in the LOOIC of -48.54, $SE = 4.8$, supported the school-level model with covariates over school-level model without. Although the inclusion of the grade by SWDe improved model fit, it is important to note that without probing the interaction between Grade and SWDe, this analysis cannot confirm that students identified with a disability grew at a different rate for the entirety of the study (Bauer & Curran, 2010; Stevens & Schulte, 2016).

Figure 4a is a plot the estimates from the school-level model with covariates illustrating the average English proficiency development for female and male students who were ever or never identified with a disability against the grade-specific total English proficiency benchmarks. Additionally, the total scores from the 277 English learners were plotted. Females never identified with a disability (green solid line), on average, met the total English proficiency benchmark by fourth grade. Males never identified with a disability (orange solid line), on average, met the benchmark by fifth grade. Females ever identified with a disability (green dashed line) also, on average, met the benchmark by fifth grade. Lastly, males ever identified with a disability (orange dashed line), on average, met the benchmark by sixth grade. All female and male English learners never identified with a disability met the total English proficiency benchmark by sixth grade while some females and males ever identified with a disability had yet to meet the benchmark by the end of seventh grade.

Weakly informative prior model. The last model as shown in Table 4 was identical to the school-level model with covariates but specified weakly informative priors for the fixed effects. That is, under the school-level model with covariates, β_4, β_5, β_6 were each specified with the non-informative prior $\mathcal{U}(-\infty, \infty)$. The weakly informative prior model, however, employed the priors $\beta_4 \sim \mathcal{N}(0, 5)$, $\beta_5 \sim \mathcal{N}(0, 2.5)$, $\beta_6 \sim \mathcal{N}(0, 5)$. Application of this model, however, resulted in parameter estimates that were unchanged. Therefore, I concluded that the school-level model with covariates was not sensitive to the choice of reasonable priors.

Domain-specific proficiency models. The same model fitting process described for the AZELLA total score was also used to analyze the AZELLA domain scores. For brevity, only the final models for each outcome were presented in Table 5. Results for the preceding steps in model fitting produced similar results as those found and just described for the AZELLA total scores. Complete results for AZELLA reading, writing, and oral proficiency growth models are shown in Tables C.1, C.2, and C.3 of Appendix C.

Reading proficiency. The models fit to the AZELLA reading domain score found that, unlike the total score models, growth in reading was linear. That is, neither the 95% *HPD* for the quadratic effect of time nor the LOOIC of the quadratic growth model favored the additional fixed effect. For the school-level model, the 95% *HPD* for the between-school intercept and slope variances did not contain zero, nor did the values of the LOOIC suggest that either model with or without school-level variances predicted new data better, $\Delta\text{LOOIC} = -1.57$, $SE = 4.6$. Like the total score models, the between-school intercept and slope correlation was imprecise and included zero in the 95% *HPD*, $\tau_{21}^{(3)}/\sqrt{\tau_{11}^{(3)}\tau_{22}^{(3)}} = -.24$, 95% $HPD = [-0.93, 0.80]$.

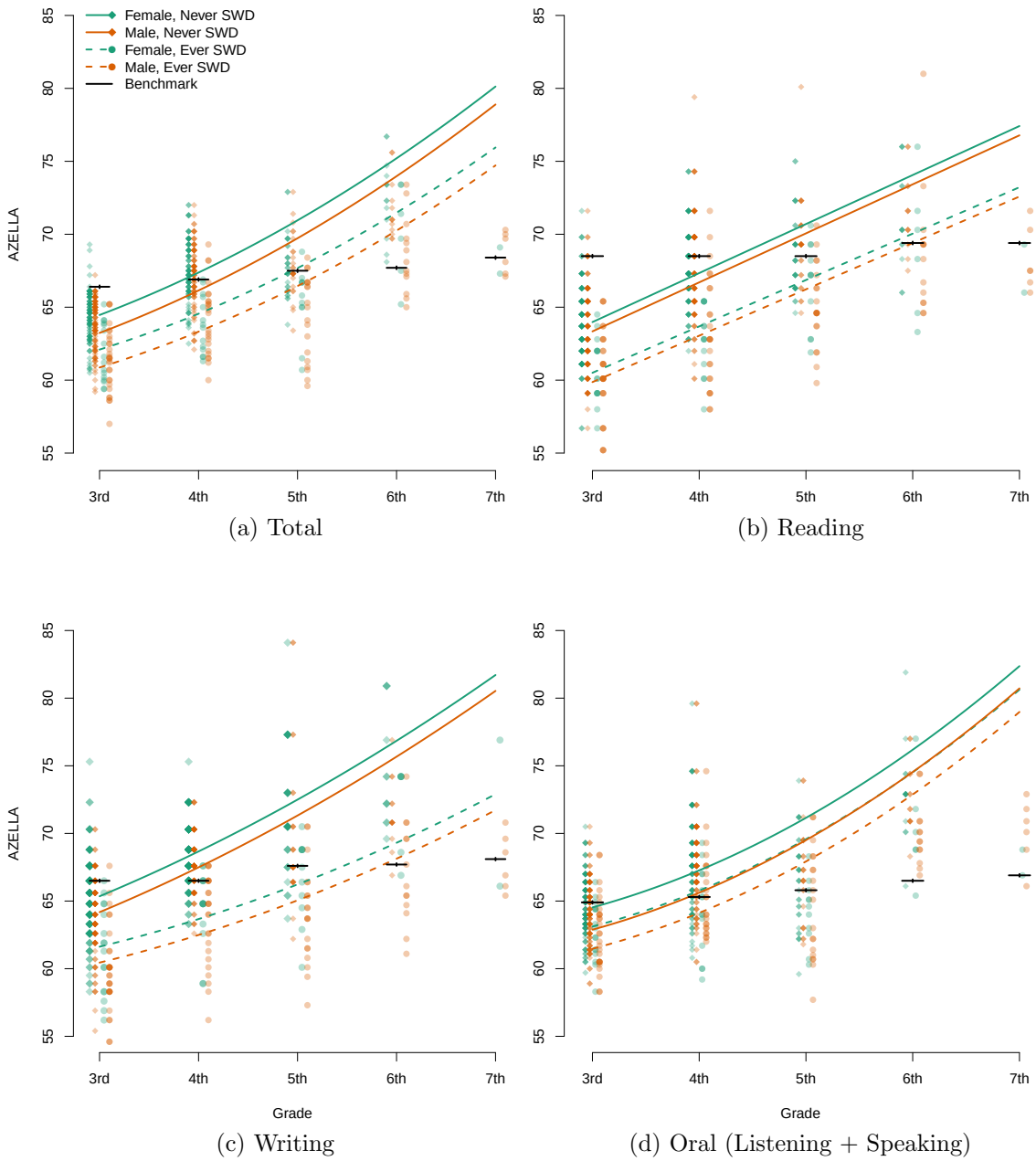
Table 4.
Parameter estimates for the total English language proficiency growth models

	Model 1			Model 2			Model 3			Model 4			Model 5		
	M	95% *HPD*		*M*	95% *HPD*		*M*	95% *HPD*		*M*	95% *HPD*		*M*	95% *HPD*	
Fixed Effects															
β_1 [Initial Status]	63.72	[63.48,	63.97]	63.81	[63.57,	64.05]	63.88	[63.56,	64.26]	64.46	[64.14,	64.80]	64.46	[64.12,	64.80]
β_2 [Grade]	3.00	[2.78,	3.22]	2.44	[2.09,	2.79]	2.37	[1.91,	2.82]	2.57	[2.13,	3.00]	2.57	[2.14,	2.98]
β_3 [Grade^2]				0.32	[0.17,	0.46]	0.36	[0.21,	0.50]	0.34	[0.19,	0.48]	0.34	[0.19,	0.48]
β_4 [SWDe]										−2.37	[−2.93,	−1.80]	−2.37	[−2.92,	−1.81]
β_5 [Grade] × [SWDe]										−0.45	[−0.85,	−0.05]	−0.45	[−0.86,	−0.05]
β_6 [Male]										−1.23	[−1.68,	−0.78]	−1.24	[−1.69,	−0.81]
Variance Components															
$\sqrt{\tau_{11}^{(2)}}$	1.54	[1.31,	1.77]	1.46	[1.24,	1.69]	1.45	[1.23,	1.67]	1.14	[0.93,	1.36]	1.14	[0.92,	1.37]
$\sqrt{\tau_{22}^{(2)}}$	0.62	[0.44,	0.83]	0.89	[0.65,	1.13]	0.87	[0.64,	1.12]	0.66	[0.43,	0.90]	0.66	[0.44,	0.89]
$\tau_{21}^{(2)}/\sqrt{\tau_{11}^{(2)}\tau_{22}^{(2)}}$	0.75	[0.43,	0.97]	0.77	[0.49,	0.97]	0.78	[0.51,	0.97]	0.67	[0.28,	0.96]	0.67	[0.27,	0.96]
$\sqrt{\tau_{11}^{(3)}}$							0.39	[0.03,	0.88]	0.24	[0.01,	0.60]	0.23	[0.01,	0.60]
$\sqrt{\tau_{22}^{(3)}}$							0.54	[0.15,	1.01]	0.45	[0.11,	0.87]	0.45	[0.13,	0.87]
$\tau_{21}^{(3)}/\sqrt{\tau_{11}^{(3)}\tau_{22}^{(3)}}$							0.49	[−0.77,	0.79]	0.52	[−0.83,	0.86]	0.51	[−0.81,	0.85]
$\sqrt{\sigma}$	1.48	[1.37,	1.60]	1.40	[1.29,	1.52]	1.37	[1.26,	1.49]	1.38	[1.26,	1.50]	1.38	[1.27,	1.50]

Model Fit	Estimate	(*SE*)	Estimate	(*SE*)	Estimate	(*SE*)	Estimate	(*SE*)	Estimate	(*SE*)
$\text{elpd}_{\text{psis-loo}}$	−1261.22	(7.40)	−1245.82	(7.05)	−1266.31	(6.55)	−1242.04	(7.47)	−1241.57	(7.48)
p_{loo}	42.06	(2.02)	37.52	(1.92)	28.24	(1.43)	34.56	(1.83)	34.57	(1.84)
LOOIC	2522.44	(14.80)	2491.65	(14.10)	2532.62	(13.10)	2484.08	(14.94)	2483.13	(14.96)

Table 5.
Parameter estimates for final reading, writing, and oral English language proficiency growth models

	Reading 4			Writing 4			Oral 4		
	M	95% *HPD*		*M*	95% *HPD*		*M*	95% *HPD*	
Fixed Effects									
β_1 [Initial Status]	63.98	[63.50,	64.48]	65.35	[64.85,	65.90]	64.53	[63.89,	65.15]
β_2 [Grade]	3.36	[2.89,	3.82]	3.06	[2.42,	3.68]	2.16	[1.51,	2.79]
β_3 [Grade^2]				0.26	[0.05,	0.46]	0.57	[0.31,	0.83]
β_4 [SWDe]	−3.47	[−4.33,	−2.63]	−3.73	[−4.58,	−2.89]	−1.41	[−2.18,	−0.68]
β_5 [Grade] × [SWDe]	−0.18	[−0.75,	0.40]	−1.27	[−1.86,	−0.67]	−0.08	[−0.63,	0.46]
β_6 [Male]	−0.64	[−1.17,	−0.10]	−1.18	[−1.82,	−0.53]	−1.65	[−2.49,	−0.85]
Variance Components									
$\sqrt{\tau_{11}^{(2)}}$	1.69	[1.35,	2.06]	1.82	[1.42,	2.20]	0.88	[0.51,	1.25]
$\sqrt{\tau_{22}^{(2)}}$	0.64	[0.23,	1.04]	1.05	[0.69,	1.40]	1.37	[0.90,	1.86]
$\tau_{21}^{(2)}/\sqrt{\tau_{11}^{(2)}\tau_{22}^{(2)}}$	0.28	[−0.25,	0.90]	0.04	[−0.33,	0.53]	0.80	[0.42,	0.98]
$\sqrt{\tau_{11}^{(3)}}$	0.29	[0.01,	0.81]	0.41	[0.04,	0.93]	0.91	[0.45,	1.62]
$\sqrt{\tau_{22}^{(3)}}$	0.55	[0.14,	1.08]	0.65	[0.27,	1.16]	0.50	[0.03,	1.25]
$\tau_{21}^{(3)}/\sqrt{\tau_{11}^{(3)}\tau_{22}^{(3)}}$	0.39	[−0.93,	0.77]	0.28	[−0.96,	0.51]	0.44	[−0.81,	0.75]
$\sqrt{\sigma}$	2.24	[2.05,	2.42]	2.15	[1.96,	2.36]	2.33	[2.15,	2.54]
Predictive Accuracy	Estimate		(*SE*)	Estimate		(*SE*)	Estimate		(*SE*)
$\text{elpd}_{\text{psis-loo}}$	−1513.84		(11.00)	−1500.41		(9.32)	−1528.71		(11.13)
p_{loo}	64.91		(3.82)	62.21		(3.50)	47.42		(3.61)
LOOIC	3027.67		(22.00)	3000.82		(18.63)	3057.42		(22.26)



(a) Total (b) Reading

(c) Writing (d) Oral (Listening + Speaking)

Figure 4. Fitted prototypical trajectories by subgroup for the total AZELLA score and each sub-test. The solid black lines represent the test-specific proficiency benchmarks. The points represent the data used to fit the models.

The school-level model with covariates showed students who were ever identified with a disability were expected to score 3.5 points lower on the Grade 3 AZELLA reading test than students never identified with a disability, $\beta_4 = -3.47$, 95% $HPD = [-4.33, -2.63]$. There, was little evidence, given the data, that students who were ever identified with a disability grew at a different rate than those never identified with a disability, $\beta_5 = -0.18$, 95% $HPD = [-0.75, 0.40]$. Finally, males were estimated to score 0.6 points lower on the Grade 3 reading sub-test than females, controlling for other covariates, $\beta_6 = -0.64$, 95% $HPD = [-1.17, -0.10]$.

Figure 4b is a plot of the estimates from the school-level model with covariates illustrating the average reading proficiency development for female and male students ever or never identified with a disability against the grade-specific reading proficiency benchmarks. Additionally, the reading domain scores from the 277 English learners were plotted. Females and males never identified with a disability (green solid line, orange solid line), on average, met the reading proficiency benchmark by fifth grade. Females and males ever identified with a disability (green dashed line and orange dashed line), on average, met the benchmark by sixth grade.

Writing proficiency. Like the total score models, the models fit to the AZELLA writing domain scores supported a quadratic term to describe the non-linearity in annual growth. Support for the inclusion of between-school variance components was similar to that of the total score model as well. The school-level model with covariates estimated that students who were ever identified with a disability were expected to score 3.5 points lower on the Grade 3 AZELLA writing sub-test, $\beta_4 = -3.73$, 95% $HPD = [-4.58, -2.89]$, and had an initial linear growth

of 1.3 points less than students never identified with a disability, controlling for other covariates, $\beta_5 = -1.27$, 95% $HPD = [-1.86, -0.67]$. Finally, males were estimated to score 1.2 points lower than females on the Grade 3 AZELLA writing sub-test, controlling for other covariates, $\beta_6 = -1.18$, 95% $HPD = [-1.82, -0.53]$.

Figure 4c is a plot of the estimates from the school-level model with covariates illustrating the average writing proficiency development for female and male student ever or never identified with a disability against the grade-specific writing proficiency benchmarks. Additionally, the writing domain scores from the 277 English learners were plotted. Females and males never identified with a disability (green solid line and orange solid line), on average, met the total English proficiency benchmark by fourth grade. Females and males ever identified with a disability (green dashed line add orange dashed line), on average, met the benchmark by sixth grade.

Oral proficiency. The last outcome analyzed was the AZELLA oral English proficiency sub-test, which combined speaking and listening domains. A quadratic growth model provided better fit than a linear growth model as demonstrated by both the 95% *HPD*s for β_3, [0.22, 0.74], and LOOIC comparison, $\Delta\text{LOOIC} = -68.65, SE = 7.6$. Likewise, 95% *HPD* and the LOOIC comparison between the quadratic growth model and the school-level model supported the inclusion of between-school intercept and slope variance components, $\Delta\text{LOOIC} = -14.9 SE = 7.4$. In the school-level model with covariates, students who were ever identified with a disability scored 1.4 points lower on the Grade 3 AZELLA writing test than students never diagnosed with a disability, controlling for other covariates, $\beta_4 = -1.41$, 95% $HPD = [-2.18, -0.68]$. Further, males scored

1.7 points lower on the Grade 3 oral sub-test than females, controlling for other variables, $\beta_6 = -1.65$, 95% $HPD = [-2.49, -0.85]$.

Figure 4d is a plot of the estimates from the school-level model with covariates illustrating the average oral proficiency development for female and male student ever or never identified with a disability against the grade-specific oral proficiency benchmarks. Additionally, the oral sub-test scores from the 277 English learners were plotted. Females ever and never identified with a disability and males never identified with a disability (green solid line, green dashed line, and orange solid line), on average, met the total English proficiency benchmark by fourth grade. Males ever identified with a disability (orange dashed line), on average, met the benchmark by fifth grade.

Correlation Among Domain-Specific English Proficiency Development

Three multivariate mixed effect models were fit to the AZELLA reading, writing and oral domain scores. The first model combined the three school-level models described above, the second model combined the three school-level models with covariates described above, and the third model combined the three weakly informative prior models described above. The fixed effect estimates for the multivariate models mirrored the estimates from the univariate analyses so they will not be reinterpreted here. Results may be found in Table C.4 in Appendix C.

To address Research Question 2, I focused on the variation and covariation among student-specific random effects from the multivariate school-level model with covariates. To facilitate interpretation, the six parameters that estimated between-student variation in each intercept and slope were presented as standard deviations, and the 15 parameters that estimated the covariation between those intercepts and slopes were presented as correlations (see Table 6).

Table 6.
Estimates of student-specific variances (as standard deviations) and correlations of the random intercepts and slopes for the final multivariate mixed effect model for reading, writing, and oral English proficiency growth

	1. $\zeta^{(2)}_{R1i}$	2. $\zeta^{(2)}_{R2i}$	3. $\zeta^{(2)}_{W1i}$	4. $\zeta^{(2)}_{W2i}$	5. $\zeta^{(2)}_{O1i}$	6. $\zeta^{(2)}_{O2i}$
1. $\zeta^{(2)}_{R1i}$ [Reading Intercept]	1.77 (0.17)					
2. $\zeta^{(2)}_{R2i}$ [Reading Slope]	0.01 (0.20)	0.79 (0.18)				
3. $\zeta^{(2)}_{W1i}$ [Writing Intercept]	0.82 (0.07)	0.04 (0.19)	1.87 (0.17)			
4. $\zeta^{(2)}_{W2i}$ [Writing Slope]	0.06 (0.17)	0.30 (0.20)	−0.04 (0.18)	1.13 (0.18)		
5. $\zeta^{(2)}_{O1i}$ [Oral Intercept]	0.54 (0.15)	0.34 (0.22)	0.66 (0.13)	0.00 (0.21)	0.99 (0.17)	
6. $\zeta^{(2)}_{O2i}$ [Oral Slope]	0.34 (0.14)	0.48 (0.18)	0.41 (0.13)	−0.12 (0.18)	0.63 (0.16)	1.32 (0.24)

*N*ote. The numbers are the posterior means and standard deviations (in parentheses).

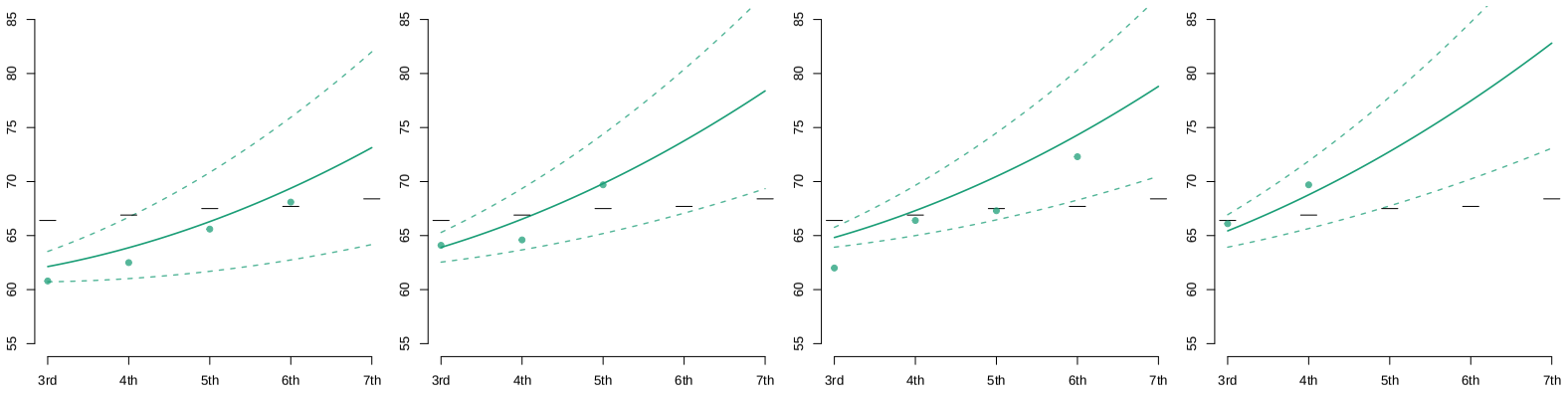
The standard deviations for the six intercepts and slopes were similar to those of their univariate counterparts. None of the 95% *HPD*s for the six variances contained 0. Like the oral English proficiency univariate model, variation in oral proficiency growth was estimated to be greater than the variation in third grade oral proficiency, $\sqrt{\tau^{(2)}_{55}} = 0.99$, 95% $HPD = [0.66, 1.33]$ and $\sqrt{\tau^{(2)}_{66}} = 1.32$, 95% $HPD = [0.87, 1.80]$, respectively.

The means of all 15 student-specific correlation posteriors were greater than 0. Only seven of the 15, however, were estimated with a 95% *HPD* that did not include zero. The correlations between intercepts and slopes for the same outcome also mirrored the findings from the univariate models. The correlation between oral proficiency intercepts and oral proficiency slopes was the only one of the three to have a 95% *HPD* that did not include zero, $\tau^{(2)}_{65}/\sqrt{\tau^{(2)}_{55}\tau^{(2)}_{66}} = .70$, 95% $HPD = [.40, .91]$.

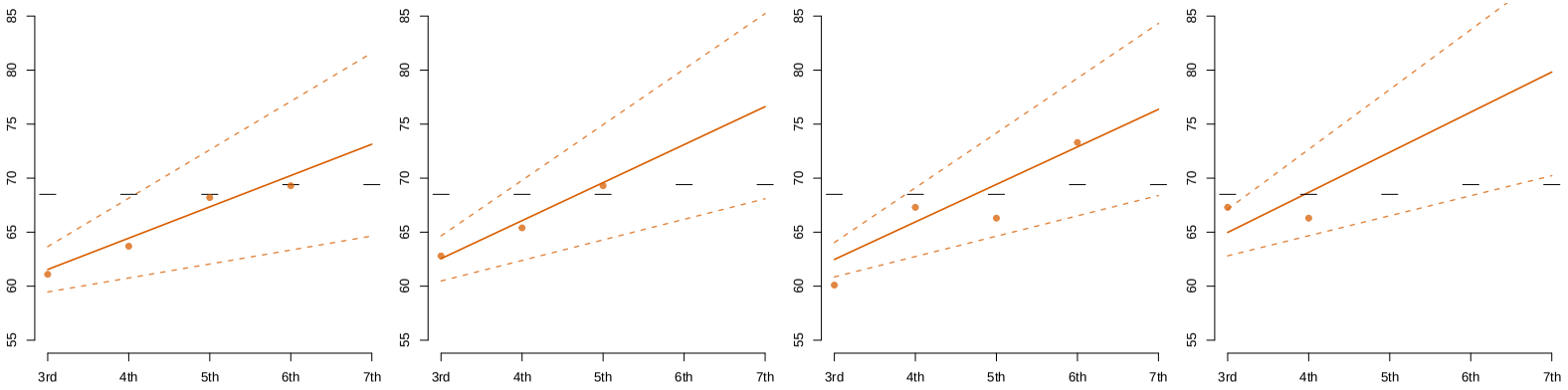
All three initial status correlation parameters were positively correlated above .5. The strongest correlation was between third grade reading proficiency and third grade writing proficiency, $\tau^{(2)}_{31}/\sqrt{\tau^{(2)}_{11}\tau^{(2)}_{33}} = .82$, 95% $HPD = [.66, .94]$, suggesting students with a high third grade reading score would also have a high third grade writing score. The next strongest correlation was between third grade writing proficiency and third grade oral proficiency, $\tau^{(2)}_{53}/\sqrt{\tau^{(2)}_{33}\tau^{(2)}_{55}} = .66$, 95% $HPD = [.39, .87]$. While still greater than .5, the correlation between the third grade reading proficiency and third grade oral proficiency was the weakest among the three, $\tau^{(2)}_{51}/\sqrt{\tau^{(2)}_{11}\tau^{(2)}_{55}} = .54$, 95% $HPD = [.24, .80]$.

The only correlation among the three linear growth random effects that was estimated with a 95% *HPD* that did not cross zero was between reading and oral proficiency, $\tau^{(2)}_{62}/\sqrt{\tau^{(2)}_{22}\tau^{(2)}_{66}} = .48$, 95% $HPD = [.08, .80]$, indicating that students who made above average growth in reading were also likely to make above average growth in listening and speaking development. There was no correlation between writing growth and oral proficiency growth, $\tau^{(2)}_{64}/\sqrt{\tau^{(2)}_{44}\tau^{(2)}_{66}} = -.12$, 95% $HPD = [-.46, .26]$, or between reading growth and writing growth, $\tau^{(2)}_{42}/\sqrt{\tau^{(2)}_{22}\tau^{(2)}_{44}} = .30$, 95% $HPD = [-.11, .66]$.

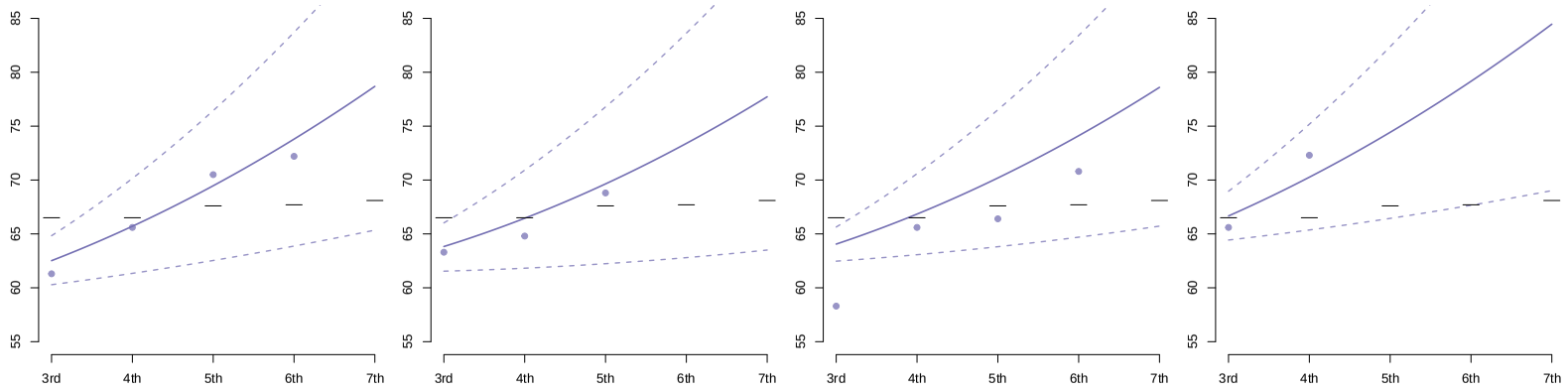
Figure 9 is a plot of the fitted subject-specific developmental trajectories (mean and 95% *HPD*), observed scores, and proficiency benchmarks for total English language proficiency and domain-specific proficiencies for a random sample of students. The first row of figures (Figures 5a to 5d) use the parameter estimates from Model 4 of Table 4 to illustrate total English language proficiency development for each student while Figures 5e, to 5p plot the subject-specific developmental trajectories for the three domains as estimated by Model 3 of Table C.4 for the same students. Focusing on Student 88, Figure 5a shows that



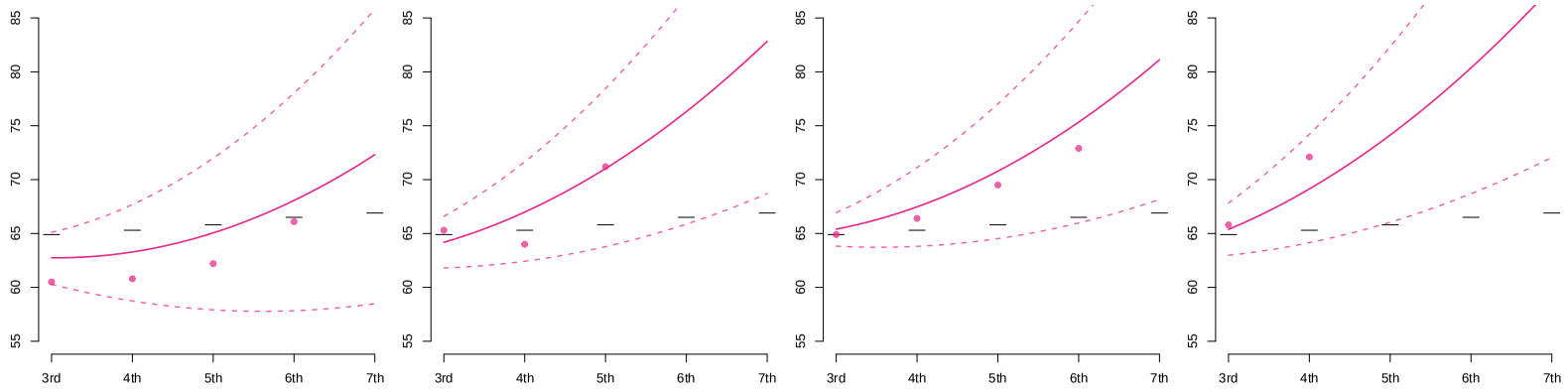
(a) Total Score, ID=88 (b) Total Score, ID=96 (c) Total Score, ID=244 (d) Total Score, ID=34



(e) Reading Score, ID=88 (f) Reading Score, ID=96 (g) Reading Score, ID=244 (h) Reading Score, ID=34



(i) Writing Score, ID=88 (j) Writing Score, ID=96 (k) Writing Score, ID=244 (l) Writing Score, ID=34



(m) Oral Score, ID=88 (n) Oral Score, ID=96 (o) Oral Score, ID=244 (p) Oral Score, ID=34

Figure 5. Fitted AZELLA trajectories, 95% *HPD*s, and observed scores for a random sample of students. The AZELLA total score fitted trajectories are based on Model 4 of Table 4 and AZELLA sub-test fitted trajectories are based on Model 3 of Table C.4.

the student met the benchmark in grade 6 which corresponded to when the model predicted Student 88 would meet the benchmark. Figure 5e shows the student met (and was predicted to meet) the reading benchmark in Grade 6. Figure 5i shows the student met (and was predicted to meet) the benchmark for writing in Grade 5. Finally, Figure 5m shows the student did not meet the benchmark for oral proficiency while classified as an English learner but the model had predicted they had.

