Exploratory Data Analysis with Clustered Data: Simulation and Application with Oregon's Statewide Longitudinal Data System using Generalized Linear Mixed-Effects Model Trees

by

Christopher M. Loan

A dissertation accepted and approved in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in Quantitative Research Methods in Education

Dissertation Committee:

Keith Zvoch, Chair

Kathleen Scalise, Core Member

Gerald Tindal, Core Member

Emily Tanner-Smith, Institutional Representative

University of Oregon

Spring 2024

DISSERTATION ABSTRACT

Christopher M. Loan

Doctor of Philosophy in Quantitative Research Methods in Education

Title: Exploratory Data Analysis with Clustered Data: Simulation and Application with Oregon's Statewide Longitudinal Data System using Generalized Linear Mixed-Effects Model Trees

Simulations were conducted to establish best practice in hyperparameter optimization and accounting for clustering in Generalized Linear Mixed-Effects Model Trees (GLMM trees). Using data-driven best practices, the relationship between a 9th Grade On-Track to Graduate (9G-OTG) indicator and observed high school graduation within four years was explored. Data originated from two cohorts of the Oregon State Longitudinal Data System (SLDS) and were joined with external datasets. Restricted to complete cases, the data were comprised of more than 58,000 observations, each with more than 1500 variables measured at student, school, district, and zip code levels. GLMM trees explored heterogeneity in a cross-classified multilevel logistic regression which regressed observed graduation on 9G-OTG, accounting for variance in school- and zip-code-level random intercepts. Subgroups were identified for whom the probability of graduating among on- and-off track students were systematically heterogeneous, relative to the supraordinate group. Results suggest that for most students, 9G-OTG is a potent early warning indicator of graduation, but systematic variation in the indicator's effectiveness was found along all levels except district. Subgroups were defined by combinations of alternative schools, absences, transferring schools, being enrolled in more than one instructional program, neighborhood unemployment, and sex. Implications and recommendations to measurement, practice, and evaluation are discussed.

Keywords: Generalized Linear Mixed Effects Model Trees; Model-based Recursive Partitioning; State Longitudinal Data Systems; Hyperparameter Optimization; Multilevel Modeling; Graduation.

ACKNOWLEDGMENTS

Research funding was partially supported by IES Grant Award Number R305S210005. Additionally, this project would not be possible without extensive data collection, storage, and sharing made possible by several U.S. State and Federal Agencies—particularly the Oregon Department of Education.

DEDICATION

This work is dedicated to my soon-to-be wife, Jennah Maier, without whom this journey would have been incomprehensibly difficult.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

Context

Paper 1

Paper 2

Paper 3

LIST OF TABLES

Table Page

Paper 1

Paper 2

Paper 3

Summative Conclusion

Preface

This dissertation first contains a *Context* section in which I present the overarching contextual information that underlies and unites the three papers of the dissertation. The *Context* section serves to minimize redundant presentation of both theoretical and methodological information across the three papers. In this dissertation, *Paper 1* and *Paper 2* describe and present robustness checks to ensure the validity of applying a relatively new method—the generalized linear mixed-effect model regression [GLMM] tree (Fokkema et al., 2018)—to data contained in the Oregon State Longitudinal Data System (SLDS). *Paper 3* follows with a practical example, the prediction of four-year graduation in Oregon as a function of students' 9th grade on-track to graduation status (9G-OTG) and a multitude of student, school, and zip code factors. From a statistical perspective, methods need to be optimized to ensure the model most accurately represents the data, and their validity needs to be tested formally before extending their use to new data structures. From a practical perspective, methods need to be optimized and tested for statistical validity in context because subsequent policy recommendations affect the day-to-day lives of students, teachers, and communities.

In the *Context* section that follows, I provide a detailed roadmap for how the three papers fit together and explain how the knowledge generated from *Paper 1* and *Paper 2* can ensure the valid application of the GLMM tree to the Oregon SLDS. In-depth discussion of the 9G-OTG literature is relegated to *Paper 3*, and only an abridged description is given in the *Context*. After *Paper 3*, the *Conclusion* summarizes how methodologists can apply a similar context-driven framework to method dissemination, testing, and development. The modular three-paper format is useful to understand GLMM trees because the method requires passing familiarity with machine learning techniques (e.g., recursive partitioning), model optimization (e.g.,

hyperparameter tuning), techniques to prevent overfitting (e.g., cross-validation), statistical techniques used across multiple fields (e.g., the parameter stability test common to economics; the GLMM common to education), and other adjacent topics. Appendix A, a table of abbreviations and acronyms used throughout, can be used as a reference. Appendix B provides a glossary of selected terms meant to help readers familiar with some parts of the covered literature but not all.

> [I]t should be our number one priority to get young people to the "starting line" of adulthood with all the hope and confidence they can muster. As a society we expect young people to take on the responsibilities and consequences of adulthood at age eighteen. Not coincidentally, age eighteen is also the point when we consider our obligation for providing a free public education to be over. For this reason, it is suggested that educators see high school graduation (or its equivalent) as simultaneously the finishing line for mandatory schooling and the entry point for full democratic citizenship. (Richard Sagor, 1999, p. 74)

Context: Appraisal and Discovery in Social Science

According to John Gerring's *Social Science Methodology, a Unified Framework*, social science has two overarching goals—appraisal and discovery—and "*these primal goals inform every methodological endeavor"* (Gerring, 2011, p. 32). Gerring's appraisal includes any method that is falsifiable such as formal mathematical proofs, statistical tests, evaluations, demonstrations, and the like. Thus, falsifiability is the hallmark of appraisal and a core component of knowledge generation in social science. To be falsifiable, Gerring says an argument must be

> operational, parsimonious, general in purview (offering a large territory for empirical testing), well bounded (so that the population of an inference is clear, and defensible), coherent (internally consistent), clear with respect to counterfactuals and comparisons, and relying on as few assumptions as possible. (2011, p. 31)

Discovery, by contrast, is not defined by falsifiability, but novelty (Gerring, 2011). Methods of discovery range from conjecture to observations and exploratory data analysis. Neither goal of social science is superior to the other, and the choice of goal depends on many factors, including research question, amount of knowledge in the field, and the relative impact of a Type I versus Type II error in the context (Gerring, 2011; Tukey, 1977; Albert, 2022; Andrews, 2021).

This dissertation primarily focuses on quantitative approaches to knowledge generation and the distinction requires different terminology. Drawing on Tukey (1977), I focus on a component of Gerring's (2011) taxonomy. Instead of focusing on appraisal and discovery broadly, I focus on confirmatory data analysis (CDA) and exploratory data analysis (EDA), which are quantitative instantiations of the goals of social science. In this way, all methods of EDA are means of discovery, but not all means of discovery are EDA; the same is true of CDA and appraisal, respectively. Because falsifiability is the hallmark of appraisal, and thus CDA,

intent alone can move someone from confirmation to exploration (Gerring, 2011). In other words, a t-test could be applied agnostically across all grouping variables in a dataset to find *any* difference between two groups without no respect to parsimony, falsifiability, coherence, or Gerring's (2011) other requirements of appraisal. However, the same t-test used to compare differences across randomized groups is a potent form of appraisal.

As with appraisal, CDA is used to support pre-existing theory or test a hypothesis based on prior observations (Tukey, 1977; Hartwig & Dearing, 1979). Common CDA techniques are useful in estimating (relatively) easy-to-understand magnitudes and tests of statistical significance for a user-specified model (e.g., t-test, analysis of variance [ANOVA], etc.; Tukey 1977; Elman et al., 2020). Systematic documentation of effects—or lack thereof—relies on approaches like these to build what is known about the world (Elman et al., 2020). Replication and synthesis of information are essential late-stage aspects of CDA used to gain confidence in impactful findings and move up a hierarchy of evidence (Petticrew & Roberts, 2003; Brannen & Coram, 1992). Without important synthesis and replication work, little confidence can be built in the validity or generalizability of the research, as well as more precise and accurate estimates of the impact of important effects (Petticrew & Roberts, 2003; Brannen & Coram, 1992).

In contrast to CDA, this work defines EDA as visualizations, algorithmic (including machine learning) or correlative analyses, and other quantitative approaches that can provide insights beyond user specification or current theory. Importantly, findings from EDA make no claims about causality, and in many cases, EDA does not make claims about distribution of the underlying data or relate the sample to a larger population as is with CDA (Albert, 2022; Fokkema et al., 2018).

As Gerring (2011, p. 33) stated *"theoretical development could not occur, or would occur only very slowly and haltingly, if researchers [limit] themselves to pre-formed hypotheses and yes/no empirical tests. A constructive methodology should enable researchers to think about problems in new ways.*" CDA and EDA are both critical steps of theory building with EDA specifically acting as a driver of hypothesis generation (Gerring, 2011; Fokkema et al., 2018). EDA expands our understanding of what may be occurring in the world around us and aligns our focus to trends in the data. Just as confirmatory research can vary widely (e.g., a randomized controlled trial assessed with a t-test, an instrumental variable design assessed with multiple regression), exploratory research can range from simple descriptive statistics to machine learning (ML) approaches, each tailored for their own purposes.

Bringing Discovery to the Light

Despite appraisal and discovery both being drivers of knowledge generation, Gerring argues that appraisal is "virtually the sum total of the field of methodology, as traditionally conceived" (Gerring, 2011, p. 30). The reason for appraisal's majority share is discussed throughout his work. Gerring (2011, p. 30) argues the incentive structure of journals is particularly at fault, as they "*frequently insist upon the presentation of a priori hypotheses ('suggested by the literature'), which will then […] be 'tested against the data,' even when the procedures actually followed in the course of the research are blithely exploratory.*"

Thus, by Gerring's own admission, discovery is occurring behind-closed-doors, obfuscating the true balance of appraisal and discovery. In fact, other authors have clearly articulated the use of exploration prior to confirmation, at least since Tukey's 1977 book *Exploratory Data Analysis* (Albert, 2022; Andrews, 2021). Relatively new publications continue to elaborate on the argument implied by Tukey nearly fifty years ago. For example, Andrews

writes "*probabilistic modelling* [i.e., CDA] *is the ultimate goal and even the raison d'être of data analysis [… and] this probabilistic modelling can only be done thoroughly and well if we understand our data" (*Andrews, 2021, Chapter 5 p. 2). Some authors go as far as to as *"[i]n a typical data analysis, we will use EDA methods to discover basic patterns or structure in the data. Then we may later use inferential methods to make statements about underlying populations" (*Albert, 2022, Chapter 2 Section 5, para. 3).

Hence, confirmatory researchers may explore their data, but in a way that is both a process of discovery and a means of checking assumptions of the match of data to model assumptions (Andrews 2021, Albert, 2022). Although it is appropriate to gain the benefits of pairing appraisal with discovery, doing so behind closed doors obfuscates and de-systematizes the research process at best (Maxwell et al., 2015; Gerring, 2011), and can lead to "*p*-hacking" (Head et al., 2015) and/or a "replication crisis" (Shrout & Rogers, 2018) at worse.

In the past two decades, a few research groups have created a prolific family of models which make it possible to work along the boundary of CDA/EDA systematically and reproducibly (e.g., Hothorn et al., 2006, Zeileis et al., 2008, Brandmaier et al., 2013-a; Fokkema et al., 2018). The models from said groups pair familiar inferential frameworks with machine-learning approaches, and others have demonstrated their ability of such models to incorporate EDA into CDA frameworks (Fokkema & Zeileis, 2023; Fokkema et al., 2021; Brandmaier et al., 2016; Brandmaier et al., 2013-b). Such models are both presented and evaluated for validity throughout this dissertation with the intent of increasing the systematicity and reproducibility of discovery and EDA in social science.

Instantiating Appraisal and Discovery in this Dissertation

Just as Tukey (1977) compiled, described, and presented useful methods less than a decade old with hopes of greater application of the method, this dissertation documents a method that is less than a decade old with the same goal. In contrast to Tukey, though, I focus on only one method in great depth rather than focusing on breadth of topics to emphasize the strong fit for this particular method to social science and educational research.

My choice of where to apply appraisal versus discovery aligns with Gerring's description of when researchers may choose one method over another. The application of this paper to a novel merging of the Oregon Statewide Longitudinal Dataset with external data sources leads to a high-dimensional space which has not been previously explored (i.e., ~58,000 students after limiting sample to those with no missingness across >1,550 variables). With such a large dataset, an open-minded discovery-driven approach is warranted because too many factors are present for a clear understanding of which effects may emerge (Gerring, 2011).

A Reproducible and (Mostly) Familiar Approach to Discovery

In their development of Generalized Linear Mixed-Effect Model (GLMM) trees, Fokkema et al. (2018) emphasize the utility for hypothesis generation. GLMM trees were made to explore heterogeneity within the bounds of a parametric model. Statistically, this type of "model-based" recursive partitioning improves precision and accuracy of the regression coefficients by identifying subgroups in the data without analyst specification (Fokkema et al., 2018; Zeileis et al., 2008). Returning unspecified subgroups of a parametric model inspires successive iterations of research. Then, systematic methods of appraisal can explore why a subgroup may exist in the population in the next cycle of research. Fokkema et al., describe the logic and its impetus concisely (2018, p. 2019):

> [A] single global [parametric model] may not describe the data well, and when additional covariates are available it may be possible to partition the dataset with respect to these covariates, and find better-fitting models in each cell of the partition. For example, to assess the effect of treatment, we may first fit a global [model] where the treatment indicator has the same effect/coefficient on the outcome for all observations. Subsequently, the data may be partitioned recursively with respect to other covariates, leading to separate models with different treatment effects/coefficients in each subsample.

Moving from Theory to Practice: Falsifiable and Confirmatory Aspects of this Dissertation

The broad design components of *Paper 3* (e.g., context and data) lend themselves to GLMM trees for several reasons discussed below, but the method is less than a decade old with only three robust investigations of their effectiveness and validity (Fokkema et al., 2018; Jorink, 2018; Fokkema & Zeileis, 2023). Considering the number of individuals impacted by findings from the SLDS, though, this project requires a falsificationist approach to test statistical nuances of GLMM trees in this exact context. Although neither paper is a power analysis—using simulation to determine the proportion of the time an effect will be identified if present (e.g., Lakens & Caldwell, 2021; Arend & Schafer, 2019)—the broad idea is a useful parallel for understanding the impetus of *Paper 1* and *Paper 2* before leveraging *Paper 3*. For those familiar with the academic ML literature, the approach of *Paper 1* is secondhand, as optimization techniques are an essential component of valid and high performing ML (Zhang et al., 2021-a; Zhang et al., 2021-b; Alibrahim & Ludwig, 2021; Falkner et al., 2018; Yang & Shami, 2020). As such, the validity of the optimization can be supported or falsified in this context prior to trusting the resultant (possibly unexpected) discoveries.

Though statistical, the process underlying *Paper 1* and *Paper 2* is analogous to that of anyone checking their oil and tire pressure before embarking on a long drive. In doing so, motorists can have more confidence in the ability of the vehicle to make it to the destination. Walking away from the metaphor, this project offers small, concrete contributions to the

leveraged methodology and aims to contribute impactful—or at least unexpected—information to the theoretical context which led to the application of the method (i.e., on-time graduation). Because of the nuances of GLMM trees discussed below, if no impactful or unexpected relationships are identified, there is evidence of invariance in the relationship between predictor (a 9th grade "early warning indicator") and outcome (on-time graduation), which would be an important finding from perspectives of equality and equity in Oregon schools. As Tukey (1977, p. 6) writes, researchers conducting EDA must have "*willingness to find some happenchance phenomena, as well as happiness in finding phenomena with some continuing reality.*"

Perhaps the primary draw of using GLMM trees to explore on-time graduation is that the method is an extension of a familiar framework, logistic regression, simplifying communication with stakeholders familiar with educational literature (Fokkema et al., 2018). However, the interaction of the method with several aspects of the context at hand—mostly arising from the dimensionality of the available data—are currently unknown. With the explosion in prevalence of large models and datasets with an excess of covariates, advanced ML practitioners have documented a breakdown of traditional statistical principles as dimensionality of the dataset increases, including such fundamental concepts of the bias-variance tradeoff (Zhang et al., 2021; Neal et al., preprint; Zagoruyko & Komodakis, 2016, etc.). Belkin et al. (2019) has shown even extensions of regression and random forest (e.g., boosted L2 regression and boosted random forests) can display the phenomenon of being trained beyond the traditional conception of balancing bias and variance (coined "interpolation"). Despite the *relatively* massive sample size for social science projects, the SLDS and joined data are unlikely to undergo interpolation with GLMM trees. Nevertheless, the concept of interpolation highlights the possibility of unexpected

phenomenon emerging from an increase in dimensionality which should be explored through simulation before application.

Created in the context of behavioral psychology, the developers have presented GLMM trees with data sets that are orders of magnitude smaller than the SLDS, especially when paired with external datasets (Fokkema et al., 2018; Fokkema et al., 2021; Jorink, 2018; Fokkema & Zeileis, 2023). Unlike the theoretical implications with novel data joins, clearly definable and falsifiable aspects of applying the methodology can be articulated, hence justifying a falsificationist approach to appraising the statistical validity of the proposed methodology.

In response to Gerring's (2011) call for social scientists to better distinguish "*theory-testing from work that is – rightly, and justifiably – theory-generating" (p. 36).* I explicitly document *Paper 3* as a project of discovery that needs to be supported through appraisal prior to trusting the resultant model. This embodies Gerring's Unified Framework for Social Science Methodology in a single instance and adopts the cyclical approach to knowledge generation outlined by Gerring and others (e.g., Tukey, 1977, Hartwig & Dearing, 1979, etc.). *Paper 1* and *Paper 2*—means of appraisal—pass their findings to *Paper 3*—a means of discovery—which will have to be verified through subsequent means of appraisal in a relay-race of knowledge generation.

CDA is used to test two falsifiable claims about the methodology of GLMM trees, each corresponding the alternative hypotheses of *Paper 1* and *Paper 2*, respectively. First, that GLMM trees which undergo hyperparameter optimization (i.e., are "calibrated to the data") will have greater predictive accuracy, better model fit to the data, and produce a more parsimonious model. Second, that GLMM trees which use only random effects to adjust for statistical influences of clustering lead to models that more accurately recover simulated parameters and

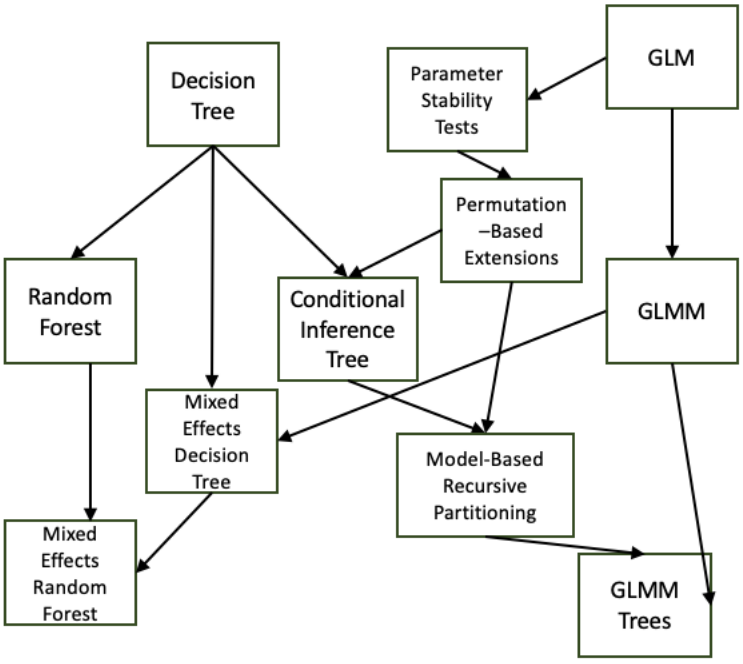
have greater predictive accuracy than models which correct for clustering two times (i.e., mixed effects followed by additional adjustments to standard errors). Both hypotheses are stated here for context (admittedly without evidence) and are elaborated in their respective papers. To understand the falsifiable aspects and to prevent presenting this information three times, the following section provides a general overview of the statistical lineage of GLMM trees and the algorithms underlying its behavior.

Precursor Models to GLMM Trees

Figure 1 provides a diagram of the more well-known models which led to the development of the final model of interest—the GLMM tree (Fokkema et al., 2018).

Figure 1

GLMM tree precursors and adjacent models.



Recursive partitioning finds splits to maximize similarity of *a single outcome with respect to observed variables* (Strobl et al., 2009). GLMM Trees use a type of recursive

partitioning which is referred to as "*model-based"* (Zeileis et al., 2008). *Model-based recursive partitioning*—referred to as "MOB" by the developers— maximizes the overall fit of a parametric model (e.g., multiple regression, logistic regression) with respect to observed variables by splitting the full sample into subgroups and comparing the fit of the single model to the sum of the fit of the partial models (Zeileis et al., 2008). Specifically, MOB models can be defined with four overarching steps (Zeileis et al., 2008):

1. Fit a parametric model to the data.
2. Test for parameter instability in the objective function over partitioning variables. If there is instability, select the variable with the highest parameter instability.
3. Find the split point in the selected variable.
4. Repeat the procedure in each daughter node until one or more splitting criterion is met.

In other words, MOB fits a parametric model (e.g., a logistic regression) and calculates the fitness of the model with the entire sample (e.g., loglikelihood). Exploration of heterogeneity in logistic regression parameters is then done by MOB, specifically by partitioning or "splitting" the data into subsets and assessing the model fit of each subset of the data. The splits are retained if the sum of the partitioned fitness functions is preferred to the fitness function of the parametric model before the split (Zeileis, et al., 2008). Splitting occurs recursively until the fit of the data is no longer improved by separating groups or when another stopping criteria is met (e.g., minimum number of observations per node, maximum number of splits, etc.). MOB is deeply founded in the lineage of EDA outlined by Tukey, as he writes EDA requires an "*emphasis on successively better fits, and on the incompleteness of all fits*" (1977, p. 6), which is the basis of the algorithm.

Thus, MOB identifies subgroups for which the specified parametric model fits differently in an exploratory way useful for generating novel insights or checking assumptions (Fokkema et al., 2018; Tukey, 1977). Model-identified variables which form said subgroups are called "partitioning" or "splitting" variables and are conceptually similar to moderators (Zeileis et al., 2008). Based in recursive partitioning, these models offer the benefits of a recursive partitioning framework, meaning they are robust to non-linear functional forms, do not need interactions to be specified, and can be fitted with categorical or continuous variables (Zeileis et al., 2008; Hothorn et al., 2006; Strobl, et al., 2009). GLMM trees return parameters that are easily interpretable to those familiar with the specified GLMM. Resultant models are also called "segmented parametric models," because the same parametric model is fitted independently to multiple segments of the data (i.e., model-identified subgroups; Zeileis, et al., 2008).

Comparing and Contrasting Familiar Models: MOB versus Logistic Regression and Classification Trees

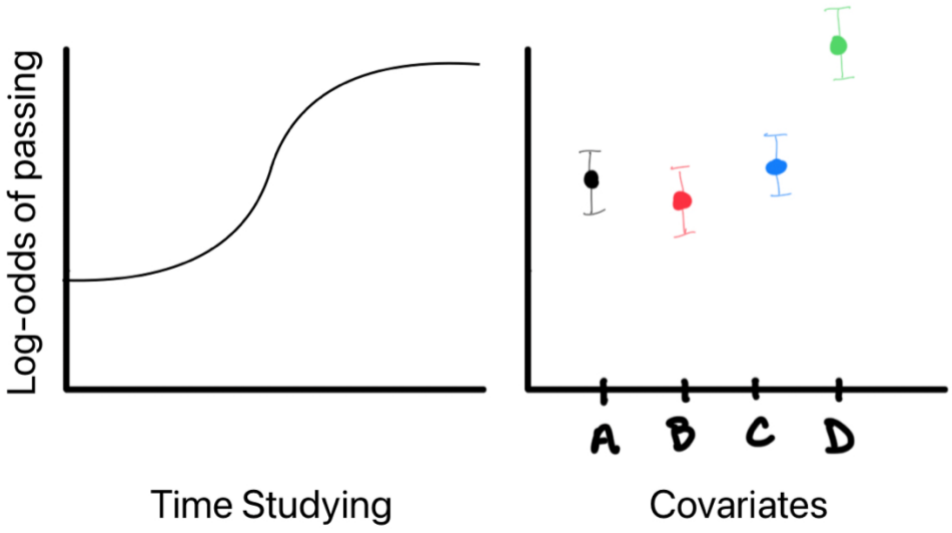
Comparing the same data assessed by logistic regression and a classification tree to that analyzed with MOB helps elucidate similarities and differences between the MOB algorithm and those more familiar models. For example, imagine a school district's analyst is tasked with exploring the impact of time studying on performance within a large district to help shape teacher recommendations on time studying in syllabi. Teachers add a question to each test asking the number of hours a student studied for the final exam (the main predictor of interest). The outcome is a dichotomized metric of passing the final exam and data is gathered for all students in the district. With results not generalized nor leaving the school's records, the analyst is given access to (1) all student-level demographics, (2) course name (e.g., Algebra I, Spanish III, etc.), (3) the primary predictor (time studying), and (4) outcome (passed exam: yes/no).

To provide stakeholders with an answer about the association of studying and passing a test, numerous models could be used. Logistic regression is useful and analyst-selected student-level factors (e.g., demographics, attendance, etc.) can be included in the model as long as statistical power allows. With logistic regression, regressing passing status on the one continuous and one categorical variable lead to (a) an intercept, (b) a slope for time studying, and (c) a fixed effect for each covariate (Figure 2).

In this example, the association of primary interest—between log-odds of passing and time studying—is interpreted as the change in log-odds of passing the test for each unit increase in studying for the reference group, controlling for other variables. The information gleaned from this model can be visualized below as a diagram of parameter estimates (Figure 2).

Figure 2

Results from the hypothetical logistic regression of passing an exam.



Thus, despite controlling for covariates, a singular slope is estimated across all subgroups and is the interpreted parameter for the research question. Of course, additional information can be gleaned by changing the parametric model, including interactions. This concept is

straightforward in cases with few covariates and/or adequate theoretical bases. However, complexity builds in areas with less defined theory or much larger number of variables, with the latter also rapidly diminishing the power to detect differences in the sample (Fokkema et al., 2018). Interactions with more than two variables are exemplars of this complexity. In the hypothetical example, imagine several "foreign" language exams—including Spanish—are included in the sample. For some students, Spanish is their home language, meaning the test itself is measuring something else in these students. Being the home language of this subgroup, a high school Spanish test may not reasonably perform the same for these students (i.e., additional studying will have diminishing returns). In other words, the underlying construct used as the universal outcome of the logistic regression—proficiency in course material—lacks the sensitivity and specificity to measure a construct in a sample that contains (a) fluent Spanish speakers and (b) second year novices. GLM trees allow this to be identified without manual specification. By pooling the subgroup into the larger sample, the estimated parameter of interest (time studying on log-odds passing) is biased (Zeileis et al., 2008). Only by including a 3-way interaction— [student home language] x [class] x [time studying slope] —is the underlying process modeled. However, with multiple student groups or classes dummy-coded, capturing each subgroup may not be viable from a perspective of statistical power (Fokkema et al., 2018). Finally, because multiple Spanish tests may be given, subgroupings which were not originally in the data would have to be created before inclusion (e.g., creating dummy-coded test-type variables, decreasing power).

By contrast, fitting a classification tree with the same outcome and covariates creates a series of results that are interpreted in a very different way (Figure 3). Specifically, the results are expressed in the percent or proportion of the group that passed the test in each terminal node

(Strobl et al., 2009). Rather than returning an association between log-odds of the outcome and a covariate, the classification tree below does not use the information about time studying to predict graduation if their household language is Spanish (Strobl et al., 2009). Rather than expressing the influence of time studying on log-odds of passing as a continuous phenomenon, the influence of time studying is grouped above or below a threshold (Strobl et al. 2009). Unlike the logistic regression, decision trees will parse unspecified relations from the data (Strobl et al., 2009; Hothorn et al., 2006), meaning the subgroup effects mentioned above (i.e., native Spanish speakers and Spanish final exams) is likely to be captured. Unfortunately, though, results from the decision tree are not guaranteed to provide information above the process of interest (the effect of time studying on the outcome). Even if the covariate of primary interest ends up in the model, results may not be easy to understand.

Figure 3

Results from the hypothetical classification tree of passing a final exam.

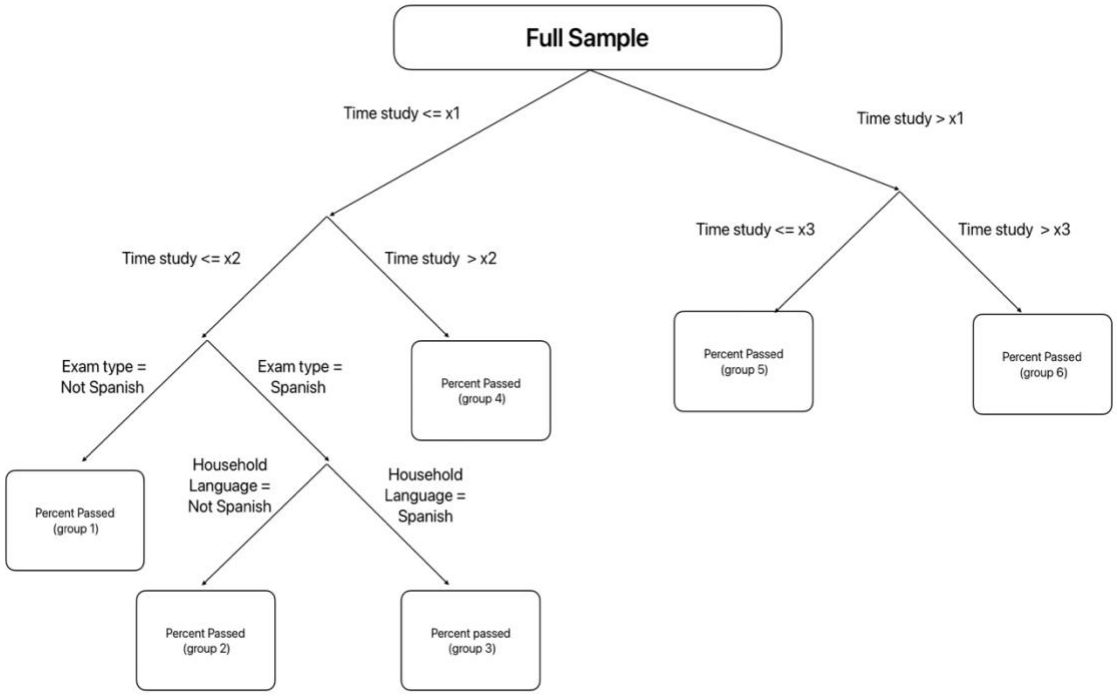
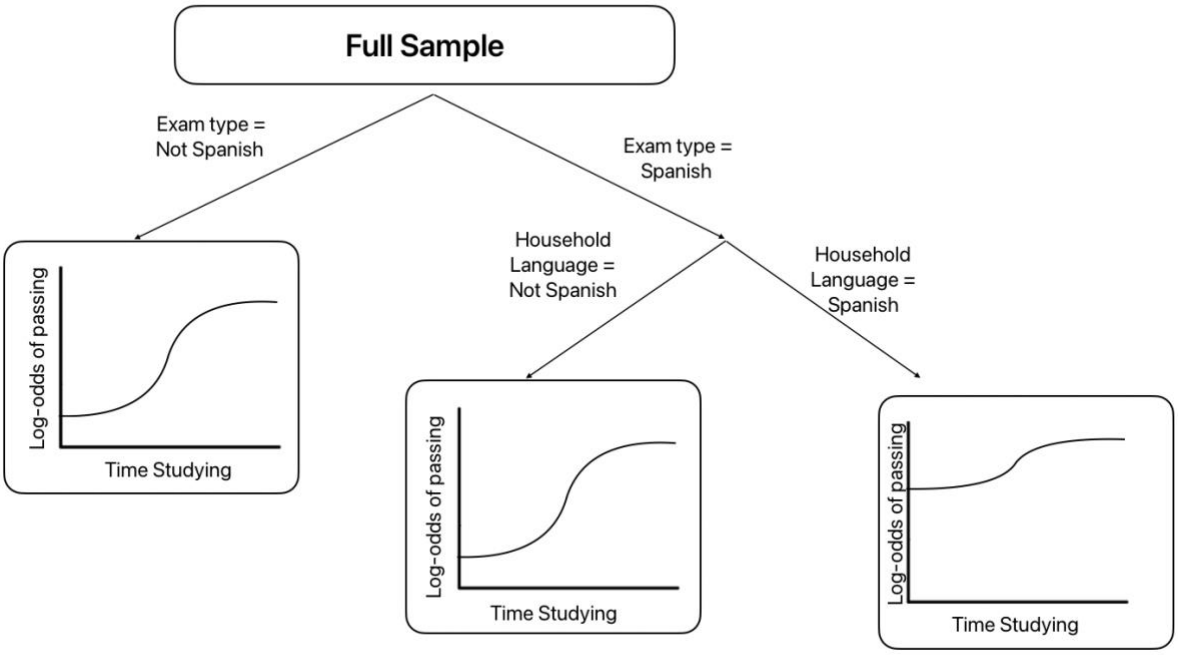


Figure 3 shows one way the subgroup effect could be expressed. To understand the sample and underlying process, a large number of groups and dichotomized decisions must be considered together. Here, the model successfully identifies the subgroup effect described above—native Spanish speakers taking the Spanish exam—among students studying below x1 and below x2. Those in Group 3 (low time studying with native language – test match) pass at a greater rate than their counterparts in Group 2 (low time studying without native language – test match). Yet, the phenomenon appears to not exist among students in any other terminal nodes. The decision tree does not differentiate between high time studying in the right half of the model, regardless of native language – test match. The decision tree is uninterested in statistical inference, meaning the average log odds of anyone in the groups on the right half of the tree—regardless of native language or test type—had similar probability of passing based on the impact of time studying alone.

Fitting GLM trees with the example data allows for the best of both worlds: strong exploratory capabilities with easy to communicate results focused on the parameter (or model) of interest. As such, the results of the GLM tree fall somewhere between the GLM (e.g., the logistic regression model) and the classification tree (Zeileis et al., 2008). Specifically, the user-specified parametric model is returned by subgroup (Figure 4). The model returns an intercept and a slope for the association of time studying to log-odds of passing. However, instead of providing a single fixed effect to account for baseline mean differences (as with covariates in the logistic regression), each group is assigned an intercept and a slope. Furthermore, groups are created by all parameters in the parametric model (i.e., the intercepts and slopes), returning groupings for which an entire logistic regression model fit well (Zeileis et al., 2008). In this way, the model

explores heterogeneity in the parametric model—the logistic regression of log-odds on time studying—among data-driven combinations of user-specified variables.

Figure 4
Results from the hypothetical GLM tree of passing a foreign language test.



In the hypothetical example, each model fitted to the data provided a means to conceptualize the observed data, and the truth in which process is occurring is up for discussion. As George Box and Alberto Luceño wrote in 1997 (p. 6):

> *It has been said that "all models are wrong but some models are useful." In other words, any model is at best a useful fiction—there never was, or ever will be, an exactly normal distribution or an exact linear relationship. Nevertheless, enormous progress has been made by entertaining such fictions and using them as approximations.*

Even if the GLM, decision tree, and GLM tree make similar predictions in aggregate, the models poorly describe the process of interest for the subgroup (Zeileis et al., 2008). Though to a lesser extent, this misspecification also introduces bias into the overall slope of a GLM, again,

because the underlying process modeled is not the same for all observations. The analyst can visually explore the entire parametric model for unexpected subgroups, making subgroup identification relevant to and in the same language as the problem motivating research.

In this example—with such an obvious reason to omit a subgroup from the analysis—an EDA approach is not quite as useful as it is in cases with more variables than power available to detect an effect (Zeileis et al., 2008; Fokkema et al., 2018). In fact, when the number of covariates is small, manual invariance testing could identify the same associations as MOB, even with plots of covariates and residuals described by Tukey (1977). As Gerring (2011) and Tukey (1977) point out, the value of EDA versus CDA depends highly on the context (Gerring, 2011; Tukey, 1997), but the simple example provides an avenue to understand MOB's utility in a higher dimensional space.

In the case above, the model suggests that overall performance (i.e., intercept) and effectiveness (i.e., slope) of studying fundamentally differs, unless the test is a high-school competency test of Spanish, and their household language is Spanish. In the immediate analysis, this allows the analyst to check in with practitioners—in this case the Spanish teachers—and share her/his thoughts about modeling these native Spanish speaking students differently, at least for their Spanish tests. The analyst can look at the Spanish tests and talk with the Spanish teachers to determine if their tests are not measuring class-specific knowledge, but generalized knowledge. If true, the analyst has both correlative evidence and expert opinion that the process differs theoretically. In our (obvious) example, the net effect is a more precise and accurate estimate of the effect of time studying on log-odds of passing through removal of a fundamentally different subgroup. In other spaces, this could inspire successive iterations of research by inspiring the generation of novel hypotheses.

Multilevel Extension of MOB: GLM Trees to GLMM Trees

Going from GLM trees to GLMM trees is a straightforward conceptual leap which intertwines multilevel modeling with MOB. GLMM trees conduct the same analysis as GLM trees, but do so iteratively, passing information between a GLM tree and a multilevel model. Fokkema et al. (2018) outline the steps of GLMM trees as follows:

1. Estimate a GLM tree with fixed-effects structure and splitting variables, assuming random effects are known from prior step (or 0 if first iteration).
2. Fit a GLMM to each terminal node, assuming fixed effects are known from prior step, and store posterior predictions of random effects.
3. Repeat until convergence (e.g., change in loglikelihood < threshold).

Thus, for the GLMM tree proposed for *Paper 3*, a logistic regression is the parametric model: regressing on-time graduation on 9th grade on track to graduation (9G-OTG) status. The 9G-OTG metric is an early warning indicator (EWI) adapted from Allensworth and Easton (2005). Allensworth and Easton (2005) proposed the EWI based on credits earned and number of *F*s in 9th grade classes, but the latter is omitted from ODE's metric (ODE, 2018-a).

However, heterogeneity will be assessed by including more than 1,500 variables as possible splitting variables. Variables are from ODE (e.g., attendance, test scores, demographics, disciplinary incidents, etc.), as well as other publicly available datasets (e.g., Civil Rights Data Collection [CRDC], 2020; Agency for Healthcare and Research Quality [AHRQ], 2020; National Center for Educational Statistics [NCES], 2020). The result will be parametric, multilevel logistic regressions, but with data-driven subgroupings extracted from more than 1,500 student-, school-, district-, and zip-level variables. The GLMM tree can explore

heterogeneity in the resultant coefficients along hundreds of variables without losing power, in a way that manually specified interactions cannot (Fokkema et al., 2018).

Falsification as a Means to Enable Discovery

GLMM trees serve as a powerful method to explore heterogeneity in a multilevel logistic regression, making it ideal as an EDA mechanism for *Paper 3*. However, Fokkema et al.'s GLMM trees are relatively new, first being published in 2018, and not yet widely studied for validity or accuracy. *Paper 1* and *Paper 2* explain the specifics as to how the respective CDA approaches are needed to trust the results for the EDA in *Paper 3*. In this way, falsificationist approaches can be used to confirm the veracity of the proposed means of discovery.

Stated broadly, *Paper 1* and *Paper 2* are motivated by similar circumstances, which can be summarized as a lack of clearly defined best practices with a few notable exceptions (Jorink, 2018; Fokkema & Zeileis, 2023). As elaborated in *Paper 1,* standard practice with algorithmic, data-driven, or machine learning models is to calibrate the model with subsets of the data or "training data" and compare performance with unseen "testing data" (Yang & Shami, 2020). Despite this practice, research with GLMM trees (and MOB more broadly) tends to either omit this step (e.g., Fokkema et al., 2018; Fokkema et al., 2021; Zeileis et al., 2008; Huber et al., 2022; Brown et al., 2022) or present little more than the optimal states (e.g., Johnson et al., 2016; Kern et al., 2021; Quan et al., 2020) which is of minimal importance, because exact settings are not typically considered to generalize in other ML models (Hertel et al., 2021; Elshawi, et al., 2019; Yang & Shami, 2020).

Though *Paper 2* is similarly focused on establishing best practice with calibration of GLMM trees, I argue that the component of interest in *Paper 2* may not be a setting to calibrate, but a statistical decision which has a clear correct specification. A problem arises because

GLMM trees are built upon the GLM tree algorithm, meaning user specifications are passed from the GLMM tree function to the GLM tree function (Fokkema et al., 2018). Without random effects, GLM trees were given the ability to account for clustering via corrections to the estimated standard errors (Zeileis et al., 2008). As the embedded GLM tree has the option to correct for clustering and GLMM trees include all arguments from a GLM tree, models estimated from GLMM trees can include a single correction (i.e., random effects) or dual correction (i.e., random effect and cluster-robust correction; Fokkema et al., 2018). Only one unpublished Master's thesis (Jorink, 2018) and an in-press article from the developers of MOB and GLMM tree (Fokkema & Zeileis, 2023) explore the impact of these statistically relevant settings—referred to as "ranefstart" and "cluster"—the focus of *Paper 2*. These papers do not discuss hyperparameter optimization literature, but both systematically vary these settings and assess predictive accuracy to choose the best model (Jorink, 2018; Fokkema & Zeileis, 2023).

Results from these papers are useful, but they cannot inform best practice in *Paper 3* given key differences with the dimensionality and the structure of the SLDS data. First, samples used in both papers are orders of magnitude smaller than SLDS data. Second, these papers both focus on modeling trajectories of individuals over time by subgroups, which differs from SLDS data in two ways. Fokkema & Zeileis (2023) and Jorink (2018) have a very small number of level-1 units for each level-2 unit—which differs dramatically from the structure of students (level-1) in schools (level-2) in SLDS. Along the same lines, repeated observations of the same individual may behave differently than multiple observations of different students in the same school. Finally—and of most importance—neither study explored the capability of these models to assess the GLMM trees' capability to explore invariance along multiple dimensions simultaneously (e.g., student-level, school-level, district-level, zip-level, etc.), and if the model is

sensitive to changing the "ranefstart" or "cluster" parameter when exploring heterogeneity across clustering units. Therefore, with the results obtained from the simulations in *Paper 1* and *Paper 2*, *Paper 3* can use GLMM trees as a means of discovery with confidence in its validity.

Context for Discovery: Exploring the Differential Effectiveness of the 9G-OTG Indicator

To this point, the information provided about the applied context (i.e., 9G-OTG and on-time graduation in Oregon) in the *Context* section of this dissertation has only covered aspects of the data needed to set up the simulation in Papers 1 and 2 (i.e., data structure, type, and dimensionality). *Paper 3*, the paper of discovery, elaborates on the context more fully, discusses what is known about the indicator, and treats the findings of *Paper 1* and *Paper 2* as a given. Thus, the remainder of the *Context* briefly outlines the applied context, focusing almost entirely on organizational, practical, and design components which influence the simulated data used for *Paper 1* and *Paper 2*. The *Context* concludes with an introduction to the history and mechanisms of GLMM trees.

9G-OTG & Graduation

Data access stemmed from the award of IES Grant Number R305S210005 (Farley et al., 2021) to ODE and UO. The funded collaboration enabled investigation of Oregon's high school success initiative on 9th grade on track to graduation rates and facilitated access to the state's SLDS. *Paper 3* presents prior evaluations of the predictivity of 9G-OTG on graduation rates conducted by the Oregon Department of Education using the SLDS. With identifiers in the data (e.g., school, district, residential zip code), Gerring's process of discovery can be undertaken with these data in a way that has not been done before. Specifically, SLDS data can be joined with multidimensional, publicly available repositories of data to either verify the invariant influence of 9G-OTG on graduation, or to identify avenues of heterogeneity warranting

additional focus. For example, GLMM can parse multilevel factors and find a subgroup with a particularly high intercept, perhaps pointing to school- or zip code-level factors which mitigate the risk of being off-track. Alternatively, subgroups with particularly high slopes may be identified, returning factors that may increase the predictivity of 9G-OTG. Regardless of the findings, these EDA results can provide impetus to probe factors which may mitigate or enhance the predictivity of 9G-OTG.

Conclusion

Presently, the use of GLMM trees to assess variation in the relationship of 9G-OTG and graduation should be met with caution. First, other implementations of GLMM trees lack careful hyperparameter tuning to optimize classification accuracy (Huber et al., 2022; Brown et al., 2022; Johnson et al. 2016, Quan et al. 2020) or only mention that it was done without discussing the impact (Rusch et al., 2013; Tiendrébéogo et al., 2019). It is unclear if this is because the algorithm underlying GLMM trees is unlikely to overfit training data—as is argued by the authors (Hothorn et al., 2006)—or if classification accuracy of GLMM trees can be improved with hyperparameter tuning, as is the norm with other machine learning approaches (Yang & Shami, 2020; Hertel et al., 2021; Elshawi et al., 2019). Additionally, the integration of MOB with GLMMs by Fokkema et al. (2018) is problematic in cases whereby the GLMM random effect structure is more complex than a single level of clustering. As of now, it is unclear if GLMM trees will consistently identify the most influential variable when measured across different clustering units, and—if so—if a specific procedure for estimating random effects is optimal.