Predicting Time-to-Reclassification

A series of discrete-time hazard models and shared random effects models were fit to understand whether initial status and growth in total English language proficiency improved predictions of time to reclassification. The fitted models were then compared based on their ability to accurately predict the time for a student to reclassify. For brevity, only the final hazard models and two hazard submodels from the shared random effects models were presented in Table 7. Interested readers may refer to Appendix C, Tables C.5 and C.6 for the complete results of the hazard models and shared random effects models, respectively.

Hazard models.

Hazard model with manifest covariates. The first model is a conventional hazard model with manifest covariates, which described the logit of the hazard of reclassification for Grades 3 through 7 conditioning on the two time-invariant dichotomous manifest covariates, SWDe and sex, and the school-level random intercept. A female English learner who was never identified with a disability had a .002 probability of reclassifying in Grade 3, $\alpha_1 = -3.70$, 95% $HPD = [-4.70, -2.83]$. The probability increased to 0.6 in Grades 4 and 5, $\alpha_2 = 0.46$, 95% $HPD = [0.02, 0.92]$; , $\alpha_3 = 0.44$, 95% $HPD = [-0.14, 1.03]$. By

Table 7.
Parameter estimates for the final time-to-reclassification hazard model and the shared random effects hazard submodels

	Model 1			Model 2			Model 3		
	M	95% *HPD*		*M*	95% *HPD*		*M*	95% *HPD*	
α_1 [Grade 3]	−3.70	[−4.70,	−2.83]	−10.59	[−13.14,	−8.35]	−8.66	[−10.95,	−6.65]
α_2 [Grade 4]	0.46	[0.02,	0.92]	−0.18	[−1.09,	0.76]	1.31	[0.23,	2.47]
α_3 [Grade 5]	0.44	[−0.14,	1.03]	2.97	[1.85,	4.27]	4.67	[3.21,	6.32]
α_4 [Grade 6]	3.11	[2.17,	4.16]	7.89	[5.93,	10.08]	10.52	[8.15,	13.19]
α_5 [Grade 7]	2.66	[1.03,	4.28]	8.80	[5.85,	11.89]	12.19	[8.87,	15.72]
α_6 [SWDe]	−2.82	[−3.65,	−2.11]				−8.35	[−10.96,	−6.15]
α_7 [Male]	−0.27	[−0.71,	0.15]				−0.83	[−2.03,	0.32]
λ_1 [Initial Status, $\zeta_{1i}^{(2)}$]				2.13	[1.00,	3.37]	2.14	[1.05,	3.32]
λ_2 [Linear Growth, $\zeta_{2i}^{(2)}$]				3.08	[1.56,	4.73]	3.01	[1.49,	4.86]
$\sqrt{\nu}$	0.57	[0.25,	1.00]	0.54	[0.04,	1.47]	0.73	[0.08,	1.77]

Model Fit	Estimate	(*SE*)	Estimate	(*SE*)	Estimate	(*SE*)
$\text{elpd}_{\text{psis-loo}}$	−290.07	(14.44)	−1672.89	(32.48)	−1637.02	(31.48)
p_{loo}	16.92	(1.26)	136.48	(7.57)	114.87	(5.68)
LOOIC	580.15	(28.87)	3345.78	(64.96)	3274.03	(62.96)

*N*ote. Models 2 and 3 are the hazard submodels from Models 2 and 3 of Tabel C.6.

middle school (Grades 6 and 7), the probability of reclassification was greater than .9, $\alpha_4 = 3.11$, 95% $HPD = [2.17, 4.16]$; $\alpha_5 = 2.66$, 95% $HPD = [1.03, 4.28]$.

The mean of the posterior for SWDe was -2.82 logits, $\alpha_6 = -2.82$, 95% $HPD = [-3.65 - 2.11]$. Expressed as an odds ratio, the odds of reclassification at any grade for an English learner never identified with a disability was 16.72 times that of an English learner ever identified with a disability. The mean of the posterior for sex was -0.27 logits, $\alpha_7 = -0.27$, 95% $HPD = [-0.71, 0.15]$. The odds of reclassification at any grade for a female English learner was 1.31 times greater than for a male English learner. While the 95% *HPD* for sex included zero, the inclusion of both predictors improved the estimated out-of-sample predictive ability of the model, $\Delta\text{LOOIC} = -80.3, SE = 17.2$. Using a latent response interpretation, the correlation among the latent responses of any two students from the same school was only

$$\rho = \frac{\nu}{\nu + \pi^2/3} = \frac{0.57^2}{0.57^2 + \pi^2/3} = .09\,.$$

Figure 6 is a plot of the estimates from the hazard model with manifest covariates, illustrating the school-specific probabilities of reclassification at each grade for female and male English learners who were ever or never identified with a disability. The primary feature of the plot is the difference in the probability of reclassification between Grades 3 and 7 for English learners who were ever and never identified with a disability, regardless of sex. The probability of reclassification for those never identified with a disability between third grade and seventh grade, regardless of sex, appeared as a step function. At the end of third grade, English learners never identified with a disability had a near 0% chance of reclassification. The probability of reclassification increased to approximately 60% at the end of fourth grade, remained steady through fifth grade, increased again

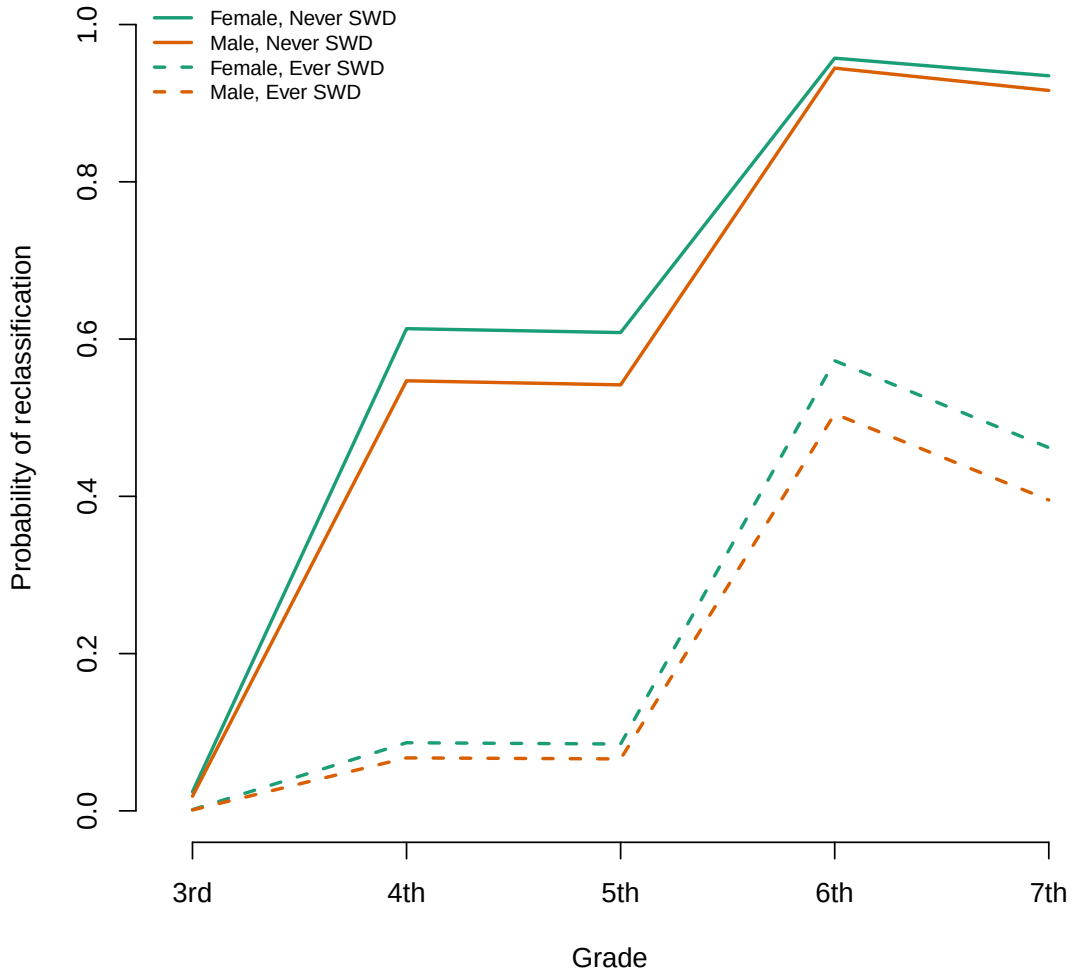


Figure 6. Probability of reclassification by subgroup for the hazard model with manifest covariates

to approximately 90% at the end of sixth grade, and remained steady through seventh grade. The probability of reclassification for English learners ever identified with a disability, regardless of sex, was nearly flat between third grade and fifth grade, starting at nearly 0% and never increasing above 10%. The probability of reclassification increased to approximately 50% at the end of sixth grade, and dropped to approximately 40% at the end of seventh grade.

The fitted model parameters were used to predict whether or not reclassification would occur for observation i at time t. Using $\pi = .5$ as a probability threshold, the hazard model with manifest covariates was 80.94% accurate when predicting if reclassification does or does not occur for the 720 observations across 277 English learners. The TPR was .76 indicating that 76% of the English learners predicted to reclassify at the end of a given grade by the hazard model with manifest covariates actually reclassified at the end of that grade. The TNR was .83 indicating that 83% of the students predicted by the model to not reclassify at the end of a given grade actually did not reclassify at the end of that grade. Figure 7 is a plot of the ROC curve for the hazard model with covariates (green solid line). The ROC curve illustrates the trade-off between true positive rate and true negative rate (1−TNR) as the probability threshold, π, increased from 0 to 1. In the context of model comparison, the ROC curve for the hazard model with covariates provided a baseline to compare the predictive power of the hazard submodels from the shared random effects models.

Shared random effects models. Both shared random effects models presented below employed the school-level model with covariates to describe AZELLA total score development. The student-specific random effects for initial status, $\zeta_{1i}^{(2)}$, and linear growth, $\zeta_{2i}^{(2)}$, were then shared with the discrete-time hazard

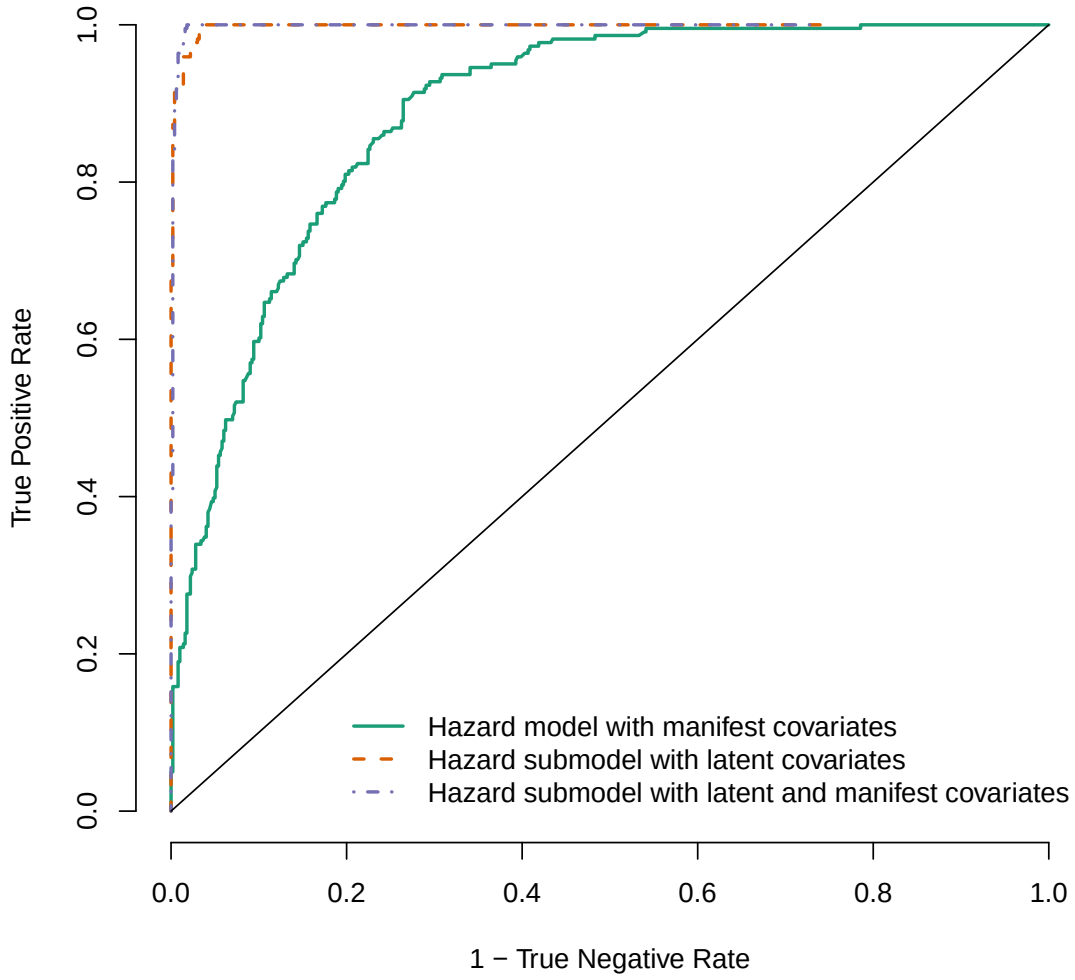


Figure 7. ROC curves for the time-to-reclassification models

model for reclassification. That is, the student-specific random effects are used as latent covariates to predict time-to-reclassification. Because the primary interest was in the prediction of time to reclassification, the following results will be limited to the hazard submodels.

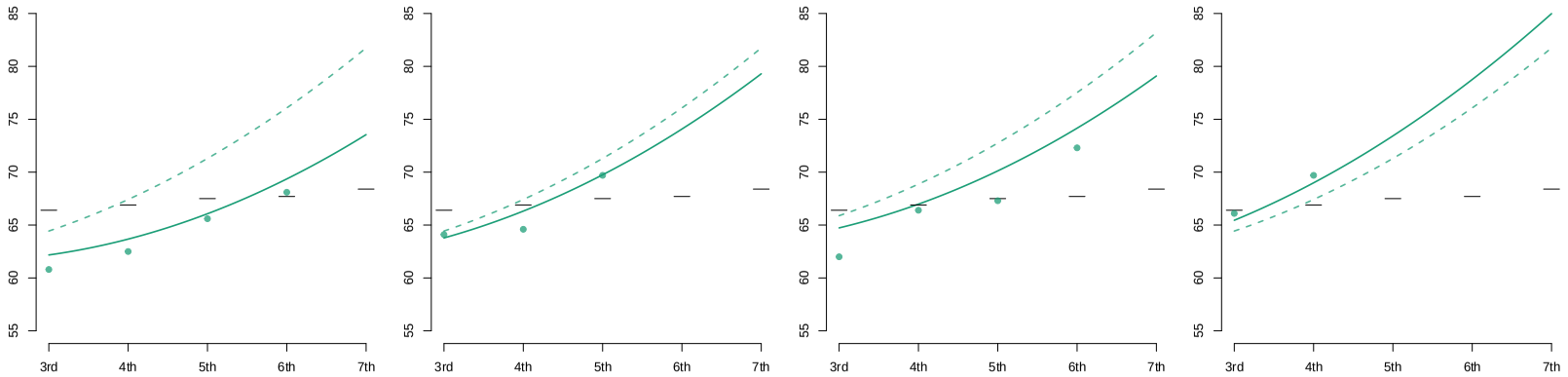
Hazard submodel with latent covariates. The hazard submodel with latent covariates shown in Table 7 estimated the probability of reclassification conditional on a students deviation from the average third grade AZELLA total score, and their deviation from the average linear growth. A student with an average third grade AZELLA total score who made average growth had a less than .001 probability of reclassifying in Grade 3, $\alpha_1 = -10.18$, 95% $HPD = [-13.14, -8.35]$. The probability increased to .45 in Grade 4, $\alpha_2 = -0.18$, 95% $HPD = [-1.09, -0.76]$. Between Grades 5 and 7, the probability of reclassification for an English learner with an average third grade AZELLA total score who grew at an average rate was greater than .95, $\alpha_3 = 2.97$, 95% $HPD = [1.85, 4.27]$; $\alpha_4 = 7.89$ 95% $HPD = [5.93, 10.08]$; $\alpha_5 = 8.80$, 95% $HPD = [5.85, 11.89]$.

Every point a student earned above the average in initial status resulted in an increased likelihood of reclassification of 2.13 logits, $\lambda_1 = 2.13$, 95% $HPD = [1.00, 3.37]$. That is, an English learner whose initial status random effect, $\zeta_{1ki}^{(2)}$, was estimated to be one —resulting in a predicted third grade AZELLA total score one point higher than average —was 8.13 times more likely to reclassify than a student with average third grade AZELLA total score. Every point above average in linear growth resulted in an increased likelihood of reclassification of 3 logits, $\lambda_2 = 3.08$, 95% $HPD = [1.56, 4.73]$. That is, an English learner whose linear growth random

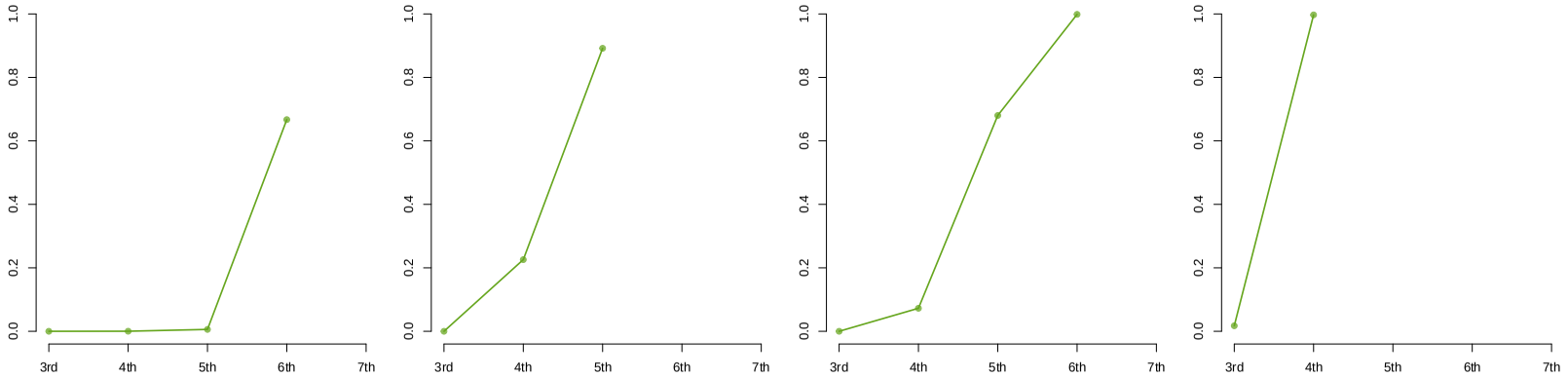
effect $\zeta_{2ki}^{(2)}$ was estimated to be one was 21.73 times more likely to reclassify than a student who was estimated to grow at an average rate.

The shared random effects model is illustrated using the submodel estimates to plot both the subject-specific total score growth trajectory and the subject-specific probability of reclassification for a single individual selected at random from the dataset. Figures 8a to 8d illustrate the average developmental trajectory (dashed line) and the subject-specific developmental trajectory (solid line) for total English proficiency for a sample of students in the dataset. Figures 8e to 8h illustrate the corresponding plot of the probability of reclassification at each grade for those students. When the growth model predicted the student to meet or exceed the benchmark at a give grade, the probability of reclassification exceeded .5. Figures 8b and 8f illustrate that as Student 96's predicted score approached the benchmark, their probability of reclassification increased. Figures 8c and 8g illustrate the situation where the growth model incorrectly predicted Student 244 to reach the benchmark in grade 5, and thus estimated their probability of reclassification at the end of Grade 5 to be greater than .6.

Using $\pi = .5$, the hazard submodel with latent covariates predicted the time-to-reclassification with 97.22% accuracy. Both the TPR and TNR were also .97 indicating that the model accurately predicted when 97% of the student would reclassify and accurately predicted when 97% of the students would not reclassify. The predictive power of the hazard submodel with latent covariates can be compared to the hazard model with manifest covariates across the range of π by assessing the ROC curves in Figure 7. The overall predictive power of the hazard submodel with latent covariates was greater than the hazard model with manifest



(a) Total Score Growth, ID=88 (b) Total Score Growth, ID=96 (c) Total Score Growth, ID=244 (d) Total Score Growth, ID=34



(e) $\mathbb{P}$(Reclassification), ID=88 (f) $\mathbb{P}$(Reclassification), ID=96 (g) $\mathbb{P}$(Reclassification), ID=244 (h) $\mathbb{P}$(Reclassification), ID=34

Figure 8. Top row: Marginal (dashed) and subject-specific (solid) AZELLA total English proficiency trajectories for a random sample of students. Bottom row: Probability of reclassification for the same random sample of students.

covariates as indicated by the area under its ROC curve (orange dashed line). The more area under the ROC curve, the better the model's predictive power.

Hazard submodel with latent and manifest covariates. The final hazard model shown in Table 7 added the two time-invariant dichotomous manifest covariates, SWDe and sex, to the hazard submodel with latent covariates. A female English learner who was never identified with a disability who had an average third grade AZELLA total score and who made average growth had a less than a .001 probability of reclassifying in Grade 3, $\alpha_1 = -8.66$, 95% $HPD = [10.95, -6.65]$. The probability increased to .79 in Grade 4, $\alpha_2 = 1.31$, 95% $HPD = [0.23, 2.47]$. Between Grades 5 and 7, the probability of reclassification for a female English learner with an average third grade AZELLA total score who made average growth was greater than .99, $\alpha_3 = 4.67$, 95% $HPD = [3.21, 6.32]$; $\alpha_4 = 10.52$, 95% $HPD = [8.15, 13.19]$; $\alpha_5 = 12.19$, 95% $HPD = [8.87, 15.72]$.

The mean of the posterior for SWDe was -8.35 logits, $\alpha_6 = -8.35$, 95% $HPD = [21.04, -13.50]$. Controlling for other covariates, the odds of reclassification at any grade for an English learner ever identified with a disability were 0.0002 times that of an English learner never identified with a disability. The mean of the posterior for sex was not different from zero, $\alpha_7 = -0.83$, 95% $HPD = [11.80, -7.38]$.

Controlling for other covariates, every point above the average in initial status resulted in an increased likelihood of reclassification was 2.14 logits, $\lambda_1 = 2.14$, 95% $HPD = [1.05, 3.32]$. That is, an English learner whose estimated initial status random effect, $\zeta^{(2)}_{1ki}$, was one —resulting in a predicted third grade AZELLA total score one point higher than average —was 8.5 times more likely to reclassify. Controlling for other covariates, every point above average in linear

growth resulted in an increased likelihood of reclassification of 3 logits, $\lambda_2 = 3.01$, 95% $HPD = [1.49, 4.86]$. That is, an English learner who was estimated to have a linear growth random effect $\zeta_{2ki}^{(2)}$ of one, was 20.33 times more likely to reclassify than a student who was estimated to grow at an average rate, controlling for other covariates. The correlation among the latent responses of any two students from the same school increased from the hazard model with manifest covariates to

$$\rho = \frac{\nu}{\nu + \pi^2/3} = \frac{0.73^2}{0.73^2 + \pi^2/3} = .14.$$

Figure 9 is a plot of the school-specific probabilities of reclassification at each grade for English learners ever or never identified with a disability who scored one standard deviation above and below average on the third grade AZELLA and who made one standard deviation above and below average linear growth. English learners who were never identified with a disability and who scored one standard deviation above average on the third grade AZELLA and whose linear growth component was one standard deviation above average growth (green solid line), on average, had a 99% change of being reclassified starting in Grade 4. English learners ever identified with a disability who scored one standard deviation above average on the third grade AZELLA and whose linear growth component was one standard deviation above average growth (green dashed line), on average, had a 30% chance of being reclassified in Grade 4, a 90% chance of being reclassified in Grade 5, and a 99% chance of being reclassified in Grades 6 and 7. English learners never identified with a disability who scored one standard deviation below average on the third grade AZELLA and whose linear growth component was one standard deviation below average growth (orange solid line), on average, had less than a 20% chance of being reclassified between Grades 3 and 5 and a 99% chance of being reclassified in Grades 6 and 7. Finally, English learners ever identified

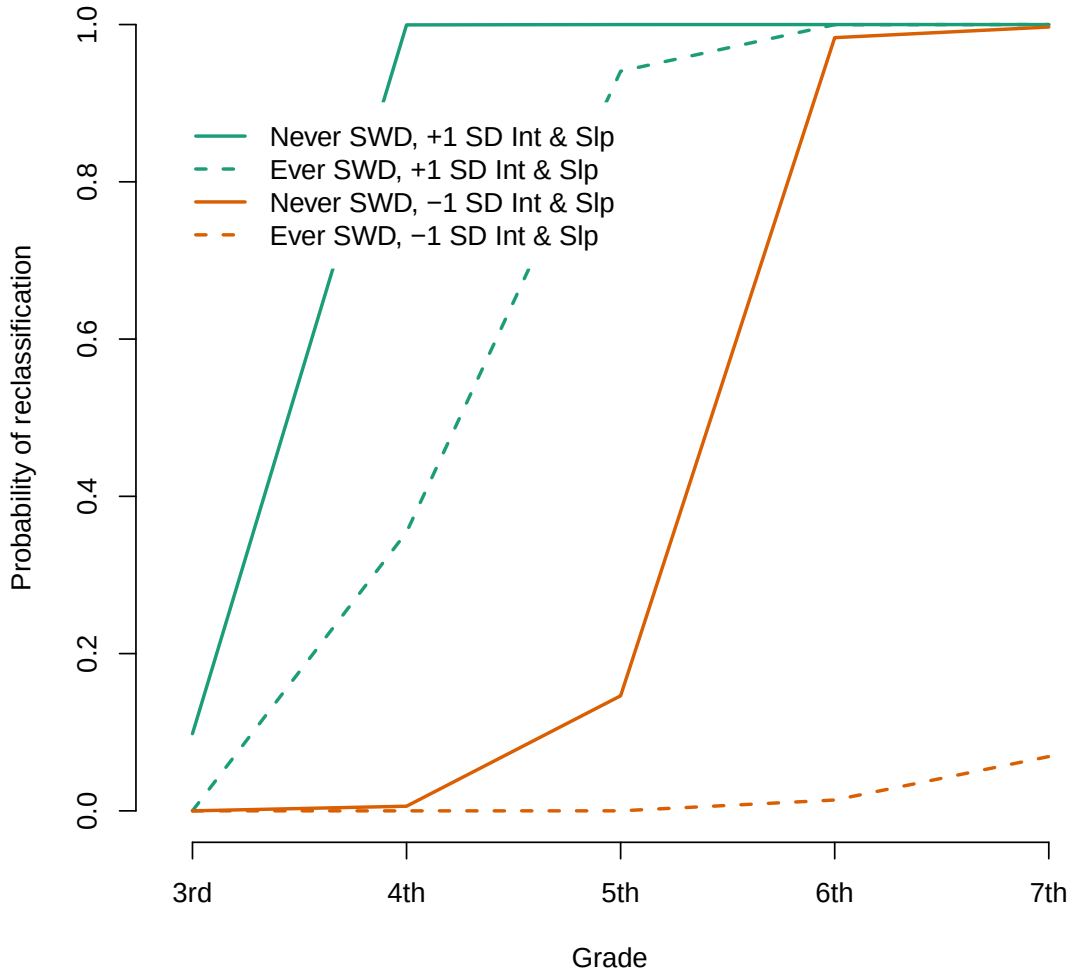


Figure 9. Probability of reclassification by subgroup for the hazard submodel with latent and manifest covariates

with a disability who scored one standard deviation below average on the third grade AZELLA and whose linear growth component was one standard deviation below average growth (orange dashed line), on average, never had more than a 15% chance of being reclassified by Grade 7.

Using the parameters to predict time-to-reclassification, the hazard submodel with latent and manifest covariates predicted the time to reclassification for the students in the sample with 98.47% accuracy. The TPR was .99 and the TNR was .98. While the ROC curve for the hazard submodel with latent and manifest covariates shown in Figure 7 strongly mirrors the ROC curve of the hazard model with latent covariates, it captures slightly more area indicating it is a better predictive model.

CHAPTER IV

DISCUSSION

The primary purpose of this dissertation was to propose a series of statistical models for studying two measures of English learner proficiency development. First, a multilevel, multivariate mixed model was developed to illustrate how each domain of academic English proficiency developed over time, and how development between language domains were correlated. Second, a multilevel shared random effects model was developed to illustrate how heterogeneity in academic English proficiency development contributed to prediction of reclassification from English learner to reclassified fluent English proficient. Secondary purposes of the dissertation included describing how academic English proficiency developed between Grade 3 through Grade 7 for a single cohort of English learners in an Arizona school district, how English proficiency development differed based on ever or never being identified with a disability and sex, as well as characterizing the degree to which academic English proficiency development, disability status, and sex predicted time to reclassification for those students.

This research contributed to the English learner literature in several ways. First, only a few studies have focused on the functional form of academic English development (e.g., Cook et al., 2008; Slama, 2012). Of these existing studies, none, to my knowledge, examined the association between language domains over time. By modeling academic English proficiency development at the domain level as a multivariate process, a more nuanced understanding of the developmental process than using total scores alone was provided.

Second, prior research on time-to-reclassification has studied the policy implementation without controlling for English proficiency development (Slama, 2014; Thompson, 2015; Umansky & Reardon, 2014), even though reclassification decisions are known to be based largely on measures of one's English proficiency (Ramsey & O'Day, 2010). Explicitly incorporating English proficiency into a statistical model, however, requires a model specification that incorporates its endogenous properties (Kalbfleisch & Prentice, 2011). This study did so by jointly modeling the longitudinal English proficiency development process and reclassification process simultaneously. The joint modeling solution acknowledged that the initial status and development of English proficiency varied across students, and the variation in that process was associated with time to reclassification.

Substantive Findings

Describing English proficiency development. The results of the univariate analyses suggested that, on average, this sample of English learners saw their overall academic English proficiency accelerate between third grade and seventh grade. Univariate and multivariate analyses of the subtest scores indicated that this quadratic functional form was not the same across language domains, however. Although average growth for the writing and oral subdomains accelerated over time, a linear model was found to fit best for reading domain development. Furthermore, the quadratic component for oral English development was more than twice as large as the quadratic component for writing development, suggesting that, on average, English learners would meet the oral English proficiency benchmark before meeting the writing benchmark.

Both the univariate analysis and multivariate analysis of sub-test scores provided evidence that ever being diagnosed with a disability was negatively associated with third grade reading, writing, and oral proficiency scores. There was also evidence that English learners who had ever been diagnosed with a disability grew more slowly than English learners who were never diagnosed with a disability for the writing domain, but there was no such evidence that this was the case for reading and oral English proficiency. One way to interpret this finding is that writing, more than reading or oral proficiency, is the domain where those English learners identified with disabilities fall farthest behind. However, not knowing specific student exceptionality classifications limits the interpretability of this finding. That is, this particular sample may happen to have more English learners with exceptionalities that impact writing over other language domains. Furthermore, male English learners were expected to score lower than female English learners in third grade scores for all three sub-tests.