Herein, I propose a project broken into three individual papers to increase applied researchers' comfortability with GLMM trees. In the first two papers, I use simulation to fill the

aforementioned gaps in knowledge when applying GLMM trees. I then apply the knowledge gained from those simulations to real-world educational datasets in the third paper. The three papers are:

1. Paper 1. Influence of Hyperparameter Tuning on GLMM Tree Performance,
2. Paper 2. Singular- versus Dual- Corrections for Clustering in GLMM trees in Nested and Cross-Classified Designs, and
3. Paper 3. Uncovering Heterogeneous Predictivity in Graduation Early Warning Indicator among Two Cohorts of the Oregon State Longitudinal Data System

The proposed dissertation integrates diverse data sources and explores multidimensional influences on the association of 9G-OTG and graduation. Neighborhood characteristics, district characteristics, or other nesting units may interact to differentially impact students in currently unknown ways. By identifying previously undocumented compounding effects, stakeholders can craft more holistic approaches to supporting students.

The results from this dissertation may shine light on where schools should allocate resources and for the value of the 9G-OTG metric in consistently predicting graduation. If models identify large differences across intercepts, this will indicate systematically lower or higher log-odds of graduating when off-track for one or more group(s) and/or circumstance(s). Qualitative research can follow up with what is working for the former and what is not in the latter. Better understanding of which could be used to create support systems, identify groups/circumstances for targeted interventions, or modify the 9G-OTG metric to reflect odds of graduating more accurately.

Paper 1: Influence of Hyperparameter Tuning on GLMM Tree Performance

Machine learning (ML) models have modifiable configurations, hyperparameters, which determine the nuances of how they function. The prefix "hyper" is used to emphasize that such parameters are *not* estimated by the model. Hyperparameters fundamentally differ from model-estimated parameters (e.g., a regression coefficient, a correlation coefficient) as parameters are determined from the data, whereas hyperparameters are pre-specified by the analyst (Yang & Shami, 2020). Hyperparameters can be thought of as boundaries in which the machine learning model must function or guidelines that the model must follow (Yang & Shami, 2020). A common example is that of a camera, whereby the resultant image is the parameter and the camera settings (e.g., focal length, aperture, etc.) are hyperparameters. Just as different cameras have different available settings, different models may have different numbers and combinations of hyperparameters. Different cameras require their own level of user-modification to capture a good image when the context varies (e.g., lighting, distance etc.). ML models are similar, meaning combinations of hyperparameters can be calibrated to match the ML task optimally (Yang & Shami, 2020). Capturing a process with an untuned machine learning algorithm is like taking a picture with an out-of-focus camera. In other words, appropriate hyperparameter-data match is important to maximize a model's validity and subsequent stakeholder confidence in the use of modeling results for high-stakes decision making. The process of calibrating hyperparameters for machine learning algorithms is more commonly called hyperparameter tuning or hyperparameter optimization (HPO; Yang & Shami, 2020).

Modeling frameworks which do not allow for hyperparameter tuning are like disposable cameras, imposing the default settings on every image. Arguably, ordinary least squares regression is a (very powerful) disposable camera by this description if used with default

configurations (e.g., stats::lm(); R Core Team, 2024). Analogue cameras fit into the analogy, too, as models which perform well after manual calibration are reliant on the user. Modern digital cameras focus on an image by automated processes, which is what automated hyperparameter tuning methods do as well.

Applying the camera analogy back to the dissertation, *Papers 1 to 3* revolve around the GLMM tree, the camera used to capture invariance in the relationship between on-time graduation and a 9th grade early warning indicator. As a relatively new camera, the user manual lacks documentation on image clarity by setting. For that reason, *Paper 1* reports on the outcomes associated with three procedures designed to investigate the "camera's" performance: (a) use a disposable camera, (b) focus the camera into one of a few positions, and (c) use algorithms to focus on the image. Herein, I simulate data structures which mimic the Oregon's 9G-OTG distribution five-hundred times, take a picture (fit model) with three ways of focusing (tuning hyperparameters) the camera (GLMM trees), average the results, and demonstrate which method results in the clearest image.

The reason for this approach can be stated in metaphor or statistically. Metaphorically, testing the camera's snapshot, manual focus, and auto focusing mechanisms is essential because humans lack the specialized structures to determine if statistical associations are "in focus," compared to the structures—the eye and brain—which allow clear perception of an out-of-focus picture. Hence, *Paper 1* is needed to ensure relationships captured in *Paper 3* are not from an out-of-focus camera. The process outlined here also serves to decrease the number of decision points and possible means by which a user can bias model selection (e.g., chasing $p < 0.05$) and builds out the user manual for GLMM trees. With this knowledge in hand, I can have confidence in the quality of the relationships captured in *Paper 3*.

Parameters & Hyperparameters

Because they are well-known, decision trees—also referred to as classification and regression trees (CARTs)—are a good place to begin understanding hyperparameters. In a decision tree, one hyperparameter is the number of splits the model is allowed to make (Therneau and Atkinson, 2022). If a large number of splits are allowed, the data will be separated into a larger number of sparse groups. If a small number of splits are allowed, only a small number of large groups will be returned from the model. Just as a single camera cannot capture every possible image well, preferences for hyperparameter state cannot be generalized across contexts. Hyperparameter tuning is thus required.

The importance of hyperparameters and their relation to model-estimated parameters can be understood by considering the goal and boundaries of a machine learning model. The goal of a machine learning model is to estimate the optimal set of parameters to describe the data, *considering the constraints applied through hyperparameters*. In other words, hyperparameters are part of the model architecture *which the model cannot modify* during estimation (Yang & Shami, 2020). In contrast, parameters are initialized and updated during model training (Yang & Shami, 2020). Because different model structures are allowed based on the hyperparameters, the properties of the resultant terminal model (i.e., the parameters) are dependent upon the hyperparameters.

At the simplest level, estimated parameters could be the mean and variance for continuous outcomes (i.e., a regression tree) or the modal group and misspecification rate for categorical outcomes (i.e., a classification tree; Luo, 2016). Here, the number of splits is the hyperparameter and the mean or mode are model-estimated parameters, (Luo, 2016).

Hyperparameter States & Spaces

Models often have many hyperparameters; some are continuous, others are categorical (Yang & Shami, 2020). The overall configuration of hyperparameters is referred to throughout as the *hyperparameter state* and the array of possible hyperparameter states are referred to as the *hyperparameter space*. The dimensionality of the hyperparameter space is determined by the number of hyperparameters and the possible values each can hold.

The *optimal* hyperparameter space is determined *in relation to something*, which is often a measure of a model's ability to predict outcomes on an unseen subset of data (Yang and Shami, 2020). The portion of the data given to the machine learning algorithm to estimate parameters is called the *training data* and the unseen subset is known as *testing data*. Optimal state within the hyperparameter space is evaluated by:

1. fitting a model to training data across varied hyperparameter states,
2. using the trained models to make predictions on testing data,
3. comparing predictions to known outcomes from the testing data,
4. aggregating model performance (e.g., classification accuracy, mean absolute error).

Now, each hyperparameter state has a corresponding fitness value, defined as an aggregation of the model accuracy across testing cases, allowing selection of the optimal hyperparameter state (Yang & Shami, 2020; Kuhn and Johnson, 2019). The trained models have their own parameters that were estimated using both the training data and the constraints applied by the hyperparameters (Kuhn and Johnson, 2019). The use of testing data ensures that the models are evaluated on their ability to generalize to new data, which is the goal of ML (Belkin et al., 2019). Tuning is important to prevent models from overgeneralizing patterns from the training data, a process known as overfitting (Yang and Shami, 2020). By contrast, when a

model fails to incorporate important variance from the training data, the model is underfit. Tuning helps to strike a balance between over- and underfitting (Belkin, et al., 2019). This tuning process is what maximizes information extracted from the training data while minimizing ungeneralizable noise from the training data; this is the "classical" answer to the problems surrounding the bias-variance trade-off (Belkin et al., 2019). The hyperparameter state which balances overfitting and underfitting is said to minimize training risk, making it the optimal model (Yang & Shami, 2020). Because ML models tend to be tuned, inferential models tend to not be tuned, and GLMM trees are a hybrid, simulation is an essential tool to understand how models perform in new spaces, as GLMM trees are being applied in *Paper 3*.

Hyperparameter Optimization

By definition, optimization is done with respect to a metric of model performance. For models with categorical outcomes, model quality is often assessed with raw classification accuracy, area under the receiver operating characteristic curve, or Gini criterion (Kuhn & Johnson, 2019; Kuhn & Silge, 2022). When outcomes are continuous, alternate measures of fit to the data are employed such as root mean square error (RMSE) or mean absolute error (MAE; Kuhn & Silge, 2022). Regardless of a which metric of model performance, each of these metrics all provide a measure of how well the model is able to predict outcomes on unseen data, given the same set of "features" (i.e., an ML term analogous to "covariates" in statistical circles).

HPO is complex, not only because different modeling approaches have their own set of hyperparameters, but also because the optimal combination varies in each application depending on the data (Luo, 2016; Yang & Shami, 2020). Further, as dimensionality of the hyperparameter space increases, so does the difficulty of finding the global optimum. As such, various approaches have been proposed to cover the hyperparameter space efficiently (Yang & Shami,

2020; Kuhn and Silge, 2022). When the hyperparameter and objective function creates a convex, differentiable space, algorithms can leverage gradient descent to determine the optimal space. However, such methods require a continuous gradient that can only be formed from continuous features, meaning other strategies have been proposed to efficiently cover the hyperparameter. Yang & Shami (2020) explain that any technique can be used for HPO as long as it has:

> *[1] an estimator (a regressor or a classifier) with its objective function, [2] a search space (configuration space), [3] a search or optimization method used to find hyperparameter combinations, and [4] an evaluation function to compare the performance of different hyperparameter configurations.*

In *Paper 1*, Yang and Shami's (2020) four components of HPO are represented as

1. GLMM trees,
2. Available hyperparameter space,
3. Three different methods (discussed below),
4. Classification accuracy; Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC).

By holding components 1, 2, and 4 as constants, I systematically assess the performance of GLMM trees across the 3 HPO approaches discussed below. To root this paper within the overall dissertation again, *Paper 1* creates a protocol which *Paper 3* follows. Before comparing the performance of these HPO approaches, I describe the three approaches to tuning, discuss their benefits versus drawbacks, and outline why I am testing these HPO approaches instead of other contenders.

Simple Approaches to Hyperparameter Optimization

Tuning approaches vary widely, and tuning is simply a modification of hyperparameters which includes—but is not limited to—algorithmic HPO (Yang & Shami, 2020; Yu and Zhu,

2020). Non-systematic tuning practices are inefficient and less reproducible, and so they are not discussed further.

Yu and Zhu (2020) write "*Grid search is the most straightforward search algorithm that leads to the most accurate predictions as long as sufficient resources are given, the user can always find the optimal combination*" (p. 14). Exhaustive grid search is thus a trade-off between having greater certainty that you covered the entire hyperparameter space for a dedicated amount of computational time (Alibrahim and Ludwig, 2021; Yu and Zhu, 2020). Therefore, its useful to systematically decrease the number of hyperparameters tested (Kuhn & Silge, 2022). Algorithms can decrease runtime through selection hyperparameter states that mathematically cover the largest amount of the hyperparameter space. This algorithmic approach to filling the hyperparameter space is referred to as a space-filling design (Kuhn & Silge, 2022; Santer, Williams, and Notz, 2018).

Maximum entropy grids and Latin-hypercubes are two commonly used examples of space-filling algorithms, both of which use conceptually similar but mathematically distinct approaches to sample as much of the space as possible with a user-specified number of points (Kuhn and Silge, 2022). See Chapter 5 of Santer, Williams, and Notz (2018) for a review of space-filling designs for computer experiments as differential performance of the space-filling algorithms is beyond the scope of this dissertation.

Automated Hyperparameter Optimization

A movement towards more reproducible and systematic hyperparameter tuning through algorithms—also known as an automated machine learning or "AutoML" framework—is gaining popularity (Elshawi, Maher & Sakr, 2019; Hertel, Baldi, & Gillen, 2021). AutoML approaches create a reproducible approach wherein the user does not directly choose all testable

hyperparameter states (Elshawi, Maher, & Sakr, 2019). Returning to the camera analogy presented above, AutoML approaches can be thought of as digital cameras which are focused without user specification.

There are many AutoML approaches to HPO, with applicability of a given approach determined by the hyperparameter space (Yang & Shami, 2020). Models with all continuous hyperparameters can leverage some HPO approaches that cannot be applied directly to categorical or discrete hyperparameters (Yang & Shami, 2020). Batch gradient descent and many of its variants require a continuous gradient of fit values that can be optimized (Yang & Shami, 2020; Ruder, 2017). This means models with categorical or conditional hyperparameters (i.e., those with values that vary based on another hyperparameter) gradient descent is not appropriate (Yang & Shami, 2020; Ruder, 2017; Nguyen, et al., 2020).

Meta-heuristic or population-based optimization algorithms offer many benefits over long-standing optimization approaches such as gradient-descent in tree-based models (Alibrahim & Ludwig, 2021; Yang & Shami, 2020). Meta-heuristic algorithms draw inspiration from biological processes (Yang & Shami, 2020). The genetic algorithm (GA) is one such metaheuristic algorithm. Drawing from evolution, favorable genetic traits are more likely to be passed on to the next generation. In hyperparameter tuning, hyperparameters are traits or genes and the hyperparameter state is the entire chromosome. The favorability of a hyperparameter state is defined by the fitness function, which can be some measure of fit to the data (e.g., classification accuracy, RMSE, MAE, etc.).

The possible shape of the hyperparameter space (i.e., the population's total composition of chromosomes) is determined by providing the algorithm with a range of possible values or upper and lower bounds. Random initializations are unlikely to produce chromosomes which

optimize the fitness function, so the algorithm modifies the chromosomes in many ways. Each iteration of the algorithm represents a generation of the population of genes. Chromosomes with higher fitness functions are retained in the population at higher rates, known as *selection* (Kuhn & Johnson, 2019). In addition to selection, crossover and mutation are used to increase odds of finding the global maximum. Crossover exchanges genes of chromosomes which were selected as promising candidates from the next generation, allowing new combinations to be created (Kuhn & Johnson, 2019).

The fitness of individuals in the population is evaluated by training the machine learning model with the corresponding set of hyperparameters and a subset of the data, then evaluating its performance on another portion of the data set. As high performing hyperparameter states are established, information is passed to the next batch of possible values or *generations* (Kuhn and Johnson, 2019). Generations continue to be created until improvement in performance plateaus. As implied by the name, *mutations* are random changes to the hyperparameter state (Nguyen et al., 2020). Mutations are introduced into generations to retain diversity in the population and decrease the odds of convergence on local minimum by continuously exploring different areas within the hyperparameter space. By mimicking nature's desire to optimize a population's reproductive success, the genetic algorithm can optimize complex hyperparameter spaces (Nguyen et al., 2020).

HPO is an extensive field in and of itself, meaning the aforementioned methods are nowhere near the entirety of possible approaches. Grid search and genetic algorithms represent two viable means to improve GLMM trees, as this model has categorical and continuous hyperparameters. Other algorithms include particle swarm optimization (Yang & Shami, 2020) as well as Bayesian Optimization and variants (Falkner et al., 2018; Alibrahim & Ludwig, 2021),

and Tree-Structured Parzen Estimators (Bergstra et al., 2011; Nguyen et al., 2020). These methods are beyond the scope of this dissertation but offer similar promise as genetic algorithms as means of optimizing GLMM trees.

State of the Field: HPO in Model-Based Recursive Partitioning

As discussed in the *Context* section, model-based recursive partitioning (MOB) is a general method encompassing the generalized linear mixed-effects model (GLMM) trees. The state of the field for GLMM trees and hyperparameter tuning was so small that the was broadened to include tuning in MOB generally. Even then, hyperparameter tuning has been ignored or done poorly thus far in MOB. For example, Huber et al. (2022) discuss hyperparameters with MOB, but choose values directly instead of evaluating various hyperparameter combinations. Brown et al. (2022) also use manual selection with a liberal alpha (0.9), allowing more splits to be identified, but did not report results from other values of alpha or any other tuned hyperparameters.

Other analysts fit their MOB with a handful of hyperparameter values, conducting a grid search of one or more hyperparameters. Johnson et al. (2016) only tuned one hyperparameter, the minimum number of observations per terminal node (minsize). They selected one of five tested values (10%, 20%, 30%, 40%, and 50% of the original sample) which minimized mean absolute error, demonstrating a grid search approach to HPO, but their coverage of the hyperparameter space was very low. In their application, Kern et al. (2021) also only tuned one parameter—maxdepth, the parameter indicating the number of splits the data can make. They explored integers between 2 and 10 with a grid search and mention two other hyperparameters (alpha and minsplit), but they did not vary these.

Quan et al. (2020) conducted hyperparameter tuning using a grid search, but only report the optimal hyperparameter values, without explaining the impact of these on any aspect of the model. Rusch et al. (2013) conduct hyperparameter tuning but use a footnote to explain that the results of tuning were not of main interest and were not explained or discussed. Tiendrébéogo et al. (2019) explain in text that they tuned minimum node size and the alpha level, but do not describe their methodology or findings from tuning. Two unpublished articles—one Master's thesis (Jorink, 2018) and one preprint (Fokkema & Zeileis, 2023)—deserve mention as they borrow influence from the HPO literature. However, these approaches differ notably from standard tuning procedures and deal with the primary research questions in *Paper 2.*

Taken together, the state of the field for hyperparameter tuning in MOB is mostly an area of inadequate exploration. At best, MOB is tuned with poor documentation, prohibitively small hyperparameter spaces, and little evidence-based selection of why a given tuning approach was taken. As stated before, HPO is an integral part of ML (Belkin et al., 2019; Yang & Shami, 2020), but a pervasive lack of systematicity in tuning opens the door for implementing poorly fitted models and poor practice in implementing ML (Hertel et al., 2021; Elshawi 2019). Hertel et al. (2021) explain how poor reporting and documentation of tuning leads to hyperparameter settings being copied from previously published research, despite the fact those hyperparameter states have no reason to generalize to the new context. To prevent misapplication of GLMM trees and optimize their validity, data-driven protocols must be established for tuning.

Gap in the Field

So far, MOB authors who conduct any form of hyperparameter tuning have not publicly reported a comparison across tuning approaches. In other words, it is unclear if hyperparameter tuning is omitted accidentally or intentionally. Perhaps those who have tried hyperparameter

tuning have found no added value and omitted the section from their publications (i.e., an instance of the "file drawer problem"; Rosenthal, 1979). Filling this gap helps to establish standard protocol when applying a new model and clarify if the GLMM trees can be improved with tuning.

An alternative possibility for the lack of hyperparameter tuning in GLMM trees is based in its lineage, specifically in a quote from the invention of unbiased recursive Hothorn et al. (2006) writes about their unbiased conditional inference tree (ctree):

> *Although it is possible to choose α in a data-dependent way when prediction accuracy is the main focus, the empirical experiments in Section 6 show that the classical convention of α = 0.05 performs well compared to tree models optimizing the prediction accuracy directly.*

In the Section 6, Hothorn et al. (2006) compared optimized classification and regression trees to untuned ctrees to demonstrate equivalent performance. This comparison was useful to show that ctree could match the performance of these well-known models when tuned. However, Hothorn et al. (2006) did not compare the performance of a tuned ctree to an untuned ctree. To increase uptake of their method, Hothorn et al. (2006) instead suggest that applied researchers would be more comfortable with *α*—one of the primary hyperparameters—as a *"pre-defined nominal level of hypothesis tests rather than as a fine-tuned hyperparameter*." This quote provides a historical explanation as to why models in the lineage of ctree—including MOB and GLMM trees—often forego the practice and/or a discussion of tuning when applied.

A third reason why HPO might not be conducted regards the training and lineage of those using these methods. GLMM trees were first published in *Behavioral Research* (Fokkema et al. 2018), a journal with an audience comprised of far more practitioners and applied researchers, as opposed to applied statisticians or data scientists familiar with machine learning. In sum, hyperparameter tuning has been given inadequate focus in GLMM trees and MOB, and even in

cases when analysts have reported that a model was tuned, a systematic discussion of how hyperparameters impact predictive and inferential aspects of the model has been lacking.

Research Question & Hypotheses

The purpose of this investigation is to determine if GLMM trees can be improved by hyperparameter tuning, a standard practice for improving generalizability and parsimony in algorithmic modeling frameworks (Yang & Shami, 2020; Hertel et al., 2019; Elshawi, et al., 2019). The research question asks if tuned GLMM trees outperform default GLMM trees across three dimensions:

1. Akaike (AIC) and Bayesian (BIC) Information Criteria,
2. classification accuracy,
3. number of terminal nodes.

This paper leverages two approaches to tuning hyperparameters—maximum entropy grid search (MEGS) and GA—both of which were chosen for their ability to work with discrete, categorical, and continuous hyperparameters. Each of these metrics have their own utility. Information criteria provide a means of assessing the relative fit of these models to the data, classification accuracy provides an assessment of the model's external validity, and number of terminal nodes demonstrates model parsimony. Because GLMM trees are built upon a variation of decision tree, GLMM trees with HPO are expected to outperform untuned GLMM trees. Specifically, higher classification accuracy, lower information criteria, and fewer terminal nodes are expected in the tuned models compared to the untuned model.

Method

Three conditions were used to test the performance of GLMM trees, each with increasing complexity and computational cost. These represented a spectrum of simple to complex

hyperparameter tuning approaches: no consideration (i.e., default/untuned), tuning hyperparameters with a MEGS, and tuning hyperparameters with genetic algorithm (GA). Each of the three approaches—untuned, GS-tuned, and GA-tuned GLMM trees—were repeated 500 times starting with a seed to ensure reproducibility.

Simulating Data

Outcome, Covariates, & Coefficients

Properties of the data were established in the process of drafting Scalise et al. (2023), which predicted on-time graduation with 9G-OTG among two cohorts from the Oregon SLDS. Small tweaks to the model, including additional covariates, random slopes, and an interaction account for the slight differences between the simulated values and those reported by Scalise et al. (2023).

The log-odds of on-track to graduate status was simulated to be dependent upon a 9G-OTG stand-in and four other effects. The 9G-OTG stand-in was a dichotomous level-1 covariate simulated to occur at the same frequency as 9G-OTG in Oregon (84.5% on-track) and impact the outcome as seen in the observed association of 9G-OTG on on-time graduation (an impact corresponding to an odds ratio of 4.95; Scalise et al. 2023 reported 4.86).

The other four variables were: a level-1 continuous variable (*mean* = 0, *standard deviation* [*SD*] = 1), a level-1 dichotomous variable (*probability* = 0.5), a level-2 continuous variable (*mean* = 0, *SD* = 1), and a level-2 dichotomous variable (*probability* = 0.5). Magnitudes of the association between the outcome and these four coefficients varied across each data set; their magnitudes were uncorrelated and were drawn from a random normal distribution (*mean* = 0, *SD* = 0.2). To mitigate any possible effect of sampling bias, the same simulated dataset was passed to all three conditions (untuned, GS-tuned, and GA-tuned) each of the 500 times. For the

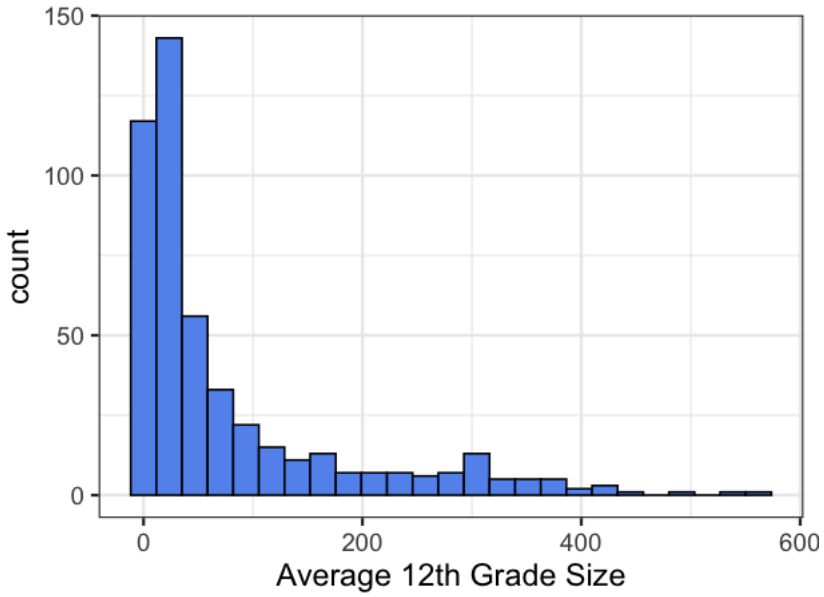
same reason, the same observations were partitioned into training (75%) and testing (25%) across the conditions.

Sample Size

The number of level-1 units per level-2 units were made to mirror the number of high school students in a graduating class observed in Oregon. This was done by sampling (with replacement) the counts of seniors per year from all years of observed data. The observed distribution is shown in Figure 1.

Figure 1

Histogram displaying the distribution of average 12th grade class size expressed in 25 bins.



Due to prohibitive runtimes, the number of level-2 units were constrained to $j = 50$ in each iteration. The GA made it through less than 5 of the desired 500 iterations in roughly 96 hours (~4 days), even distributing the process in parallel across 10 cores. After, j = 200 was attempted unsuccessfully (roughly 64 iterations completed after 192 hours), before settling on j = 50. Though important to mimic *Paper 3*'s data structure as much as possible, the focus of *Paper 1* is to evaluate differential performance of two HPO methods against the default metric making

use of identical data sets across condition most important for *Paper 1*. Stated formally, the data simulated were as follows:

$$outcome_i \sim Binomial\left(n = 1, prob_1 = \widehat{\mathrm{P}}\right)$$

$$\log\left[\frac{\widehat{\mathrm{P}}}{1-\widehat{\mathrm{P}}}\right] =$$

$$\alpha_{j[i]} + \beta_1 * (\text{main\_predictor\_lv1}) + \beta_2 * (continuous_a\_lv1) + \beta_3 * (dicotomous_b\_lv1)$$

$$\alpha_j \sim N\left(\gamma_0^{\alpha} + \beta_4 * (continuous\_lv2) \ + \beta_5 * (dichotomous\_lv2), \sigma^2_{\alpha_j} = 1\right),$$

$$\text{for School}_j \ = [1, 50]$$

$$\gamma_0^{\alpha} \ = \ 2.4 \text{ (Scalise et al., 2023) report 2.4}$$

$$\beta_1 \ = \ 1.6 \text{ (Scalise et al., 2023) report 1.58}$$

$$\beta_2, \beta_3, \beta_4, \beta_5 \sim N(\mu \ = \ 0, \sigma^2 = 0.04)$$

Such a distribution of regression coefficients β was chosen empirically. Each variable in *Paper 3*'s dataset was correlated with on-time graduation, then random sampling of this distribution was plotted. Data were then simulated to match the average distribution across samples. From such a distribution of β, the maximum average effect of the covariates was much smaller than that from the main predictor of interest—*main_predictor_lv1*— the dichotomous stand in for 9G-OTG with a probability of roughly 0.85 (the observed value for the entire sample in *Paper 3*). Four more variables—two dichotomous and two continuous—spread across level 1 and level 2 were included with zero influence on the outcome and are not shown in the equation.

Conditions

For the default condition, models were fitted using default hyperparameters for each of the 500 simulated training data sets and their respective classification accuracies were assessed on unseen data. MEGS covered the hyperparameter space with only 10 distinct states. Each

hyperparameter state was assessed over the k-folds and the classification accuracy of unseen testing data was averaged.

Finally, for each of the 500 data sets, hyperparameter tuning was conducted using a genetic algorithm to optimize the average fit across the k-folds based on training data. As with MEGS, the optimal hyperparameter state found by the GA was then used to assess the classification accuracy of the entire sample. The same 500 simulated datasets (with identical training/testing splits) were provided to each of the models to prevent differences in sampling bias. Table 1 presents an outline of the conditions tested, and all inputs and outputs.

Table 1

Conditions tested and corresponding processes in assessment of hyperparameter tuning on GLMM trees.

Condition	**Default Hyperparameters**	**Tuned via Maximum Entropy Bounded Grid Search (MEGS)**	**Tuned via Genetic Algorithm (GA)**
Input	500 Simulated Data Sets, pre-split (75% for model training; 25% for testing unseen classification accuracy)	500 Simulated Data Sets, pre-split (75% for model training; 25% for testing unseen classification accuracy) A Defined Hyperparameter Space: Possible ranges (continuous) and values (categorical) hyperparameters	500 Simulated Data Sets, pre-split (75% for model training; 25% for testing unseen classification accuracy) A Defined Hyperparameter Space: Possible ranges (continuous) and values (categorical) hyperparameters.
Choosing Hyperparameter States	Use Model Defaults	Initialize a maximum entropy grid of 10 hyperparameter states from hyper parameter space. Create 5-fold Cross-Validation Object from training data. For each fold of the cross-validated object { Fit a model with the hyperparameter grid and k-fold training data.	Randomly initialize 16 hyperparameter states (“chromosomes”) from hyperparameter space as first generation of the genetic algorithm. Create 5-fold Cross-Validation Object from training data. For each fold of the cross-validated object {

Condition	Default Hyperparameters	Tuned via Maximum Entropy Bounded Grid Search (MEGS)	Tuned via Genetic Algorithm (GA)
		Assess classification accuracy with k-fold testing data. } Take mean of classification accuracy for each hyperparameter state across the five folds Select hyperparameter state with best mean classification accuracy across 5 folds.	Fit a model with the hyperparameter grid and k-fold training data. Assess classification accuracy with k-fold testing data. } Take mean classification accuracy for each chromosome (hyperparameter state) across the five folds Select the best performing hyperparameter state from the initial population and use these as the starting values in the next generation. Conduct random variations ("mutations") in each of these hyperparameter states ("chromosomes"). Allow new chromosomes to cross-over (i.e., randomly shuffle some pieces of hyperparameters to create new combinations). Repeat this process up to 100 times or until no improvement has occurred in within 10 generations (i.e., no improvement in 16 x 10 = 160 consecutive models). Select hyperparameters which met the stopping criteria above.
Process	Using default hyperparameters: 1. Fit model with training data, extract AIC, BIC, and number of terminal nodes 2. Predict outcomes of withheld final testing data (25% of original sample)	Using hyperparameters selected after MEGS: 1. Fit model with training data, extract AIC, BIC, and number of terminal nodes 2. Predict outcomes of withheld final testing data (25% of original sample)	Using hyperparameters selected after genetic algorithm: 1. Fit model with training data, extract AIC, BIC, and number of terminal nodes 2. Predict outcomes of withheld final testing

Condition	Default Hyperparameters	Tuned via Maximum Entropy Bounded Grid Search (MEGS)	Tuned via Genetic Algorithm (GA)
	3. Calculate final classification accuracy with simulated values vs. predictions	3. Calculate final classification accuracy with simulated values vs. predictions	data (25% of original sample) 3. Calculate final classification accuracy with simulated values vs. predictions
Outputs	Classification accuracy; AIC; BIC; number of terminal nodes	Classification accuracy; AIC; BIC; number of terminal nodes	Classification accuracy; AIC; BIC; number of terminal nodes

Hyperparameter Space

The hyperparameter space of GLMM trees is complex. It contains categorical and continuous hyperparameters. Table 2 presents the boundaries of the hyperparameter space explored in these conditions. See Zeileis et al. (2008) and Fokkema et al. (2018) for greater explanation of hyperparameters. Because the paper focuses on HPO with cross-validation to decrease overfitting, the prune parameter can be confusing. Here is how Zeileis et al. (2008) define the *prune* parameter The documentation for MOB explains the prune parameter (from calling function `*?mob_control()*`, no page number; Zeileis et al., 2008)

> *In mob-based model trees, pre-pruning based on p-values is used by default and often no post-pruning is necessary in such trees. However, if pre-pruning is switched off (by using a large alpha) or does is not sufficient (e.g., possibly in large samples) the prune method can be used for subsequent post-pruning based on information criteria.*

Cross-Validation & Hyperparameter Tuning

In each of the 500 repetitions, the 75% of data used for training were split with k-fold cross validation (k = 5). K-fold cross validation is a type of resampling used to prevent overfitting of model training data (Kuhn & Johnson, 2019). By using 5-fold cross validation, each 20% portion of the training data was used to both estimate the model (four times) and test the model fit (one time), thus using all data maximally during this process (Kuhn & Johnson, 2019).

Table 2

Hyperparameters selected for tuning GLMM trees.

Name	Meaning	Boundaries of Tested Hyperparameter Space
alpha	Minimum threshold of significance in parameter instability tests required for a split to be taken.	[0.01, 0.33]
minsize	Minimum number of observations in a node.	[5%, 25%]
maxdepth	Maximum depth of the tree (number of total splits the model is allowed to find).	[2, 20]
trim	The proportion of outliers removed in assessing instability of a given split (i..e., in calculating instability statistic and p-value)	[0.01, 0.3]
prune	Should models be pruned by AIC, BIC, or do not prune?	c(AIC, BIC, no pruning)
bonferroni	Use Bonferroni corrections to adjust threshold of instability statistic?	TRUE/FALSE
breakties	Should mathematical ties be broken randomly?	TRUE/FALSE
restart	Restart model estimation with NULL starting values?	Re-estimate, use priors
ranefstart	Should the algorithm initialize with the tree or with random effects	Tree First / Random Effects First

Table 2 outlines the conditions tested. For models tuned with MEGS, the hyperparameter state which resulted in the highest average proportion of accurately classified testing data across the k-folds was chosen. The genetic algorithm was allowed to initialize a group of 16 hyperparameter states—also referred to as chromosomes—and their classification accuracies were tested. The best fitting chromosomes were selected, random mutations were made to the genes (i.e., allowing individual hyperparameters to vary randomly), and chromosome portions were exchanged randomly to introduce more variation in the next generation of chromosomes (i.e., the next 16 hyperparameter states tested). When the average classification accuracy of tuned models across 5-folds could not be improved in 10 generations (i.e., 10 x 16 = 160 hyperparameter states), the genetic algorithm converged and returned these best fitting hyperparameters. In addition to final classification accuracy of unseen testing data, the

information criteria (i.e., Akaike, Bayesian), and the number of terminal nodes were determined for these models.

Estimated Model

All models were specified equivalently with a logistic regression of a dichotomous outcome on a dichotomous predictor with 8 splitting variables, comprised of a pair of continuous and pair of categorical variables at both level-1 and level-2. Stated formally, the GLMM was:

$$outcome_i \sim Binomial\left(n = 1, prob_1 = \widehat{\mathrm{P}}\right)$$

$$\log\left[\frac{\widehat{\mathrm{P}}}{1-\widehat{\mathrm{P}}}\right] = \alpha_{j[i]} + \beta_1 * (\text{main\_predictor\_lv1})$$

$$\alpha_j \sim N\left(\gamma_0^{\alpha}, \sigma_{\alpha_j}^2\right)$$

A total of 12 splitting variables were included in the model, the 4 simulated to have influence on the outcome (see above) and the 8 nuisance variables. Thus, GLMM trees explored variation in the above multilevel logistic regression parameters across 12 level-1 and level-2 covariates.

Software

All analyses were conducted using R version 4.3.1 (R Core Team, 2024). All data manipulation and visualization were done with the *tidyverse* suite of packages (Wickham et al., 2019). The *glmertree* package was used for fitting GLMM trees (Fokkema et al., 2018), and the *gardenr* package (Loan, 2023) was used for tuning hyperparameters on GLMM trees. The *rsample* package (Frick, et al., 2023) was used to split training/testing data, create cross-validation datasets from training data, and create the maximum entropy grid used in grid search. The *GA* package was used for genetic algorithms (Scrucca, 2013).

Evaluation of Results

Simulation results were evaluated by regressing model performance on the following fixed-effects: an intercept (α_j; default model performance), a dummy-code indicating MEGS-tuned (β_1), and a dummy-code indicating GA-tuned (β_2). The same 500 simulated datasets (and training/testing splits) were fitted to each of the three conditions to decreasing the influence of sampling bias on the results. Likelihood ratio tests (LRTs) were used to assess if random variation in model was sufficient to be modeled as a random effect by simulated data index (i.e., set 1 to 500). Significant LRTs and lower information criteria (AIC, BIC) dictate evaluation of models with the following multilevel model:

$$Model\ Performance_i = \alpha_{j[i]} + \beta_1(grid\ search) + \beta_2\ (genetic\ algorithm)$$

$$\alpha_j \sim N\left(\mu_{\alpha_j}, \sigma^2_{\alpha_j}\right), \text{ for data index}_j = [1, 500]$$

Non-significant LRTs and higher information criteria (AIC, BIC) in the multilevel model would support the use of the single-level variant of this model to evaluate performance, meaning:

$$Model\ Performance_i = \alpha + \beta_1(grid\ search) + \beta_2\ (genetic\ algorithm)$$

Results

In all conditions—classification accuracy, AIC, and BIC—a multilevel model allowing a random intercept by simulated dataset was preferred to single-level models which estimated one intercept for all datasets (Table 3). BIC, AIC, and LRTs used to compare the validity of fixed versus random intercept models preferred the multilevel model (Table 3). Three multilevel models had continuous outcomes and were estimated with identity link functions, but number of terminal nodes was assessed with a Poisson link function to handle the skewed count distribution. In all cases, the effect of tuning approach on number of terminal nodes was

evaluated with a multilevel model because all metrics—ΔBIC, ΔAIC, LRTs—suggested preference for multilevel models over a single level model.

Table 3

Information criteria and results from LRTs comparing single- versus multilevel-models to evaluate model performance.

Outcome	ΔBIC	ΔAIC	Δχ2	Δdf	p
Classification accuracy	5419.1	5424.5	5426.50	1	< 0.001
AIC	7317	7323	7325.10	1	< 0.001
BIC	3784	3789	3791.10	1	< 0.001
Number Terminal Nodes	87.4	92.7	94.67	1	<0.001

Δ *are reported as the improvement observed by the multilevel model over the single-level mode. df = degrees of freedom.*

Multilevel models were thus used to compare model performance by regressing the model fit metric (i.e., classification accuracy, AIC, BIC, number of terminal nodes) on the dummy-coded condition. Table 4 outlines properties of random intercepts for each of the three models for each performance metric.

Table 4

Description of random intercepts from models comparing performance metrics.

Performance Metric	Variance	Standard Deviation	Range
Classification Accuracy	0.00026	0.016	[-0.054, 0.034]
Akaike Information Criterion	87873.41	296.43	[-632.46, 953.72]
Bayesian Information Criterion	93079.91	305.09	[-643.40, 1031.82]

Performance of GLMM Trees by HPO Method

On average, GLMM trees with the default hyperparameters were able to accurately classify 0.942 of the proportion of the sample or ~94.2%, and no statistically significant differences were observed between the classification accuracies of either MEGS or GA (Table 5). As suggested by comparing their estimates and standard errors, swapping the dummy-coded references to estimate a difference between MEGS and GA demonstrated a null difference between the conditions across all metrics (additional results available upon request). Information criteria (AIC and BIC) and the final number of terminal nodes were similarly regressed onto the dummy coded conditions (Table 5).

Table 5 shows statistically lower BIC for tuned models relative to defaults, null differences in AIC, and significantly fewer terminal nodes in tuned models, relative to defaults. Taken together, it appears that (a) GLMM trees can be trained to predict unseen data with great accuracy, (b) classification accuracy is not significantly improved by hyperparameter tuning, and (c) hyperparameter tuning can find more parsimonious specifications of GLMM trees with equivalent predictive accuracies to default models according to BIC and number of terminal nodes.

Table 5

Parameter Estimates for 3 Multilevel Models Comparing Model Properties by Condition

	Estimate	Standard Error	t-statistic	Degrees of freedom[1]	p	95% Lower CI	95% Upper CI
Classification Accuracy (Unseen Data)							
(Intercept)	**0.942**	**<0.001**	**1305.692**	**501**	**<0.001**	**0.940**	**0.943**
GA	-0.000	<0.001	-0.679	998	0.498	-0.000	0.000
MEGS	-0.000	<0.001	-0.100	998	0.92	-0.000	0.000
BIC Final							
(Intercept)	**1210.749**	**13.724**	**88.224**	**506**	**<0.001**	**1183.787**	**1237.711**
GA	**-25.275**	**2.018**	**-12.525**	**998**	**<0.001**	**-29.235**	**-21.315**
MEGS	**-25.621**	**2.018**	**-12.696**	**998**	**<0.001**	**-29.581**	**-21.661**
AIC Final							
(Intercept)	**1152.829**	**1.259**	**86.944**	**499**	**<0.001**	**1126.778**	**1178.881**
GA	-0.007	0.365	-0.019	998	0.985	-0.723	0.709
MEGS	-0.649	0.365	-1.779	998	0.075	-1.366	0.067
Final Number of Terminal Nodes[2]							
(Intercept)	**3.230**	**0.081**	**39.777**	**—**	**—**	**3.071**	**3.389**
GA	**-1.406**	**0.107**	**-13.115**	**—**	**—**	**-1.616**	**-1.196**
MEGS	**-1.388**	**0.107**	**-12.947**	**—**	**—**	**-1.598**	**-1.178**

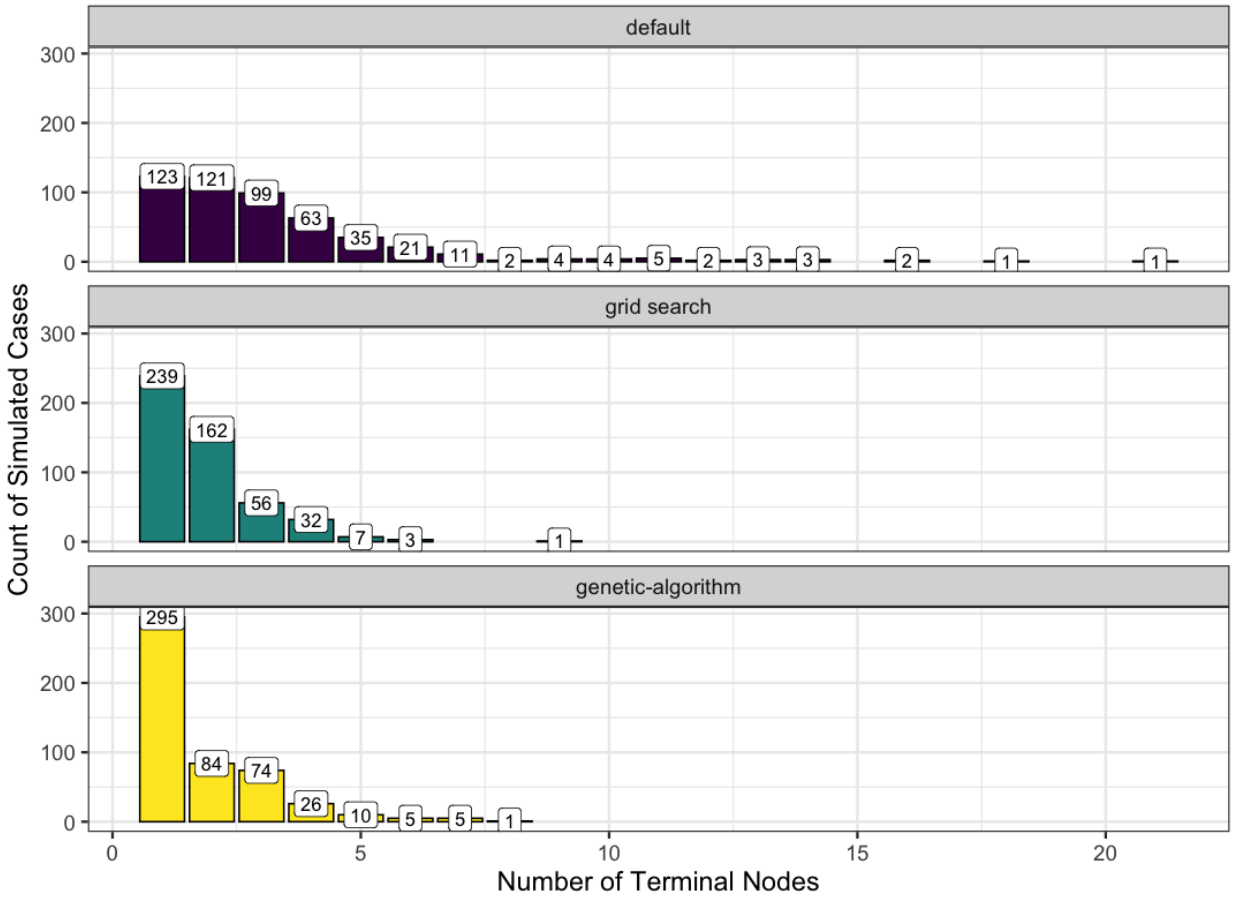
[1]Satterwaite degrees of freedom estimates are non-integer but have been rounded.
[2]Satterwaite degrees of freedom approximations not available for Poisson models; see Bates et al. (2015) for a discussion of the complexity in determining degrees of freedom in multilevel models.
CI = confidence interval.

The distributions of the terminal nodes varied by tuning approach. All were highly skewed, being count variables (Figure 2). Models with default hyperparameters had a mean of 3.23 terminal nodes (SE = 0.120), models trained through MEGS had a mean number of terminal nodes of 1.84 (SE = 0.048), and models trained with genetic algorithm had a mean of 1.82 (SE =

0.056). The default model was either the largest or tied for largest 79.4% of cases, whereas the genetic algorithm was largest or tied for largest 34.8% of the time.

Figure 2

Terminal node counts by approach for each tuning method.