The multivariate analysis provided estimates of how the development of a given domain correlated with development in another domain. There was a large positive association between third grade English proficiency for each domain. That is, if a student scored higher than average on the reading sub-test, they were also likely to score higher than average on both the writing and oral domain sub-tests in Grade 3. The only association found among the linear growth components of the three language domains was among reading and oral proficiency development. This is not to say other associations did not exist for these students. The relative sparseness of within-student information due to high proportions of reclassification in the early grades could have been one reason for the imprecision in the random effects variance-covariance matrix.

A subject-specific domain-level analysis based on a multivariate random effects model has the potential to provide states and districts with additional information to evaluate reclassification criteria. For example, Arizona used only the total score for reclassification. As Figure 5 indicated, however, some students had been reclassified to fluent English proficient without reaching proficiency in each domain. This finding suggests that some students may be reclassified prior to be fully proficient. Such an analysis can also be used to evaluate more stringent reclassification criteria. For example, California requires, among other criteria, that English learners reach proficiency in all domains to be reclassified. Subject-specific trajectories may be used to evaluate the probability that a specific student will meet the proficiency benchmark between the time the test was taken and the end of the year. Such an analysis can indicate if too many students are being retained as English learners when they would likely be fluent English proficient and could be reclassified sooner.

Predicting time-to-reclassification. The final discrete-time hazard model predicted time to reclassification based on one's ever being diagnosed with a disability and sex, and was able to predict the time of a students reclassification with 80% accuracy. The shared random effects model that used one's deviation from the average third grade English proficiency scores and their deviation from average English proficiency growth was able to predict the time of a students reclassification with 97% accuracy. The predictive accuracy of the final shared random effects model with disability status, sex and random effects as predictors improved classification accuracy to 98%. While the predictive accuracy of the discrete-time hazard model met most conventional criteria for diagnostic models

(Steyerberg et al., 2010), using the information from one's English proficiency was exceptionally accurate in this sample.

This finding suggests that for this particular cohort and district, the district adhered closely to Arizona's reclassification policies that required a student to be reclassified when his or her total English proficiency score reached the benchmark. This particular model, using total English proficiency scores, may need to be altered for those states that use a wider range of reclassification criteria. Furthermore, the predictive accuracy would likely be lower for districts that do not adhere as closely to the reclassification policies, or for states that make reclassification decisions based on additional criteria (e.g., teacher recommendations).

Although this study used the shared random effect models for prediction, the statistical framework is also useful for descriptive and inferential analyses of reclassification. To model time to reclassification using survival analysis without including students' English proficiency development ignores the direct connection between English proficiency development and reclassification. Not controlling for English proficiency in a time-to-reclassification model suggests that any two students who are equal on all covariates being controlled for in the model are equally likely to be reclassified at time t, regardless of their unique English proficiency development. Furthermore, not controlling for English proficiency can bias the estimates of other covariates that may be of interest. For example, in this study, the SWDe parameter from the hazard model with manifest covariates differed in size and interpretation from the SWDe parameter from the hazard submodel that also included latent covariates.

Limitations

The first set of limitations for this study pertained to the research design. The primary limitation is that the data used in the analysis were from a single cohort from a single district in one state. Because Arizona's English learner policies are unique, these findings may not be generalizable to other locales, testing systems, or policy contexts. Furthermore, the data came from a single Arizona district and is unlikely to be representative of other districts or the entire state. With only a single cohort used in the analysis, the results presented here may be subject to differences from one cohort to another. Finally, because limited testing occurs after a student reclassifies, alternative functional forms for language development based on the time-of-reclassification could not be adequately assessed.

A related limitation is that this research spanned Grades 3 through 7, but all prior research on reclassification had used Kindergarten as the first measurement occasion. Had this research extended back to Kindergarten, there would have likely been greater within-student information that could have warranted the application of a broader range of growth models. Furthermore, this study used a strict cohort design where those students who entered the school district after Grade 3 were excluded. This decision was made because late-arriving students lacked information about prior schooling which could have biased hazard model parameter estimates (Guo, 1993). These decisions limit interpretation of parameter estimates in the current study —rather than applying to all English learners in the district, study findings do not apply to those English learners who reclassify prior to, or enter the district after Grade 3.

Yet another study limitation pertains to the measures of English language proficiency analyzed. It is important to note that all analyses used observed

AZELLA scores, which contained error. Modeling the observed score results in a residual that contains both measurement error and the deviation in the true score from the fitted trend line (Skrondal & Rabe-Hesketh, 2004). Furthermore, this study used oral English proficiency rather than separate measures for speaking and listening. Using the speaking and listening measures in the multivariate growth model would have provided a more complete description of academic English development.

The relatively small number of schools, and a small number of students within some schools, also limited the precision of the school-level variance components. That said, a noisy estimate, particularly within a Bayesian framework, results in uncertainty, not bias. Regarding variance components, the residual variances for each submodel of the multivariate mixed effect models were assumed independent. Future research could test this independence assumption by specifying more complex variance-covariance structures for the residuals.

Another limitation, common in most educational research, surrounds the reliability and validity of the free/reduced lunch (FRL) and the student with disabilities status (SWD) variables. The FRL covariate is a dichotomous proxy for the more complex construct of economic status. Furthermore, the actual indicator is time-varying in that a student may qualify one year and not the next. The time-invariant version used in this study, where any student who ever received free/reduced lunch is defined as FRL assumed that those students who received the benefit in any year are representative of the FRL group overall.

A similar approach was used for students who ever received services due to their disability status. Like FRL, this assumes students who ever received services are more like those who had received services throughout their schooling. This

indicator also did not indicate that those who are categorized as SWD in the district testing system actually receive services for that disability.

Finally, limitations of this research include the possibility that the assumption pertaining to the missing data mechanism were incorrect, and the potential omission of other variables that may be related to proficiency development and reclassification. Regarding missing data, subsequent analysis should assess the sensitivity of the parameter estimates to competing missing data assumptions (Xu & Blozis, 2011). As for potential omitted variables, generational status and information about the instructional program and its implementation are two in particular. Slama (2012) found that the average English development of foreign-born English learners differed from US-born English learners. As for program information, Hakuta (2011) pointed out, "[t]here are well implemented and poorly implemented programs..." (p.166). While all students in this sample were enrolled in an SEI program, the implementation of the program likely varied from one school to another which may have impacted rates of English proficiency development. Exploration of program instructional design and fidelity of program implementation are important issues for future study.

Conclusions and Future Directions

Understanding how academic English proficiency develops and its role in time-to-reclassification is critical for the development of effective policies and programs for the improvement of outcomes for English learners. This study fit a multivariate growth model to estimate the association between reading, writing and oral English proficiency development for a single cohort of students in one Arizona school district. It also fit a shared random effects model to estimate the association between academic English development and time to reclassification. Modeling

academic English language development and reclassification simultaneously enabled a more nuanced understanding of the time required to attain English proficiency. While the data used in this study cannot support causal inferences, nor can the results be broadly generalized, the models presented in this dissertation provide an analytical foundation for future research aimed at prediction, description, and testing the development of academic English proficiency and reclassification.

An avenue for future research is to extend and assess the utility of the shared random effect model for states that use multiple reclassification criteria. Future research can also focus on the role of school-level variation in academic English development and reclassification. For example, incorporating fidelity measures at the classroom/school level may better estimate how program quality associates with academic English development and time-to-reclassification. Additionally, future research could extend the multilevel shared parameter model to test how school-level random effects for academic English development impact time-to-reclassification.

A new requirement in the recent ESSA legislation requires schools to report disaggregated data for English learners who are also students with disabilities. Findings from this dissertation indicate that this specific subgroup may be particularly challenged when it comes to attaining academic English proficiency. Future research should continue to analyze the English proficiency development for English learners with disabilities, and when possible, by specific exceptionality category.

This study demonstrated that the inclusion of academic English proficiency in the survival process resulted in improved predictions of time to reclassification. There has been much work related to SREM for dynamic participant-specific

predictions in the medical literature (Fieuws et al., 2008; Rizopoulos, 2011; Rizopoulos, Hatfield, Carlin, & Takkenberg, 2014; Yu, Taylor, & Sandler, 2008). These extensions can be used to develop an early warning system to identify those students at risk of becoming long-term English learners and which students may be in need of additional language support. Such a system would provide practitioners with information to aid decisions regarding resource allocation for English learners.

Finally, within the English learner research there has been much interest in understanding the quality of reclassification criteria (e.g., Robinson, 2011). A joint modeling approach could be adopted to acknowledge that not all English learners reach the reclassification threshold in the same way, and how they get there may provide valuable information for evaluating reclassification criteria. To that end, a joint modeling approach could also be used to better understand which students succeed and which struggle academically after reclassification.

The dynamic nature of English learner populations and the complex policy surrounding their academic achievement requires researchers to utilize more complex analytical frameworks. Joint models are still in their relative infancy and this study sought to illustrate their use for studying English language development and reclassification. With continued research into their strengths and limitations, such models may provide key insights into those policies and programs that help English learners succeed.

APPENDIX A

MODEL SPECIFICATION

This section provides the details of the models estimated to answer Research Questions 1, 2 and 3. I first review those models used to answer Research Question 1, followed by the models used for Research Question 2, and conclude with the models used for Research Question 3.

Univariate AZELLA Models

Research Question 1 fit a series of mixed models, some that estimate only student-specific random effects and others that estimate school- and student-specific random effects. For the student-specific random effects models, let there be $i = 1, 2, \ldots, n$ students and $j = 1, 2, \ldots, n_j$ longitudinal measures for student j. Define $\boldsymbol{y}_j$ as a $n_j \times 1$ vector containing the longitudinal measures for student j. I can then specify the mixed model generally using the notation of Skrondal and Rabe-Hesketh (2004),

$$\boldsymbol{y}_j = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\boldsymbol{\zeta}_j + \boldsymbol{\epsilon}_j \tag{A.1}$$

where $\epsilon_{ij} \sim \mathcal{N}(0, \sigma)$, $\boldsymbol{\zeta}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{T})$ and

$$\mathbf{T} = \begin{bmatrix} \tau_{11} & \tau_{12} & \cdots & \tau_{1q} \\ \tau_{21} & \tau_{22} & \cdots & \tau_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \tau_{q1} & \tau_{q2} & \cdots & \tau_{qq} \end{bmatrix}. \tag{A.2}$$

For Equation A.1, $\mathbf{X}_j$ is a known $n_j \times p$ design matrix corresponding to the $p \times 1$ vector of fixed effects $\boldsymbol{\beta}$, $\mathbf{Z}_j$ is a known $n_j \times q$ design matrix corresponding to the $q \times 1$ vector of random effects, $\boldsymbol{\zeta}_j$, and $\boldsymbol{\epsilon}_j$ is an $n_j \times 1$ vector of residuals.

Extending Equation A.1 to include school-specific random effects requires extended notation. Let there be $k = 1, 2, \ldots, K$ schools, $j = 1, 2, \ldots, n_k$ students

in school k, and $i = 1, 2, \ldots, n_{jk}$ longitudinal measures for student j in school k. Further, let $N_k = \sum_{j=1}^{n_k} n_{jk}$ be the total number of measures in school k. Define $\boldsymbol{y}_{jk}$ as a $N_k \times 1$ vector containing the longitudinal measures for student j in school k and $\boldsymbol{y}_k \equiv \left[\boldsymbol{y}'_{1k}, \ldots, \boldsymbol{y}'_{n_k k}\right]'$ as an $N_k \times 1$ vector containing all observations in school k. I can then specify the three-level mixed model for school k generally as:

$$\boldsymbol{y}_k = \mathbf{X}_k\boldsymbol{\beta} + \mathbf{Z}_k\boldsymbol{\zeta}_k + \boldsymbol{\epsilon}_k \tag{A.3}$$

where $\mathbf{X}_k$ is a known $N_k \times p$ design matrix corresponding to the $p \times 1$ vector of fixed effects $\boldsymbol{\beta}$. Because this is a so-called three-level model, $\mathbf{Z}_k \equiv \left[\mathbf{Z}_k^{(3)},\ \text{diag}\left[\mathbf{Z}_{1k}^{(2)}, \ldots, \mathbf{Z}_{n_k k}^{(2)}\right]\right]$ and $\boldsymbol{\zeta}_k \equiv \left[\boldsymbol{\zeta}_k^{(3)\prime}, \boldsymbol{\zeta}_{1k}^{(2)\prime}, \ldots, \boldsymbol{\zeta}_{n_k k}^{(2)\prime}\right]'$, where $\mathbf{Z}_k^{(3)}$ is a known $N_k \times q^{(3)}$ design matrix corresponding to the $q^{(3)} \times 1$ vector of school-specific random effects, $\boldsymbol{\zeta}_k^{(3)}$, and $\mathbf{Z}_{jk}^{(2)}$ is a known $n_{jk} \times q^{(2)}$ design matrix corresponding to the $q^{(2)} \times 1$ vector of student-specific random effects, $\boldsymbol{\zeta}_{jk}^{(2)}$. For clarity, diag $\left[\mathbf{Z}_{k1}^{(2)}, \ldots, \mathbf{Z}_{n_k k}^{(2)}\right]$ is a matrix of size $N_k \times (n_k\, q^{(2)})$. This notation is essentially equivalent to the 3-level mixed model notation presented by Hedeker and Gibbons (2006). Furthermore, we assume $\boldsymbol{\epsilon}_{jk} \sim \mathcal{N}(0, \sigma\mathbf{I})$, $\boldsymbol{\zeta}_{jk}^{(2)} \sim \mathcal{N}(\mathbf{0},\, \mathbf{T}^{(2)})$, and $\boldsymbol{\zeta}_k^{(3)} \sim \mathcal{N}(\mathbf{0},\, \mathbf{T}^{(3)})$, where

$$\mathbf{T}^{(2)} = \begin{bmatrix} \tau_{11}^{(2)} & \tau_{12}^{(2)} & \cdots & \tau_{1q^{(2)}}^{(2)} \\ \tau_{21}^{(2)} & \tau_{22}^{(2)} & \cdots & \tau_{2q^{(2)}}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ \tau_{q^{(2)}1}^{(2)} & \tau_{q^{(2)}2}^{(2)} & \cdots & \tau_{q^{(2)}q^{(2)}}^{(2)} \end{bmatrix} \text{ and } \mathbf{T}^{(3)} = \begin{bmatrix} \tau_{11}^{(3)} & \tau_{12}^{(3)} & \cdots & \tau_{1q^{(3)}}^{(3)} \\ \tau_{21}^{(3)} & \tau_{22}^{(3)} & \cdots & \tau_{2q^{(3)}}^{(3)} \\ \vdots & \vdots & \ddots & \vdots \\ \tau_{q^{(3)}1}^{(3)} & \tau_{q^{(3)}2}^{(3)} & \cdots & \tau_{q^{(3)}q^{(3)}}^{(3)} \end{bmatrix}. \tag{A.4}$$

For subsequent notation, the two level model is notated using $\mathbf{Z}_j^{(2)}$ and $\boldsymbol{\zeta}_j^{(2)}$ such that Equation A.1 becomes

$$\boldsymbol{y}_j = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j^{(2)}\boldsymbol{\zeta}_j^{(2)} + \boldsymbol{\epsilon}_j \tag{A.5}$$

For the prior specification, let $\mathbf{T}$ be either $\mathbf{T}^{(2)}$ or $\mathbf{T}^{(3)}$ and $\boldsymbol{\zeta}_j$ be the corresponding vector of random effects. Define diag$(\mathbf{T}) \equiv \mathbf{D}$ where $\mathbf{D}$ is a

vector of standard deviations for the student-specific random effects. Next, let $\mathbf{R} = \mathbf{D}^{-1}\mathbf{T}\left(\mathbf{D}^{-1}\right)'$ be the correlation matrix associated with the variance-covariance matrix $\mathbf{T}$. Lastly, let $[\text{diag}(\mathbf{D})\mathbf{S}']\,\boldsymbol{z}_i = \boldsymbol{\zeta}_i$ where $\mathbf{S}'$ is the Cholesky decomposition of the correlation matrix $\mathbf{R}$. By doing this, I was able to specify priors for the specified priors for the uncorrelated standard normal random effects, $\boldsymbol{z}_i$, standard deviations, $\mathbf{D}$, and correlation matrix $\mathbf{R}$ rather than the variance covariance matrix $\mathbf{T}$ directly.

AZELLA total score models.

AZELLA total score model 1.

$$y_{Tij} = \beta_{T1} + \zeta^{(2)}_{Tj1} + \left(\beta_{T2} + \zeta^{(2)}_{Tj2}\right) \text{GRADE}_{ij} + \epsilon_{Tij}$$

$$\begin{aligned}
&p(\mathbf{D}) \sim \mathcal{U}(0,\, \infty),\\
&p(\mathbf{R}) \sim \mathcal{LKJ}(1.5),\\
&p(\beta_{T1}) \sim \mathcal{U}(-\infty,\, \infty), \;\; p(\beta_{T2}) \sim \mathcal{U}(-\infty,\, \infty),\\
&p(\boldsymbol{z}^{(2)}_j) \sim \mathcal{N}(\mathbf{0},\, \mathbf{I}),\\
&p(\sigma_{Tijk}) \sim \mathcal{U}(0,\, \infty).
\end{aligned}$$

where $\mathcal{U}(a, b)$ is a uniform distribution over the support $x \in [a,\, b]$, $\mathcal{LKJ}(\eta)$ is an LKJ correlation distribution where $\eta \in [0, \infty]$, $\mathcal{N}(\mu, \sigma)$ is a normal distribution where $\mu \in \mathbb{R}$ and $\sigma^2 \in (0, \infty]$, and $\mathbf{I}$ is the identity matrix.

AZELLA total score model 2.

$$y_{Tij} = \beta_{T1} + \zeta^{(2)}_{Tj1} + \left(\beta_{T2} + \zeta^{(2)}_{Tj2}\right) \text{GRADE}_{ij} + \beta_{T3} \text{GRADE}^2_{ij} + \epsilon_{Tij}$$

$$p(\mathbf{D}) \sim \mathcal{U}(0, \infty),$$

$$p(\mathbf{R}) \sim \mathcal{LKJ}(1.5),$$

$$p(\beta_{T1}) \sim \mathcal{U}(-\infty, \infty), \;\; p(\beta_{T2}) \sim \mathcal{U}(-\infty, \infty), \;\; p(\beta_{T3}) \sim \mathcal{U}(-\infty, \infty),$$

$$p(\boldsymbol{z}^{(2)}_j) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

$$p(\sigma_{Tijk}) \sim \mathcal{U}(0, \infty).$$

AZELLA total score model 3.

$$y_{Tijk} = \beta_{T1} + \zeta^{(2)}_{Tjk1} + \zeta^{(3)}_{Tk1} + \left(\beta_{T2} + \zeta^{(2)}_{Tjk2} + \zeta^{(3)}_{Tk2}\right) \text{GRADE}_{ijk} + \beta_{T3} \text{GRADE}^2_{ijk} + \epsilon_{Tijk}$$

$$p(\mathbf{D}^{(2)}) \sim \mathcal{U}(0, \infty), \;\; p(\mathbf{D}^{(3)}) \sim \mathcal{U}(0, \infty),$$

$$p(\mathbf{R}^{(2)}) \sim \mathcal{LKJ}(1.5), \;\; p(\mathbf{R}^{(3)}) \sim \mathcal{LKJ}(1.5),$$

$$p(\beta_{T1}) \sim \mathcal{U}(-\infty, \infty), \;\; p(\beta_{T2}) \sim \mathcal{U}(-\infty, \infty), \;\; p(\beta_{T3}) \sim \mathcal{U}(-\infty, \infty),$$

$$p(\boldsymbol{z}^{(2)}_{jk}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \;\; p(\boldsymbol{z}^{(3)}_{k}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

$$p(\sigma_{Tijk}) \sim \mathcal{U}(0, \infty).$$

AZELLA total score model 4.

$$y_{Tijk} = \beta_{T1} + \zeta^{(2)}_{Tjk1} + \zeta^{(3)}_{Tk1} + \beta_{T4}\,\mathrm{SWDe}_{jk} + \beta_{T6}\,\mathrm{MALE}_{jk} + \left(\beta_{T2} + \zeta^{(2)}_{Tjk2} + \zeta^{(3)}_{Tk2} + \beta_{T5}\,\mathrm{SWDe}_{jk}\right)\,\mathrm{GRADE}_{ijk} + \beta_{T3}\,\mathrm{GRADE}^2_{ijk} + \epsilon_{Tijk}$$

$$p(\mathbf{D}^{(2)}) \sim \mathcal{U}(0,\,\infty),\;\; p(\mathbf{D}^{(3)}) \sim \mathcal{U}(0,\,\infty),$$

$$p(\mathbf{R}^{(2)}) \sim \mathcal{LKJ}(1.5),\;\; p(\mathbf{R}^{(3)}) \sim \mathcal{LKJ}(1.5),$$

$$p(\beta_{T1}) \sim \mathcal{U}(-\infty,\,\infty),\;\; p(\beta_{T2}) \sim \mathcal{U}(-\infty,\,\infty),\;\; p(\beta_{T3}) \sim \mathcal{U}(-\infty,\,\infty),$$

$$p(\beta_{T4}) \sim \mathcal{U}(-\infty,\,\infty),\;\; p(\beta_{T5}) \sim \mathcal{U}(-\infty,\,\infty),\;\; p(\beta_{T6}) \sim \mathcal{U}(-\infty,\,\infty),$$

$$p(\boldsymbol{z}^{(2)}_{jk}) \sim \mathcal{N}(\mathbf{0},\,\mathbf{I}),\;\; p(\boldsymbol{z}^{(3)}_{k}) \sim \mathcal{N}(\mathbf{0},\,\mathbf{I}),$$

$$p(\sigma_{Tijk}) \sim \mathcal{U}(0,\,\infty).$$

AZELLA total score model 5.

$$y_{Tijk} = \beta_{T1} + \zeta^{(2)}_{Tjk1} + \zeta^{(3)}_{Tk1} + \beta_{T4}\,\mathrm{SWDe}_{jk} + \beta_{T6}\,\mathrm{MALE}_{jk} + \left(\beta_{T2} + \zeta^{(2)}_{Tjk2} + \zeta^{(3)}_{Tk2} + \beta_{T5}\,\mathrm{SWDe}_{jk}\right)\,\mathrm{GRADE}_{ijk} + \beta_{T3}\,\mathrm{GRADE}^2_{ijk} + \epsilon_{Tijk}$$

$$p(\mathbf{D}^{(2)}) \sim \mathcal{U}(0,\,\infty),\;\; p(\mathbf{D}^{(3)}) \sim \mathcal{U}(0,\,\infty),$$

$$p(\mathbf{R}^{(2)}) \sim \mathcal{LKJ}(1.5),\;\; p(\mathbf{R}^{(3)}) \sim \mathcal{LKJ}(1.5),$$

$$p(\beta_{T1}) \sim \mathcal{N}(60,\,10),\;\; p(\beta_{T2}) \sim \mathcal{N}(3,\,6),\;\; p(\beta_{T3}) \sim \mathcal{N}(0,\,5),$$

$$p(\beta_{T4}) \sim \mathcal{N}(0,\,2.5),\;\; p(\beta_{T5}) \sim \mathcal{N}(0,\,5),\;\; p(\beta_{T6}) \sim \mathcal{N}(0,\,2.5),$$

$$p(\boldsymbol{z}^{(2)}_{jk}) \sim \mathcal{N}(\mathbf{0},\,\mathbf{I}),\;\; p(\boldsymbol{z}^{(3)}_{k}) \sim \mathcal{N}(\mathbf{0},\,\mathbf{I}),$$

$$p(\sigma_{Tijk}) \sim \mathcal{U}(0,\,\infty).$$

AZELLA reading score models.

AZELLA reading score model 1.

$$y_{Rij} = \beta_{R1} + \zeta^{(2)}_{Rj1} + \left(\beta_{T2} + \zeta^{(2)}_{Rj2}\right) \text{GRADE}_{ij} + \epsilon_{Rij}$$

$$p(\mathbf{D}) \sim \mathcal{U}(0, \infty),$$

$$p(\mathbf{R}) \sim \mathcal{LKJ}(1.5),$$

$$p(\beta_{R1}) \sim \mathcal{U}(-\infty, \infty), \;\; p(\beta_{R2}) \sim \mathcal{U}(-\infty, \infty),$$

$$p(\boldsymbol{z}^{(2)}_j) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

$$p(\sigma_{Rij}) \sim \mathcal{U}(0, \infty).$$

AZELLA reading score model 2.

$$y_{Rij} = \beta_{R1} + \zeta^{(2)}_{Rj1} + \left(\beta_{R2} + \zeta^{(2)}_{Rj2}\right) \text{GRADE}_{ij} + \beta_{R3} \text{GRADE}^2_{ij} + \epsilon_{Rij}$$

$$p(\mathbf{D}) \sim \mathcal{U}(0, \infty),$$

$$p(\mathbf{R}) \sim \mathcal{LKJ}(1.5),$$

$$p(\beta_{R1}) \sim \mathcal{U}(-\infty, \infty), \;\; p(\beta_{R2}) \sim \mathcal{U}(-\infty, \infty), \;\; p(\beta_{R3}) \sim \mathcal{U}(-\infty, \infty),$$

$$p(\boldsymbol{z}^{(2)}_j) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

$$p(\sigma_{Rij}) \sim \mathcal{U}(0, \infty).$$

AZELLA reading score model 3.

$$y_{Rijk} = \beta_{R1} + \zeta^{(2)}_{Rjk1} + \zeta^{(3)}_{R1k} + \left(\beta_{R2} + \zeta^{(2)}_{Rjk2} + \zeta^{(3)}_{Rk2}\right) \mathrm{GRADE}_{ijk} + \epsilon_{Rijk}$$

$$p(\mathbf{D}^{(2)}) \sim \mathcal{U}(0, \infty), \;\; p(\mathbf{D}^{(3)}) \sim \mathcal{U}(0, \infty),$$

$$p(\mathbf{R}^{(2)}) \sim \mathcal{LKJ}(1.5), \;\; p(\mathbf{R}^{(3)}) \sim \mathcal{LKJ}(1.5),$$

$$p(\beta_{R1}) \sim \mathcal{U}(-\infty, \infty), \;\; p(\beta_{R2}) \sim \mathcal{U}(-\infty, \infty), \;\; p(\beta_{R3}) \sim \mathcal{U}(-\infty, \infty),$$

$$p(\boldsymbol{z}^{(2)}_{jk}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \;\; p(\boldsymbol{z}^{(3)}_{k}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

$$p(\sigma_{Rijk}) \sim \mathcal{U}(0, \infty).$$

AZELLA reading score model 4.

$$y_{Rij} = \beta_{R1} + \zeta^{(2)}_{Rjk1} + \zeta^{(3)}_{Rk1} + \beta_{R4}\,\mathrm{SWDe}_{jk} + \beta_{R6}\,\mathrm{MALE}_{jk} + \left(\beta_{R2} + \zeta^{(2)}_{Rjk2} + \zeta^{(3)}_{Rk2} + \beta_{R5}\,\mathrm{SWDe}_{jk}\right) \mathrm{GRADE}_{ijk} + \epsilon_{Rijk}$$

$$p(\mathbf{D}^{(2)}) \sim \mathcal{U}(0, \infty), \;\; p(\mathbf{D}^{(3)}) \sim \mathcal{U}(0, \infty),$$

$$p(\mathbf{R}^{(2)}) \sim \mathcal{LKJ}(1.5), \;\; p(\mathbf{R}^{(3)}) \sim \mathcal{LKJ}(1.5),$$

$$p(\beta_{R1}) \sim \mathcal{U}(-\infty, \infty), \;\; p(\beta_{R2}) \sim \mathcal{U}(-\infty, \infty), \;\; p(\beta_{R3}) \sim \mathcal{U}(-\infty, \infty),$$

$$p(\beta_{R4}) \sim \mathcal{U}(-\infty, \infty), \;\; p(\beta_{R5}) \sim \mathcal{U}(-\infty, \infty), \;\; p(\beta_{R6}) \sim \mathcal{U}(-\infty, \infty),$$

$$p(\boldsymbol{z}^{(2)}_{jk}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \;\; p(\boldsymbol{z}^{(3)}_{k}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

$$p(\sigma_{Rijk}) \sim \mathcal{U}(0, \infty).$$

AZELLA reading score model 5.

$$y_{Rij} = \beta_{R1} + \zeta^{(2)}_{Rjk1} + \zeta^{(3)}_{Rk1} + \beta_{R4}\,\mathrm{SWDe}_{jk} + \beta_{R6}\,\mathrm{MALE}_{jk} + \left(\beta_{R2} + \zeta^{(2)}_{Rjk2} + \zeta^{(3)}_{Rk2} + \beta_{R5}\,\mathrm{SWDe}_{jk}\right)\,\mathrm{GRADE}_{ijk} + \epsilon_{Rijk}$$

$$p(\mathbf{D}^{(2)}) \sim \mathcal{U}(0,\,\infty),\;\; p(\mathbf{D}^{(3)}) \sim \mathcal{U}(0,\,\infty),$$

$$p(\mathbf{R}^{(2)}) \sim \mathcal{LKJ}(1.5),\;\; p(\mathbf{R}^{(3)}) \sim \mathcal{LKJ}(1.5),$$

$$p(\beta_{R1}) \sim \mathcal{N}(60,\,10),\;\; p(\beta_{R2}) \sim \mathcal{N}(3,\,6),\;\; p(\beta_{R3}) \sim \mathcal{N}(0,\,5),$$

$$p(\beta_{R4}) \sim \mathcal{N}(0,\,2.5),\;\; p(\beta_{R5}) \sim \mathcal{N}(0,\,5),\;\; p(\beta_{R6}) \sim \mathcal{N}(0,\,2.5),$$

$$p(\boldsymbol{z}^{(2)}_{jk}) \sim \mathcal{N}(\mathbf{0},\,\mathbf{I}),\;\; p(\boldsymbol{z}^{(3)}_{k}) \sim \mathcal{N}(\mathbf{0},\,\mathbf{I}),$$

$$p(\sigma_{Rijk}) \sim \mathcal{U}(0,\,\infty).$$

AZELLA writing score models.