Across simulations, models trained with default hyperparameters had ≥5 or ≥6 terminal nodes 18.8% and 11.7% of the time, respectively. Both tuned approaches had relatively few terminal nodes, compared to the default. Only 2.2% and 0.8% of MEGS simulations resulted in solutions with ≥5 or ≥6 terminal nodes, respectively. When tuned with genetic algorithm, roughly 4.2% and 2.2% of simulations had ≥5 or ≥6 terminal nodes, respectively.

For the majority of the 500 simulations, default-hyperparameter models resulted in more terminal nodes compared to models trained with MEGS (54.8%) and GA (59.2%). Roughly a

third of simulations resulted in the same number of terminal nodes across untuned and tuned models. The average deviation of an approach from the data index's average (i.e., mean), the number of terminal nodes, was assessed by taking within dataset averages, subtracting the group mean number of terminal nodes from the observation's mean, and then averaging the deviation from the mean number of terminal nodes. For default models, the average deviation from the mean by data index was 0.93 (SD = 1.78; SE = 0.08). For models trained using MEGS, deviation from the data index mean was -0.457 (SD = 1.10; SE = 0.05), and for those with genetic algorithms the deviation from the data index mean of terminal nodes was -0.475 (SD = 1.17; SE = 0.05). Figure 3 presents number of terminal nodes by tuning approach for each data index, ordered by average number of terminal nodes.

Figure 3

Number of terminal nodes by tuning approach for all data indices; points are made semi-transparent and jittered by 0.25-points horizontally to increase visibility.

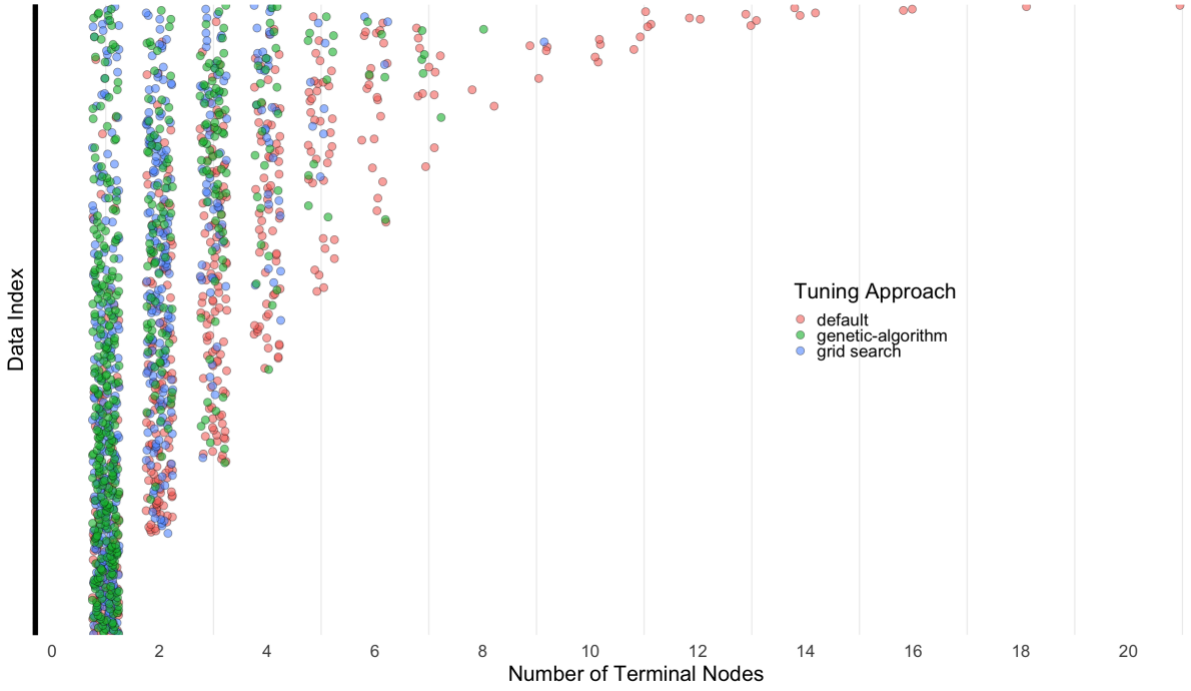


Figure 3 adds to the information from Figure 2 by showing the difference in model size for tuned versus untuned models. Figure 3 demonstrates the frequency with which default models are larger than tuned models. In the most extreme examples—where tuning picks a hyperparameter state that prevents any splitting—the default model finds 16 to 18 times more splits than tuned models. Figure 3 demonstrates the variance in distribution of number of terminal nodes by approach (Table 5).

Figure 4 shows the classification accuracy for all data indices, ordered by the average classification accuracy of data index across tuning approaches. In total, 341 of 500 data sets resulted in identical classification accuracy across tuning condition, and those identical accuracies were removed to improve the interpretability of Figure 4. With a relatively even distribution of approaches within a data index, Figure 4 demonstrates the lack of a direct difference in classification accuracy by tuning approach.

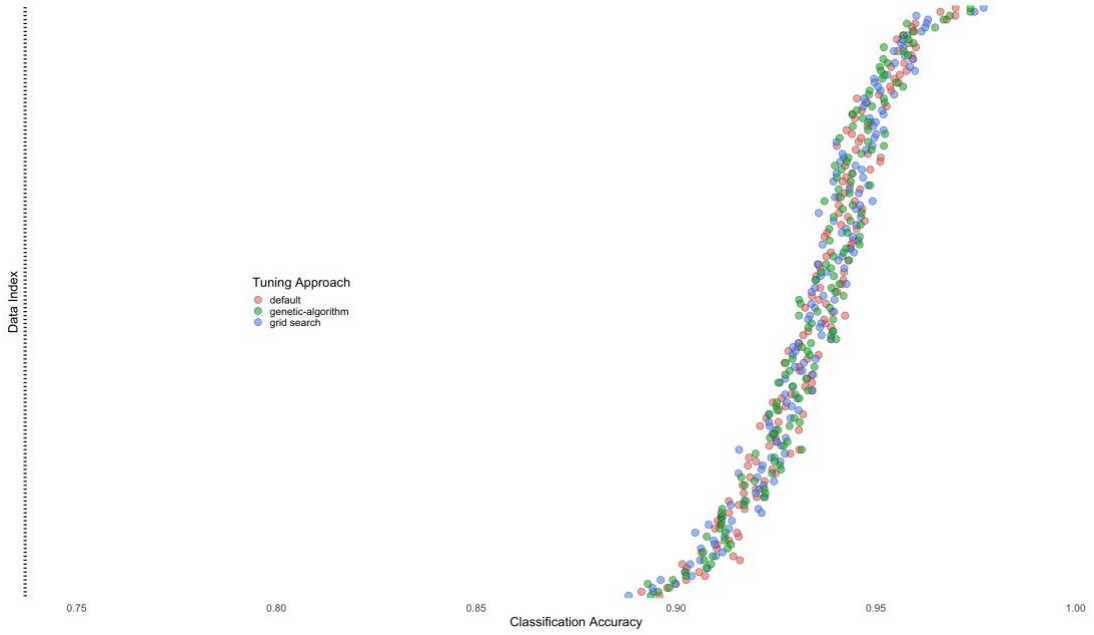
Association between Classification Accuracy & Number of Terminal Nodes

Across conditions, there was a negative correlation between number of terminal nodes and classification accuracy (r = -0.23, t = -9.36, df = 1498, $p < 0.001$). This bivariate correlation provides a rough assessment of the relationship between the variables. Proper specification of the exact relationship is difficult within a linear modeling framework, though. In other words, the effect of going from 1 to 2 terminal nodes may not reasonably be the same as 2 to 3, and so on. Furthermore, the effect may vary across groups, meaning a GLMM or related model can provide robust results only if piecewise-dummy codes were fitted by approach and number of terminal nodes. In the final dataset, this includes 51 regression coefficients between the intercept (b = 1), main effects (b = 23), and interactions (p = 27). Though possible, such a model is extremely

overparameterized and requires comparison of conditions to be done over dozens of parameters.

Figure 4

Classification accuracy by tuning approach for all data indices; points are made semi-transparent and jittered by 0.005.



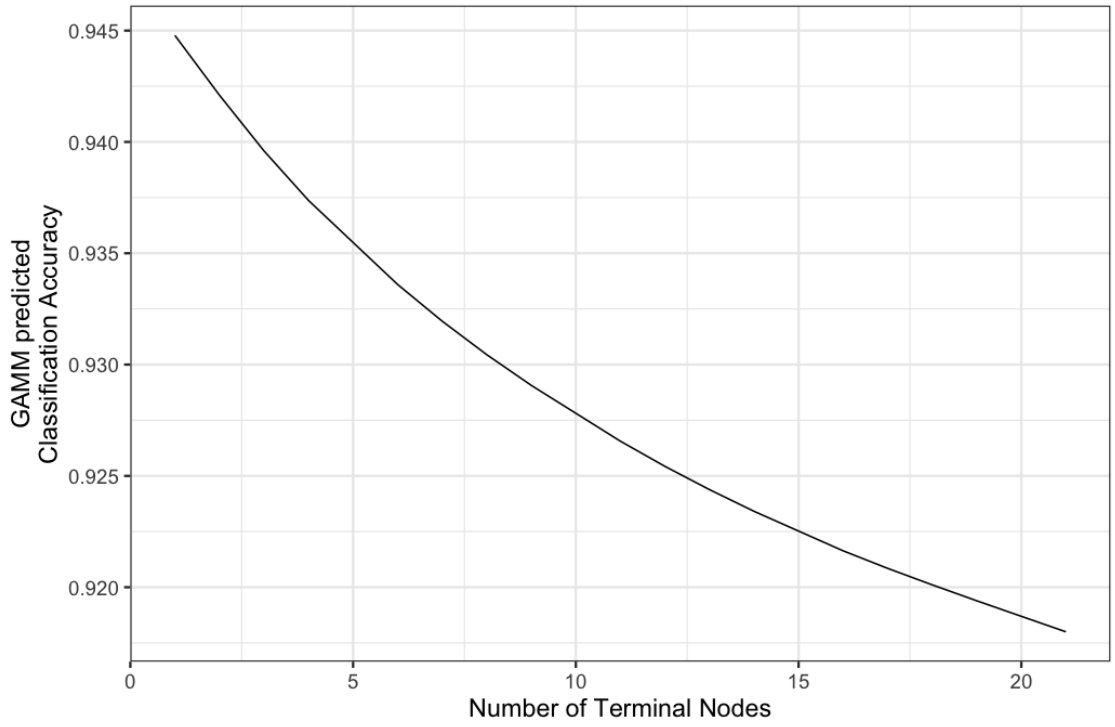
Rather, the generalized additive mixed-effects model (GAMM) framework was used to confirm the effect documented by the bivariate correlation at the start of this paragraph (Lin & Zhang, 1999). The GAMM use "*additive nonparametric functions*" to create a smoothed, parameterizable functions "*to model covariate effects while accounting for overdispersion and correlation [with] random effects*" (Lin & Zhang, 1999, p. 382). The smoothed functions are referred to as splines and these models contain knots or curves in the relationship of the predictor and outcome (Wood, 2003; Wood et al., 2016; Wood, 2017). Thus, splines convert the unknown functional form between number of terminal nodes and classification accuracy into a smoothed function. The results allowing for better inference and prediction than occurs when imposing an overly simplistic functional form on the data (Wood, 2003; Wood, Pya, & Safken., 2016; Wood,

2017). However, it is also unclear if splines should be fit by group or not (i.e., an interaction of approach x non-linear influence of number of terminal nodes on classification accuracy). Fokkema (n.d.) has integrated GAMMs into the MOB framework, which is uniquely suited for regressing GAMMs and conducting MOB on this parametric model because it is unclear if splines should be fit by approach (i.e., interactions with spline coefficient) or not.

Parameter instability tests did not return evidence for instability across the groups (*Instability Statistic* = 26.85, $p = 0.14$), meaning the estimated model had no evidence of subgroup effects by tuning approach (i.e., no partitions). As with other instances of MOB, a lack of instability returns the baseline model, here a traditional GAMM. The GAMM estimated 9 knots in the spline, resulting in a shallow, non-linear descent in classification accuracy of GLMM trees. Rather than providing spline coefficients, a partial dependency plot was constructed to visualize the estimated decrease (Figure 5; Greenwell, 2017). Partial dependency plots (PDPs) are "*low-dimensional graphical renderings of the prediction function so that the relationship between the outcome and predictors of interest can be more easily understood*," which are especially useful when "*large observational databases [do not] adhere to the strict assumptions imposed by traditional statistical techniques*" (Greenwell, 2017, p. 421). See Greenwell (2017) for more information on PDPs.

Figure 5

PDP from GAMM regressing classification accuracy on number of terminal nodes.

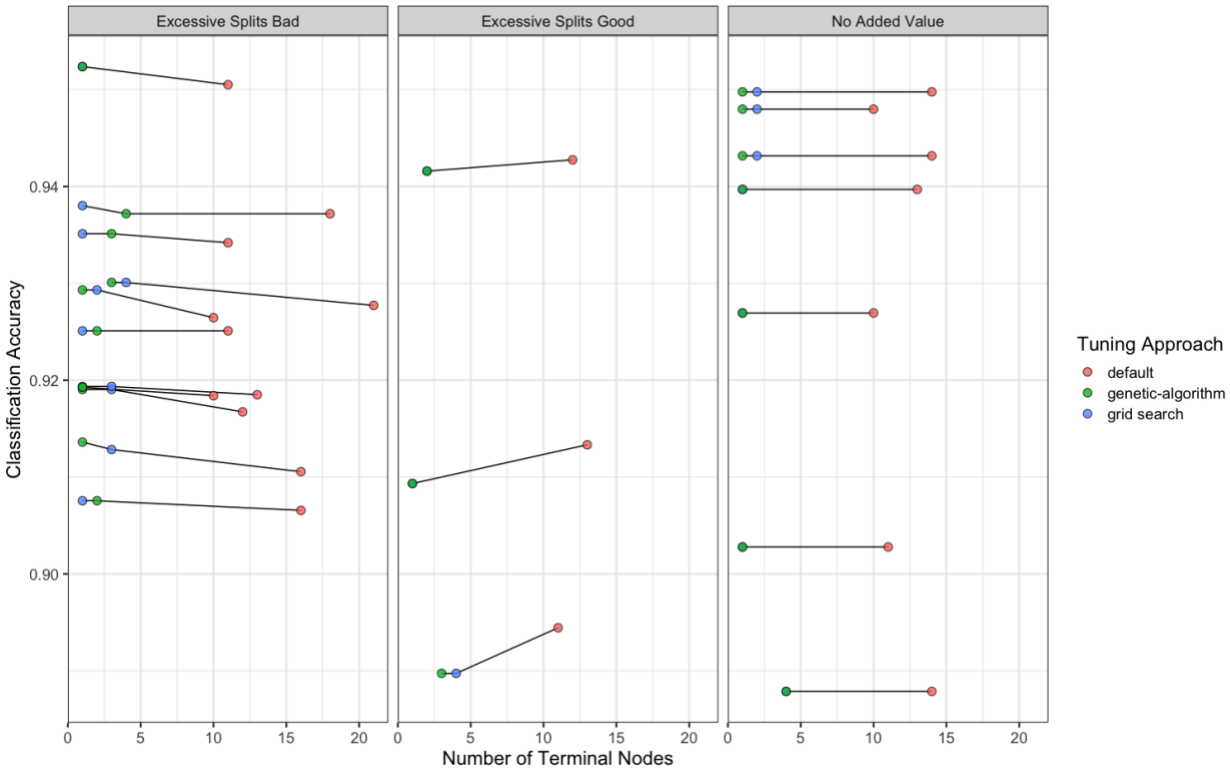


Number of Terminal Node Outliers

In addition to these diagnostics, cases at the tail of the distribution of number of terminal nodes were visually inspected to ensure they were consistent with the average result (i.e., greater performance with fewer terminal nodes). Data indices which resulted in untuned models having more than 9 terminal nodes (i.e., the maximum observed in either tuning approach) were plotted as classification accuracy (y-axis) by number of terminal nodes (x-axis) in Figure 6.

Figure 6

Classification accuracy by number of terminal nodes and tuning approach among the largest models (≥10 splits); black lines connect observations with same training data.



In most cases, classification accuracy of unseen data was nearly identical for smaller (tuned) models when fit to the same data as larger (untuned) models. In 32 models, the untuned model found 10 or more splits (Figure 6). Only 3 of those 32 models displayed higher performance of the larger untuned models, compared to the smaller, tuned models (Figure 6, Middle Panel). Such cases represent 3 of 500 total cases or 0.6% of the cases.

Discussion

The purpose of this paper was to determine if the performance of GLMM trees is improved by HPO—the norm for ML models (Yang & Shami, 2020, Hertel et al. 2021; Elshawi et al., 2019)—or if the lack of systematic HPO in the literature was due to its lack of utility in

this hybrid inferential-ML framework, suggested by Hothorn et al. (2006) for *ctree* (the recursive partitioning algorithm underlying MOB). The primary goal of pure ML models is prediction (Belkin et al., 2019). MOB, and by extension GLMM trees, use both ML and inferential techniques to balance ML's goal of prediction with the goal of inference—interpretation (Zeileis et al., 2008; Fokkema et al., 2018). Interpretable models—barring differences in accuracy—should be as parsimonious as possible, not only because they are easier to interpret and communicate, but because they allow generation of more falsifiable hypotheses (Gerring, 2011). Together, these two goals dictated which metrics were used to evaluate GLMM trees: classification accuracy, AIC, BIC, and number of terminal nodes.

Model Performance by Tuning Approach

The influence of HPO on GLMM tree performance varied by metric, but together a clear narrative was formed. Overall, the results suggested HPO leads to equivalent predictive accuracy on unseen data despite using fewer parameters than untuned GLMM trees. With models sometimes 16 to 18 times smaller, theorists can generate more parsimonious, falsifiable theory in subsequent iterations of research. No evidence supported differences in classification accuracy or AIC between untuned GLMM trees and either HPO approach. Despite no evidence of differences in classification accuracy, tuned models were more parsimonious than untuned models, as shown by the fewer terminal nodes and lower BIC, compared to untuned models. Displaying equivalent predictive accuracy with greater parsimony, GLMM trees appear to be improved through HPO.

Model Size & Performance

Investigation into the relationship between classification accuracy and number of terminal nodes identified a negative correlation between classification accuracy and size of the

estimated model in a non-linear fashion. Recursive partitioning of the non-linear model with GAMM trees (Fokkema, n.d.) demonstrated that the decrease was invariant across tuning approach, demonstrating adherence to the classical paradigm of the bias-variance tradeoff (Belkin et al., 2019).

Furthermore, minimal differences were found comparing (a) 10 well-spread hyperparameter states (i.e., MEGS) to (b) data-based traversal of the hyperparameter space (i.e., GA). For large datasets like this process modeled by the SLDS, exploring a relatively small hyperparameter space with grid search appears to be adequate. Substantial improvements are not seen by implementing many more states with a GA, an important finding for computational complexity. With these implementations, maximum entropy grid search matched the performance of genetic algorithms despite the former only trying 10 values and the latter trying 160 (at minimum). From a purely predictive standpoint, little value may be gained by tuning. From a theory-generating standpoint, however, these results suggest GLMM trees are in fact improved by hyperparameter tuning.

Having the same classification accuracy with fewer splits suggests that something differs in the grouping of the data by HPO approach. Likely, the combination of splits and the location of splits on continuous variables improve when tuning the model, meaning they better represent the underlying process estimated by the parametric model with observed data. Regardless, under these conditions HPO appears to be uncorrelated with differential classification accuracy in a direct way. The indirect impact of parsimonious models correlates with greater accuracy, though, supporting HPO with GLMM trees.

Limitations

This study was intentionally designed to mimic the data structures observed in the Oregon SLDS. Thus, limitations are presented in the context of the SLDS as well as to other social science research questions. In the context of the SLDS and *Paper 3*, the greatest limitation in implementing these findings were the use of only 50 level-2 units to model the Oregon SLDS. The exact numbers vary by year (e.g., new schools opening, schools closing, small schools with no students in a certain class), but 429 and 431 schools had seniors in the student-level data sent from ODE in *Paper 3* (i.e., high schools, K-12, etc), and cases were limited to 265 and 263 schools after removing missing data. The data analyzed above had 50 level-2 units per simulated dataset. The decrease was necessary due to computational complexity of repeating the genetic algorithm 500 times on such a large sample. Underestimating the sample size is less problematic than overestimating because precision in estimates of increases with sample size. In other words, true SLDS data—with many more level-2 observations—should be able to predict level-2 effects on unseen data more accurately than our process.

Considering the relatively unusual size of the data in the SLDS, hesitation should be taken in generalizing these findings to other data. Simulated data were tailored to the SLDS and the underlying process of *Paper 3*, which has unusual properties in social science individually, let alone in combination. First, the outcome is extremely unbalanced, with the probability of 1 (i.e., graduating) being much higher than 0 (i.e., not graduating). This fact is so true that a model which predicted everyone graduates would score at least an 80%, depending on the year. Second, the main covariate (9G-OTG) is observed at a similarly high probability. Intertwining these problems is the magnitude of association between these variables, as the predictivity of other variables pale in comparison to the effect of the simulated predictor. Sample size, clustering

structure, and dependencies within this data further separate the process modeled in *Paper 3* from much of social science research.

Generalization of these results to MOB models other than GLMM trees should be met with caution. In MOB, the fitness of the parametric model defines the outcome for recursive partitioning (Zeileis et al., 2008). As such, models with fitness functions robust to partitioned effects may not see as large of a benefit from HPO, compared to those whose fitness functions vary widely. To provide an analogy, imagine the fitness function estimated by MOB during HPO is a field and we want a view from the highest point in the field. Instantiations of MOB which result in very stable estimates of fitness across subgroups are like flat fields. In flat fields, the view is similar at all points in the field. An obvious example of this would be tuning a process with absolutely no subgroup effects, meaning the field is perfectly flat and the view is always identical. Other parametric models, though, might result in very large variations in fitness, making a landscape of hills and valleys. In such variable terrain, HPO would be able to find locations with extremely different viewpoints. Analysts working with models, data, and processes that are extremely variable should pay additional attention to HPO, though, our highly stable process with very small effects occasionally returned very large differences, highlighting its utility even when fields are relatively flat.

As with all ML models, adequate coverage of the hyperparameter space is a primary concern in practice (Yang & Shami, 2020). This project only assessed one implementation of GA and MEGS. Any process used to sample the hyperparameter space has its own hyperparameters. In other words, GA, MEGS, and any algorithm besides an *exhaustive* search can also be tuned (Yang & Shami, 2020). This project used one implementation of GA and MEGS. With GA, for example, the amount of random chance in the model (mutation rate), the number of

hyperparameter states (chromosomes), and the way new hyperparameter states are chosen from existing states (e.g., crossover, selection, elitism) are all hyperparameters which can be tuned. MEGS has fewer hyperparameters, but still has several of its own including the number of states to cover the space, the number of iterations before the final states are chosen from the space, and the likelihood of returning regions of empty space (the variogram range). Possibly, other implementations of GA and MEGS would return different hyperparameter states. With GA, however, the probability of better solutions might be unlikely in data like those in *Paper 1* and *Paper 3*, because all simulated results were terminated from the early stopping criteria (i.e., no improvement after 10 iterations or 10 x 16 = 160 states) rather than the maximum number of iterations (i.e., 50 iterations or 50 x 16 = 800 hyperparameter states).

Next Directions

Additional research is needed before generalizing results to other data and modeling frameworks within MOB. Each property of the parametric model should be varied systematically before generalizing to other GLMM tree implementations. Such properties include number of level-2 units, number of level-1 units per level-2 unit, the magnitude of the effect, the link function of the GLMM, the variance-covariance matrix, clustering structure, and more.

Future work should assess the differential validity of a given HPO across its own hyperparameter states (e.g., mutation, cross-over, elitism, etc.). For GA in particular, the model typically found the correct solution in the first or second iteration. Researchers must consider their research question and the variance in their fitness functions to find the comfortable balance between computational resources and likelihood of finding a better hyperparameter state. As others have stated, exhaustive grid search is the only way to be certain to identify the best fit, though, this is rarely done in practice (Yang & Shami, 2020).

Contributions to Paper 3

These results clarify the best practice for *Paper 3*, specifically showing that HPO is important for theory generation due to its ability to match predictive accuracy with more parsimonious models. Without evidence of significant differences between GA and MEGS, MEGS was selected due to the decrease in run-time to estimate a small number (e.g., 10) states, compared to (e.g., ≥ 160).

Singular- versus Dual- Corrections for Clustering in GLMM trees in Nested and Cross-Classified Designs

As with Paper 1, this paper starts with an analogy. Here, GLMM trees are like an organism that is the byproduct of biological evolution. As Charles Darwin observed in 1859: "*Organs or parts in this strange condition [rudimentary, atrophied, or aborted], bearing the stamp of inutility, are extremely common throughout nature*" (p. 418). Since Darwin, many others have documented such atrophied structures, which are now referred to as "vestigial" (Fong et al., 1995). Several evolutionary theories as to how vestigialization may occur are outlined by Fong, Kane, and Culver (1995). Put broadly, circumstantial or environmental factors change, and prior functionality is no longer ideal or necessary. A striking example given by Fong et al. (1995) is that of male, adult mayflies who live only briefly to mate, and thus many mouthparts are vestigial and nonfunctional. Here, the adult male mayfly has no need for a mouth, yet mouthparts are observable. With a species' driving factor being reproduction, trying to use its mouthparts distracts the mayfly from its evolved purpose—reproduction. By relying on the necessary tools and ignoring superfluous tools, the male mayfly is better suited for the needs of its environment than it would be otherwise (Fong et al., 1995).

In the same way, I believe that GLMM trees have their own vestigial structure—cluster-robust corrections to parameter stability tests (described below)—which was required for proper functioning of its predecessor model (i.e., the non-multilevel general linear model [GLM] tree) and are not necessary in GLMM trees. As happens to species undergoing vestigialization, a combination of external factors and responses to those factors create the circumstances which allowed GLMM trees to retain a functionality it no longer needs. Close examination of the development of GLMM trees elucidates which factors may have led GLMM trees to have this

optional parameter. In particular, three factors deserve discussion as primary contributors to vestigialization. First, disagreement about the nuances of cluster-robust corrections to standard errors are common even among experts discussing much more established modeling frameworks than these hybrid inferential and algorithmic models (see Abadie et al.'s [2017] heated response to Cameron and Miller's [2015] rules of thumb with clustered data). Second, the driving philosophy of the authors is one of innovation, advancement, and growth through action over inaction. Finally, a modular paradigm is both directly articulated by the research group (for decades) and was implemented in the underlying R code of GLMM trees (i.e., after estimating and offsetting for random effects, the GLMM tree function passes all arguments directly to its "parent" model, GLM tree, the single level model implemented in Zeileis et al., 2008).

The culmination of the three aforementioned factors is that it was easier to offer the option of the secondary correction and leave the discretion to the user of the model. Though flexible, this approach opens the door for confusion in best practice. According to the software documentation for GLMM trees, cluster-robust corrections are an optional part of the model, but random effects (as specified in the GLMM) are required components. Fokkema et al. (2018) do not use cluster-robust corrections in their vignette demonstrating the method, nor do they address the optional parameter, though it is included in the software documentation.

Jorink (2018)—an unpublished paper which explored this topic—argued for using random effects without additional corrections because it optimized predictive accuracy on unseen data. However, Jorink omits deeper discussion of the implication of their findings on other samples, implying that corrections are a hyperparameter that can be optimized. As important context, Jorink's thesis was overseen and approved by the method developer, Marjoleon Fokkema, suggesting the method developer had prior knowledge about the

performance of singular versus dual correction around the time of the Fokkema et al. (2018) publication. Otherwise, Jorink (2018) likely would have relied on theory to justify one approach or the other. Another limitation of Jorink (2018), is that the author manually selected a hyperparameter state, forcing the results to be from the perspective of a tree having fewer than 4 partitions.

In September of 2023, Fokkema and Zeileis published a preprint which extends the work of Jorink (2018) with simulated data, rather than observed data. This pre-print was not available when my dissertation was proposed, and Part I of their paper conducts a roughly equivalent analysis to that presented in this paper (see their Table 1, middle section). There are numerous important differences however, including context, sample size, link function, simulation procedures, random effects structures that need to be considered. In both Jorink (2018) and Fokkema and Zeileis (2023), this vestigial hyperparameter is tested alongside another which determines if the model initializes with the tree (default) or random effects in the first iteration. *Paper 2* does the same, testing both combinations of algorithm initialization to ensure the (supposedly) vestigial structure is not simply a "conditional" hyperparameter that is important depending on the state used to initialize (Yang & Shami, 2020). When non-nested covariates are measured for the sample (e.g., zip-code-level covariates and school-level covariates), variables measured across clustering levels can compete for their relative influence on the association of 9G-OTG and graduation.

Therefore, the incremental progress of Paper 2 is to make a rule of thumb which extends to datasets similar to those tested via simulation. If—like Jorink (2018)—my findings support singular correction as more appropriate than dual corrections, there will be an additional piece of evidence that suggests the structure may be vestigial. In their investigation, Fokkema & Zeileis

(2023) report recovering the effect correctly 100% of the time in 3 of the 4 specifications used in this paper (i.e., their Table 1, middle section), with the worst performing of the 4 models identifying the effect correctly 99.6% of the time.

Though developed completely independently, *Paper 2* follows a procedure analogous to Fokkema and Zeileis (2023), but Paper 2 uses data similar to the SLDS across nested and cross-classified data. Fokkema and Zeileis do not include splitting variables across multiple levels simulataneously. Using *Paper 2* to determine *Paper 3*'s parameterization, the following questions are assessed:

1. Can GLMM trees accurately recover the most influential effect across more than one clustering level when nested (2-level) or not nested (cross-classified)?
2. Is the effectiveness (from #1) invariant across the 2 x 2 design settings tested by Jorink (2018) and Fokkema and Zeileis (2023)?

If the dual corrections lead to better performing models, GLMM trees will be restricted to exploring nesting along a single clustering dimension in Paper 3. Further, whichever condition of the 2x2 design performs best will be chosen. Through replication, synthesis, and formal mathematical proof, generalized answers can be established in time, building a literature alongside the prior investigations into parameterizing GLMM trees (Jorink, 2018; Fokkema & Zeileis, 2023). The results presented here should be true across similar datasets and models, including *Paper 3*, but should be interpreted with caution in new contexts, data, and models.

Accounting for Clustering in Education & Social Science

Students attending the same school share a pool of resources, ranging from physical structures to faculty, staff, and budget. Within a school, further similarities can be drawn among students in the same grade or classroom. Statistically speaking, these similarities can be seen as

"*correlation[s] among observations resulting from the multi-level structure*" (Aitkin & Longford, 1986, p. 2). The educational measurement and effectiveness literature has discussed the impact of multilevel structures on estimated parameters—and even demonstrated methods to account for clustering—for nearly forty years (Goldstein, 1986; Aitkin & Longford, 1986). When a researcher fails to properly account for the dependencies due to clustering, the estimated parameters and their standard errors can be biased (Aitkin & Longford, 1986; Baayen et al., 2008; Zeileis et al., 2020).

"The Consensus"

Many publications on clustering make concrete claims about how bias presents itself (e.g., Aitkin & Longford, 1986; Baayen et al., 2008; Zeileis et al., 2020; Moulton, 1990; Moulton 1986; Cameron & Miller, 2015), claiming that ignoring clustering in inferential models (always) artificially *deflates* standard errors, p-values, and the size of confidence intervals while (always) inflating the t-value, relative to what is representative of the population. In their review, Cameron and Miller (2015) outline three conditions under which parameters tend to have more bias when clustering is ignored, including if there are: (1) high within-cluster correlation of error terms, (2) high within-clusters correlation of included covariates, and (3) large numbers of observations per cluster. Each of Cameron and Miller's (2015) factors are present across large-scale assessments, clarifying why education research has advocated for multilevel solutions for so long (Goldstein, 1986) and why journals still publish papers that directly compare cluster-corrected to their uncorrected counterparts in education (e.g., Shirilla et al., 2022; Ker, 2014).

Cameron and Miller (2015) offer their understanding of the "consensus" of the field, and argue for the continuation of the practice: *"The consensus is to be conservative and avoid bias and to use bigger and more aggregate clusters when possible, up to and including the point at*

which there is concern about having too few clusters" (p. 17). In fact, Cameron and Miller (2015) prefer an overly conservative approach across any clustered data, advocating for calculation of standard errors both with and without cluster-robust corrections, and *"if there is an appreciable difference, then use cluster-robust standard errors"* (p. 17). The affinity of the field to this simple rule of thumb is evident by the more than 5000 citations Cameron and Miller's 2015 work has received (Google Scholar, 2024).

A Problem with "The Consensus," A Nuanced Statistical Perspective or Difference in Epistemology?

Despite the popularity of Cameron and Miller's (2015) propositions, a vocal group led by Abadie et al. (2017) believes this rule of thumb is incorrect and harmful. Countering the "*consensus*" viewpoint of applied methodologists, Abadie et al.'s argument received its fair share of attention by the field with over 2500 citations since 2017, according to Google Scholar (2024). Repeatedly, Abadie et al. (2017) directly quote Cameron & Miller (2015) to disagree with their most basic assertions regarding under what conditions researchers can expect greater deviations (i.e., the three conditions outlined above). In response to Cameron and Miller's (2015) supposition of a preference for being overly conservative, Abadie et al. (2017) simulate cases which directly contradict Cameron and Miller's (2015) assertions point-for-point after directly quoting Cameron and Miller's (2015) assertions one-by-one.

Abadie et al. (2017) argues that data are *"only partially informative about whether one should adjust the standard errors for clustering. A consequence is that in general clustering at too aggregate a level is not innocuous, and can lead to standard errors that are unnecessarily conservative, even in large samples" (p. 2).* Abadie et al.'s (2017) argument is that standard

errors should be adjusted only if (1) units in the sample were selected in a clustered way, (2) clusters in the population are not in the sample, or (3) assignment to treatment is clustered.

Full interpretation of Cameron and Miller's (2015) argument demonstrates their assumption of large-data spaces, stated clearly in their abstract and throughout the paper, but Abadie et al. (2017) explicitly state they disagree with the generalization even in large spaces. The arguments outlined in Abadie et al.'s and Cameron and Miller's paper have nuanced truths in the specifics of their statistical examples. However, below the discussion of correcting standard errors, there are tones of a pervasive and fundamental differences in epistemology between the authors. Years before this argument, Gerring (2011) points out what Abadie et al. (2017) is doing in this instance, acting as a "falsificationist." From *Social Science Methodology: A Unified Framework*, Gerring (2011, p. 33):

> *The falsificationist considers the greatest sins of social science to be those of commission, rather than omission. The virtue of good science is to keep quiet when the truth is ambiguous – not to say more than one knows with a reasonable level of certainty. Only in this fashion will the products of science be distinguishable from conjectures, the stock-in-trade of politicians, journalists, and cocktail-party prognosticators. […] Many social scientists have embraced this austere, taciturn view of science (at least rhetorically). Here, the primary job of the methodologist is to vigilantly guard the gates of science, ensuring that no unauthorized entrants are admitted.*

Thus, Gerring (2011) argued for two underlying, often implicit, beliefs as to which is worse, action or inaction, decades before this example of the phenomenon. Those in alignment with Abadie prefer inaction to an improper action, but those like Cameron and Miller believe that the truth will be made evident eventually through cumulative action followed by synthesis and pruning. Those more pained by improper action see growth as discovering an objective truth through a direct trajectory of science. Gerring (2011, p.33) argues that those most averse to improper action believe "*[o]nly if the field is clear of nonsense will the long, slow process of scientific cumulation occur."* Abadie et al.'s (2017) paper is a caricature of what Gerring (2011)

describe. By Abadie et al.'s (2017) own admission, Cameron and Miller's (2015) argument—that the "nest when the data are nested" rule of thumb—is correct more often than it is not. Unlike Abadie et al. (preferring inaction to improper action), Cameron and Miller (2015) offer a practical solution that decreases the barrier to entry in a practitioner focused journal. The valuation of action over inaction—even when errors may occur—implies knowledge generation through iterative process of trial-and-error, replication, and synthesis (Gerring, 2011).

Cameron and Miller's (2015) arguments imply one of three underlying beliefs all with the same outcome. First, perhaps Cameron and Miller (2015) believe a rule of thumb is required for applied researchers in nested settings to feel comfortable with using the method. Even by Abadie et al.'s (2017) admission, it is more common for the data and model to both be nested than for the data to be nested when the model should not be. Such a dichotomy among well-respected methodologists is not purely mathematical, but is biased by perceptions of the world in practice and the process of knowledge generation. Cameron and Miller (2015) advocate for action that works *most of the time,* in an attempt to decrease the barrier to entry, instill confidence in readers, and encourage continual well-intentioned action over inaction. Abadie et al. (2017) carry their own biases towards accuracy over pace of progress. In fact, Abadie et al. (2017) is a *National Bureau of Economic Research (NBER)* publication, an organization whose website boasts *"NBER's greatest asset is its reputation for scholarly integrity. Affiliated researchers are expected to conduct their research in ways that adhere to the highest standards of scientific conduct and that will not reflect adversely on the integrity of the NBER"* (NBER, 2024).

Second, Cameron and Miller (2015) implicitly may believe that the increase in power from the large sample could be considered enough to cancel out any artificial deflation of the

standard errors caused by unnecessarily clustering (the unlikely but possible outcome argued by Abadie et al.). Alternatively, these two camps may have an epistemological bias towards false negatives or false positives, with the other camp having the opposite bias. It is unclear based on the available information, but any of those options (or some combination) led to the same outcome: a divided philosophical camp amongst methodologists whereby rapid action (and implied pruning later) is pitted against slow deliberate growth.

Recontextualizing the conversation in *Paper 2* of my dissertation, the models discussed by both Cameron and Miller (2015) and Abadie et al. (2017) are models common to social science research, unlike GLMM trees. With such prominent methodologists arguing over the use of corrections in niche edge-cases, a (statistical) software developer would be incentivized to maximize flexibility in a novel statistical method. Perhaps this is why the secondary correction is optional and is discussed in the software manual but not in the original publication. As with an evolving organism in a harsh environment, a statistical model which is overly assertive may be met with harsh consequences, incentivizing the developer to offer both options for flexibility.

Modular Development

Although this paper is not aimed at programmers, a surface-level understanding of one programming paradigm—modular development—is important to continuing the analogy for GLMM trees containing a vestigial structure. See Jee (2021) for an explanation of modularization to a non-programmer). Jee defines modularization as *"the process of separating the functionality of a program into independent, interchangeable modules [or subprograms], such that each contains everything necessary to execute only one aspect of the desired functionality."*

As Zeileis wrote in 2004 (p. 2) "*applied researchers typically cannot wait until a certain procedure becomes available in the software package of their choice but are often forced to program new techniques themselves.*" This is why Zeileis and colleagues vocally argued for a modular approach for statistical software, e.g.,

> *just as suitable covariance estimators are routinely plugged into formulas in theoretical work, programmers should be enabled to plug in implementations of such estimators in computational work. Hence, the aim of this paper is to present an econometric computing approach [… of] reusable components which can be used as modular building blocks in implementing new inferential techniques and in applications*. (Zeileis, 2004 p. 2)

Zeileis' "economic computing approach" is therefore a modular one that allows researchers of varied expertise apply the method to their data and research question. Such an approach led to powerful developments by Zeileis and downstream research teams, including *mo*del-*b*ased (MOB) recursive partitioning, the general framework which subsumes GLMM trees (Fokkema et al., 2018; Zeileis et al., 2008), which is analogous to how the general linear model subsumes the t-test and other statistical models. Figure 1 in the *Context* section (above) shows a simplified confluence of algorithmic and statistical lineages required for the development of GLMM trees.

In the publication standing alongside the release of the *{partykit}* package—that which houses the MOB function—Hothorn and Zeileis (2015) explain why modular statistical programming approaches are an improvement over standalone algorithms:

> *[Many] algorithms and software implementations have to deal with very similar challenges. However, due to the fragmentation of the communities in which the corresponding research is published – ranging from statistics over machine learning to various applied fields – many discussions of the algorithms do not reuse established theoretical results and terminology. Similarly, there is no common "language" for the software implementations and different solutions are provided by different packages (even within R) with relatively little reuse of code. The partykit tries to address the latter point and improve the computational situation by providing a common unified infrastructure for recursive partytioning in the R system for statistical computing. In*

> *particular, partykit provides tools for representing fitted trees along with printing, plotting, and computing predictions. (pp. 1-2)*

The first step of MOB is to fit the data to a parametric model, meaning, the first module in the MOB algorithm is any pre-made statistical method which has been formatted in a specified way. For example, in the paper proposing MOB, linear regression trees are demonstrated, which relies on the *glm.fit()* function from R's built-in {*stats*} package (R Core Team, 2024; Zeileis et al., 2008). In another example, the authors demonstrate a type of survival analysis called a Weibull regression by calling upon functions (or modules) written in the {*survival*} package as the first step of the MOB algorithm (Therneau, 2015).

Any parametric model can be used as a module to start MOB, as long as it uses the expected inputs (e.g., formatted data, model specification) and provides the expected outputs (e.g., residuals, a fit function) to the next module. Since the development, other parametric frameworks have been incorporated into MOB including Rasch modeling (Strobl et al., 2013) structural equation modeling (Brandmaier et al., 2013; Arnold et al., 2021), and more (Jones et al., 2020; Lang, et al., 2020; Karapetyan et al., preprint).

Modularity in GLMM Trees

GLMM trees are yet another example of a parametric framework built as an instance of MOB, rather than as a standalone program. The modularity of *{partykit}* allowed GLMM trees to function by merging extant functions for multilevel modeling (e.g., glmer() and lmer() from *{lme4},* Bates et al., 2015) with the necessary infrastructure for MOB (Fokkema et al., 2018). The *glmertree()* function used for fitting GLMM trees spends 74 lines preparing data for the *lme4* model and the *mob* model. Then, predictions of random effects are made with data fitted to the *glmer()*function (used for multilevel GLMMs; Bates et al., 2015). By line 91 of the

glmertree() function, the transformed data and the random effects are passed to the *glmtree*() function, developed by Zeileis et al. (2008) for single-level data. As stated above, the *glmtree()* function passes the models' best prediction of the random effects as statistical offsets to the *glm.fit()* function. In *glm.fit()*, statistical offsets are values that are treated as fact, meaning they do not have standard errors. For linear models, identity link functions are used, meaning offsets are directly subtracted from the observed outcome and the transformed outcome before the *[outcome – offset]* is used as the new outcome of the linear model (R Core Team, 2024). For other (e.g., logistic) regression models, offsets are handled analogously but with transformations to account for the link function (R Core Team, 2024). The results from the offset GLM tree are used to improve predictions of the offsets (i.e., random effects) and the process continues recursively. In this way, GLMM trees are fully dependent upon GLM trees, allowing this complex method to be written in 182 lines of code. This modularity simplifies higher-level understanding of code and has other benefits (Jee, 2021), but also obfuscates the fact that some specifications of GLMM trees correct for standard errors two times, with one step being unnecessary at best and an avenue to inject bias in estimates at worst. After establishing some terminology, I present pseudocode which shows how GLMM trees can account for clustering two times.

Evaluating which Aspects of GLMM trees are Clustered

This paper assumes familiarity with the GLM, and defines terms in the GLM as done by Fokkema et al. (2018), specifically a GLM is defined as:

$$g(\mu_i)=x_i^T\beta$$

and

$$E[y_i|x_i] = \mu_i$$

where

g = link function (e.g., normal/identity, binomial, Poisson, etc.)

μ_i = the expectation of an outcome y_i , given regressors x_i

β = vector of fixed-effects regression coefficients

The GLM can be extended to incorporate cluster-specific random effects (i.e., a GLMM) as:

$$g(\mu_i) = x_i^T\beta + z_j^T b$$

where

z_i = is a unit vector of length M which is 1 for the M-th element and 0 for all others.

b = random vector of length M, corresponding to random coefficient of each cluster.

i = level-1 element

j = level-2 element

As implemented in the {*lme4*} (and therefore{*glmertree*} package), the GLMM assumes b is normally distributed with mean 0, variance σ_b^2 and that the errors ε have constant variance across clusters (Bates et al., 2015; Fokkema et al., 2018). In fact, transformed residuals are used to calculate the estimating functions against which GLMMs are optimized (e.g., log-likelihood, Bates et al., 2015). With optimizers striving to remove the clustered effect from residuals, it is difficult to understand why the cluster robust correction would have been added to GLMM trees if they were developed without the GLM tree.

GLMM trees estimate *global random effects* (i.e., for all data) and local fixed effect (i.e., by terminal node; Fokkema et al., 2018; Fokkema et al., 2021; Fokkema & Zeileis, 2023). This "global-local" approach was first proposed by Hajjem et al. (2011) and has been applied in other contexts before GLMM trees (e.g., Sela and Simonoff, 2012). GLMM trees treat the random effect as a *known, statistically exact* offset (i.e., directly subtracted from the link-function-

adjusted outcome with no imprecision) while determining group membership (Fokkema & Zeileis, 2023, Fokkema et al., 2018). It reasons that if clustering dimensions are accounted for by the random effects, there is no need to adjust standard errors along the same dimensions. If the outcome is adjusted by this offset, some or all cluster-level variance no longer contributes to the estimating function GLMM trees are optimizing.