AZELLA writing score model 1.

$$y_{Wij} = \beta_{W1} + \zeta^{(2)}_{Wj1} + \left(\beta_{W2} + \zeta^{(2)}_{Wj2}\right)\,\mathrm{GRADE}_{ij} + \epsilon_{Wij}$$

$$p(\mathbf{D}) \sim \mathcal{U}(0,\,\infty),$$

$$p(\mathbf{R}) \sim \mathcal{LKJ}(1.5),$$

$$p(\beta_{W1}) \sim \mathcal{U}(-\infty,\,\infty),\;\; p(\beta_{W2}) \sim \mathcal{U}(-\infty,\,\infty),$$

$$p(\boldsymbol{z}^{(2)}_{j}) \sim \mathcal{N}(\mathbf{0},\,\mathbf{I}),$$

$$p(\sigma_{Wij}) \sim \mathcal{U}(0,\,\infty).$$

AZELLA writing score model 2.

$$y_{Wij} = \beta_{W1} + \zeta^{(2)}_{Wj1} + \left(\beta_{W2} + \zeta^{(2)}_{Wj2}\right) \text{GRADE}_{ij} + \beta_{W3} \text{GRADE}^2_{ij} + \epsilon_{Wij}$$

$$p(\mathbf{D}) \sim \mathcal{U}(0, \infty),$$

$$p(\mathbf{R}) \sim \mathcal{LKJ}(1.5),$$

$$p(\beta_{W1}) \sim \mathcal{U}(-\infty, \infty), \;\; p(\beta_{W2}) \sim \mathcal{U}(-\infty, \infty), \;\; p(\beta_{W3}) \sim \mathcal{U}(-\infty, \infty),$$

$$p(\boldsymbol{z}^{(2)}_j) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

$$p(\sigma_{Wij}) \sim \mathcal{U}(0, \infty).$$

AZELLA writing score model 3.

$$y_{Wijk} = \beta_{W1} + \zeta^{(2)}_{Wjk1} + \zeta^{(3)}_{Wk1} + \left(\beta_{W2} + \zeta^{(2)}_{Wjk2} + \zeta^{(3)}_{Wk2}\right) \text{GRADE}_{ijk} + \beta_{W3} \text{GRADE}^2_{ijk} + \epsilon_{Wijk}$$

$$p(\mathbf{D}^{(2)}) \sim \mathcal{U}(0, \infty), \;\; p(\mathbf{D}^{(3)}) \sim \mathcal{U}(0, \infty),$$

$$p(\mathbf{R}^{(2)}) \sim \mathcal{LKJ}(1.5) \;\; p(\mathbf{R}^{(3)}) \sim \mathcal{LKJ}(1.5),$$

$$p(\beta_{W1}) \sim \mathcal{U}(-\infty, \infty), \;\; p(\beta_{W2}) \sim \mathcal{U}(-\infty, \infty), \;\; p(\beta_{W3}) \sim \mathcal{U}(-\infty, \infty),$$

$$p(\boldsymbol{z}^{(2)}_{jk}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \;\; p(\boldsymbol{z}^{(3)}_{k}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

$$p(\sigma_{Wijk}) \sim \mathcal{U}(0, \infty).$$

AZELLA writing score model 4.

$$
\begin{aligned}
y_{Wijk} = {} & \beta_{W1} + \zeta^{(2)}_{Wjk1} + \zeta^{(3)}_{Wk1} + \beta_{W4}\,\mathrm{SWDe}_{jk} + \beta_{W6}\,\mathrm{MALE}_{jk} + \\
& \left(\beta_{W2} + \zeta^{(2)}_{Wjk2} + \zeta^{(3)}_{Wk2} + \beta_{W5}\,\mathrm{SWDe}_{jk}\right)\,\mathrm{GRADE}_{ijk} + \\
& \beta_{W3}\,\mathrm{GRADE}^2_{ijk} + \epsilon_{Wijk}
\end{aligned}
$$

$$
\begin{aligned}
& p(\mathbf{D}^{(2)}) \sim \mathcal{U}(0,\,\infty), \;\; p(\mathbf{D}^{(3)}) \sim \mathcal{U}(0,\,\infty), \\
& p(\mathbf{R}^{(2)}) \sim \mathcal{LKJ}(1.5), \;\; p(\mathbf{R}^{(3)}) \sim \mathcal{LKJ}(1.5), \\
& p(\beta_{W1}) \sim \mathcal{U}(-\infty,\,\infty), \;\; p(\beta_{W2}) \sim \mathcal{U}(-\infty,\,\infty), \;\; p(\beta_{W3}) \sim \mathcal{U}(-\infty,\,\infty), \\
& p(\beta_{W4}) \sim \mathcal{U}(-\infty,\,\infty), \;\; p(\beta_{W5}) \sim \mathcal{U}(-\infty,\,\infty), \;\; p(\beta_{W6}) \sim \mathcal{U}(-\infty,\,\infty), \\
& p(\boldsymbol{z}^{(2)}_{jk}) \sim \mathcal{N}(\mathbf{0},\,\mathbf{I}), \;\; p(\boldsymbol{z}^{(3)}_{k}) \sim \mathcal{N}(\mathbf{0},\,\mathbf{I}), \\
& p(\sigma_{Wijk}) \sim \mathcal{U}(0,\,\infty).
\end{aligned}
$$

AZELLA writing score model 5.

$$
\begin{aligned}
y_{Wijk} = {} & \beta_{W1} + \zeta^{(2)}_{Wjk1} + \zeta^{(3)}_{Wk1} + \beta_{W4}\,\mathrm{SWDe}_{jk} + \beta_{W6}\,\mathrm{MALE}_{jk} + \\
& \left(\beta_{W2} + \zeta^{(2)}_{Wjk2} + \zeta^{(3)}_{Wk2} + \beta_{W5}\,\mathrm{SWDe}_{jk}\right)\,\mathrm{GRADE}_{ijk} + \\
& \beta_{W3}\,\mathrm{GRADE}^2_{ijk} + \epsilon_{Wijk}
\end{aligned}
$$

$$
\begin{aligned}
& p(\mathbf{D}^{(2)}) \sim \mathcal{U}(0,\,\infty), \;\; p(\mathbf{D}^{(3)}) \sim \mathcal{U}(0,\,\infty), \\
& p(\mathbf{R}^{(2)}) \sim \mathcal{LKJ}(1.5), \;\; p(\mathbf{R}^{(3)}) \sim \mathcal{LKJ}(1.5), \\
& p(\beta_{W1}) \sim \mathcal{N}(60,\,10), \;\; p(\beta_{W2}) \sim \mathcal{N}(3,\,6), \;\; p(\beta_{W3}) \sim \mathcal{N}(0,\,5), \\
& p(\beta_{W4}) \sim \mathcal{N}(0,\,2.5), \;\; p(\beta_{W5}) \sim \mathcal{N}(0,\,5), \;\; p(\beta_{W6}) \sim \mathcal{N}(0,\,2.5), \\
& p(\boldsymbol{z}^{(2)}_{jk}) \sim \mathcal{N}(\mathbf{0},\,\mathbf{I}), \;\; p(\boldsymbol{z}^{(3)}_{k}) \sim \mathcal{N}(\mathbf{0},\,\mathbf{I}), \\
& p(\sigma_{Wijk}) \sim \mathcal{U}(0,\,\infty).
\end{aligned}
$$

AZELLA oral score models.

AZELLA oral score model 1.

$$y_{Oij} = \beta_{O1} + \zeta^{(2)}_{Oj1} + \left(\beta_{O2} + \zeta^{(2)}_{Oj2}\right) \text{GRADE}_{ij} + \epsilon_{Oij}$$

$$p(\mathbf{D}) \sim \mathcal{U}(0, \infty),$$

$$p(\mathbf{R}) \sim \mathcal{LKJ}(1.5),$$

$$p(\beta_{O1}) \sim \mathcal{U}(-\infty, \infty), \;\; p(\beta_{O2}) \sim \mathcal{U}(-\infty, \infty),$$

$$p(\boldsymbol{z}^{(2)}_j) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

$$p(\sigma_{Oij}) \sim \mathcal{U}(0, \infty).$$

AZELLA oral score model 2.

$$y_{Oij} = \beta_{O1} + \zeta^{(2)}_{Oj1} + \left(\beta_{O2} + \zeta^{(2)}_{Oj2}\right) \text{GRADE}_{ij} + \beta_{O3} \text{GRADE}^2_{ij} + \epsilon_{Oij}$$

$$p(\mathbf{D}) \sim \mathcal{U}(0, \infty),$$

$$p(\mathbf{R}) \sim \mathcal{LKJ}(1.5),$$

$$p(\beta_{O1}) \sim \mathcal{U}(-\infty, \infty), \;\; p(\beta_{O2}) \sim \mathcal{U}(-\infty, \infty), \;\; p(\beta_{O3}) \sim \mathcal{U}(-\infty, \infty),$$

$$p(\boldsymbol{z}^{(2)}_j) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

$$p(\sigma_{Oij}) \sim \mathcal{U}(0, \infty).$$

AZELLA oral score model 3.

$$
\begin{aligned}
y_{Oijk} = \beta_{O1} + \zeta^{(2)}_{Ojk1} + \zeta^{(3)}_{Ok1} + \\
\left(\beta_{O2} + \zeta^{(2)}_{Ojk2} + \zeta^{(3)}_{Ok2}\right) \mathrm{GRADE}_{ijk} + \\
\beta_{O3}\, \mathrm{GRADE}^2_{ijk} + \epsilon_{Oijk}
\end{aligned}
$$

$$
\begin{aligned}
&p(\mathbf{D}^{(2)}) \sim \mathcal{U}(0,\, \infty), \;\; p(\mathbf{D}^{(3)}) \sim \mathcal{U}(0,\, \infty), \\
&p(\mathbf{R}^{(2)}) \sim \mathcal{LKJ}(1.5), \;\; p(\mathbf{R}^{(3)}) \sim \mathcal{LKJ}(1.5), \\
&p(\beta_{O1}) \sim \mathcal{U}(-\infty,\, \infty), \;\; p(\beta_{O2}) \sim \mathcal{U}(-\infty,\, \infty), \;\; p(\beta_{O3}) \sim \mathcal{U}(-\infty,\, \infty), \\
&p(\boldsymbol{z}^{(2)}_{jk}) \sim \mathcal{N}(\mathbf{0},\, \mathbf{I}), \;\; p(\boldsymbol{z}^{(3)}_{k}) \sim \mathcal{N}(\mathbf{0},\, \mathbf{I}), \\
&p(\sigma_{Oijk}) \sim \mathcal{U}(0,\, \infty).
\end{aligned}
$$

AZELLA oral score model 4.

$$
\begin{aligned}
y_{Oijk} = \beta_{O1} + \zeta^{(2)}_{Ojk1} + \zeta^{(3)}_{Ok1} + \beta_{O4}\, \mathrm{SWDe}_{jk} + \beta_{O6}\, \mathrm{MALE}_{jk} + \\
\left(\beta_{O2} + \zeta^{(2)}_{Ojk2} + \zeta^{(3)}_{Ok2} + \beta_{O5}\, \mathrm{SWDe}_{jk}\right) \mathrm{GRADE}_{ijk} + \\
\beta_{O3}\, \mathrm{GRADE}^2_{ijk} + \epsilon_{Oijk}
\end{aligned}
$$

$$
\begin{aligned}
&p(\mathbf{D}^{(2)}) \sim \mathcal{U}(0,\, \infty), \;\; p(\mathbf{D}^{(3)}) \sim \mathcal{U}(0,\, \infty), \\
&p(\mathbf{R}^{(2)}) \sim \mathcal{LKJ}(1.5), \;\; p(\mathbf{R}^{(3)}) \sim \mathcal{LKJ}(1.5), \\
&p(\beta_{O1}) \sim \mathcal{U}(-\infty,\, \infty), \;\; p(\beta_{O2}) \sim \mathcal{U}(-\infty,\, \infty), \;\; p(\beta_{O3}) \sim \mathcal{U}(-\infty,\, \infty), \\
&p(\beta_{O4}) \sim \mathcal{U}(-\infty,\, \infty), \;\; p(\beta_{O5}) \sim \mathcal{U}(-\infty,\, \infty), \;\; p(\beta_{O6}) \sim \mathcal{U}(-\infty,\, \infty), \\
&p(\boldsymbol{z}^{(2)}_{jk}) \sim \mathcal{N}(\mathbf{0},\, \mathbf{I}), \;\; p(\boldsymbol{z}^{(3)}_{k}) \sim \mathcal{N}(\mathbf{0},\, \mathbf{I}), \\
&p(\sigma_{Oijk}) \sim \mathcal{U}(0,\, \infty)
\end{aligned}
$$

AZELLA oral score model 5.

$$
\begin{aligned}
y_{Oijk} = \beta_{O1} + \zeta^{(2)}_{Ojk1} + \zeta^{(3)}_{Ok1} + \beta_{O4}\,\mathrm{SWDe}_{jk} + \beta_{O6}\,\mathrm{MALE}_{jk} + \\
\left(\beta_{O2} + \zeta^{(2)}_{Ojk2} + \zeta^{(3)}_{Ok2} + \beta_{O5}\,\mathrm{SWDe}_{jk}\right)\,\mathrm{GRADE}_{ijk} + \\
\beta_{O3}\,\mathrm{GRADE}^2_{ijk} + \epsilon_{Oijk}
\end{aligned}
$$

$$
\begin{aligned}
&p(\mathbf{D}^{(2)}) \sim \mathcal{U}(0,\,\infty),\;\; p(\mathbf{D}^{(3)}) \sim \mathcal{U}(0,\,\infty),\\
&p(\mathbf{R}^{(2)}) \sim \mathcal{LKJ}(1.5),\;\; p(\mathbf{R}^{(3)}) \sim \mathcal{LKJ}(1.5),\\
&p(\beta_{O1}) \sim \mathcal{N}(60,\,10),\;\; p(\beta_{O2}) \sim \mathcal{N}(3,\,6),\;\; p(\beta_{O3}) \sim \mathcal{N}(0,\,5),\\
&p(\beta_{O4}) \sim \mathcal{N}(0,\,2.5),\;\; p(\beta_{O5}) \sim \mathcal{N}(0,\,5),\;\; p(\beta_{O6}) \sim \mathcal{N}(0,\,2.5),\\
&p(\boldsymbol{z}^{(2)}_{jk}) \sim \mathcal{N}(\mathbf{0},\,\mathbf{I}),\;\; p(\boldsymbol{z}^{(3)}_{k}) \sim \mathcal{N}(\mathbf{0},\,\mathbf{I}),\\
&p(\sigma_{Oijk}) \sim \mathcal{U}(0,\,\infty).
\end{aligned}
$$

Multivariate AZELLA Growth Models

Research Question 2 extended Equations A.3 to model the multiple English language proficiency sub-tests simultaneously. Let there be $g = 1, 2, \ldots, m$ outcomes, each measured longitudinally, such that $\boldsymbol{y}_{gjk}$ is a vector of length n_{gjk} that contains the longitudinal measures of outcome g for student j in school k and $\boldsymbol{y}_{gk}$ is a length $N_{gk} = \sum_{j=1}^{n_{gk}} n_{gjk}$ vector that contain all observations for outcome g in school k. Next, let $\boldsymbol{y}_{Mjk} \equiv \left[\boldsymbol{y}'_{1jk}, \ldots, \boldsymbol{y}'_{mjk}\right]'$ be the length $n_{Mjk} = \sum_{g=1}^{m} n_{gjk}$ vector that contains all outcomes measured longitudinally for subject j. Finally, let $\boldsymbol{y}_{Mk} \equiv [\boldsymbol{y}'_{1k}, \ldots, \boldsymbol{y}'_{mk}]'$ be the length $N_{Mk} = \sum_{g=1}^{m}\sum_{j=1}^{n_{gk}} n_{gjk}$ vector of all observations for all outcomes for school k. I can express the three-level multivariate mixed model generally by expanding Equation A.3,

$$
\boldsymbol{y}_{Mk} = \mathbf{X}_{Mk}\boldsymbol{\beta}_M + \mathbf{Z}_{Mk}\boldsymbol{\zeta}_{Mk} + \epsilon_{Mk} \tag{A.6}
$$

where $\mathbf{X}_{Mk} \equiv \text{diag}[\mathbf{X}_{1k}, \ldots, \mathbf{X}_{mk}]$ is a known $N_{Mk} \times p$ design matrix corresponding to the $p \times 1$ vector of fixed effects $\boldsymbol{\beta}_M \equiv [\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_m']'$. Define $\mathbf{Z}_{Mk} \equiv \text{diag}[\mathbf{Z}_{1k}, \ldots, \mathbf{Z}_{mk}]$ and $\boldsymbol{\zeta}_M \equiv [\boldsymbol{\zeta}_1', \ldots, \boldsymbol{\zeta}_m']'$ where $\mathbf{Z}_{gk} \equiv \left[\mathbf{Z}_{gk}^{(3)}, \text{diag}\left[\mathbf{Z}_{g1k}^{(2)}, \ldots, \mathbf{Z}_{gn_kk}^{(2)}\right]\right]$ and $\boldsymbol{\zeta}_{gk} \equiv \left[\boldsymbol{\zeta}_{gk}^{(3)\prime}, \boldsymbol{\zeta}_{gk1}^{(2)\prime}, \ldots, \boldsymbol{\zeta}_{gn_kk}^{(2)\prime}\right]'$. Following Equation A.3, $\mathbf{Z}_{gk}^{(3)}$ is an $N_{gk} \times q_g^{(3)}$ design matrix corresponding to the school-specific random effects for outcome g, $\boldsymbol{\zeta}_{gk}^{(3)}$ and $\mathbf{Z}_{gjk}^{(2)}$ is an $n_{gjk} \times q_g^{(2)}$ design matrix corresponding to the $q_g^{(2)} \times 1$ vector of random effects for student j in school k for outcome g, $\boldsymbol{\zeta}_{gjk}^{(2)}$. Given this notation, $\boldsymbol{\zeta}_{Mjk}^{(2)} \equiv \left[\boldsymbol{\zeta}_{1jk}^{(2)\prime}, \ldots, \boldsymbol{\zeta}_{mjk}^{(2)\prime}\right]'$ is the vector of student-specific random effects for all outcomes for student j in school k and $\boldsymbol{\zeta}_{Mk}^{(3)} \equiv \left[\boldsymbol{\zeta}_{1k}^{(3)\prime}, \ldots, \boldsymbol{\zeta}_{mk}^{(3)\prime}\right]'$ is the vector of school-specific random effects for all outcomes for school k. Finally, $\boldsymbol{\epsilon}_{Mk}$ is the $N_{Mk} \times 1$ vector of residuals.

I assume ϵ_{gijk} are independent $\mathcal{N}(0, \sigma_g)$, $\boldsymbol{\zeta}_{Mki}^{(2)} \sim \mathcal{N}(\mathbf{0}, \mathbf{T}^{(2)})$, $\boldsymbol{\zeta}_{Mk}^{(3)} \sim \mathcal{N}(\mathbf{0}, \mathbf{T}^{(3)})$, where $\mathbf{T}^{(2)}$ and $\mathbf{T}^{(3)}$ are defined by Equation A.4. Using this generalization, $\mathbf{D}\,\mathbf{R}$, $\boldsymbol{z}_{jk}$, and $\boldsymbol{z}_k$ have the same definitions as in Research Question 1.

Multivariate AZELLA model 1.

$$y_{Rijk} = \beta_{R1} + \zeta^{(2)}_{Rjk1} + \zeta^{(3)}_{Rk1} + \left(\beta_{R2} + \zeta^{(2)}_{Rjk1} + \zeta^{(3)}_{Rk2}\right) \text{GRADE}_{ijk} + \epsilon_{Rijk}$$

$$y_{Wijk} = \beta_{W1} + \zeta^{(2)}_{Wjk1} + \zeta^{(3)}_{Wk1} + \left(\beta_{W2} + \zeta^{(2)}_{Wjk1} + \zeta^{(3)}_{Wk2}\right) \text{GRADE}_{ijk} + \beta_{W3} \text{GRADE}^2_{ijk} + \epsilon_{Wijk}$$

$$y_{Oijk} = \beta_{O1} + \zeta^{(2)}_{Ojk1} + \zeta^{(3)}_{Ok1} + \left(\beta_{O2} + \zeta^{(2)}_{Ojk1} + \zeta^{(3)}_{Ok2}\right) \text{GRADE}_{ijk} + \beta_{O3} \text{GRADE}^2_{ijk} + \epsilon_{Oijk}$$

$$p(\mathbf{D}^{(2)}) \sim \mathcal{U}(0, \infty), \;\; p(\mathbf{D}^{(3)}) \sim \mathcal{U}(0, \infty),$$

$$p(\mathbf{R}^{(2)}) \sim \mathcal{LKJ}(1.5), \;\; p(\mathbf{R}^{(3)}) \sim \mathcal{LKJ}(1.5),$$

$$p(\beta_{R1}) \sim \mathcal{U}(-\infty, \infty), \;\; p(\beta_{R2}) \sim \mathcal{U}(-\infty, \infty),$$

$$p(\beta_{W1}) \sim \mathcal{U}(-\infty, \infty), \;\; p(\beta_{W2}) \sim \mathcal{U}(-\infty, \infty), \;\; p(\beta_{W3}) \sim \mathcal{U}(-\infty, \infty),$$

$$p(\beta_{O1}) \sim \mathcal{U}(-\infty, \infty), \;\; p(\beta_{O2}) \sim \mathcal{U}(-\infty, \infty), \;\; p(\beta_{O3}) \sim \mathcal{U}(-\infty, \infty),$$

$$p(\boldsymbol{z}^{(2)}_{Rjk}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \;\; p(\boldsymbol{z}^{(3)}_{Rk}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

$$p(\boldsymbol{z}^{(2)}_{Wjk}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \;\; p(\boldsymbol{z}^{(3)}_{Wk}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

$$p(\boldsymbol{z}^{(2)}_{Ojk}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \;\; p(\boldsymbol{z}^{(3)}_{Ok}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

$$p(\sigma_{Rijk}) \sim \mathcal{U}(0, \infty), \;\; p(\sigma_{Wijk}) \sim \mathcal{U}(0, \infty), \;\; p(\sigma_{Oijk}) \sim \mathcal{U}(0, \infty).$$

It is important to note that $[\zeta^{(2)}_{Rjk1}, \zeta^{(2)}_{Rjk2}, \zeta^{(2)}_{Wjk2}, \zeta^{(2)}_{Wjk2}, \zeta^{(2)}_{Ojk1}, \zeta^{(2)}_{Ojk2}]$ and $[\zeta^{(3)}_{Rjk1}, \zeta^{(3)}_{Rjk2}, \zeta^{(3)}_{Wjk2}, \zeta^{(3)}_{Wjk2}, \zeta^{(3)}_{Ojk1}, \zeta^{(3)}_{Ojk2}]$ are both assumed to be distributed multivariate normal with mean zero and 6×6 unstructured covariance matrices $\mathrm{T}^{(2)}$ and $\mathrm{T}^{(3)}$.

Multivariate AZELLA model 2.

$$
\begin{aligned}
y_{Rijk} = {} & \beta_{R1} + \zeta^{(2)}_{Rjk1} + \zeta^{(3)}_{Rk1} + \beta_{R4}\,\mathrm{SWDe}_{jk} + \beta_{R6}\,\mathrm{MALE}_{jk} + \\
& \left(\beta_{R2} + \zeta^{(2)}_{Rjk2} + \zeta^{(3)}_{Rk2} + \beta_{R5}\,\mathrm{SWDe}_{jk}\right)\,\mathrm{GRADE}_{ijk} + \epsilon_{Rijk} \\
y_{Wijk} = {} & \beta_{W1} + \zeta^{(2)}_{Wjk1} + \zeta^{(3)}_{Wk1} + \beta_{W4}\,\mathrm{SWDe}_{jk} + \beta_{W6}\,\mathrm{MALE}_{jk} + \\
& \left(\beta_{W2} + \zeta^{(2)}_{Wjk2} + \zeta^{(3)}_{Wk2} + \beta_{W5}\,\mathrm{SWDe}_{jk}\right)\,\mathrm{GRADE}_{ijk} + \\
& \beta_{W3}\,\mathrm{GRADE}^2_{ijk} + \epsilon_{Wijk} \\
y_{Oijk} = {} & \beta_{O1} + \zeta^{(2)}_{Ojk1} + \zeta^{(3)}_{Ok1} + \beta_{O4}\,\mathrm{SWDe}_{jk} + \beta_{O6}\,\mathrm{MALE}_{jk} + \\
& \left(\beta_{O2} + \zeta^{(2)}_{Ojk2} + \zeta^{(3)}_{Ok2} + \beta_{O5}\,\mathrm{SWDe}_{jk}\right)\,\mathrm{GRADE}_{ijk} + \\
& \beta_{O3}\,\mathrm{GRADE}^2_{ijk} + \epsilon_{Oijk}
\end{aligned}
$$

$$
\begin{aligned}
& p(\mathbf{D}^{(2)}) \sim \mathcal{U}(0,\,\infty),\;\; p(\mathbf{D}^{(3)}) \sim \mathcal{U}(0,\,\infty), \\
& p(\mathbf{R}^{(2)}) \sim \mathcal{LKJ}(1.5),\;\; p(\mathbf{R}^{(3)}) \sim \mathcal{LKJ}(1.5), \\
& p(\beta_{R1}) \sim \mathcal{U}(-\infty,\,\infty),\;\; p(\beta_{R2}) \sim \mathcal{U}(-\infty,\,\infty), \\
& p(\beta_{R4}) \sim \mathcal{U}(-\infty,\,\infty),\;\; p(\beta_{R5}) \sim \mathcal{U}(-\infty,\,\infty),\;\; p(\beta_{R6}) \sim \mathcal{U}(-\infty,\,\infty), \\
& p(\beta_{W1}) \sim \mathcal{U}(-\infty,\,\infty),\;\; p(\beta_{W2}) \sim \mathcal{U}(-\infty,\,\infty),\;\; p(\beta_{W3}) \sim \mathcal{U}(-\infty,\,\infty), \\
& p(\beta_{W4}) \sim \mathcal{U}(-\infty,\,\infty),\;\; p(\beta_{W5}) \sim \mathcal{U}(-\infty,\,\infty),\;\; p(\beta_{W6}) \sim \mathcal{U}(-\infty,\,\infty), \\
& p(\beta_{O1}) \sim \mathcal{U}(-\infty,\,\infty),\;\; p(\beta_{O2}) \sim \mathcal{U}(-\infty,\,\infty),\;\; p(\beta_{O3}) \sim \mathcal{U}(-\infty,\,\infty), \\
& p(\beta_{O4}) \sim \mathcal{U}(-\infty,\,\infty),\;\; p(\beta_{O5}) \sim \mathcal{U}(-\infty,\,\infty),\;\; p(\beta_{O6}) \sim \mathcal{U}(-\infty,\,\infty), \\
& p(\boldsymbol{z}^{(2)}_{Rjk}) \sim \mathcal{N}(\mathbf{0},\,\mathbf{I}),\;\; p(\boldsymbol{z}^{(3)}_{Rk}) \sim \mathcal{N}(\mathbf{0},\,\mathbf{I}), \\
& p(\boldsymbol{z}^{(2)}_{Wjk}) \sim \mathcal{N}(\mathbf{0},\,\mathbf{I}),\;\; p(\boldsymbol{z}^{(3)}_{Wk}) \sim \mathcal{N}(\mathbf{0},\,\mathbf{I}), \\
& p(\boldsymbol{z}^{(2)}_{Ojk}) \sim \mathcal{N}(\mathbf{0},\,\mathbf{I}),\;\; p(\boldsymbol{z}^{(3)}_{Ok}) \sim \mathcal{N}(\mathbf{0},\,\mathbf{I}), \\
& p(\sigma_{Rijk}) \sim \mathcal{U}(0,\,\infty),\;\; p(\sigma_{Wijk}) \sim \mathcal{U}(0,\,\infty),\;\; p(\sigma_{Oijk}) \sim \mathcal{U}(0,\,\infty).
\end{aligned}
$$

Multivariate AZELLA model 3.

$$
\begin{aligned}
y_{Rijk} = {} & \beta_{R1} + \zeta^{(2)}_{Rjk1} + \zeta^{(3)}_{Rk1} + \beta_{R4}\,\mathrm{SWDe}_{jk} + \beta_{R6}\,\mathrm{MALE}_{jk} + \\
& \left(\beta_{R2} + \zeta^{(2)}_{Rjk2} + \zeta^{(3)}_{Rk2} + \beta_{R5}\,\mathrm{SWDe}_{jk}\right)\,\mathrm{GRADE}_{ijk} + \epsilon_{Rijk} \\
y_{Wijk} = {} & \beta_{W1} + \zeta^{(2)}_{Wjk1} + \zeta^{(3)}_{Wk1} + \beta_{W4}\,\mathrm{SWDe}_{jk} + \beta_{W6}\,\mathrm{MALE}_{jk} + \\
& \left(\beta_{W2} + \zeta^{(2)}_{Wjk2} + \zeta^{(3)}_{Wk2} + \beta_{W5}\,\mathrm{SWDe}_{jk}\right)\,\mathrm{GRADE}_{ijk} + \\
& \beta_{W3}\,\mathrm{GRADE}^2_{ijk} + \epsilon_{Wijk} \\
y_{Oijk} = {} & \beta_{O1} + \zeta^{(2)}_{Ojk1} + \zeta^{(3)}_{Ok1} + \beta_{O4}\,\mathrm{SWDe}_{jk} + \beta_{O6}\,\mathrm{MALE}_{jk} + \\
& \left(\beta_{O2} + \zeta^{(2)}_{Ojk2} + \zeta^{(3)}_{Ok2} + \beta_{O5}\,\mathrm{SWDe}_{jk}\right)\,\mathrm{GRADE}_{ijk} + \\
& \beta_{O3}\,\mathrm{GRADE}^2_{ijk} + \epsilon_{Oijk}
\end{aligned}
$$

$$
\begin{aligned}
& p(\mathbf{D}^{(2)}) \sim \mathcal{U}(0,\,\infty), \;\; p(\mathbf{D}^{(3)}) \sim \mathcal{U}(0,\,\infty), \\
& p(\mathbf{R}^{(2)}) \sim \mathcal{LKJ}(1.5), \;\; p(\mathbf{R}^{(3)}) \sim \mathcal{LKJ}(1.5), \\
& p(\beta_{R1}) \sim \mathcal{N}(60,\,10), \;\; p(\beta_{R2}) \sim \mathcal{N}(3,\,6), \\
& p(\beta_{R4}) \sim \mathcal{N}(0,\,5), \;\; p(\beta_{R5}) \sim \mathcal{N}(0,\,5), \;\; p(\beta_{R6}) \sim \mathcal{N}(0,\,2.5), \\
& p(\beta_{W1}) \sim \mathcal{N}(60,\,10), \;\; p(\beta_{W2}) \sim \mathcal{N}(3,\,6), \;\; p(\beta_{W3}) \sim \mathcal{N}(0,\,2.5), \\
& p(\beta_{W4}) \sim \mathcal{N}(0,\,5), \;\; p(\beta_{W5}) \sim \mathcal{N}(0,\,5), \;\; p(\beta_{W6}) \sim \mathcal{N}(0,\,2.5), \\
& p(\beta_{O1}) \sim \mathcal{N}(60,\,10), \;\; p(\beta_{O2}) \sim \mathcal{N}(3,\,6), \;\; p(\beta_{O3}) \sim \mathcal{N}(0,\,2.5), \\
& p(\beta_{O4}) \sim \mathcal{N}(0,\,5), \;\; p(\beta_{O5}) \sim \mathcal{N}(0,\,5), \;\; p(\beta_{O6}) \sim \mathcal{N}(0,\,2.5), \\
& p(\boldsymbol{z}^{(2)}_{Rjk}) \sim \mathcal{N}(\mathbf{0},\,\mathbf{I}), \;\; p(\boldsymbol{z}^{(3)}_{Rk}) \sim \mathcal{N}(\mathbf{0},\,\mathbf{I}), \\
& p(\boldsymbol{z}^{(2)}_{Wjk}) \sim \mathcal{N}(\mathbf{0},\,\mathbf{I}), \;\; p(\boldsymbol{z}^{(3)}_{Wk}) \sim \mathcal{N}(\mathbf{0},\,\mathbf{I}), \\
& p(\boldsymbol{z}^{(2)}_{Ojk}) \sim \mathcal{N}(\mathbf{0},\,\mathbf{I}), \;\; p(\boldsymbol{z}^{(3)}_{Ok}) \sim \mathcal{N}(\mathbf{0},\,\mathbf{I}), \\
& p(\sigma_{Rijk}) \sim \mathcal{U}(0,\,\infty), \;\; p(\sigma_{Wijk}) \sim \mathcal{U}(0,\,\infty), \;\; p(\sigma_{Oijk}) \sim \mathcal{U}(0,\,\infty).
\end{aligned}
$$

Time-to-Reclassification Models

Research Question 3 estimated a series of discrete-time hazard models and a series of shared random effects models.

Discrete-time hazard models. Let there be $k = 1, 2, \ldots, K$ schools, and $j = 1, 2, \ldots, n_k$ students in school k. Event occurrence can only happen at discrete time points, $t = 1, 2, \ldots, T$ with a student's survival time is indicated by $T_{jk} = t$. Define $h_{tjk} = \mathbb{P}[t_{jk} = t | t_{jk} \geq t]$ as the probability of reclassification for student j in school k at time t given reclassification has not occurred prior to time t. Those students who do not experience the event while under observation are considered censored. Using grouped-time survival parametrization (Allison, 1982; D'Agostino et al., 1990; Hedeker, Siddiqui, & Hu, 2000; Singer & Willett, 1993), each student contributes a $t_{jk} \times 1$ vector of of dichotomous indicators of event status for each discrete time point. Students who experience reclassification at time t_{jk} will have $t_{jk} - 1$ zeros followed by a one indicating event occurrence. Students who are censored have a $t_{jk} \times 1$ vector of zeros.

Using the logit link function, I can express log odds of reclassification for student j at school k at time t as

$$\log\left[h_{tjk}/1 - h_{tjk}\right] = \text{logit}(h_{tjk}) = \mathbf{w}'_{tjk}\boldsymbol{\alpha} + \nu_k \tag{A.7}$$

and the probability or reclassification as

$$1/\left(1 + \exp\left[-\text{logit}(h_{tjk})\right]\right) = 1/\left(1 + \exp\left[-(\mathbf{w}'_{tjk}\boldsymbol{\alpha} + \nu_k)\right]\right) \tag{A.8}$$

where $\boldsymbol{\alpha}$ is a $p \times 1$ vector of fixed effects that contains T intercept terms for the T discrete time points and additional covariates, and $\mathbf{w}_{tjk}$, a $p \times 1$ known vector of fixed effects. Because students are nested in schools, ν_k is a school-level random intercept assumed to be normally distributed with mean zero and variance ψ.

Hazard model 1.

$$\begin{aligned}\text{logit}(h_{tjk}) = &\alpha_1 \text{GRADE3}_{tjk} + \alpha_2 \text{GRADE4}_{tjk} + \alpha_3 \text{GRADE5}_{tjk} + \alpha_4 \text{GRADE6}_{tjk} + \\ &\alpha_5 \text{GRADE7}_{tjk}\end{aligned}$$

$$\begin{aligned}&p(\alpha_1) \sim \mathcal{U}(-\infty, \infty), \;\; p(\alpha_2) \sim \mathcal{U}(-\infty, \infty), \;\; p(\alpha_3) \sim \mathcal{U}(-\infty, \infty), \\ &p(\alpha_4) \sim \mathcal{U}(-\infty, \infty), \;\; p(\alpha_5) \sim \mathcal{U}(-\infty, \infty).\end{aligned}$$

Hazard model 2.

$$\begin{aligned}\text{logit}(h_{tjk}) = &\alpha_1 \text{GRADE3}_{tjk} + \alpha_2 \text{GRADE4}_{tjk} + \alpha_3 \text{GRADE5}_{tjk} + \alpha_4 \text{GRADE6}_{tjk} + \\ &\alpha_5 \text{GRADE7}_{tjk} + \nu_k\end{aligned}$$

$$\begin{aligned}&p(\alpha_1) \sim \mathcal{U}(-\infty, \infty), \;\; p(\alpha_2) \sim \mathcal{U}(-\infty, \infty), \;\; p(\alpha_3) \sim \mathcal{U}(-\infty, \infty), \\ &p(\alpha_4) \sim \mathcal{U}(-\infty, \infty), \;\; p(\alpha_5) \sim \mathcal{U}(-\infty, \infty), \\ &p(\psi) \sim \mathcal{U}(0, \infty).\end{aligned}$$

Hazard model 3.

$$\begin{aligned}\text{logit}(h_{tjk}) = &\alpha_1 \text{GRADE3}_{tjk} + \alpha_2 \text{GRADE4}_{tjk} + \alpha_3 \text{GRADE5}_{tjk} + \alpha_4 \text{GRADE6}_{tjk} + \\ &\alpha_5 \text{GRADE7}_{tjk} + \alpha_6 \text{SWDe}_{tjk} + \alpha_6 \text{MALE}_{tjk} + \nu_k\end{aligned}$$

$$\begin{aligned}&p(\alpha_1) \sim \mathcal{U}(-\infty, \infty), \;\; p(\alpha_2) \sim \mathcal{U}(-\infty, \infty), \;\; p(\alpha_3) \sim \mathcal{U}(-\infty, \infty), \\ &p(\alpha_4) \sim \mathcal{U}(-\infty, \infty), \;\; p(\alpha_5) \sim \mathcal{U}(-\infty, \infty), \;\; p(\alpha_6) \sim \mathcal{U}(-\infty, \infty), \\ &p(\alpha_7) \sim \mathcal{U}(-\infty, \infty), \\ &p(\psi) \sim \mathcal{U}(0, \infty).\end{aligned}$$

Shared random effects models. The final set of models combines Equations A.3 and A.8 and links them by specifying the student-specific random

effects estimated by Equations A.3 as latent covariates in Equation A.8.

$$\boldsymbol{y}_{jk} = \mathbf{X}_{jk}\boldsymbol{\beta} + \mathbf{Z}_{jk}\boldsymbol{\zeta}_{jk} + \boldsymbol{\epsilon}_{jk}$$

$$h_{tjk} = \mathbf{w}'_{tjk}\boldsymbol{\alpha} + \boldsymbol{\zeta}_{jk}^{(2)\prime}\,\boldsymbol{\lambda} + \nu_k$$

Shared random effects model 1. Note that the hazard submodel estimates are also presented in Table 7 as Model 2.

$$\begin{aligned} y_{Tijk} = {} & \beta_{T1} + \zeta_{Tjk1}^{(2)} + \zeta_{Tk1}^{(3)} + \beta_{T4}\,\mathrm{SWDe}_{jk} + \beta_{T6}\,\mathrm{MALE}_{jk} + \\ & \left(\beta_{T2} + \zeta_{Tjk2}^{(2)} + \zeta_{Tk2}^{(3)} + \beta_{T5}\,\mathrm{SWDe}_{jk}\right)\,\mathrm{GRADE}_{ijk} + \\ & \beta_{T3}\,\mathrm{GRADE}_{ijk}^2 + \epsilon_{Tijk} \end{aligned}$$

$$\begin{aligned} \mathrm{logit}(h_{tjk}) = {} & \alpha_1 \mathrm{GRADE3}_{tjk} + \alpha_2 \mathrm{GRADE4}_{tjk} + \alpha_3 \mathrm{GRADE5}_{tjk} + \alpha_4 \mathrm{GRADE6}_{tjk} + \\ & \alpha_5 \mathrm{GRADE7}_{tjk} + \zeta_{jk1}^{(2)}\,\lambda_1 + \zeta_{jk2}^{(2)}\,\lambda_2 + \nu_k \end{aligned}$$

$$p(\mathbf{D}^{(2)}) \sim \mathcal{U}(0,\,\infty),\;\; p(\mathbf{D}^{(3)}) \sim \mathcal{U}(0,\,\infty),$$

$$p(\mathbf{R}^{(2)}) \sim \mathcal{LKJ}(1.5),\;\; p(\mathbf{R}^{(3)}) \sim \mathcal{LKJ}(1.5),$$

$$p(\beta_{T1}) \sim \mathcal{N}(60,\,10),\;\; p(\beta_{T2}) \sim \mathcal{N}(3,\,6),\;\; p(\beta_{T3}) \sim \mathcal{N}(0,\,2.5),$$

$$p(\beta_{T4}) \sim \mathcal{N}(0,\,5),\;\; p(\beta_{T5}) \sim \mathcal{N}(0,\,2.5),\;\; p(\beta_{T6}) \sim \mathcal{N}(0,\,5),$$

$$p(\boldsymbol{z}_{jk}^{(2)}) \sim \mathcal{N}(\mathbf{0},\,\mathbf{I}),\;\; p(\boldsymbol{z}_{k}^{(3)}) \sim \mathcal{N}(\mathbf{0},\,\mathbf{I}),$$

$$p(\sigma_T) \sim \mathcal{U}(0,\,\infty),$$

$$p(\alpha_1) \sim \mathcal{N}(0,\,10),\;\; p(\alpha_2) \sim \mathcal{N}(0,\,10),\;\; p(\alpha_3) \sim \mathcal{N}(0,\,10),$$

$$p(\alpha_4) \sim \mathcal{N}(0,\,10),\;\; p(\alpha_5) \sim \mathcal{N}(0,\,10),$$

$$p(\lambda_1) \sim \mathcal{N}(5,\,2),\;\; p(\lambda_2) \sim \mathcal{N}(5,\,2),$$

$$p(\psi) \sim \mathcal{U}(0,\,\infty).$$

Shared random effects model 2. Note that the hazard submodel estimates are also presented in Table 7 as Model 2.

$$
\begin{aligned}
y_{Tijk} = {} & \beta_{T1} + \zeta^{(2)}_{Tjk1} + \zeta^{(3)}_{Tk1} + \beta_{T4}\,\mathrm{SWDe}_{jk} + \beta_{T6}\,\mathrm{MALE}_{jk} + \\
& \left(\beta_{T2} + \zeta^{(2)}_{Tjk2} + \zeta^{(3)}_{Tk2} + \beta_{T5}\,\mathrm{SWDe}_{jk}\right)\,\mathrm{GRADE}_{ijk} + \\
& \beta_{T3}\,\mathrm{GRADE}^2_{ijk} + \epsilon_{Tijk}
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{logit}(h_{tjk}) = {} & \alpha_1 \mathrm{GRADE3}_{tjk} + \alpha_2 \mathrm{GRADE4}_{tjk} + \alpha_3 \mathrm{GRADE5}_{tjk} + \alpha_4 \mathrm{GRADE6}_{tjk} + \\
& \alpha_5 \mathrm{GRADE7}_{tjk} + \alpha_6 \mathrm{SWDe}_{jk} + \alpha_6 \mathrm{MALE}_{jk} + \nu_k
\end{aligned}
$$

$$
\begin{aligned}
& p(\mathbf{D}^{(2)}) \sim \mathcal{U}(0,\,\infty), \;\; p(\mathbf{D}^{(3)}) \sim \mathcal{U}(0,\,\infty), \\
& p(\mathbf{R}^{(2)}) \sim \mathcal{LKJ}(1.5), \;\; p(\mathbf{R}^{(3)}) \sim \mathcal{LKJ}(1.5), \\
& p(\beta_{T1}) \sim \mathcal{N}(60,\,10), \;\; p(\beta_{T2}) \sim \mathcal{N}(3,\,6), \;\; p(\beta_{T3}) \sim \mathcal{N}(0,\,2.5), \\
& p(\beta_{T4}) \sim \mathcal{N}(0,\,5), \;\; p(\beta_{T5}) \sim \mathcal{N}(0,\,2.5), \;\; p(\beta_{T6}) \sim \mathcal{N}(0,\,5), \\
& p(\boldsymbol{z}^{(2)}_{jk}) \sim \mathcal{N}(\mathbf{0},\,\mathbf{I}), \;\; p(\boldsymbol{z}^{(3)}_{k}) \sim \mathcal{N}(\mathbf{0},\,\mathbf{I}), \\
& p(\sigma_T) \sim \mathcal{U}(0,\,\infty), \\
& p(\alpha_1) \sim \mathcal{N}(0,\,10), \;\; p(\alpha_2) \sim \mathcal{N}(0,\,10), \;\; p(\alpha_3) \sim \mathcal{N}(0,\,10), \\
& p(\alpha_4) \sim \mathcal{N}(0,\,10), \;\; p(\alpha_5) \sim \mathcal{N}(0,\,10), \;\; p(\alpha_6) \sim \mathcal{N}(0,\,10), \\
& p(\alpha_7) \sim \mathcal{N}(0,\,10), \\
& p(\psi) \sim \mathcal{U}(0,\,\infty).
\end{aligned}
$$

Shared random effects model 3. Note that the hazard submodel estimates are also presented in Table 7 as Model 3.

$$
\begin{aligned}
y_{Tijk} = {} & \beta_{T1} + \zeta^{(2)}_{Tjk1} + \zeta^{(3)}_{Tk1} + \beta_{T4}\,\mathrm{SWDe}_{jk} + \beta_{T6}\,\mathrm{MALE}_{jk} + \\
& \left(\beta_{T2} + \zeta^{(2)}_{Tjk2} + \zeta^{(3)}_{Tk2} + \beta_{T5}\,\mathrm{SWDe}_{jk}\right)\,\mathrm{GRADE}_{ijk} + \\
& \beta_{T3}\,\mathrm{GRADE}^2_{ijk} + \epsilon_{Tijk}
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{logit}(h_{tjk}) = {} & \alpha_1 \mathrm{GRADE3}_{tjk} + \alpha_2 \mathrm{GRADE4}_{tjk} + \alpha_3 \mathrm{GRADE5}_{tjk} + \alpha_4 \mathrm{GRADE6}_{tjk} + \\
& \alpha_5 \mathrm{GRADE7}_{tjk} + \alpha_6 \mathrm{SWDe}_{jk} + \alpha_6 \mathrm{MALE}_{jk} + \\
& \zeta^{(2)}_{jk1}\,\lambda_1 + \zeta^{(2)}_{jk2}\,\lambda_2 + \nu_k
\end{aligned}
$$

$$
\begin{aligned}
& p(\mathbf{D}^{(2)}) \sim \mathcal{U}(0,\,\infty), \;\; p(\mathbf{D}^{(3)}) \sim \mathcal{U}(0,\,\infty), \\
& p(\mathbf{R}^{(2)}) \sim \mathcal{LKJ}(1.5), \;\; p(\mathbf{R}^{(3)}) \sim \mathcal{LKJ}(1.5), \\
& p(\beta_{T1}) \sim \mathcal{N}(60,\,10), \;\; p(\beta_{T2}) \sim \mathcal{N}(3,\,6), \;\; p(\beta_{T3}) \sim \mathcal{N}(0,\,2.5), \\
& p(\beta_{T4}) \sim \mathcal{N}(0,\,5), \;\; p(\beta_{T5}) \sim \mathcal{N}(0,\,2.5), \;\; p(\beta_{T6}) \sim \mathcal{N}(0,\,5), \\
& p(\boldsymbol{z}^{(2)}_i) \sim \mathcal{N}(\mathbf{0},\,\mathbf{I}), \;\; p(\boldsymbol{z}^{(3)}_i) \sim \mathcal{N}(\mathbf{0},\,\mathbf{I}), \\
& p(\sigma_T) \sim \mathcal{U}(0,\,\infty), \\
& p(\alpha_1) \sim \mathcal{N}(0,\,10), \;\; p(\alpha_2) \sim \mathcal{N}(0,\,10), \;\; p(\alpha_3) \sim \mathcal{N}(0,\,10), \\
& p(\alpha_4) \sim \mathcal{N}(0,\,10), \;\; p(\alpha_5) \sim \mathcal{N}(0,\,10), \;\; p(\alpha_6) \sim \mathcal{N}(0,\,10), \\
& p(\alpha_7) \sim \mathcal{N}(0,\,10), \\
& p(\lambda_1) \sim \mathcal{N}(5,\,2), \;\; p(\lambda_2) \sim \mathcal{N}(5,\,2), \\
& p(\psi) \sim \mathcal{U}(0,\,\infty).
\end{aligned}
$$

APPENDIX B

CONVERGENCE

This section describes the methods used to assess convergence of the Markov chains specified for each model. It is important to note that approximate convergence cannot be tested empirically. Thus the assessment of mixing and stationarity of the chains was assessed following the guidelines of Gelman, Carlin, et al. (2014).

Each model was estimated by specifying four Markov chains, each for 2500 iterations. The conservative choice of discarding the first half of each Markov chain was adopted, leaving 1250 simulations from the target distribution per chain, or a total of 5000 simulated values. Visual inspection of the chains using trace plots provided the initial evidence of both mixing and stationarity. Convergence for each key parameter was further evaluated using the $\hat{R}$ statistic. $\hat{R}$ is defined as an estimate of the factor by which the scale of the current posterior for a given parameter might be reduced if each chain continued to ∞ (Gelman, Carlin, et al., 2014). Finally, because the simulations within a chain are subject to autocorrelation, the effective number of simulations, n_{eff}, was computed for each parameter using the post-warm-up simulations. Gelman, Carlin, et al. (2014) recommended $\hat{R} < 1.1$ and $n_{eff} = 10$ per parameter.

Table B.1 provides the $\hat{R}$ and n_{eff} estimates for key parameter from the AZELLA total score growth models. In addition, Figures B.1 to B.15 provide trace and density plots of those parameter. Chains were run for all models used for this dissertation until $\hat{R} < 1.1$ and $n_{eff} \geq 400$. Readers interested in convergence evidence for the other models are encouraged to contact the author.

Table B.1.
Quantitative convergence evidence for the total English language proficiency growth models

	Model 1		Model 2		Model 3		Model 4		Model 5	
Parameter	$\hat{R}$	n_{eff}	$\hat{R}$	n_{eff}	$\hat{R}$	n_{eff}	$\hat{R}$	n_{eff}	$\hat{R}$	n_{eff}
β_1 [Initial Status]	1.00	2240.60	1.00	3184.26	1.00	2142.97	1.00	3510.37	1.00	3380.88
β_2 [Grade]	1.00	2975.54	1.00	5000.00	1.00	2805.45	1.00	5000.00	1.00	5000.00
β_3 [Grade^2]			1.00	5000.00	1.00	3050.33	1.00	5000.00	1.00	5000.00
β_4 [SWDe]							1.00	3865.01	1.00	3509.17
β_5 [Grade] $\times$ [SWDe]							1.00	3276.29	1.00	3304.35
β_6 [Male]							1.00	5000.00	1.00	5000.00
$\sqrt{\tau_{11}^{(2)}}$	1.00	2154.47	1.00	1636.20	1.00	1448.96	1.00	1760.98	1.00	2021.60
$\sqrt{\tau_{22}^{(2)}}$	1.00	1091.43	1.01	668.75	1.00	971.59	1.00	629.77	1.00	958.45
$\rho_{21}^{(2)}$	1.00	567.80	1.01	489.11	1.01	490.01	1.01	493.57	1.01	573.17
$\sqrt{\tau_{11}^{(3)}}$					1.01	745.25	1.01	1270.43	1.00	1251.81
$\sqrt{\tau_{22}^{(3)}}$					1.00	1096.29	1.00	1319.30	1.00	1625.57
$\rho_{21}^{(3)}$					1.00	1258.03	1.00	1194.89	1.00	1396.82
$\sqrt{\sigma}$	1.00	1967.68	1.00	991.38	1.00	1144.23	1.00	1254.02	1.00	1183.33

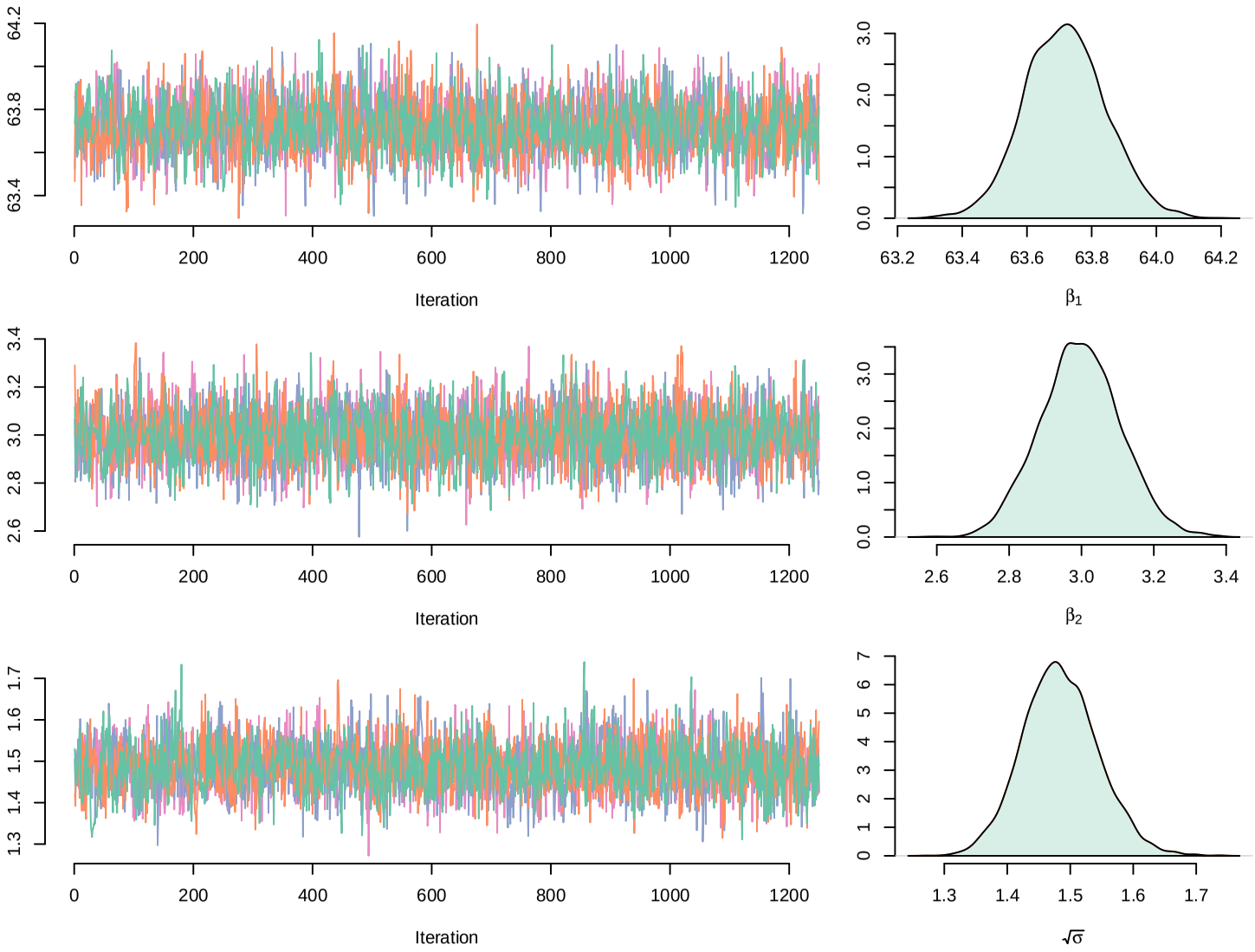


Figure B.1. Trace and density plots for the linear growth model fixed effects

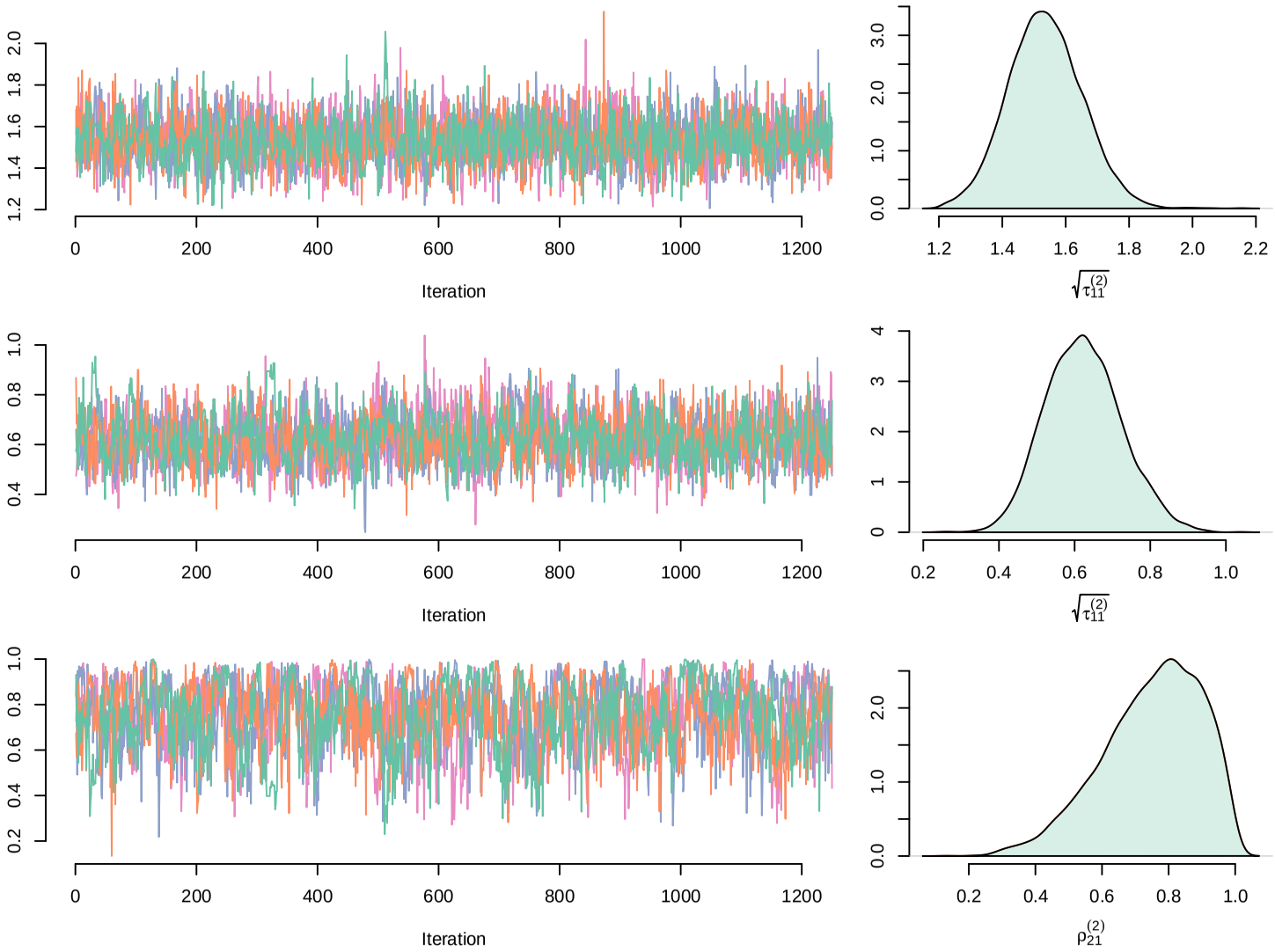


Figure B.2. Trace and density plots for the linear growth model student-specific variance components

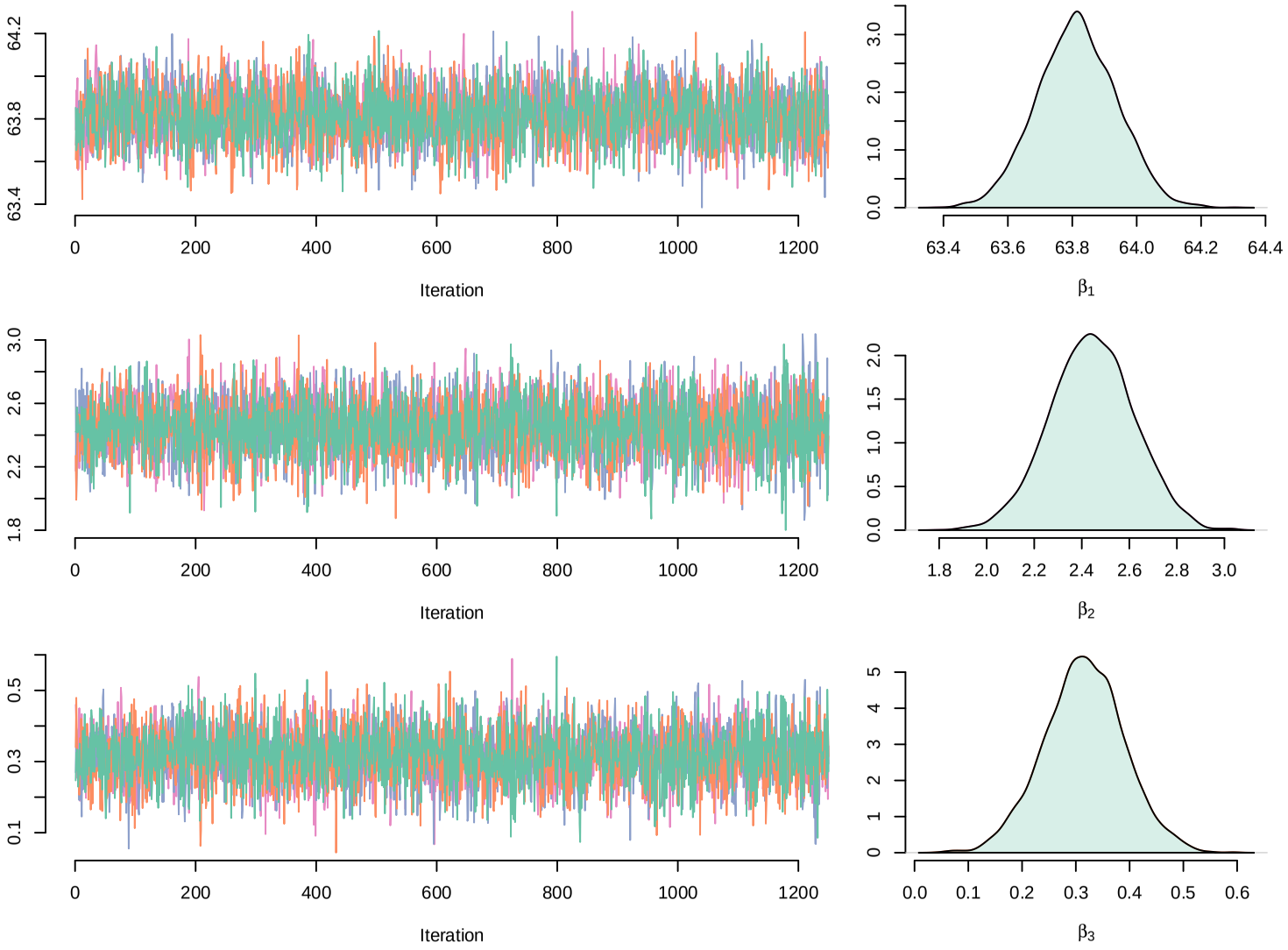


Figure B.3. Trace and density plots for the quadratic growth fixed effects

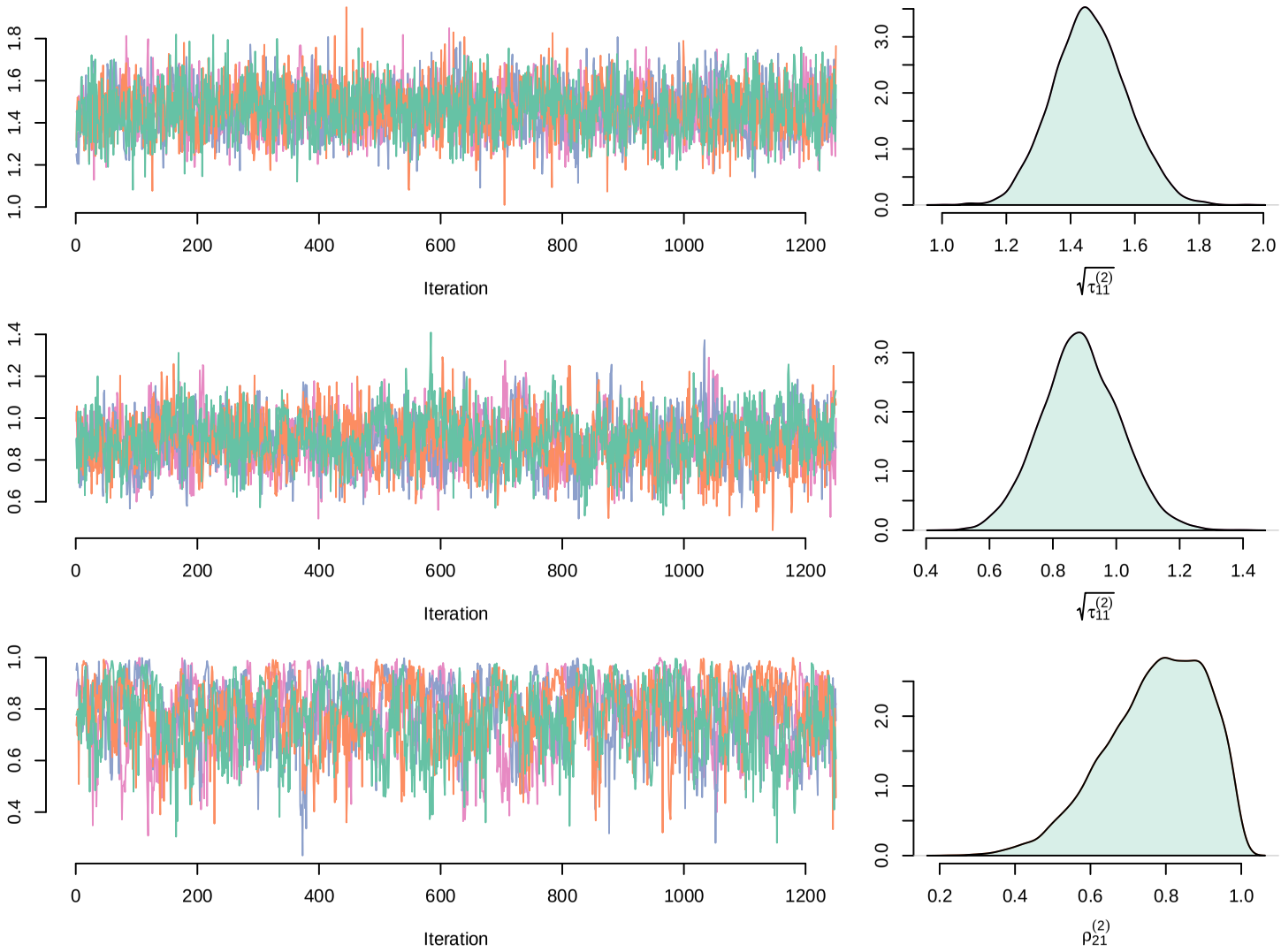


Figure B.4. Trace and density plots for the quadratic growth model student-specific variance components