Though outside the context of GLMM trees specifically, Abadie et al. (2017) emphasize and demonstrate problems from improperly correcting for clustering. From both camps (i.e., that of Abadie et al. [2017] and Cameron and Miller [2015]), the effect of students in schools should be clustered. Cameron and Miller's (2015) broad rule of thumb apply by the simple presence of clustering in a large sample. Abadie et al.'s (2017) camp would focus on the underlying population and how that relates to the sample. Specifically, new (statistically unrepresented) schools have emerged since this testing window and missing data leads to unrepresented cases, and GLMM trees lack methods for pooling multiple estimates presently. Though a small number of cases, ODE's institution lookup table (ODE, 2024-b) shows schools entering and exiting the population over time. Hence, the value of the GLMM to model dependencies of students in school is apparent from either perspective.

The second means of handling clustering, however, is not a matter of students in schools or anything else observable in the real world directly. The second means of clustering in GLMM trees applies corrections to the standard errors *of the test which drive variable selection* in the tree-based aspect of GLMM trees, *not directly on the observed data*. Specifically, cluster robust corrections are applied to the parameter stability tests, the mechanism of variable selection in GLMM trees. As a refresher, parameter stability tests assess variation in a coefficient (Hothorn et al., 2007)—which is the fit function of a GLM or GLMM in this context (e.g., log-likelihood;

Zeileis et al., 2008; Fokkema et al., 2018). Figure 1 provides pseudocode for the GLMM tree which simplifies the algorithm into (made-up) modules with intuitive names.

Figure 1

GLMM tree pseudocode, showing how GLM trees are offset by random effects (line 22) and an optional secondary correction (line 24).

```
1  # assumes default settings
2
3  function(data, input_formula){
4    ## initialize
5    loglik = -Inf
6    tolerance = 1e-4
7    random_effects = 0 # default algorithm initialization
8    improvement = 1
9
10   GLM_tree_formula ←
11     convert_to_GLM_tree_formula(input_formula)
12
13   GLMM_formula ←
14     convert_to_GLMM_formula(input_formula)
15
16   while (improvement > tolerance){
17
18     # estimate GLM tree
19     GLM_tree ←
20       glmtree(
21         GLM_tree_formula,
22         offset = random_effects,
23         ## OPTIONAL cluster parameter
24         # cluster = cluster_level
25       )
26
27     ## extract node for each observation
28     tree_node ←
29       extract_node_from_tree(GLM_tree)
30
31     ## modify the GLMM equation to account for nodes
32     GLMM_formula_by_node ←
33       update_glmm_formula_by_node(GLMM_formula, tree_node)
34
35     ## estimate the GLMM
36     segmented_GLMM ← glmer(GLMM_formula_by_node)
37
38     ## extract random effects
39     random_effects ← extract_random_effects(segmented_GLMM)
40
41     ## extract loglik
42     newloglik ← logLik(segmented_GLMM)
43
44     ## calculate improvment
45     improvement ← newloglik - oldloglik
46
47     ## replace old loglik
48     oldloglik ← newloglik
49   }
50 }
```

By default, GLMM trees initialize assuming all random effects are 0, then fit the GLM tree after offsetting for the random effects (which are all 0 in the first iteration; Figure 1). Such a process means the results from the first iteration of GLMM tree is simply applying a cluster-agnostic GLM tree. If the algorithm is set to initialize with estimated random effects, the order of events is reversed and estimated random effects used in the estimation of the GLM tree. Line 22 of Figure 1 shows where the variance from clustering is partitioned from the GLMM (Figure 1) and how the GLM tree is estimated with the variance from random effects removed from observations (Figure 1; Fokkema et al., 2018; Fokkema & Zeileis, 2023). Line 24 shows the location of the *optional* parameter stability test (Figure 1; Fokkema et al., 2018; Fokkema & Zeileis, 2023). Figure 1 emphasizes why Abadie's camp may argue this correction is not needed from a statistical or theoretical standpoint. Statistically, the effect of clustering in students is accounted for by random effects, are subtracted from the outcome when estimating the GLM in the GLM tree. Thus, the fixed effects estimated by the GLM in the GLM tree have removed the effect of clustering, meaning the residuals were estimated to be orthogonal to the effect of clustering on the outcome (Fokkema et al., 2018; Fokkema et al., preprint). These (unbiased) residuals are transformed to a measure of model fit (e.g., log-likelihood) and used as an outcome in parameter stability tests (the mechanism of variable selection). One assumption of the underlying generalized parameter stability tests is that the residuals are independent (Zeileis et al., 2007), which is why GLM tree (and many single-level MOB algorithms) adjustment of standard errors in parameter stability tests (Zeileis et al., 2008). Unlike the residuals of GLMM trees, the residuals of other MOB models (e.g., GLM trees, the single level equivalent to GLMM trees) do not have an explicit means of handling the clustering. Therefore, clustering in the

population is passed through the residuals to the parameter stability tests and subsequent split points identified by MOB, which introduces bias unless corrected (Zeileis et al., 2008).

In agreement about the clustering in the population, the random effects of GLMM trees are likely necessary in the opinion of both camps. Unlike the process modeled in the fixed effects—i.e., the GLMM—the process underlying the parameter stability test does not meet any of the criteria of Abadie et al. (2017). With random effects removed from the outcome, the samples passed to parameter stability tests are not students in schools but transformed residuals from a regression of graduation on 9G-OTG which has been estimated to make these coefficients orthogonal to the (previously removed) random effects (Fokkema et al., 2018; Fokkema et al., preprint).

The second objection of Abadie et al. (2017)—the presence of unrepresented clusters in the population—deserves discussion here, as it is less clear. By conducting a statistical test (i.e., parameter stability tests), we are using observations to describe an underlying phenomenon (i.e., the stability of a coefficient), though, the phenomenon in question is directly an estimate of a fit function. In this case, should the population be so broad that corrections are always justified? If so, the only stance is that a population of possible parameter stability tests of fit functions exists, and we are only testing a subset of these. However, this is an extremely restrictive definition of the population and sample which is too theoretical for applied work and should instead be left for formal proofs by pure mathematicians.

In Cases of Uncertainty, Offer Both Options

In summary, why then, is the option available and lacking in-depth theoretical explanations of proper data-model-design match like those in Abadie et al. (2017)? I believe it is

because Zeileis and the associated research camp are in alignment with the ideologies of discovery outlined by Cameron and Miller (2015), Gerring (2011), and Tukey (1977).

From a practical standpoint, adding the secondary clustering requires nominal additional work on top of passing other hyperparameters from the *glmertree()* function to the *glmtree()* function. Returning to the context of a divided field, it does not make sense to limit the use of their model. If Zeileis et al. (2008) and Fokkema et al. (2018) are in alignment with the ideology of Cameron and Miller (2015), they stand with those favoring discovery and an iterative approach to knowledge generation. In an ecosystem of modular software and rapidly developing methodology, implementing new algorithms as extensions of prior algorithms requires constant validity checking, which is part of what Paper 2 is designed to do. Each of these independently may lead to a pragmatic pass-through of the option for a secondary correction, let alone when cooccurring.

Prior Systematic Investigations into Clustering in GLMM Trees

Dual versus singular corrections had not been evaluated with simulated data until the preprint by Fokkema and Zeileis (Fokkema et al., 2018; Fokkema et al., 2021). Without clear direction on the use of cluster-robust corrections over-and-above random effects, Jorink (2018) investigated this issue in an unpublished manuscript, selecting the specification which most accurately predicted unseen data. Jorink (2018) specified GLMM trees as recursively partitioned growth curve models (GCM), using Early Childhood Longitudinal Study-Kindergarten (ECLS-K) class of 1998-99 data. Random intercept GCMs were fit with and without random slopes using reading, science, and math scores. Jorink included several time-invariant partitioning variables including gender, race, socioeconomic status, gross motor skills, fine motor skills, interpersonal skills, self-control, internalizing/externalizing problem behavior, baseline age, and

a dichotomous measure of if the student was a first-time kindergartener. Jorink (2018) compared model performance with mean square error on unseen data with various model specifications. Despite all covariates being time-invariant—at the individual / cluster level in this GCM—the preferrable model varied based on specification of the random effects.

Jorink (2018) varied one other aspect of the model, initialization of the algorithm with the random effects versus the tree. For reading and math, no differences were found between the results of the two models. Science showed minimal differences. Further supporting a lack of utility in the double-correction is that Fokkema et al. (2018) did not employ cluster-robust corrections in any tests in their original or second (Fokkema et al., 2021) demonstration of the method. Relying only on (simulated) level 1 splitting variables, the first demonstration would not reasonably correct for clustering in stability tests. However, in Fokkema's second publication of the method, Fokkema et al. (2021) apply this to the mixed-effects growth model and incorporate all level-2 covariates (i.e., time-variant and time-invariant covariates). Despite this, Fokkema et al. (2021) do not adjust for clustering in standard errors or even discuss the topic.

After searching the published literature to no avail, a gray-literature search—i.e., one of unpublished and semi-formal research—opened two investigations into the use of cluster-robust standard error corrections in GLMM trees (Jorink 2018; Fokkema & Zeileis, 2023). Unsurprisingly among a relatively new research method, Jorink's evaluation was supervised by its developer—Fokkema—meaning the research was approved in completion of the master's thesis. Using similarly parameterized growth models, it is my assumption that Fokkema et al., (2021) parameterized their model in a data driven way based on the results of Jorink (2018), however, Fokkema et al. (2021) do not discuss the choice to omit using cluster-robust corrections

to standard errors. Fokkema and Zeileis's in-press article explains the secondary correction and provide a hypothesis for effect of the dual correction (p. 10):

> *When partitioning longitudinal data, covariates will often be measured at the subject level (i.e., time-invariant covariates), which should be accounted for in computing the estimated covariance matrix. […] By summing the scores within clusters prior to computation of the covariances, so-called clustered covariances are obtained, which account for dependence between observations within the same cluster (Zeileis, Koll, & Graham, 2020). This resembles a GEE-type approach with an independence correlation structure. Our expectation is that in partitioning LGCMs, use of clustered covariances in the parameter stability tests will improve subgroup recovery.*

Though Fokkema and Zeileis (2023) directly follow Jorink (2018), they do not cite the paper anywhere. Directly copying the results of Jorink for their hypotheses would not have led Fokkema and Zeileis to the quoted hypothesis, as Jorink's findings support the preprint article without random slopes, but not in models with random slopes.

Fokkema & Zeileis (2023) go on to empirically test differences across condition, finding equivalently perfect (i.e., 100% accurate) performance in selecting the most influential simulated variable across clustered versus non-clustered covariances for all models (Table 1 in Fokkema & Zeileis, 2023). Only one model-data condition tested by Fokkema & Zeileis (2023) performed poorly: models simulated to have random intercepts *and* random slopes, initialized with random effects (without clustered covariances). Peculiarly, this means when random slopes were fitted, the worst performing model in Fokkema & Zeileis (2023) was the best performing model in Jorink (2018). In Fokkema & Zeileis (2023), initializing with random effects and using only singular corrections had a 49.9% chance of being correct, a 48.5% chance of returning a false negative (i.e., reporting no heterogeneity), and less than 1% reported a false positive (i.e., a less-important variable).

Without simulated data, Jorink (2018) could not recover known effects, meaning models were compared in a very different way. Fokkema & Zeileis (2023) recover rank-order of

simulated splitting variables over simulated data, whereas Jorink (2018) evaluates predictive accuracy and model size with cross-validation of observed data. In Jorink (2018), the best parameterization varied by specification of random effects (random intercept vs. random intercept + random slope). More complex random-effects specifications performed notably better when random effects were estimated first, which differs from the results in Fokkema & Zeileis (2023), unless dual corrections are used in conjunction with random effect initialization. The previous George Box quote—rephrased as "all models are wrong, but some are useful"—acts as Occam's Razor. Meaning, despite accurate predictions in Jorink (2018), there is no way to know if the variance is adequately parsed across fixed effects, random effects, and tree-structure without simulation. Jorink (2018) may have recovered a spiderweb of effects which result in good predictions, but poorly model the effects (i.e., underlying process) in the model. *Paper 2* works to clarify best practice in the face of conflicting results, especially considering the novel data, models, and conditions comparing the SLDS to prior investigations (Fokkema & Zeileis, 2023; Jorink, 2018).

Therefore, a shortcoming of Jorink (2018) is the lack of a true known effect; Jorink's conclusions about GLMM trees are drawn based upon real-world data. Without simulating data with known effects, it is unclear if the random effects or rank-order of splits are recovered properly. Jorink's (2018) data-driven solution showed support for only correcting with random effects, providing a singular piece of evidence for the lack of utility of both corrections using only one dataset. Fokkema & Zeileis (2023) demonstrate all models can recover rank-order of simulated effects greater than 99% of the time when data do not have random slopes.

Despite past research, prior investigations are very different than SLDS data structure, hindering generalization of findings. For Jorink (2018) the data is a *much smaller* than the SLDS

(i.e., the class of 1998-99 ECLS-K). Importantly, the ECLS-K data differs in other ways, including clustering structure, number of available variables, and measurement level of the splitting variables. With so many differences from ECLS-K, projects built upon the SLDS suggest extending the work of Jorink (2018) with data at hand. Similarly, considerations of data, model, and context complicate extrapolating results from Fokkema & Zeileis (2023).

Fokkema and Zeileis's (preprint) also focus on widely different models and data compared to that used in *Paper 3*. A few differences include (a) different parametric models (growth models with 4 level-1 units vs. logistic regression with dozens or hundreds of level-1 units), (b) small samples, (c) weaker relationship between main predictor and outcome (growth model slopes of b = -1, 0, 1 for the trajectory over time vs. odds ratio of ~4.9), (c) simulation of overly optimistic conditions, whereby data are tailor-made for recursive partitioning growth models (i.e., three dichotomous variables dictate low, medium, and high intercept for negative, zero, and positive slope), and (d) simulation of splitting variables at a single level.

I expect the model herein to recover the simulated random-effects structure when using only a single correction more accurately than when both corrections are used. Based on my assertion that the secondary corrections are vestigial, I propose that models with singular corrections will still accurately recover the rank-order of splitting variables across clustering units whether nested or cross-classified.

Contextualizing Paper 2 in the Dissertation

Before GLMM trees can inform theory or high-stakes decision making, an understanding of how their misspecification influences recovery of effects must be established in the context of the SLDS and *Paper 3*. Bringing the technical assertion back to an analogy of a vestigial structure, it is unclear if applying both corrections simultaneously:

a. Is required when fitting a GLMM tree, meaning the functionality is not vestigial.
b. If iterative corrections of the data (i.e., once by each method) introduces bias to estimates (discussed below), meaning the functionality is vestigial and use of both corrections actively organism (i.e., model) performance.
c. If the resultant model will be unchanged, meaning the vestigial structure does not harm organism performance.
d. If (a) to (c) vary by algorithm initialization (i.e., estimate tree or random effects first).

I take a practical perspective to testing the preference of using mixed effects alone or also using corrections to standard errors in variable selection by simulating data and recovering known effects across multiple conditions. Using data with known characteristics allows clear demonstration of (a) if GLMM trees can function on an exploratory level across multiple levels and (b) if both mixed effects and cluster robust offsets in parameter stability tests are required to accurately choose the most influential variable.

If (a) happens to be true, GLMM trees must be restricted to a single level of nesting (e.g., students in schools) because the cluster-robust corrections to parameter stability tests—as implemented in GLMM trees—can only handle one level of nesting. If (b) happens to be true, using only one correction is preferable to correcting iteratively. If (c) happens to be true, the user choose will not matter. By testing the influence of initialization approach (d), this project ensures the results of (a) to (c) are invariant across the conditional hyperparameter,

In summary, because the GLMM has accounted for the dependencies of nesting, it is possible that no corrections need to be applied to the assessment of instability across the loglikelihood. If that were true, as Abadie et al (2017) demonstrated, improper adjustment can

bias subsequent interpretations. The practical implications of Paper 3 are therefore straightforward and bifurcate into two possible paths for Paper 3:

A. If simulation shows an ability to accurately recover simulated effects in data similar to the SLDS without dual corrections, GLMM trees can be applied across the available clustering dimensions in the SLDS in one model. This is the preferred goal, as it allows for interactive effects to be identified across level and a comparison of magnitude of effects across levels.
B. If simulation shows that the model more accurately recovers simulated effects in data similar to the SLDS only when corrected along a given clustering dimension, GLMM trees can only be applied to one level of clustering at a time. Although less preferable, this model allows the major factors to be explored in an ad-hoc parametric model which can more flexibly incorporate the random-effects structure.

Research Questions and Hypotheses

- Are GLMM trees differentially effective across single-versus dual corrections?

GLMM trees are expected to perform better when clustering is accounted for only by the random effects based on the information above.

- Is the prior effect conditional upon algorithm initialization?

GLMM trees are expected to perform worse when initialized with random effects, as treating them as an offset prevents their influence from contributing to the optimizing function.

- Are GLMM trees able to recover the most influential variable across multiple levels?

Because the GLMM underlying GLMM trees is capable of handling both nested and cross-classified models, GLMM trees are expected to accurately recover an effect from nested and

cross-classified when specified properly (initialized with the tree and omitting cluster-robust corrections to instability tests).

Method

Simulating Data

Data were simulated to correspond to the multilevel structure observed in the SLDS, with fixed and random effects corresponding to conditions matching the research questions. As with *Paper 1*, number of level 1 units within a level-2 unit were determined by randomly sampling the observed distribution. The same report established the intercept and slope of simulated data (Scalise et al., 2023). Simulated data included 481 units in the nested condition, which is slightly larger than the number of schools which had 12th graders (i.e., high schools, K-12, etc.) in the full student-level SLDS data (j = 431 and j = 429).

Nested Data

Nested Data were simulated as

$$outcome_i \sim Binomial\left(n = 1, prob_1 = \widehat{\mathrm{P}}\right)$$

$$\log\left[\frac{\widehat{\mathrm{P}}}{1-\widehat{\mathrm{P}}}\right] =$$

$$\alpha_{j[i]} + \beta_1 * (main\_predictor) + \beta_2 * (dichotomous\_lv1)$$

$$\alpha_j \sim N\left(\gamma_0^{\alpha} + \beta_3(dichotomous\_lv2), \sigma^2_{\alpha_j} = 1\right),$$

$$\text{for School}_j = [1, 481]$$

$$\gamma_0^{\alpha} = 2.4$$

$$\beta_1 = 1.6$$

The magnitude of effects for β_2 and β_3 depended on the condition being tested. Specifically, three conditions were repeated 500 times to simulate nested data sets:

- Nested Condition 1. More instability in Level-1: $\beta_2 = \beta_{large}$; $\beta_3 = \beta_{small}$
- Nested Condition 2. More instability in Level-2: $\beta_2 = \beta_{small}$; $\beta_3 = \beta_{large}$
- Nested Condition 3. Equal instability across Level-1 and Level-2: $\beta_2 = \beta_3 = \beta_{large}$

In each of the 500 repetitions, their magnitudes were drawn from the following distributions corresponding to the large and small effect. Use of a distribution of effect sizes was done to ensure findings generalized across a range of small versus large effects while ensuring the smaller effect remains smaller and the larger effect remains larger.

$$\beta_{large} \sim N(\mu = 0.8, \sigma^2 = 0.0025)$$

$$\beta_{small} \sim N(\mu = 0.2, \sigma^2 = 0.0025)$$

Where μ and σ^2 are the mean and standard deviation of the distribution of possible effects.

Equal instability, though unlikely in practice, is an important metric in determining the models' bias towards selecting one level versus another across specifications.

Cross-Classified Data

Assessment in cross-classified models requires three variables, one for each of two nesting units and one for the individual, to be tested as splitting variables. Cross classified data were simulated similarly to nested, with the equation and conditions shown below:

$$outcome_i \sim Binomial(n = 1, prob_1 = \hat{\mathrm{P}})$$

$$\log\left[\frac{\hat{\mathrm{P}}}{1-\hat{\mathrm{P}}}\right] =$$

$$\alpha_{j[i],k[i]} + \beta_1 * (main\_predictor) + \beta_2 * (dichotomous\_lv1)$$

$$\alpha_j \sim N\left(\gamma_0^{\alpha} + \beta_3(dichotomous\_lv\_A), \sigma^2_{\alpha_j} = 1\right),$$

$$\alpha_k \sim N\left(\gamma_0^{\alpha} + \beta_4(dichotomous\_lv\_B), \sigma^2_{\alpha_k} = 1\right),$$

$$\text{for Nesting Level A}_j = [1, 481]$$

$$\text{for Nesting Level B}_k = [1, 293]$$

$$\gamma_0^{\alpha} = 2.4$$

$$\beta_1 = 1.6$$

$$\beta_{large} \sim N(\mu = 0.8, \sigma^2 = 0.0025)$$

$$\beta_{small} \sim N(\mu = 0.2, \sigma^2 = 0.0025)$$

Level 1 (i.e., the "student-level" observations nested both within levels A and B) was simulated to be comprised of n = 40000 observations, roughly the number of 12th graders in the student-level SLDS data without any deletion (n = 38,399 and n = 40,275). Magnitude of effects for β_2 and β_3 depended on the condition being tested. As with nested models, cross-classified models were estimated with one large and small effect at each level, as well as an edge-case whereby all models have equally large instability. Specifically, four conditions were repeated 500 times to simulate nested data sets:

- Cross-Classified Condition 1. More instability in Level-1:
    - $\beta_2 = \beta_{large}$
    - $\beta_3 = \beta_4 = \beta_{small}$
- Cross-Classified Condition 2. More instability in Level-A:
    - $\beta_3 = \beta_{large}$
    - $\beta_2 = \beta_4 = \beta_{small}$
- Cross-Classified Condition 3. More instability in Level-A:
    - $\beta_4 = \beta_{large}$
    - $\beta_2 = \beta_3 = \beta_{small}$
- Cross-Classified Condition 4. Equal instability:

- $\beta_2 = \beta_3 = \beta_4 = \beta_{large}$

Table 1

Design conditions to be tested.

Data Structure	Simulation Conditions	Specification Conditions	
		Clustering Specification	Initialization
Nested (2 levels) (e.g., student in school)	1. More instability at level 1 2. More instability at level 2 3. Equal Instability	• Random Effects Only • REs + level 2 correction	• Tree • Random Effects
Cross-Classified (e.g., students in residential districts and school districts)	• More instability along level 1 • More instability along Grouping Variable A • More instability along Grouping Variable B • Equal Instability	• Random Effects Only • REs + level A correction • REs + level B correction	• Tree • Random Effects

RE = Random Effect

Estimated Model

All models were specified equivalently with a logistic regression of a dichotomous outcome on a dichotomous predictor with 8 splitting variables, comprised of a pair of continuous and pair of categorical variables at both level-1 and level-2. Stated formally, the GLMM was

$$outcome_i \sim Binomial\left(n = 1, prob_1 = \widehat{\mathrm{P}}\right)$$

$$\log\left[\frac{\widehat{\mathrm{P}}}{1-\widehat{\mathrm{P}}}\right] = \alpha_{j[i]} + \beta_1 * (\text{main\_predictor\_lv1})$$

$$\alpha_j \sim N\left(\gamma_0^{\alpha}, \sigma^2_{\alpha_j}\right)$$

Best-fit parameter coefficients were estimated with penalized iteratively reweighted least squares (Bates et al., 2015). The model used Powell's (2009) bounded by quadratic approximation (BOBYQA) optimizer.