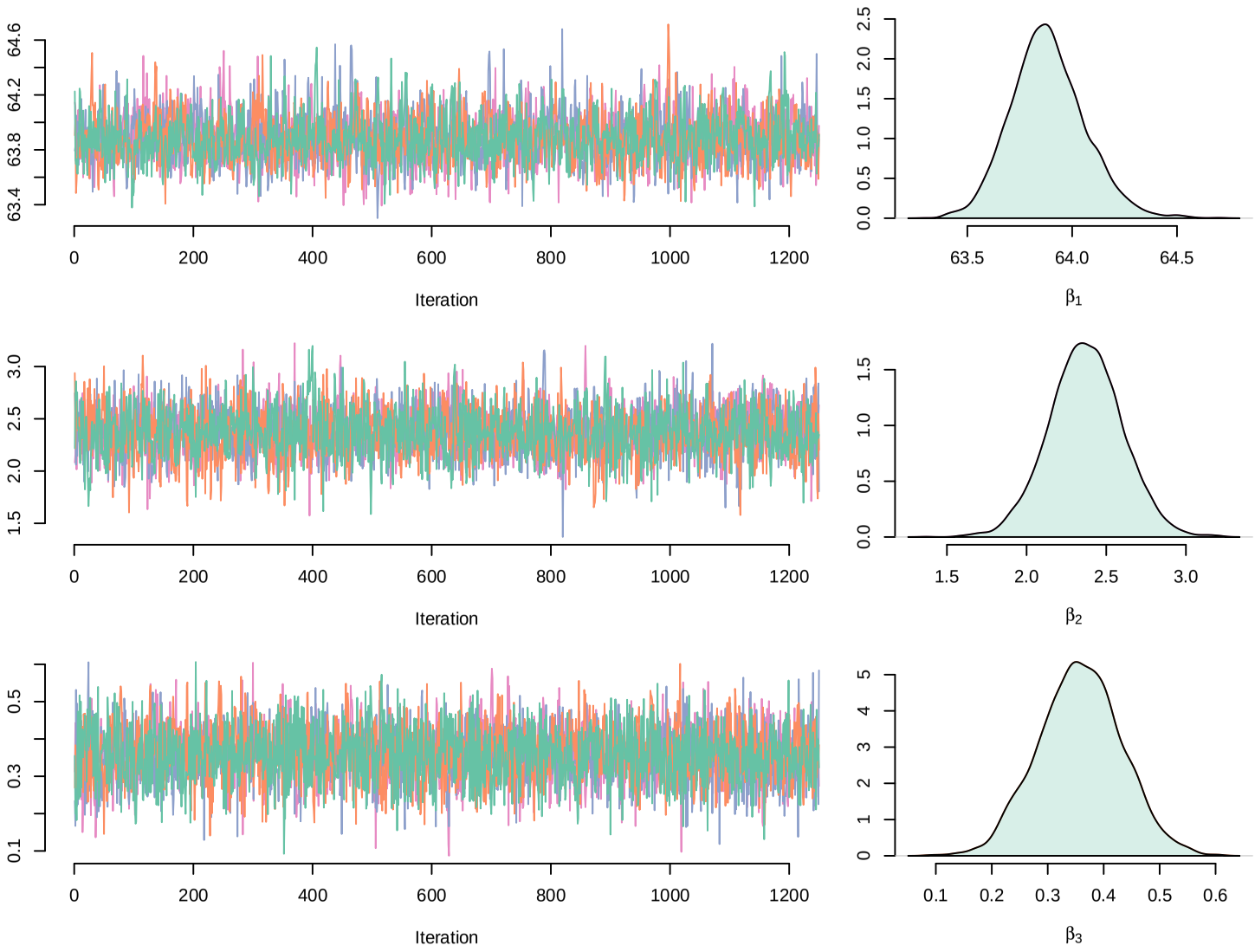


Figure B.5. Trace and density plots for the school-level model fixed effects

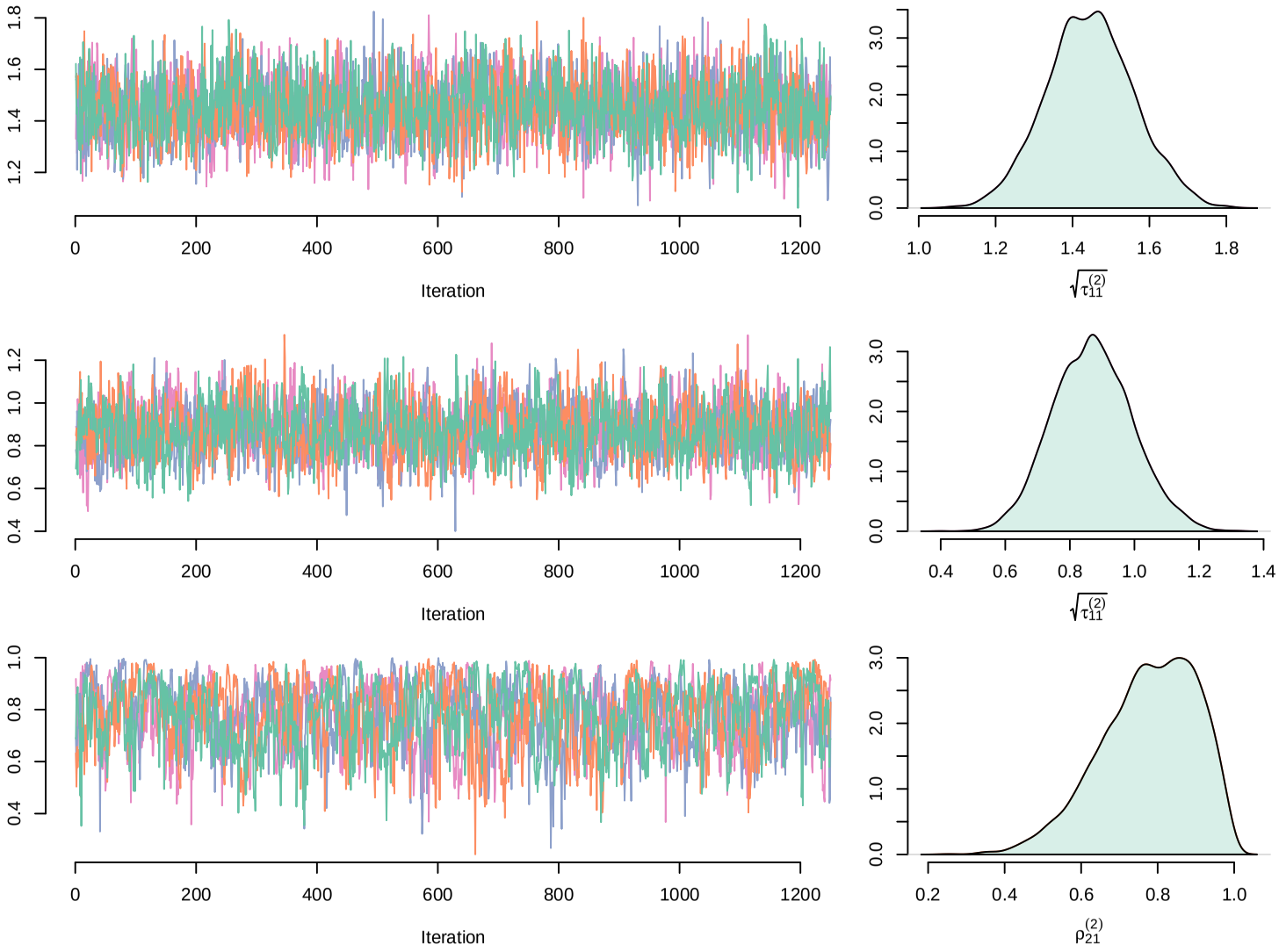


Figure B.6. Trace and density plots for the school-level model student-specific variance components

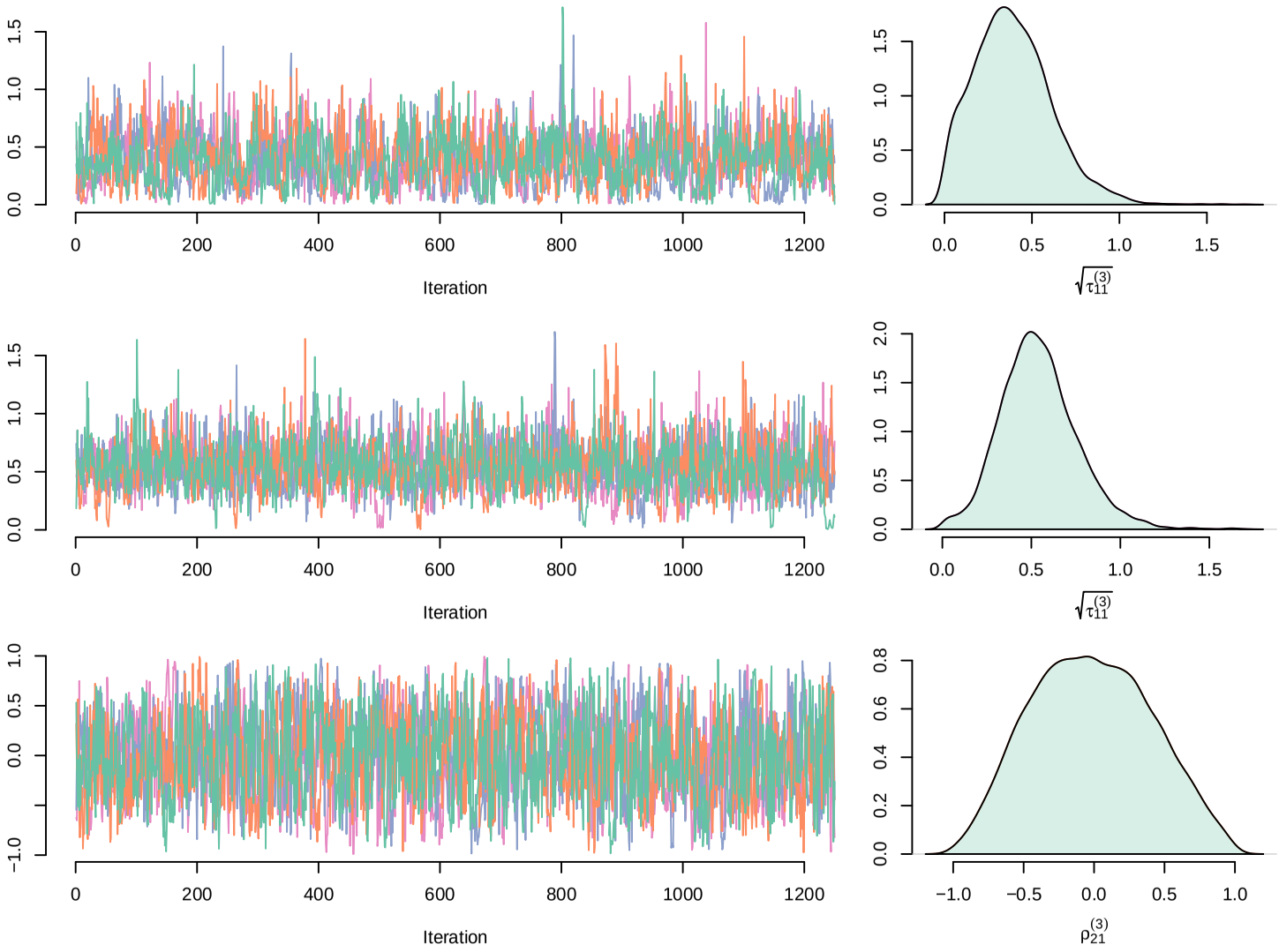


Figure B.7. Trace and density plots for the school-level model school-specific variance components

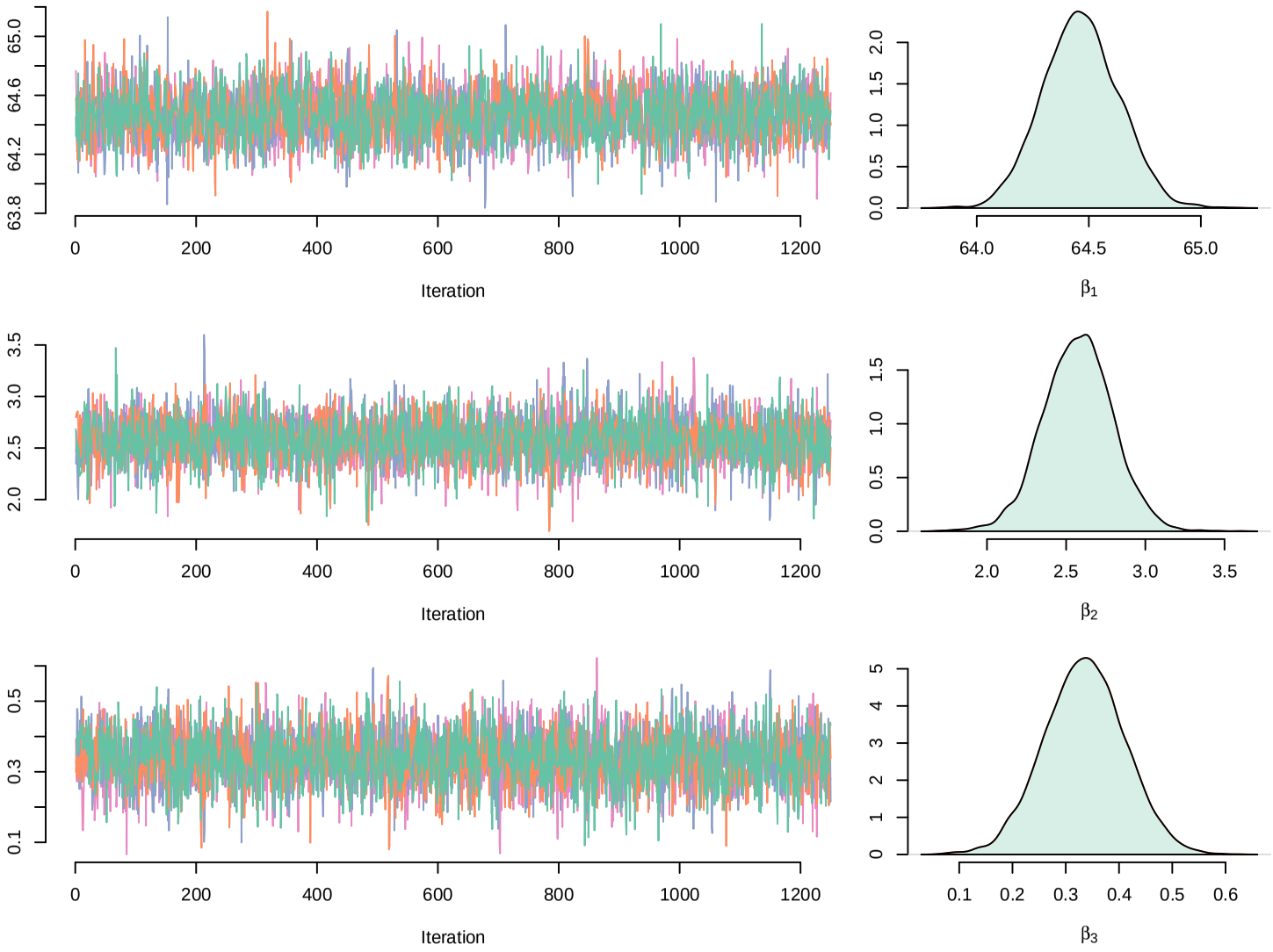


Figure B.8. Trace and density plots for the school-level model with covariates fixed effects, 1 of 2

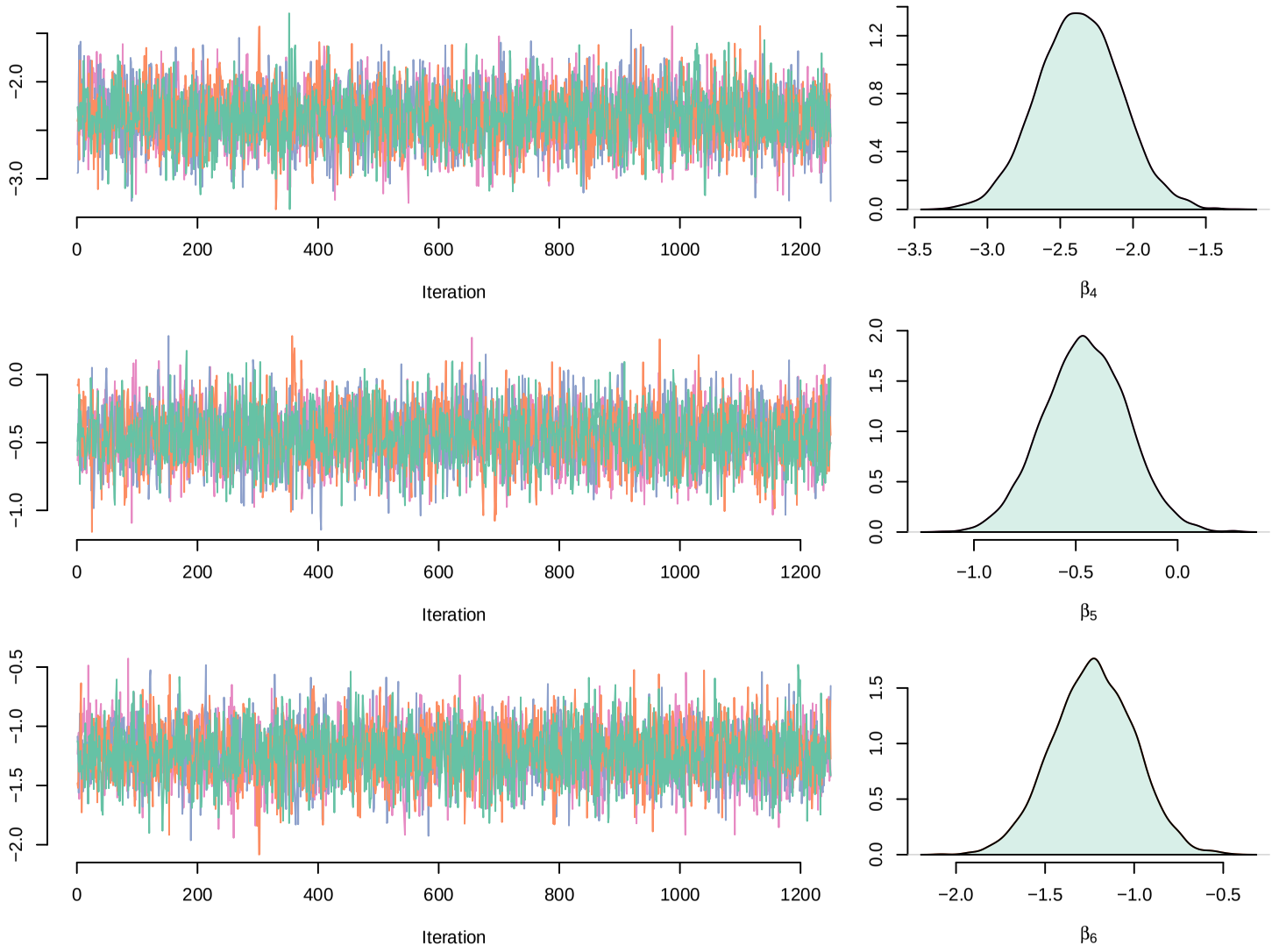


Figure B.9. Trace and density plots for the school-level model with covariates fixed effects, 2 of 2

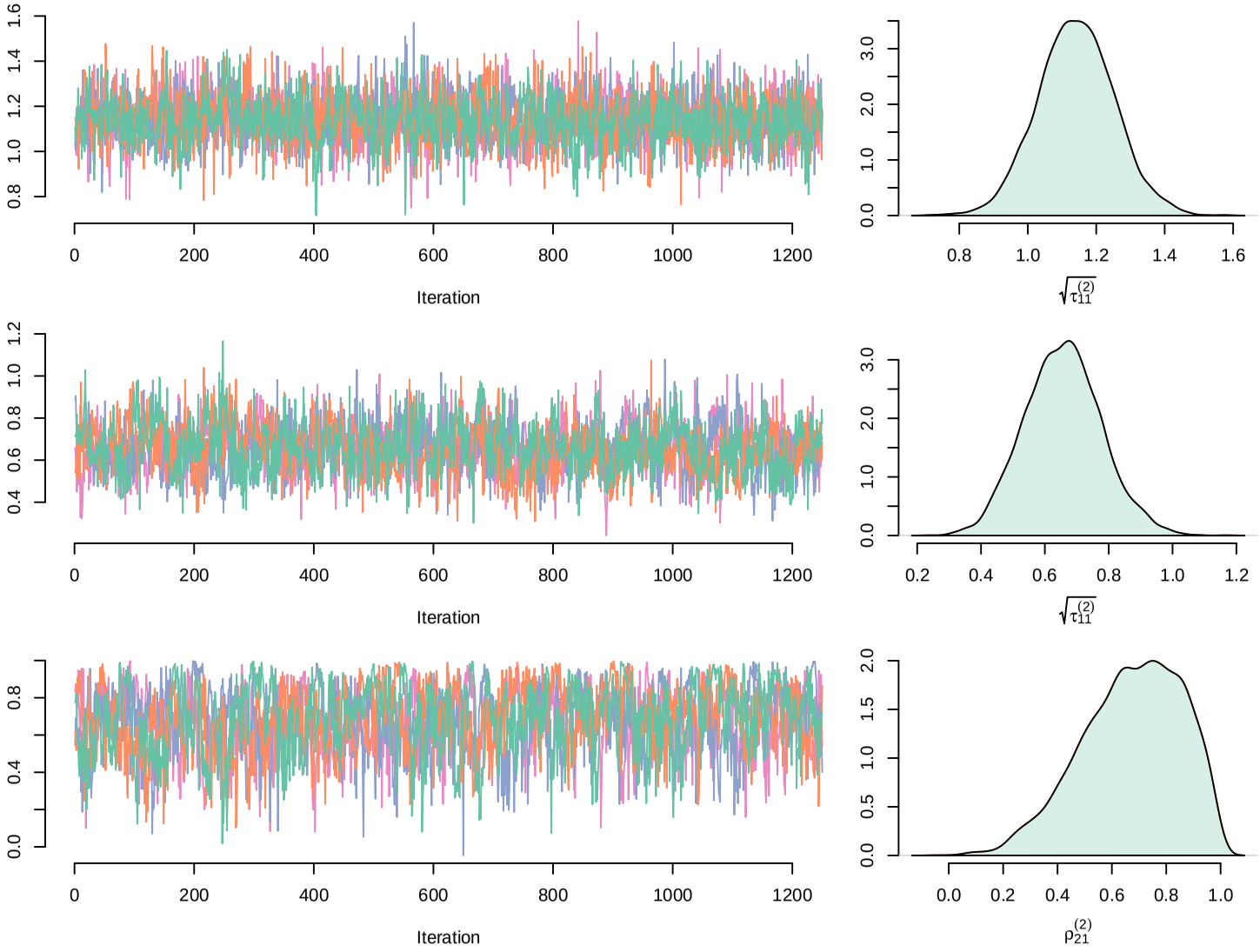


Figure B.10. Trace and density plots for the school-level model with covariates student-specific variance components

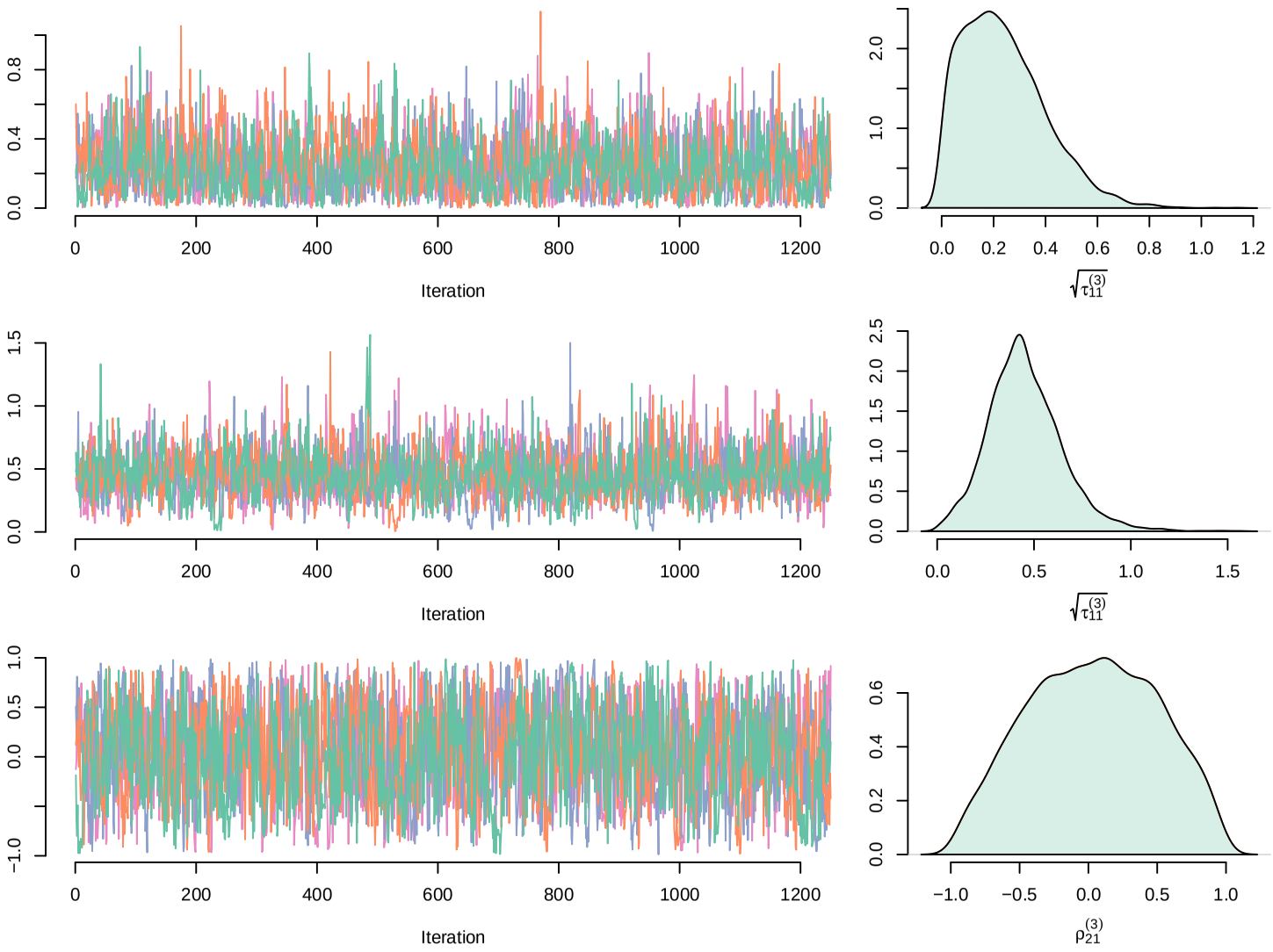


Figure B.11. Trace and density plots for the school-level model with covariates school-specific variance components

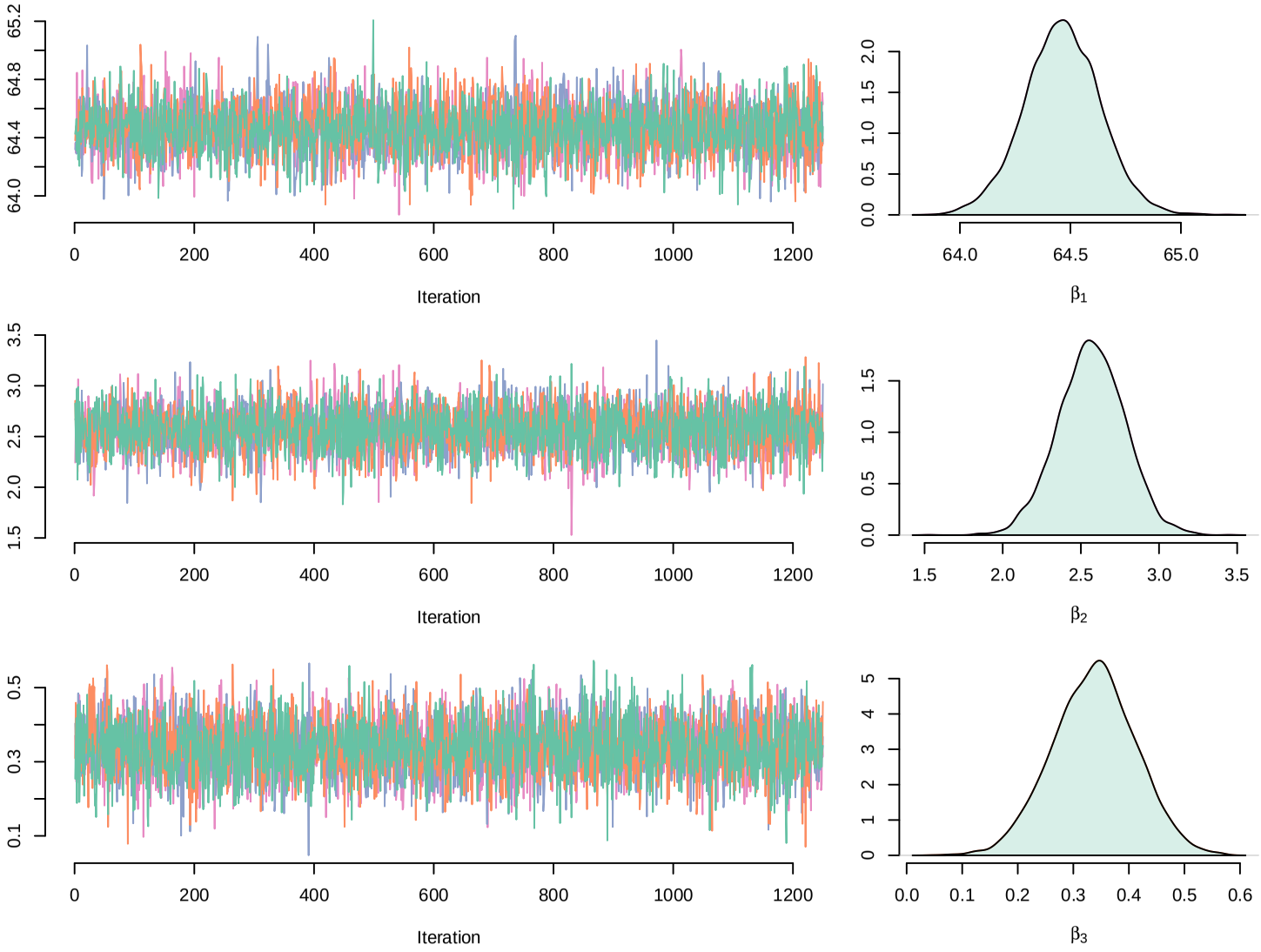


Figure B.12. Trace and density plots for the weakly informative prior model fixed effects, 1 of 2

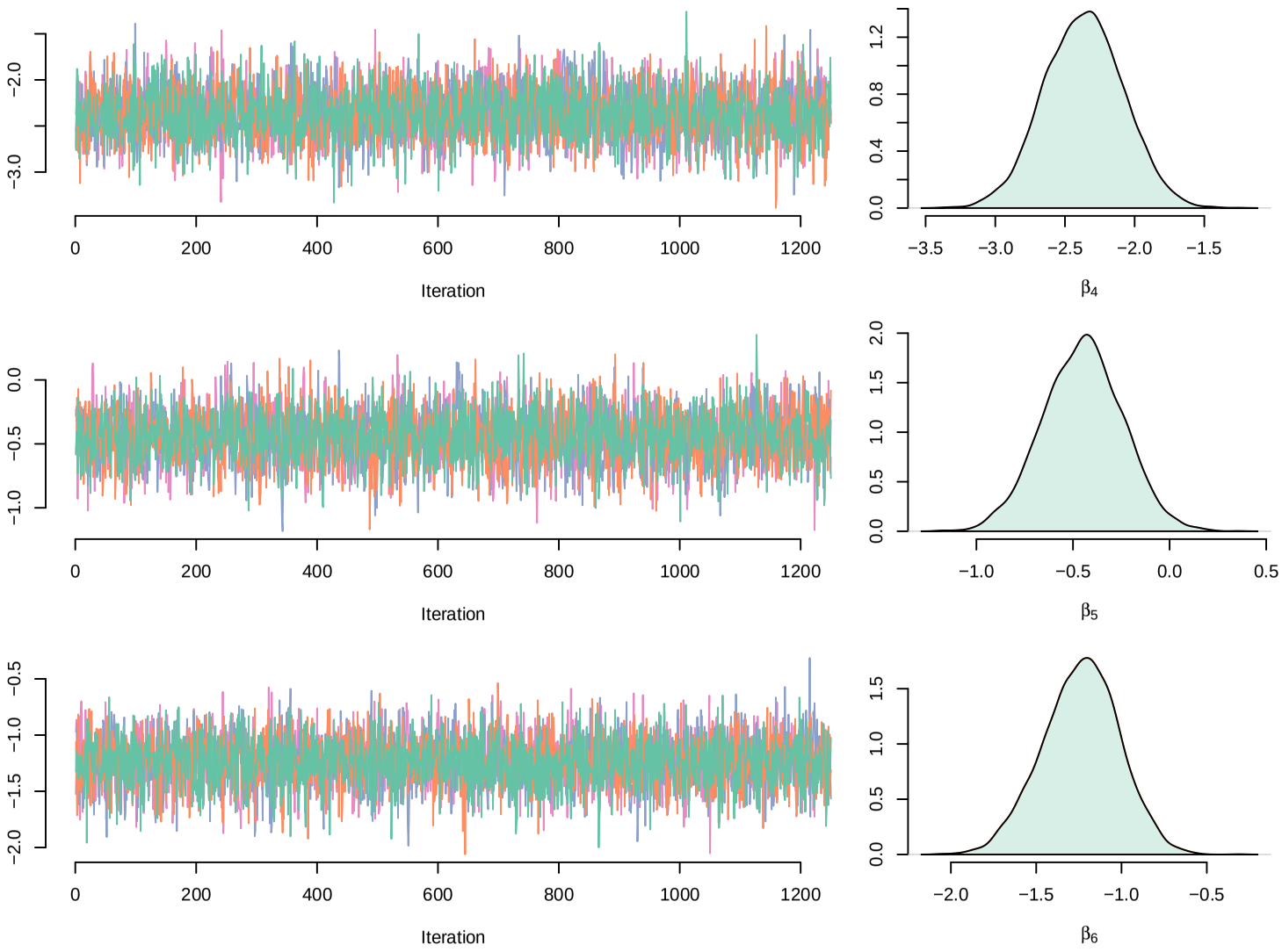


Figure B.13. Trace and density plots for the weakly informative prior model fixed effects, 2 of 2

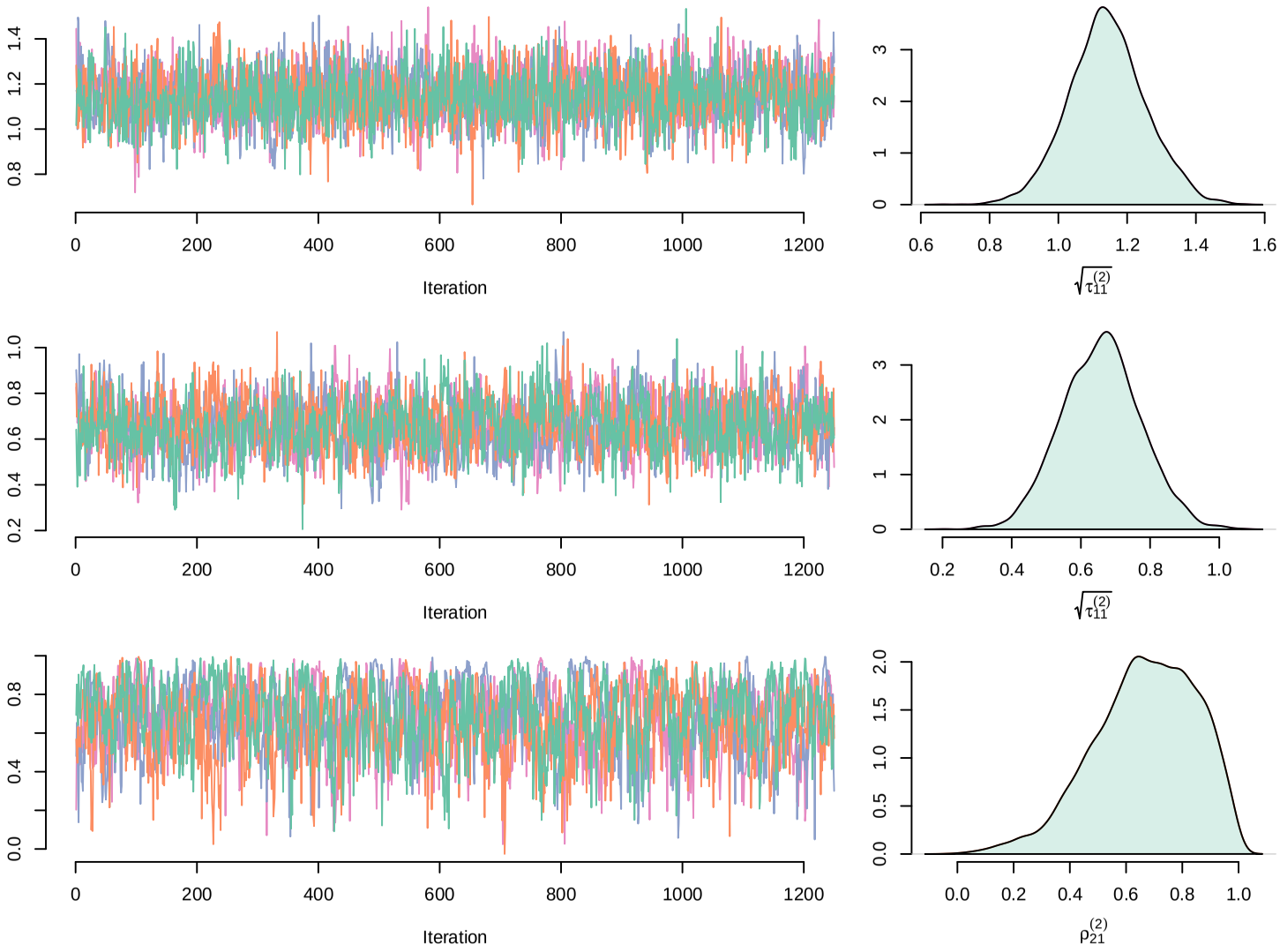


Figure B.14. Trace and density plots for the weakly informative prior model student-specific variance components

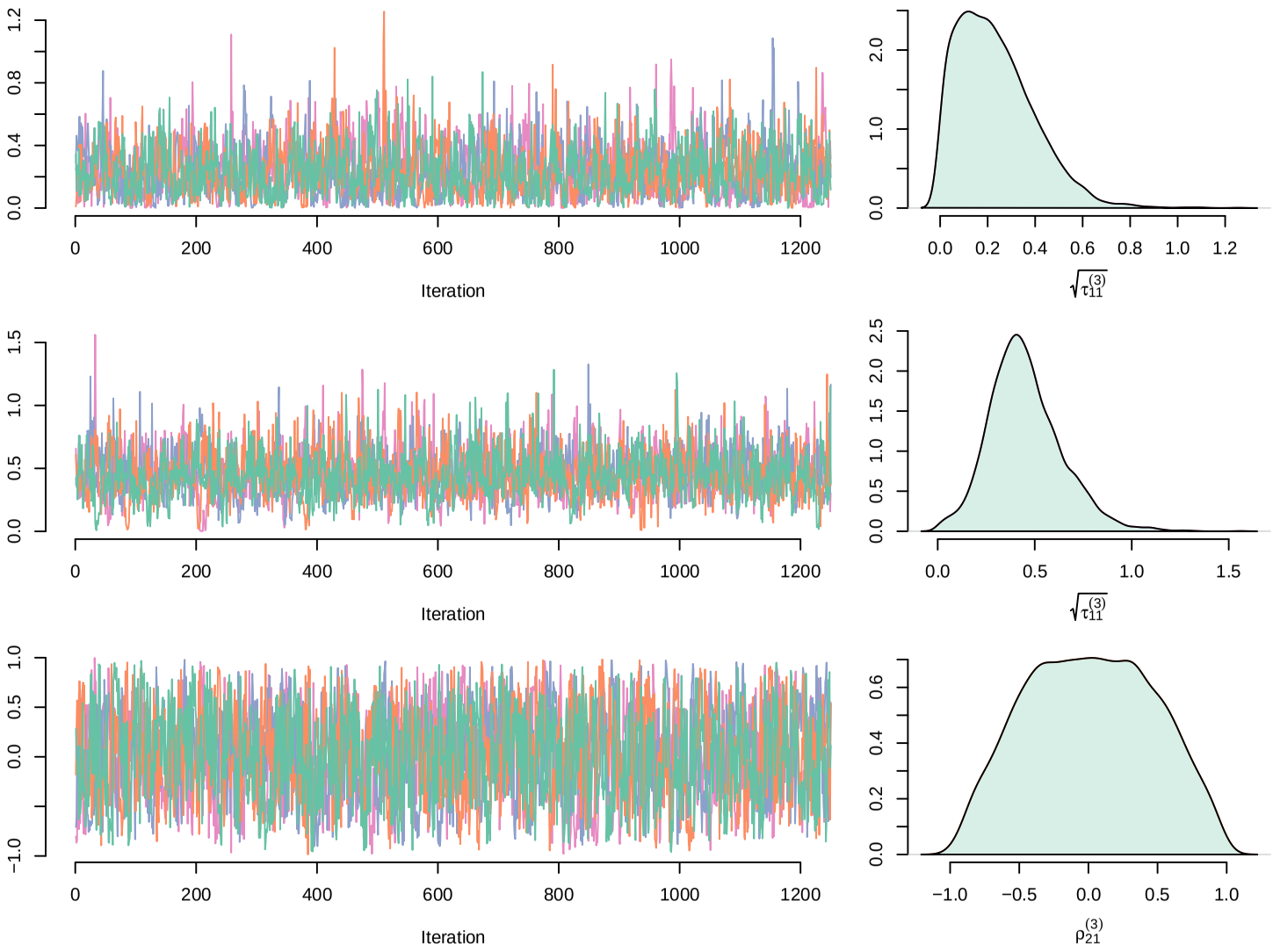


Figure B.15. Trace and density plots for the weakly informative prior model school-specific variance components

APPENDIX C

FULL RESULTS

Table C.1.
Parameter estimates for the reading English language proficiency growth models

	Model 1			Model 2			Model 3			Model 4			Model 5		
	M	95% *HPD*		*M*	95% *HPD*		*M*	95% *HPD*		*M*	95% *HPD*		*M*	95% *HPD*	
Fixed Effects															
β_1 [Initial Status]	63.31	[62.97,	63.66]	63.23	[62.86,	63.59]	63.35	[62.92,	63.80]	63.98	[63.50,	64.48]	63.98	[63.51,	64.50]
β_2 [Grade]	3.28	[2.97,	3.59]	3.71	[3.19,	4.23]	3.25	[2.78,	3.72]	3.36	[2.89,	3.82]	3.36	[2.89,	3.81]
β_3 [Grade^2]				−0.22	[−0.43,	0.00]									
β_4 [SWDe]										−3.47	[−4.33,	−2.63]	−3.47	[−4.29,	−2.63]
β_5 [Grade] × [SWDe]										−0.18	[−0.75,	0.40]	−0.18	[−0.75,	0.40]
β_6 [Male]										−0.64	[−1.17,	−0.10]	−0.63	[−1.18,	−0.09]
Variance Components															
$\sqrt{\tau_{11}^{(2)}}$	2.13	[1.76,	2.50]	2.18	[1.82,	2.54]	2.13	[1.78,	2.49]	1.69	[1.35,	2.06]	1.69	[1.33,	2.04]
$\sqrt{\tau_{22}^{(2)}}$	0.91	[0.55,	1.26]	0.77	[0.37,	1.17]	0.78	[0.43,	1.15]	0.64	[0.23,	1.04]	0.66	[0.25,	1.06]
$\tau_{21}^{(2)}/\sqrt{\tau_{11}^{(2)}\tau_{22}^{(2)}}$	0.31	[−0.10,	0.78]	0.25	[−0.22,	0.80]	0.39	[−0.06,	0.87]	0.28	[−0.25,	0.90]	0.26	[−0.27,	0.85]
$\sqrt{\tau_{11}^{(3)}}$							0.36	[0.01,	1.04]	0.29	[0.01,	0.81]	0.29	[0.01,	0.86]
$\sqrt{\tau_{22}^{(3)}}$							0.60	[0.14,	1.16]	0.55	[0.14,	1.08]	0.55	[0.13,	1.08]
$\tau_{21}^{(3)}/\sqrt{\tau_{11}^{(3)}\tau_{22}^{(3)}}$							0.38	[−0.93,	0.80]	0.39	[−0.93,	0.77]	0.38	[−0.93,	0.78]
$\sqrt{\sigma}$	2.23	[2.04,	2.43]	2.26	[2.07,	2.46]	2.22	[2.03,	2.41]	2.24	[2.05,	2.42]	2.24	[2.05,	2.43]

Model Fit	Estimate	(*SE*)	Estimate	(*SE*)	Estimate	(*SE*)	Estimate	(*SE*)	Estimate	(*SE*)
$\text{elpd}_{\text{psis-loo}}$	−1516.92	(10.05)	−1521.59	(10.37)	−1520.80	(9.26)	−1513.84	(11.00)	−1514.87	(11.10)
p_{loo}	71.04	(3.87)	73.20	(3.99)	62.08	(3.33)	64.91	(3.82)	65.97	(3.95)
LOOIC	3033.84	(20.11)	3043.17	(20.74)	3041.60	(18.52)	3027.67	(22.00)	3029.75	(22.21)

Table C.2.
Parameter estimates for the writing English language proficiency growth models

	Model 1			Model 2			Model 3			Model 4			Model 5		
	M	95% *HPD*		*M*	95% *HPD*		*M*	95% *HPD*		*M*	95% *HPD*		*M*	95% *HPD*	
Fixed Effects															
β_1 [Initial Status]	63.95	[63.57,	64.33]	64.04	[63.65,	64.42]	64.12	[63.66,	64.65]	65.35	[64.85,	65.90]	65.33	[64.82,	65.83]
β_2 [Grade]	3.41	[3.09,	3.73]	2.98	[2.44,	3.53]	2.84	[2.20,	3.47]	3.06	[2.42,	3.68]	3.05	[2.42,	3.68]
β_3 [Grade^2]				0.21	[0.00,	0.42]	0.26	[0.05,	0.47]	0.26	[0.05,	0.46]	0.26	[0.05,	0.47]
β_4 [SWDe]										−3.73	[−4.58,	−2.89]	−3.70	[−4.56,	−2.86]
β_5 [Grade] × [SWDe]										−1.27	[−1.86,	−0.67]	−1.24	[−1.83,	−0.65]
β_6 [Male]										−1.18	[−1.82,	−0.53]	−1.18	[−1.81,	−0.52]
Variance Components															
$\sqrt{\tau_{11}^{(2)}}$	2.46	[2.08,	2.85]	2.43	[2.05,	2.81]	2.38	[2.02,	2.74]	1.82	[1.42,	2.20]	1.80	[1.38,	2.18]
$\sqrt{\tau_{22}^{(2)}}$	1.26	[0.93,	1.62]	1.33	[0.98,	1.67]	1.19	[0.84,	1.54]	1.05	[0.69,	1.40]	1.02	[0.63,	1.40]
$\tau_{21}^{(2)}/\sqrt{\tau_{11}^{(2)}\tau_{22}^{(2)}}$	0.11	[−0.21,	0.49]	0.15	[−0.16,	0.55]	0.29	[−0.05,	0.72]	0.04	[−0.33,	0.53]	0.07	[−0.33,	0.67]
$\sqrt{\tau_{11}^{(3)}}$							0.47	[0.03,	1.12]	0.41	[0.04,	0.93]	0.41	[0.03,	0.90]
$\sqrt{\tau_{22}^{(3)}}$							0.65	[0.27,	1.18]	0.65	[0.27,	1.16]	0.66	[0.29,	1.16]
$\tau_{21}^{(3)}/\sqrt{\tau_{11}^{(3)}\tau_{22}^{(3)}}$							0.34	[−0.93,	0.62]	0.28	[−0.96,	0.51]	0.28	[−0.95,	0.53]
$\sqrt{\sigma}$	2.19	[1.99,	2.40]	2.16	[1.96,	2.38]	2.15	[1.97,	2.36]	2.15	[1.96,	2.36]	2.17	[1.97,	2.38]

Model Fit	Estimate	(SE)	Estimate	(SE)	Estimate	(SE)	Estimate	(SE)	Estimate	(SE)
$\text{elpd}_{\text{psis-loo}}$	−1514.52	(8.63)	−1509.89	(8.52)	−1516.84	(7.97)	−1500.41	(9.32)	−1503.10	(9.45)
p_{loo}	66.60	(3.53)	65.94	(3.54)	55.93	(3.04)	62.21	(3.50)	62.34	(3.55)
LOOIC	3029.04	(17.27)	3019.78	(17.04)	3033.68	(15.94)	3000.82	(18.63)	3006.21	(18.90)

Table C.3.
Parameter estimates for the oral English language proficiency growth models

	Model 1			Model 2			Model 3			Model 4			Model 5		
	M	95% *HPD*		*M*	95% *HPD*		*M*	95% *HPD*		*M*	95% *HPD*		*M*	95% *HPD*	
Fixed Effects															
β_1 [Initial Status]	64.09	[63.77,	64.40]	64.22	[63.91,	64.52]	64.31	[63.74,	64.91]	64.53	[63.89,	65.15]	64.53	[63.88,	65.18]
β_2 [Grade]	2.75	[2.38,	3.11]	1.96	[1.40,	2.53]	1.90	[1.19,	2.62]	2.16	[1.51,	2.79]	2.17	[1.52,	2.78]
β_3 [Grade^2]				0.49	[0.22,	0.74]	0.57	[0.30,	0.82]	0.57	[0.31,	0.83]	0.57	[0.32,	0.82]
β_4 [SWDe]										−1.41	[−2.18,	−0.68]	−1.42	[−2.18,	−0.67]
β_5 [Grade] × [SWDe]										−0.08	[−0.63,	0.46]	−0.07	[−0.60,	0.45]
β_6 [Male]										−1.65	[−2.49,	−0.85]	−1.67	[−2.55,	−0.85]
Variance Components															
$\sqrt{\tau_{11}^{(2)}}$	1.30	[0.93,	1.68]	1.17	[0.80,	1.52]	0.98	[0.63,	1.34]	0.88	[0.51,	1.25]	0.89	[0.52,	1.26]
$\sqrt{\tau_{22}^{(2)}}$	0.90	[0.52,	1.30]	1.43	[0.96,	1.90]	1.59	[1.11,	2.09]	1.37	[0.90,	1.86]	1.38	[0.91,	1.87]
$\tau_{21}^{(2)}/\sqrt{\tau_{11}^{(2)}\tau_{22}^{(2)}}$	0.80	[0.45,	0.98]	0.84	[0.56,	0.99]	0.83	[0.52,	0.98]	0.80	[0.42,	0.98]	0.78	[0.42,	0.98]
$\sqrt{\tau_{11}^{(3)}}$							0.97	[0.49,	1.65]	0.91	[0.45,	1.62]	0.91	[0.46,	1.53]
$\sqrt{\tau_{22}^{(3)}}$							0.65	[0.03,	1.52]	0.50	[0.03,	1.25]	0.50	[0.03,	1.29]
$\tau_{21}^{(3)}/\sqrt{\tau_{11}^{(3)}\tau_{22}^{(3)}}$							0.43	[−0.81,	0.69]	0.44	[−0.81,	0.75]	0.45	[−0.80,	0.75]
$\sqrt{\sigma}$	2.57	[2.38,	2.77]	2.40	[2.23,	2.60]	2.32	[2.13,	2.53]	2.33	[2.15,	2.54]	2.33	[2.15,	2.54]

Model Fit	Estimate	(*SE*)	Estimate	(*SE*)	Estimate	(*SE*)	Estimate	(*SE*)	Estimate	(*SE*)
$\text{elpd}_{\text{psis-loo}}$	−1573.72	(14.66)	−1539.40	(12.54)	−1531.94	(10.23)	−1528.71	(11.13)	−1528.34	(11.10)
p_{loo}	71.51	(5.31)	63.53	(4.73)	45.34	(3.23)	47.42	(3.61)	47.22	(3.62)
LOOIC	3147.44	(29.32)	3078.79	(25.08)	3063.89	(20.45)	3057.42	(22.26)	3056.69	(22.21)

Table C.4.
Parameter estimates for the multivariate reading, writing, and oral English language proficiency growth models

	Model 1		Model 2		Model 3	
	M	95% *HPD*	*M*	95% *HPD*	*M*	95% *HPD*
Fixed Effects						
Reading						
β_{R1}, [Initial Status]	63.26	[62.87, 63.67]	63.91	[63.42, 64.39]	63.89	[63.42, 64.36]
β_{R2}, [Grade]	3.46	[3.03, 3.89]	3.57	[3.13, 4.01]	3.58	[3.14, 4.01]
β_{R4}, [SWDe]			−3.50	[−4.33,−2.66]	−3.46	[−4.30,−2.62]
β_{R5}, [Grade]×[SWDe]			−0.16	[−0.74, 0.44]	−0.16	[−0.73, 0.41]
β_{R6}, [Male]			−0.73	[−1.28,−0.17]	−0.74	[−1.29,−0.19]
Writing						
β_{W1}, [Initial Status]	64.07	[63.64, 64.53]	65.34	[64.83, 65.86]	65.33	[64.81, 65.85]
β_{W2}, [Grade]	2.92	[2.30, 3.53]	3.10	[2.49, 3.70]	3.12	[2.53, 3.72]
β_{W3}, [Grade^2]	0.28	[0.07, 0.49]	0.28	[0.07, 0.49]	0.28	[0.07, 0.48]
β_{W4}, [SWDe]			−3.71	[−4.57,−2.86]	−3.68	[−4.54,−2.81]
β_{W5}, [Grade]×[SWDe]			−1.31	[−1.89,−0.71]	−1.31	[−1.91,−0.72]
β_{W6}, [Male]			−1.26	[−1.91,−0.59]	−1.26	[−1.93,−0.58]
Oral						
β_{O1}, [Initial Status]	64.22	[63.63, 64.82]	64.51	[63.85, 65.15]	64.51	[63.86, 65.15]
β_{O2}, [Grade]	1.86	[1.18, 2.53]	2.14	[1.50, 2.75]	2.14	[1.49, 2.79]
β_{O3}, [Grade^2]	0.63	[0.38, 0.88]	0.62	[0.37, 0.87]	0.62	[0.36, 0.87]
β_{O4}, [SWDe]			−1.37	[−2.14,−0.61]	−1.36	[−2.14,−0.59]
β_{O5}, [Grade]×[SWDe]			−0.14	[−0.68, 0.41]	−0.15	[−0.68, 0.41]
β_{O6}, [Male]			−1.75	[−2.57,−0.98]	−1.74	[−2.55,−0.97]

Continued on next page

Table C.4 continued

	Model 1		Model 2		Model 3	
	M	95% HPD	M	95% HPD	M	95% HPD
Variance Components						
Student-level						
$\sqrt{\tau_{11}^{(2)}}$	2.21	[1.90, 2.54]	1.77	[1.46, 2.11]	1.77	[1.44, 2.11]
$\sqrt{\tau_{22}^{(2)}}$	0.90	[0.57, 1.23]	0.79	[0.42, 1.14]	0.80	[0.44, 1.15]
$\sqrt{\tau_{33}^{(2)}}$	2.43	[2.10, 2.77]	1.87	[1.53, 2.21]	1.87	[1.53, 2.22]
$\sqrt{\tau_{44}^{(2)}}$	1.31	[0.99, 1.65]	1.13	[0.78, 1.47]	1.13	[0.79, 1.49]
$\sqrt{\tau_{55}^{(2)}}$	1.10	[0.78, 1.43]	0.99	[0.66, 1.33]	0.99	[0.66, 1.33]
$\sqrt{\tau_{55}^{(2)}}$	1.52	[1.06, 1.99]	1.32	[0.87, 1.80]	1.33	[0.87, 1.81]
$\tau_{21}^{(2)}/\sqrt{\tau_{11}^{(2)}\tau_{22}^{(2)}}$	0.14	[−0.18, 0.50]	0.01	[−0.34, 0.45]	0.01	[−0.34, 0.43]
$\tau_{31}^{(2)}/\sqrt{\tau_{11}^{(2)}\tau_{33}^{(2)}}$	0.87	[0.76, 0.96]	0.82	[0.66, 0.94]	0.82	[0.66, 0.94]
$\tau_{41}^{(2)}/\sqrt{\tau_{11}^{(2)}\tau_{44}^{(2)}}$	0.23	[−0.05, 0.52]	0.06	[−0.26, 0.39]	0.06	[−0.26, 0.39]
$\tau_{51}^{(2)}/\sqrt{\tau_{11}^{(2)}\tau_{55}^{(2)}}$	0.64	[0.39, 0.86]	0.54	[0.24, 0.80]	0.54	[0.22, 0.80]
$\tau_{61}^{(2)}/\sqrt{\tau_{11}^{(2)}\tau_{66}^{(2)}}$	0.49	[0.25, 0.70]	0.34	[0.04, 0.61]	0.34	[0.05, 0.61]
$\tau_{32}^{(2)}/\sqrt{\tau_{22}^{(2)}\tau_{33}^{(2)}}$	0.14	[−0.17, 0.46]	0.04	[−0.31, 0.42]	0.04	[−0.31, 0.41]
$\tau_{42}^{(2)}/\sqrt{\tau_{22}^{(2)}\tau_{44}^{(2)}}$	0.41	[0.06, 0.71]	0.30	[−0.11, 0.66]	0.31	[−0.11, 0.67]
$\tau_{52}^{(2)}/\sqrt{\tau_{22}^{(2)}\tau_{55}^{(2)}}$	0.38	[−0.01, 0.73]	0.34	[−0.12, 0.73]	0.34	[−0.12, 0.73]
$\tau_{62}^{(2)}/\sqrt{\tau_{22}^{(2)}\tau_{66}^{(2)}}$	0.54	[0.19, 0.81]	0.48	[0.08, 0.80]	0.48	[0.08, 0.80]
$\tau_{43}^{(2)}/\sqrt{\tau_{33}^{(2)}\tau_{44}^{(2)}}$	0.17	[−0.12, 0.49]	−0.04	[−0.34, 0.35]	−0.05	[−0.35, 0.33]
$\tau_{53}^{(2)}/\sqrt{\tau_{33}^{(2)}\tau_{55}^{(2)}}$	0.73	[0.50, 0.90]	0.66	[0.39, 0.87]	0.66	[0.39, 0.87]
$\tau_{63}^{(2)}/\sqrt{\tau_{33}^{(2)}\tau_{66}^{(2)}}$	0.53	[0.31, 0.73]	0.41	[0.14, 0.66]	0.42	[0.14, 0.67]
$\tau_{54}^{(2)}/\sqrt{\tau_{44}^{(2)}\tau_{55}^{(2)}}$	0.15	[−0.22, 0.52]	0.00	[−0.41, 0.42]	−0.01	[−0.42, 0.40]
$\tau_{64}^{(2)}/\sqrt{\tau_{44}^{(2)}\tau_{66}^{(2)}}$	0.06	[−0.27, 0.39]	−0.12	[−0.46, 0.26]	−0.13	[−0.47, 0.24]
$\tau_{65}^{(2)}/\sqrt{\tau_{55}^{(2)}\tau_{66}^{(2)}}$	0.70	[0.40, 0.91]	0.63	[0.28, 0.89]	0.63	[0.28, 0.89]

Continued on next page

Table C.4 continued

	Model 1		Model 2		Model 3	
	M	95% HPD	M	95% HPD	M	95% HPD
Variance Components						
School-level						
$\sqrt{\tau_{11}^{(3)}}$	0.28	[0.01, 0.81]	0.26	[0.01, 0.74]	0.25	[0.01, 0.73]
$\sqrt{\tau_{22}^{(3)}}$	0.51	[0.14, 0.94]	0.50	[0.17, 0.91]	0.50	[0.17, 0.92]
$\sqrt{\tau_{33}^{(3)}}$	0.38	[0.02, 0.92]	0.36	[0.02, 0.85]	0.37	[0.02, 0.85]
$\sqrt{\tau_{44}^{(3)}}$	0.57	[0.17, 1.05]	0.61	[0.24, 1.08]	0.60	[0.26, 1.06]
$\sqrt{\tau_{55}^{(3)}}$	0.97	[0.49, 1.65]	0.93	[0.46, 1.57]	0.92	[0.47, 1.56]
$\sqrt{\tau_{55}^{(3)}}$	0.60	[0.04, 1.33]	0.50	[0.03, 1.18]	0.51	[0.04, 1.22]
$\tau_{21}^{(3)}/\sqrt{\tau_{11}^{(3)}\tau_{22}^{(3)}}$	−0.16	[−0.76, 0.57]	−0.15	[−0.75, 0.57]	−0.14	[−0.75, 0.56]
$\tau_{31}^{(3)}/\sqrt{\tau_{11}^{(3)}\tau_{33}^{(3)}}$	0.06	[−0.63, 0.72]	0.03	[−0.64, 0.69]	0.03	[−0.65, 0.69]
$\tau_{41}^{(3)}/\sqrt{\tau_{11}^{(3)}\tau_{44}^{(3)}}$	0.03	[−0.64, 0.66]	0.05	[−0.62, 0.68]	0.04	[−0.64, 0.68]
$\tau_{51}^{(3)}/\sqrt{\tau_{11}^{(3)}\tau_{55}^{(3)}}$	0.02	[−0.64, 0.66]	−0.03	[−0.67, 0.62]	−0.04	[−0.67, 0.64]
$\tau_{61}^{(3)}/\sqrt{\tau_{11}^{(3)}\tau_{66}^{(3)}}$	0.00	[−0.68, 0.66]	0.02	[−0.65, 0.67]	0.02	[−0.66, 0.68]
$\tau_{32}^{(3)}/\sqrt{\tau_{22}^{(3)}\tau_{33}^{(3)}}$	−0.01	[−0.65, 0.61]	−0.07	[−0.69, 0.57]	−0.07	[−0.67, 0.56]
$\tau_{42}^{(3)}/\sqrt{\tau_{22}^{(3)}\tau_{44}^{(3)}}$	0.40	[−0.22, 0.85]	0.41	[−0.19, 0.84]	0.41	[−0.19, 0.84]
$\tau_{52}^{(3)}/\sqrt{\tau_{22}^{(3)}\tau_{55}^{(3)}}$	0.27	[−0.34, 0.76]	0.28	[−0.30, 0.76]	0.29	[−0.29, 0.77]
$\tau_{62}^{(3)}/\sqrt{\tau_{22}^{(3)}\tau_{66}^{(3)}}$	0.28	[−0.42, 0.80]	0.24	[−0.45, 0.77]	0.24	[−0.45, 0.77]
$\tau_{43}^{(3)}/\sqrt{\tau_{33}^{(3)}\tau_{44}^{(3)}}$	−0.22	[−0.77, 0.49]	−0.30	[−0.80, 0.39]	−0.29	[−0.81, 0.40]
$\tau_{53}^{(3)}/\sqrt{\tau_{33}^{(3)}\tau_{55}^{(3)}}$	−0.06	[−0.68, 0.57]	−0.17	[−0.73, 0.49]	−0.17	[−0.75, 0.49]
$\tau_{63}^{(3)}/\sqrt{\tau_{33}^{(3)}\tau_{66}^{(3)}}$	0.05	[−0.61, 0.68]	0.06	[−0.64, 0.67]	0.06	[−0.61, 0.68]
$\tau_{54}^{(3)}/\sqrt{\tau_{44}^{(3)}\tau_{55}^{(3)}}$	0.23	[−0.36, 0.74]	0.21	[−0.36, 0.71]	0.22	[−0.36, 0.73]
$\tau_{64}^{(3)}/\sqrt{\tau_{44}^{(3)}\tau_{66}^{(3)}}$	0.12	[−0.53, 0.69]	0.10	[−0.54, 0.68]	0.10	[−0.53, 0.68]
$\tau_{65}^{(3)}/\sqrt{\tau_{55}^{(3)}\tau_{66}^{(3)}}$	−0.11	[−0.66, 0.52]	−0.08	[−0.65, 0.57]	−0.08	[−0.66, 0.56]

Continued on next page

Table C.4 continued

	Model 1		Model 2		Model 3	
	M	95% *HPD*	*M*	95% *HPD*	*M*	95% *HPD*
Residual Variances						
$\sqrt{\sigma_R}$	2.17	[2.00, 2.35]	2.19	[2.02, 2.37]	2.19	[2.02, 2.37]
$\sqrt{\sigma_W}$	2.09	[1.91, 2.28]	2.11	[1.93, 2.29]	2.10	[1.93, 2.29]
$\sqrt{\sigma_O}$	2.31	[2.12, 2.50]	2.33	[2.14, 2.52]	2.32	[2.14, 2.53]
Model Fit	Estimate	(*SE*)	Estimate	(*SE*)	Estimate	(*SE*)
$\text{elpd}_{\text{psis-loo}}$	−4497.32	(20.51)	−4491.78	(21.47)	−4491.56	(21.57)
p_{loo}	186.89	(7.86)	190.83	(7.83)	191.71	(7.99)
LOOIC	8994.65	(41.02)	8983.56	(42.95)	8983.12	(43.15)

Table C.5.
Parameter estimates for the time-to-reclassification hazard models

	Model 1		Model 2		Model 3	
	M	95% *HPD*	*M*	95% *HPD*	*M*	95% *HPD*
Fixed Effects						
α_1 [Baseline]	−3.89	[−4.82, −3.14]	−3.92	[−4.86, −3.12]	−3.70	[−4.70, −2.83]
α_2 [Grade 4]	−0.07	[−0.32, 0.18]	−0.01	[−0.33, 0.33]	0.46	[0.02, 0.92]
α_3 [Grade 5]	−0.50	[−0.88, −0.11]	−0.39	[−0.83, 0.12]	0.44	[−0.14, 1.03]
α_4 [Grade 6]	0.93	[0.35, 1.55]	1.14	[0.50, 1.81]	3.11	[2.17, 4.16]
α_5 [Grade 7]	−0.45	[−1.88, 0.90]	−0.30	[−1.73, 1.02]	2.66	[1.03, 4.28]
α_6 [Male]					−2.82	[−3.65, −2.11]
α_7 [SWDe]					−0.27	[−0.71, 0.15]
Variance Components						
$\sqrt{\nu}$			0.40	[0.11, 0.78]	0.57	[0.25, 1.00]

Model Fit	Estimate	(*SE*)	Estimate	(*SE*)	Estimate	(*SE*)
$\text{elpd}_{\text{psis-loo}}$	−332.63	(12.95)	−330.21	(13.10)	−290.07	(14.44)
p_{loo}	5.47	(0.62)	12.81	(0.86)	16.92	(1.26)
LOOIC	665.26	(25.90)	660.41	(26.20)	580.15	(28.87)

Table C.6.
Parameter estimates for the time-to-reclassification shared random effects models

	Model 1		Model 2		Model 3	
	M	95% *HPD*	*M*	95% *HPD*	*M*	95% *HPD*
Growth Model						
β_1 [Initial Status]	64.46	[64.12, 64.80]	64.12	[63.81, 64.43]	64.43	[64.11, 64.76]
β_2 [Grade]	2.57	[2.14, 2.96]	2.34	[1.97, 2.70]	2.54	[2.16, 2.89]
β_3 [Grade^2]	0.34	[0.19, 0.48]	0.46	[0.33, 0.59]	0.45	[0.31, 0.58]
β_4 [SWDe]	−2.36	[−2.91,−1.81]	−0.87	[−1.40,−0.38]	−2.20	[−2.74,−1.68]
β_5 [Grade] × [SWDe]	−0.44	[−0.86,−0.04]	−0.29	[−0.60, 0.01]	−0.48	[−0.87,−0.08]
β_6 [Male]	−1.24	[−1.68,−0.82]	−0.32	[−0.69, 0.05]	−1.45	[−1.89,−1.03]
Variance Components						
$\sqrt{\tau_{11}^{(2)}}$	1.14	[0.92, 1.36]	1.26	[1.04, 1.49]	1.17	[0.97, 1.37]
$\sqrt{\tau_{22}^{(2)}}$	0.66	[0.43, 0.91]	0.99	[0.74, 1.24]	0.81	[0.60, 1.03]
$\tau_{21}^{(2)}/\sqrt{\tau_{11}^{(2)}\tau_{22}^{(2)}}$	0.68	[0.28, 0.97]	0.79	[0.53, 0.98]	0.75	[0.43, 0.98]
$\sqrt{\tau_{11}^{(3)}}$	0.24	[0.01, 0.62]	0.25	[0.02, 0.58]	0.21	[0.01, 0.53]
$\sqrt{\tau_{22}^{(3)}}$	0.45	[0.12, 0.87]	0.27	[0.02, 0.61]	0.30	[0.03, 0.65]
$\tau_{21}^{(3)}/\sqrt{\tau_{11}^{(3)}\tau_{22}^{(3)}}$	0.52	[−0.80, 0.86]	0.29	[−0.96, 0.67]	0.32	[−0.96, 0.68]
$\sqrt{\sigma}$	1.38	[1.27, 1.50]	1.30	[1.21, 1.41]	1.31	[1.21, 1.41]

Continued on next page

Table C.6 continued

	Model 1		Model 2		Model 3	
	M	95% *HPD*	*M*	95% *HPD*	*M*	95% *HPD*
Hazard Model						
α_1 [Grade 3]	−3.64	[−4.64,−2.81]	−10.59	[−13.14,−8.35]	−8.66	[−10.95,−6.65]
α_2 [Grade 4]	0.46	[0.01, 0.93]	−0.18	[−1.09, 0.76]	1.31	[0.23, 2.47]
α_3 [Grade 5]	0.43	[−0.15, 1.05]	2.97	[1.85, 4.27]	4.67	[3.21, 6.32]
α_4 [Grade 6]	3.05	[2.11, 4.07]	7.89	[5.93, 10.08]	10.52	[8.15, 13.19]
α_5 [Grade 7]	2.54	[0.92, 4.13]	8.80	[5.85, 11.89]	12.19	[8.87, 15.72]
α_6 [SWDe]	−2.77	[−3.58,−2.05]			−8.35	[−10.96,−6.15]
α_7 [Male]	−0.27	[−0.69, 0.16]			−0.83	[−2.03, 0.32]
λ_1 [Initial Status, $\zeta_{1i}^{(2)}$]			2.13	[1.00, 3.37]	2.14	[1.05, 3.32]
λ_2 [Linear Growth, $\zeta_{2i}^{(2)}$]			3.08	[1.56, 4.73]	3.01	[1.49, 4.86]
$\sqrt{\nu}$	0.56	[0.25, 0.99]	0.54	[0.04, 1.47]	0.73	[0.08, 1.77]
Model Fit	Estimate	(*SE*)	Estimate	(*SE*)	Estimate	(*SE*)
$\text{elpd}_{\text{psis-loo}}$	−1555.99	(29.41)	−1672.89	(32.48)	−1637.02	(31.48)
p_{loo}	37.98	(2.05)	136.48	(7.57)	114.87	(5.68)
LOOIC	3111.99	(58.81)	3345.78	(64.96)	3274.03	(62.96)

REFERENCES CITED

Abedi, J. (2008). Classification system for English language learners: Issues and recommendations. *Educational Measurement: Issues and Practice*, *27*(3), 17–31. doi: 10.1111/j.1745-3992.2008.00125.x

Abedi, J., & Herman, J. (2010). Assessing English language learners' opportunity to learn mathematics : Issues and limitations. *Teacher College Record*, *112*(3), 723–746.

Allison, P. D. (1982). Discrete-time methods for the analysis of event histories. *Sociological Methodology*, *13*(1982), 61–98.

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly Harmless Econometrics : An Empiricist ' s Companion*. Princeton, NJ: Princeton University Press. doi: 10.1017/CBO9781107415324.004

August, D., & Hakuta, K. (1997). *Improving schooling for language-minority children: A research agenda*. Washington, DC: The National Academies Press.

August, D., & Shanahan, T. (2006). *Developing literacy in second-language learners*. Mahwah, NJ: Lawrence Erlbaum Associates.

Bailey, A. L., & Carroll, P. (2015). Assessment of English language learners in the era of new academic content standards. *Review of Research in Education*, *39*(1), 253–294. doi: 10.3102/0091732X14556074

Bailey, A. L., & Huang, B. H. (2011). Do current English language development/proficiency standards reflect the English needed for success in school? *Language Testing*, *28*(3), 343–365. doi: 10.1177/0265532211404187

Bates, D. M., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. , 1–27.

Bauer, D. J., & Curran, P. J. (2010). Probing interactions in fixed and multilevel regression: Inferential and graphical techniques. *Multivariate Behavioral Research*, *40*(3), 331–349. doi: 10.1207/s15327906mbr4003

Beckett, L. A., Tancredi, D. J., & Wilson, R. S. (2004). Multivariate longitudinal models for complex change processes. *Statistics in Medicine*, *23*(2), 231–239. doi: 10.1002/sim.1712

Blozis, S. A. (2004). Structured latent curve models for the study of change in multivariate repeated measures. *Psychological methods*, *9*(3), 334–353. doi: 10.1037/1082-989X.9.3.334

Blozis, S. A. (2007). On fitting nonlinear latent curve models to multiple variables measured longitudinally. *Structural Equation Modeling*, *14*(2), 179–201. doi: 10.1080/10705510709336743

Bock, R. D., & Bargmann, R. E. (1966). Analysis of covariance structures. *Psychometrika*, *31*(4), 81–87.

Bollen, K. A., & Curran, P. J. (2006). *Latent Curve Models: A structural equation perspective*. Hoboken, NJ: Wiley-Interscience. doi: 10.1353/sof.0.0084

Callahan, R. (2005). Tracking and high school English learners: Limiting opportunity to learn. *American Educational Research Journal*, *42*(2), 305–328. doi: 10.3102/00028312042002305

Callahan, R., Wilkinson, L., & Muller, C. (2010). Academic achievement and course taking among language minority uouth in U.S. schools: Effects of ESL placement. *Educational Evaluation and Policy Analysis*, *32*(1), 84–117. doi: 10.3102/0162373709359805

Cheung, A. C. K., & Slavin, R. E. (2012). Effective reading programs for Spanish-dominant English language learners (ELLs) in the elementary grades: A synthesis of research. *Review of Educational Research*, *82*(4), 351–395. doi: 10.3102/0034654312465472

Conboy, B. T., & Thal, D. J. (2006). Ties between the lexicon and grammar: Cross-sectional and longitudinal studies of bilingual toddlers. *Child Development*, *77*(3), 712–735.

Cook, G. H., Boals, T., Wilmes, C., & Santos, M. (2008). *Issues in the development of Annual Measurable Achievement Objectives for WIDA consortium states (WCER Working Paper No. 2008-2)*. Madison, WI: Wisconsin Center for Education Research School.

Council of Chief State School Officers. (2012). *Framework for English language proficiency development standards corresponding to the Common Core State Standards and the Next Generation Science Standards*. Washington, DC: Council of Chief State School Officers.

D'Agostino, R. B., Lee, M. L., Belanger, A. J., Cupples, I. A., Anderson, K., & Kannel, W. B. (1990). Relation of pooled logitstic regression to the time dependent cox regression analysis: The Framingham heart study. *Statistics in Medicine*, *9*, 1501–1515.

De Gruttola, V., & Tu, X. M. (1994). Modelling progression of CD4-lymphocyte count and its relationship to survival time. *Biometrics*, *50*(4), 1003–1014.

Diggle, P., Zeger, S. L., Liang, K., & Heagerty, P. (2002). *Analysis of longitudinal data* (2nd ed.). Oxford: Oxford University Press.

Doss, B. D., Thum, Y. M., Sevier, M., Atkins, D. C., & Christensen, A. (2005). Improving relationships: mechanisms of change in couple therapy. *Journal of Consulting and Clinical Psychology*, *73*(4), 624–633. doi: 10.1037/0022-006X.73.4.624

Feldman, B. J., & Rabe-Hesketh, S. (2012). Modeling achievement trajectories when attrition is informative. *Journal of Educational and Behavioral Statistics*, *37*(6), 703 – 736. doi: 10.3102/1076998612458701

Ferrer, E., & McArdle, J. (2003). Alternative structural models for multivariate longitudinal data analysis. *Structural Equation Modeling*, *10*(4), 493–524.

Fieuws, S., & Verbeke, G. (2006). Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics*, *62*(2), 424–431. doi: 10.1111/j.1541-0420.2006.00507.x

Fieuws, S., Verbeke, G., Maes, B., & Vanrenterghem, Y. (2008). Predicting renal graft failure using multivariate longitudinal profiles. *Biostatistics*, *9*(3), 419–431. doi: 10.1093/biostatistics/kxm041

Galecki, A. T. (1994). General class of covariance structures for two or more factors in longitudinal data analysis. *Communications in Statistics*, *23*(11), 3105–3119.

Gándara, P., & Orfield, G. (2010). *A return to the Mexican room: The segregation of Arizona's English learners*. Los Angeles, CA: Civil Rights Project.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: Chapman & Hall/CRC. doi: 10.1007/s13398-014-0173-7.2

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997–1016. doi: 10.1007/s11222-013-9416-2

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y. S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*, *2*(4), 1360–1383. doi: 10.1214/08-AOAS191

Gelman, A., Lee, D., & Guo, J. (2015). Stan: A probabilistic programming language for Bayesian inference and optimization. *Journal of Educational and Behavioral Statistics*, *40*(5), 530–543. doi: 10.3102/1076998615606113

Guo, G. (1993). Event-history analysis for left-truncated data. *Sociological Methodology*, *23*, 217–243.

Hakuta, K. (2011). Educating language minority students and affirming their equal rights: Research and practical perspectives. *Educational Researcher*, *40*(4), 163–174.

Hakuta, K., Butler, Y. G., & Witt, D. (2000). *How long does it take English learners to attain proficiency?* The University of California Linguistic Minority Research Institute.

Hammer, C. S., Lawrence, F. R., & Miccio, A. W. (2008). The effect of summer vacation on bilingual preschoolers' language development. *Clinical Linguistics & Phonetics*, *22*(9), 686–702. doi: 10.1080/02699200802028033

Harcourt. (2007). *Arizona English Language Learner Assessment (AZELLA) technical manual*. San Antonio, Texas: Author.

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. Hoboken, NJ: John Wiley & Sons. doi: 10.1002/0470036486

Hedeker, D., Siddiqui, O., & Hu, F. B. (2000). Random-effects regression analysis of correlated grouped-time survival data. *Statistical Methods in Medical Research*, *9*(2), 161–179. doi: 10.1191/096228000667253473

Heilig, J. V. (2011). Understanding the interaction between high-stakes graduation tests and English learners. *Teachers College Record*, *113*(12), 2633–2669.

Henderson, R., Diggle, P., & Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, *1*(4), 465–480. doi: 10.1093/biostatistics/1.4.465

Hoffman, M., & Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623.

Joreskog, K. G. (1970). A general method for analysis of covariance structures. *Biometrika*, *57*(2), 239–251.

Joreskog, K. G., & Goldberger, A. S. (1975). Estimation of a Model with Multiple Indicators and Multiple Causes of a Single Latent Variable. *Journal of the American Statistical Association*, *70*(351), 631–639. doi: 10.2307/2285946

Kalbfleisch, J. D., & Prentice, R. L. (2011). *The statistical analysis of failure time data* (2nd ed.). Hoboken, NJ: Wiley-Interscience.

Kanno, Y., & Cromley, J. G. (2013). English language learners' access to and attainment in postsecondary education. *TESOL Quarterly*, *47*(1), 89–121. doi: 10.1002/tesq.49

Kao, G., & Thompson, J. S. (2003). Racial and ethnic stratification in educational achievement and attainment. *Annual Review of Sociology*, *29*, 417–442. doi: 10.1146/annurev.soc.29.010202.100019

Kieffer, M. J. (2008). Catching up or falling behind? Initial English proficiency, concentrated poverty, and the reading growth of language minority learners in the United States. *Journal of Educational Psychology*, *100*(4), 851–868. doi: 10.1037/0022-0663.100.4.851

Kieffer, M. J. (2011). Converging trajectories: Reading growth in language minority learners and their classmates, Kindergarten to Grade 8. *American Educational Research Journal*, *48*(5), 1187–1225. doi: 10.3102/0002831211419490

Kim, J., & Herman, J. L. (2009). A three-state study of English learner progress. *Educational Assessment*, *14*(3-4), 212–231.

Kohnert, K., & Conboy, B. T. (2010). Lexical and grammatical bilingual preschoolers. *Journal of Speech, Language, and Hearing Research*, *53*(3), 684–698.

Koretz, D., & Guo, Q. (2012). Estimating the impact of the Massachusetts English immersion law on limited English proficient students' reading achievement. *Educational Policy*, *27*(1), 121–149. doi: 10.1177/0895904812462776

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001. doi: 10.1016/j.jmva.2009.04.008

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*(404), 1198 – 1202.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, *90*(431), 1112–1121.

MacCallum, R. C., Kim, C., Malarkey, W. B., & Kiecolt-Glaser, J. K. (1997). Studying multivariate change using multilevel models and latent curve models. *Multivariate Behavioral Research*, *32*(3), 215–253. doi: 10.1207/s15327906mbr3203

Molenberghs, G., Beunckens, C., Sotto, C., & Kenward, M. G. (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society. Series B*, *70*, 371–388.

Muthen, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*(4), 557–585. doi: 10.1007/BF02296397

Muthen, B., & Masyn, K. (2005). Discrete-time survival mixture analysis. *Journal of Educational and Behavioral Statistics*, *30*(1), 27–58.

Muthén, B. O. (2002). Beyond SEM : general latent variable modelling. *Behaviormetrika*, *29*(1), 81–117. doi: 10.2333/bhmk.29.81

National Center for Education Statistics. (2015). *The condition of education 2015: English language learners*. Washington, DC: Author.

National Research Council. (2011). *Allocating federal funds for state programs for English language learners*. Washington, DC: National Academies Press.

Office of English Language Acquisition. (2015). *Profiles of English learners (ELs)*.

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

Proust-Lima, C., Séne, M., Taylor, J. M. G., & Jacqmin-Gadda, H. (2014). Joint latent class models for longitudinal and time-to-event data: a review. *Statistical Methods in Medical Research*, *23*(1), 74–90. doi: 10.1177/0962280212445839

Ramsey, A., & O'Day, J. (2010). *Title III policy : State of the states*. Washington, D.C.: U.S. Department of Education.

Raudenbush, S. W. (2001). Comparing personal trajectories and drawing causal inferences from longitudinal data. *Annual Review of Psychology*, *52*, 501–525. doi: 10.1146/annurev.psych.52.1.501

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models* (2nd ed.). Thousand Oaks, CA: Sage.

Reinsel, G. (1982). Multivariate repeated-measurement or growth curve models with multivariate random-effects covariance structure. *Journal of the American Statistical Association*, *77*(377), 190–195.

Ribaudo, H., & Thompson, S. G. (2002). The analysis of repeated multivariate binary quality of life data: a hierarchical model approach. *Statistical Methods in Medical Research*, *11*(1), 69–83.

Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, *67*(3), 819–829. doi: 10.1111/j.1541-0420.2010.01546.x

Rizopoulos, D., Hatfield, L., Carlin, B., & Takkenberg, J. (2014). Combining dynamic predictions from joint models for longitudinal and time-to-event data using Bayesian model averaging. *Journal of the American Statistical Association*, *109*(January 2016), 1385–1397. doi: 10.1080/01621459.2014.931236

Rizopoulos, D., & Lesaffre, E. (2014). Introduction to the special issue on joint modelling techniques. *Statistical Methods in Medical Research*, *23*(1), 3–10. doi: 10.1177/0962280212445800

Roberts, G., & Bryant, D. (2011). Early mathematics achievement trajectories: English-language learner and native English-speaker estimates, using the Early Childhood Longitudinal Survey. *Developmental Psychology*, *47*(4), 916–930. doi: 10.1037/a0023865

Robinson, J. P. (2011). Evaluating criteria for English learner reclassification: A causal-effects approach using a binding-score regression discontinuity design with instrumental variables. *Educational Evaluation and Policy Analysis*, *33*(3), 267–292. doi: 10.3102/0162373711407912

Rojas, R., & Iglesias, A. (2013). The language growth of Spanish-speaking English language learners. *Child Development*, *84*(2), 630–646. doi: 10.1111/j.1467-8624.2012.01871.x

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, *63*(3), 581. doi: 10.2307/2335739

Scarcella, R. (2003). *Academic English: A conceptual framework (Technical Report 2003-1)*. Santa Barbara, CA: The University of California Linguistic Minority Research Institute.

Schleppegrell, M. J. (2001). Linguistic features of the language of schooling. *Linguistics and Education*, *12*(4), 431–459. doi: 10.1016/S0898-5898(01)00073-0

Singer, J. D., & Willett, J. B. (1993). It ' s about time : Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational Statistics*, *18*(2), 155–195.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling*. Boca Raton, FL: Chapman & Hall/CRC.

Slama, R. B. (2012). A longitudinal analysis of academic English proficiency outcomes for adolescent English language learners in the United States. *Journal of Educational Psychology*, *104*(2), 265–285. doi: Doi 10.1037/A0025861

Slama, R. B. (2014). Investigating whether and when English learners are reclassified into mainstream classrooms in the United States: A discrete-time survival analysis. *American Educational Research Journal*, *51*(2), 220–252. doi: 10.3102/0002831214528277

Stevens, J. J., & Schulte, A. C. (2016). The interaction of learning disability status and student demographic characteristics on mathematics growth. *Journal of Learning Disabilities*. doi: 10.1177/0022219415618496

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., ... Kattan, M. W. (2010). Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology*, *21*(1), 128–138. doi: 10.1097/EDE.0b013e3181c30fb2.Assessing

Stoolmiller, M. (1994). Antisocial behavior, delinquent peer association, and unsupervised wandering for boys: Growth and change from childhood to early adolescence. *Multivariate Behavioral Research*, *29*(3), 263–288. doi: 10.1017/CBO9781107415324.004

Teachers of English to Speakers of Other Languages. (2006). *PreK-12 English language proficiency standards*. Alexandria, VA: Teachers of English to Speakers of Other Languages.

Thompson, K. D. (2015). English learners' time to reclassification: An analysis. *Educational Policy*, 1 – 34. doi: 10.1177/0895904815598394

Thum, Y. M. (1994). *Analysis of individual variation: A multivariate hierarchical linear model for behavioral data (Unpublished doctoral dissertation)*. Chicago, IL: University of Chicago. doi: 10.1017/CBO9781107415324.004

Thum, Y. M. (1997). Hierarchical linear models for multivariate outcomes. *Journal of Educational and Behavioral Statistics*, *22*(1), 77–108. doi: 10.3102/10769986022001077

Thum, Y. M., & Matta, T. H. (2015a). *MAP College Readiness Benchmarks: A research brief.* Portland, OR: NWEA.

Thum, Y. M., & Matta, T. H. (2015b). *Predicting College Readiness from Interim Assessment Results : Growth modeling with selection.* Chicago: Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Tsiatis, A. A., & Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Stat Sinica*, *14*, 809–834.

Tsiatis, A. A., Degruttola, V., & Wulfsohn, M. S. (1995). Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association*, *90*(429), 27–37.

Uchikoshi, Y. (2012). Development of vocabulary in Spanish-speaking and Cantonese-speaking English language learners. *Applied Psycholinguistics*, *35*(2014), 1–35. doi: 10.1017/S0142716412000264

Umansky, I. M., & Reardon, S. F. (2014). Reclassification patterns among Latino English learner students in bilingual, dual immersion, and English immersion classrooms. *American Educational Research Journal*, *51*(5), 879–912. doi: 10.3102/0002831214545110

U.S. Department of Education, Institute of Education Sciences, & National Center for Education Statistics. (2015). *NAEP data explorer.*

Vagh, S. B., Pan, B. A., & Mancilla-martinez, J. (2009). Measuring growth in bilingual and monolingual children's English productive vocabulary development: The utility of combining parent and teacher report. *Child Development*, *80*(5), 1545–1563.

Valentino, R. A., & Reardon, S. F. (2015). Effectiveness of four instructional programs designed to serve English learners: Variation by ethnicity and initial English proficiency. *Educational Evaluation and Policy Analysis*, *37*(4), 612–637. doi: 10.3102/0162373715573310

Vaupel, J. W., Manton, K. G., & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, *16*(3), 439–454. doi: 10.2307/2061224

Vehtari, A., Gelman, A., & Gabry, J. (2016). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *arXiv*. Retrieved from `http://arxiv.org/abs/1507.04544`

Verbeke, G., & Davidian, M. (2008). Joint models for longitudinal data: Introcution and overview. In G. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs (Eds.), *Longitudinal data analysis* (pp. 319–326). Boca Raton, FL: Chapman & Hall/CRC.

Verbeke, G., Fieuws, S., Molenberghs, G., & Davidian, M. (2014). The analysis of multivariate longitudinal data: A review. *Statistical Methods in Medical Research*, *23*(1), 42–59. doi: 10.1177/0962280212445834

Wu, M. C., & Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, *44*(1), 175–188. doi: 10.2307/2531905

Wulfsohn, M. S., & Tsiatis, a. a. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, *53*(1), 330–339. doi: 10.1111/j.1541-0420.2006.00719.x

Xu, S., & Blozis, S. A. (2011). Sensitivity Analysis of Mixed Models for Incomplete Longitudinal Data. *Journal of Educational and Behavioral Statistics*, *36*(2), 237–256.

Yu, M., Taylor, J. M. G., & Sandler, H. M. (2008). Individual prediction in prostate cancer studies using a joint longitudinal survivalcure model. *Journal of the American Statistical Association*, *103*(481), 178–187. doi: 10.1198/016214507000000400