Quantifying Model Capability

Data were simulated 500 times across all conditions. These 500 simulated data sets were then fit to GLMM trees with varying specifications (Table 1). In each case, the top splitting

variable was extracted. Because data were simulated to have instability, “true negatives” were not assessed but (a) true positives (correct choice), (b) false negative (no splits), (c) false positive (incorrect choice), and (d) convergence errors were calculated as a percentage of total simulations.

Software

R (Version 4.3.1) was used (R Core Team, 2024). The *tidyverse* suite of packages (Wickham et al. 2019) were used for data cleaning and visualization. GLMM trees were estimated with the *glmertree* package (Fokkema et al., 2018).

Results

Nested Instability

All nested conditions were evaluated 500 times in 2.78 hours. This equates to a given condition being fitted 500 times in roughly 13.9 minutes.

Unequal Instability

Greatest Instability along Level 1. When the true level of greatest instability was simulated at level 1, the model correctly identified this variable as the first split 100% of the time (500 of 500 simulations), regardless of specification of parameter stability test clustering or algorithm initialization (Table 2). In other words, the model was robust to any specification in correctly finding the most influential split when it was simulated at level 1.

Table 2

Results from 500 repetition of nested simulations with unequal instability across levels; best model(s) bolded.

Greater Instability Simulated at Level 1				
Initialization	Clustering Specification	Percent Correct	Percent Wrong Choice	Percent False Negative
Tree	**Only Random Effects**	**100%**	**0%**	**0%**
Tree	**REs + level-2**	**100%**	**0%**	**0%**
Random Effects	**Only Random Effects**	**100%**	**0%**	**0%**
Random Effects	**REs + level-2**	**100%**	**0%**	**0%**
Greater Instability Simulated at Level 2				
Initialization	Clustering Specification	Percent Correct	Percent Wrong Choice	Percent False Negative
Tree	**Only Random Effects**	**100%**	**0%**	**0%**
Tree	REs + level-2	91%	9%	0%
Random Effects	Only Random Effects	16.0%	44.2%	39.8%
Random Effects	REs + level-2	69.6%	30.4%	0%

RE = Random Effect

Greatest Instability along Level 2. When the larger effect was simulated to level 2, a more complex pattern emerged, depending on specification. By default, the model ignored clustering during parameter stability tests and did not offset the initial algorithm by the random effects (Fokkema et al., 2018). The random effects would be calculated within the model during the first iteration and clustering would be handled by the GLMM, instead of the parameter stability tests. In this condition, 100% of simulations correctly identified the order of the splitting.

When accounting for clustering in splitting but not initializing with an offset, the level-2 variable was correctly chosen 91% of the time. In this condition, the wrong (i.e., level-1) variable

was selected first 9% of the time. This appears to have led to more frequent selection of an incorrect choice, likely due to the inflation of standard errors from clustering.

The model displayed problematic behavior when initializing with random effects. As shown in Table 2, parameter stability tests within GLMM trees were less able to correctly rank-order the influence of splitting variables. With random effects offsets at initialization, use of clustering in parameter stability tests appeared more important, as 69.6% of cases were correct for those with matching parameter stability tests, but only 16% were correctly identified without matching. Thus, failing to include the influence of nesting into the objective function appears to lead to incorrect selection of splitting variables 30.4% of the time when the parameter stability tests were specified to match the true level of instability. When clustering in stability tests did not match specifications, 44.2% of models chose the incorrect first splitting variable and 39.8% of cases falsely asserted a lack of heterogeneity, totaling 84% improper selection.

Equal Instability

Table 3

Results from 500 repetitions of nested simulations with equal instability across levels; best model(s) bolded.

Initialization	Clustering Specification	Percent Selected Level 1	Percent Selected Level 2	Percent False Negative
Tree	**Only Random Effects**	**64.6%**	**35.4%**	**0%**
Tree	REs + level-2	67.2%	32.8%	0%
Random Effects	Only Random Effects	90.6%	9.4%	0%
Random Effects	REs + level-2	70.4%	29.6%	0%

RE = Random Effect

When level 1 and level 2 were specified to have the same magnitude of effect, GLMM trees preferred selection of level 1 to level 2 regardless of condition. In one case, though, the model only selected level 2 as the top splitting variable 9.4% of the time. This was when the

model was initialized with random effects offsets and parameter stability tests did not use clustered standard errors (Table 3). None of the models in Table 6 reported no heterogeneity. The closest model to picking the variables at their simulated rate (i.e., 50% level-1 and 50% level-2) was the default algorithm (initialized with the tree and correcting for clustering only once).

Cross-Classified Instability

All cross-classified conditions were evaluated 500 times in 15.58 hours. This equates to each of the 24 given conditions completing 500 iterations in roughly 38.95 minutes, on average.

Unequal Instability

Greatest Instability along Level-1. As with the nested design, 100% of simulated cases selected the correct splitting variable as the most influential when the instability was at the observation level (i.e., level 1), regardless of model specification (Table 4).

Greatest Instability along Cluster A (j = 481)

Cluster A always had 481 units. All models without an offset at initialization were correct in 98.2-100% of all simulations. Models with random effects offsets at initialization performed well when parameter stability tests were implemented at either cluster level (92.6% and 94.8% for clusters A and B, respectively). When clustering was not accounted for in parameter stability tests, the model was only correct 53.8% of the time. In that condition, an incorrect variable was chosen 28.2% of the time and no heterogeneity was reported at any level 18.0% of the time.

Table 4

Results from 500 repetitions of cross-classified simulations with unequal instability across levels; best model(s) bolded.

Greatest Instability Simulated at Level 1				
Offset at Initialization	Cluster Specification in Stability Tests	Percent Correct	Percent Wrong Choice	Percent False Negative
REs First	**Only REs**	**100.0%**	**0.0%**	**0.0%**
REs First	**REs + level A correction**	**100.0%**	**0.0%**	**0.0%**
REs First	**REs + level B correction**	**100.0%**	**0.0%**	**0.0%**
Tree First	**Only REs**	**100.0%**	**0.0%**	**0.0%**
Tree First	**REs + level A correction**	**100.0%**	**0.0%**	**0.0%**
Tree First	**REs + level B correction**	**100.0%**	**0.0%**	**0.0%**
Greatest Instability Simulated at Level A (j = 481)				
Offset at Initialization	Cluster Specification in Stability Tests	Percent Correct	Percent Wrong Choice	Percent False Negative
Tree First	**Only REs**	**100.0%**	**0.0%**	**0.0%**
Tree First	**REs + level A correction**	**100.0%**	**0.0%**	**0.0%**
Tree First	REs + level B correction	98.2%	1.8%	0.0%
REs First	REs + level B correction	94.8%	5.2%	0.0%
REs First	REs + level A correction	92.6%	7.4%	0.0%
REs First	Only REs	53.8%	28.2%	18.0%
Greatest Instability Simulated at Level B (k = 293)				
Offset at Initialization	Cluster Specification in Stability Tests	Percent Correct	Percent Wrong Choice	Percent False Negative
Tree First	**Only REs**	**99.4%**	**0.6%**	**0.0%**
Tree First	REs + level A correction	98.8%	1.2%	0.0%
Tree First	REs + level B correction	87.4%	12.6%	0.0%
REs First	REs + level B correction	23.0%	77.0%	0.0%
REs First	REs + level A correction	21.4%	77.4%	1.2%
REs First	Only REs	3.6%	54.8%	41.0%

RE = Random Effect

Greatest Instability along Cluster B (k = 293)

Cluster B always had 293 units. The pattern of effects was very similar for this model, as was seen when clustering occurred along a variable measured at Level A when no initializations

were used. When corrections were applied to either level *other than* level B—the level simulated to exhibit greatest instability—the model performed at 99.4% and 98.8%. When correcting for clustering at level B, the model correctly selected the first split 87.4% of the time and reported an incorrect split the remainder (12.6%).

The three conditions with worst performance are seen when the variable has the greatest instability along cluster B (the cluster with fewer units; Table 4). Initializing the algorithm with random effects as offsets consistently led to the worst models. Using clustering along the same level (i.e., along level B) led to 23.0% chance of correctly identifying the variable with the most instability and a 77% chance of selecting incorrectly. Clustering along level A was worse (21.4% correct, 77.4% wrong choice, 1.2% false negative), but not as bad as implementing no corrections in the parameter stability tests (3.6% correct, 56.9% incorrect variable selected, 40.4% false negative).

Equal Instability

When initializing offsets with random effects, the model displayed the same rank-order of first splitting variable, regardless of clustering specification in parameter stability tests: the level-1 variable, the larger cluster (cluster A), then the smaller cluster (cluster B). When the model was initialized with no offset, the model made much closer to random (i.e., equal) selection across the three levels (Table 5). Although equal instability along two variables in the sample is unlikely, the possibility increases as the number of measured variables increase (which is important for *Paper 3*).

Table 5

Results from 500 repetition of cross-classified simulations with equal instability across levels; best model(s) bolded.

Initialization	Clustering Specification	Percent Selected Level-1	Percent Selected Cluster-A	Percent Selected Cluster-B	Percent False Negative
Random Effects	Random Effects	75.8%	23.8%	0.4%	0%
Random Effects	REs + level A correction	58.3%	34.2%	7.5%	0%
Random Effects	REs + level B correction	60.7%	36.5%	2.8%	0%
Tree	**Random Effects**	**39.4%**	**25.3%**	**35.3%**	**0%**
Tree	REs + level A correction	40.3%	24.4%	35.3%	0%
Tree	REs + level B correction	40.8%	25.6%	33.6%	0%

RE = Random Effect

Discussion

In the introduction, I compared GLMM trees to an organism that was the product of biological evolution which contains vestigial organ which facilitated performance of an ancestor. Specifically, GLM trees—the single-level implementation of GLMM trees—could not offset the effects of clustering without this hyperparameter. GLMM trees main extension to the GLM tree is the use of random effects, allowing for specification of complex clustering structures. Focusing on data structures critical to the analysis in Paper 3, results from this paper support the assertion that a secondary correction for clustering is not necessary.

Optimal Specification

Across both nested and cross-classified designs, initializing the model with the tree and relying on random effects to offset the effect of clustering (i.e., the default setting of these hyperparameters) led to better performance than any other model. With this (default)

specification, identification of the most influential variable occurred at rates of 99.4% (in one case) or 100% (in all others).

Admittedly unlikely with only few variables, it is important to understand how GLMM trees perform if a sample has indistinguishable differences in effect across levels. In models with hundreds or thousands of variables—like *Paper 3*—it is possible that observed cases for multiple variables have extremely similar magnitudes of effect. Initializing with random effects strongly biases models away from selecting variables from clustering units and towards selection of observation level variables. Simulated edge-cases whereby one variable at each level had identically large influence also preferred the default specification for initialization and clustering. Although variables were not selected at perfect rates (50% for nested and 33% for cross-classified), the default condition returned the closest values to these rates compared to other specifications, with some bias towards selecting the variable at the observation level over the clustering level.

Other Specifications

When instability was simulated to occur at the observation level (i.e., level-1), the model captured the effect in all simulated cases and specifications perfectly (i.e., 100%). In all other specifications, differential model performance was not straightforward. At least three additional trends are supported by these results. First, initialization with random effects always degraded the ability of the model to accurately identify the most important variable. In fact, these models displayed a greater affinity for false negatives, with the highest false negative rate of all conditions. More false negatives suggest such a specification—initialized with random effects and only using random effects corrections—parses more variance into the random effects than

other specifications. Further, the degradation in performance from initializing with random effects appears to be buffered slightly by secondary corrections to the GLM tree.

Taken together, this result makes it clear that no offsets should be used at initialization and parameter stability tests do not need to account for clustering when assessing splits. After including their effects in the GLMM, the loglikelihood appears to be unbiased with respect to the influence of clustering, making inclusion of the clustering term in parameter stability tests unnecessary in Paper 3. The optimal specification in Paper 3—with respect to elements tested in Paper 2—can be written in code to solidify the contribution of this paper to Paper 3 (relevant portions bolded):

```
glmertree(
    formula =
      graduation ~ OTG | random_effects | splitting_variables,
    data = SLDS_data,
    cluster = NULL,
    ranefstart = NULL
)
```

Limitations

Perhaps the greatest limitation is a limited ability to generalize these results to other datasets, models, and contexts. Formal proofs and assessment of GLMM trees across all combinations of data and model specification are beyond the scope of this project but should be considered before ignoring the cluster and ranefstart parameters in other contexts. Including so few potential splitting variables is a substantial limitation. Only two (nested) and three (cross-classified) variables were included as splitting variables and their effects were drawn from one medium-to-large and one small-to-medium distribution. Another potential limitation is a failure to track splits beyond the first. Possibly, subsequent splits are more difficult for the model to parse or that bias may be more present with different magnitudes of effects. However, the presented results are roughly equivalent to those of Fokkema & Zeileis, which reports

identification of variables at second, third, and further splits as well (preprint; Table 1 middle section). Use of only one sample size is not problematic to *Paper 3*, but it does limit the generalizability of these findings to other contexts. As elaborated in *Paper 1*, sample size, design, skew of outcome, magnitude of association with primary covariate, and binomial distribution differentiate *Paper 3* from much of social science research.

Next Directions

The results of *Paper 2* provide evidence for a given specification of GLMM trees in *Paper 3*, however, the novelty of GLMM trees should be considered. As Gerring (2011, p. 32) quoted Paul Feyerabend,

> *The idea of a method that contains firm, unchanging, and absolutely binding principles for conducting the business of science meets considerable difficulty when confronted with the results of historical research. […] [T]here is not a single rule, however plausible, and however firmly grounded in epistemology, that is not violated at some time or other. It becomes evident that such violations are not accidental events, they are not results of insufficient knowledge or of inattention which might have been avoided. On the contrary, we see that they are necessary for progress. Indeed, one of the most striking features of recent discussions in the history and philosophy of science is the realization that events and developments, such as the invention of atomism in antiquity, the Copernican Revolution, the rise of modern atomism (kinetic theory; dispersion theory; sterochemistry; quantum theory), the gradual emergence of the wave theory of light, occurred only because some thinkers either decided not to be bound by certain "obvious" methodological rules, or because they unwittingly broke them.*

Hence, GLMM trees as a method may one day become extremely popular, fall completely out-of-use, or be developed into an unrecognizable variant of the model. As new methods are discovered and extended, they should be appraised repeatedly before generalizing their findings.

GLMM trees contain additional hyperparameters which were not varied in this or *Paper 1*, with one in particular standing out as likely relevant to the findings of *Paper 2*. Unlike LMMs, GLMMs must approximate the likelihood of a mixed-effects model over the random effects

space instead of being able to get the value from integration (Bates et al., 2015). The *lme4* package contains three options for this approximation, the Laplace approximation (default), iterative reweighted least squares (the fastest), and the adaptive Gauss-Hermite approximation of the log likelihood (slowest; Bates et al., 2015). In the *lme4* documentation, Bates et al. (2015) explain,

> *The most reliable approximation for GLMMs is adaptive Gauss-Hermite quadrature [(AGQ)], at present implemented only for models with a single scalar random effect. The [number of nodes in AGQ] argument controls the number of nodes in the quadrature formula. A model with a single, scalar random-effects term could reasonably use up to 25 quadrature points per scalar integral.*

Relying on both nested and cross-classified designs, this model could not use the AGQ, and used iterative reweighted least squares to estimate random effects in both models. Future research should determine if changing these terms affects GLMM trees. Because these terms influence likelihood of GLMM, and GLMM trees assess heterogeneity in GLMM trees, this hyperparameter may reasonably change ability to identify most influential splitting variable, recover a simulated threshold of instability, or otherwise improve performance.

Contribution to Paper 3

These findings suggest GLMM trees are able to accurately identify the variable with greatest instability across levels. Furthermore, the model is more likely to accurately recover the rank-order of instability when there is no offset at initialization. When random effects are part of the calculation of loglikelihood, instability in the loglikelihood appears to be unbiased to the nesting structure; thus, clustering will be handed by the GLMM portion of the GLMM tree and the parameter stability tests do not need to employ corrections for clustering.

Uncovering Heterogeneous Predictivity in Graduation Early Warning Indicator among Two Cohorts of the Oregon State Longitudinal Data System

Decades of research have demonstrated associations between dropping out of high school and negative outcomes spanning domains of chronic health conditions, life expectancy and income (Freudenberg & Ruglis, 2007; Tamborini et al., 2015). Freudenberg and Ruglis (2007) call education an "elixir that [can] increase life expectancy, reduce the burden of illness, delay the consequences of aging, decrease risky health behavior, and shrink disparities in health" in their review of the impact of high school graduation on individual and public health.

As one example, Krueger et al. (2015) examined hazard ratios for mortality in the United States across the lifetime and reported that those who did not graduate high school were at a 23% greater risk of dying in the survey window for men and 32% for women, compared to those who graduated high school. When high school dropouts were compared to college graduates, the disparity was larger, as college-graduating men had a 48% lower hazard for dying than those who did not graduate high school. For women, the relative hazard was 54% (Krueger et al., 2015). Moreover, in a nationally representative US sample, those who dropped out of high school were shown to have significantly higher odds of endorsing 4 of the 7 chronic diseases gathered–including heart disease and stroke—compared to those who graduated high school (Covariate-Adjusted Odds Ratio [OR] range: 1.18-1.55; Vaughn et al. 2014).

With respect to lifetime earnings, Tamborini et al. (2015) reported that those who graduated high school earned approximately 357,000 - 564,000 more dollars than those who did not graduate in the 50 years between age 20 and age 69. Larger gaps in income were seen between those who dropped out of high school and those who graduated college. The

aforementioned findings on educational attainment and life outcomes suggests that high school graduation is key factor underlying longevity and quality of life in the United States.

For these reasons and others, the US Federal Government has tracked educational statistics since 1870 (National Center for Educational Statistics [NCES], n.d.-a). What is tracked, how often data are collected, and who oversees reporting have changed over time however, with NCES—housed within the Institute of Education Sciences (IES)—providing oversight in recent history (NCES, n.d.-a; Conaway et al., 2015; Balfanz & Byrnes, 2019).

The Technical Assistance Act & State Longitudinal Data Systems

Prior to the mid-2000s, graduation monitoring was primarily focused on mandatory reporting (Conaway et al., 2015). Specifically, aggregated statistics funneled from districts to state agencies, which were then compiled into federal repositories. Aggregated statistics provided a broad picture of the national education system with minimal payoff for local agencies. Three state practitioners write *[n]ot much went back to districts in a format that could help them make educational decisions about their students,"* and even when it did "*[r]eports were slow to come, [t]heir content reflected the needs of the government agencies that requested them rather than of the districts […], [t]hey did not include individual student-level data [or] real-time access […] in an educator's classroom that day"* (Conaway et al., 2015, p. 17S).

Without timely access to the data or findings from the results, the primary reason to submit data to the state was to comply with mandatory reporting guidelines (Conaway et al., 2015). In 2002, the federal government intervened to improve this process for stakeholders on the ground. The Technical Assistance Act was enacted with an underlying principle of "*Better decisions require better information*" (NCES, n.d.-b). The Technical Assistance Act allowed the IES to grant up to twenty million dollars over a three-to-five-year period for the development of

a state longitudinal data system (SLDS), with the first round of funding distributed to states in 2005-2006 academic year (NCES n.d.-b). According to a Congressional Research Service report, funds were allocated to SLDS infrastructure and development since 2003 (Kuenzi & Zota, 2023). Last year, in 2023, New Mexico became the final US State to receive funding for their SLDS (NCES n.d.-b; NCES, 2023). By 2024, roughly $934 million dollars of federal funds were used to create, support, and improve SLDSs over more than two decades (IES, 2024; Kuenzi & Zota, 2023).

With systems now pervasive across all 50 states (and 6 territories), continuous allocation of funding over the last 20 years, and a cumulative budget of nearly a billion dollars, the US Federal Government has made its perception of SLDS value clear (NCES, 2024). By pairing collection of mandatory reporting variables with locally adapted needs, state educational agencies (SEAs) can support the needs of local and state agencies while complying with federal regulations, a notable improvement over the prior unidirectional system (Conaway et al., 2015).

Benefits to States & State Educational Agencies

At the start of SLDS funding, most SEAs were focused primarily on building the (a) internal capacity (i.e., hiring qualified staff), (b) computational infrastructure, and (c) instituting unique identifiers to collate data across various levels (e.g., student-, school-, district-, and state-levels; Conaway et al., 2015; NCES, n.d.-c; Balfanz & Byrnes, 2019). Centralizing and systematizing data collection into SEA-controlled data improved the timeliness of feedback from SEAs to district personnel, families, and the general public—a primary benefit from the perspective of SEA agents (Conaway et al., 2015).

Maintaining ownership of the SLDS had another large benefit: it helped change the research agenda for state agencies from "*disconnected and scattershot*" to being "*tied closely to*

the strategic policy priorities of the organization" (Conaway et al., 2015). SLDSs have enabled SEAs to take control of their research agenda, meaning research with academic institutions can be more collaborative than before. Before maintaining their own SLDSs, collaborations with external researchers "*tended to exist on the sidelines*" of state agendas and were "*nice to take part in, but separate from the central work of program development and improvement*" (Conaway et al., 2015; p. 19S). States now have the resources to ensure local priorities were incorporated into the monitoring and research process, and that these local priorities could be carried across research studies (Conaway et al., 2015).

Before SLDSs, even "*successful and productive*" collaborations between academics and SEAs created "*challenges for the state agencies*" (Conaway et al., 2015, p. 21S). Conaway et al. (2015, p. 21S) write:

> *One issue is finding common ground between the personal research agendas of external researchers and a state's research priorities. Academic institutions privilege researchers who have built a continuous stream of research in a specific topic and, particularly among quantitative researchers, have used causal analysis techniques to answer their research questions. But states' research needs tend to shift quickly and often can be satisfied with descriptive analysis. Given the priorities of the academic enterprise, it is difficult to find researchers who are willing to study what the state needs, versus what the researcher wants to study.*

Able to prioritize and adapt research based on local needs, most SEAs prioritized real-time data reporting in whichever format best clarifies critical district issues. Conaway et al. (2015, p. 18S) explain how such reports were

> *simply not possible without an SLDS, and the information they yield is crucial for districts to plan effectively for meeting their students' needs and for the public to have easy access to key school and district performance data. […] Facilitating research is certainly an additional intended purpose of this investment in SLDSs; indeed, it is listed as a requirement in every associated federal grant program. But from a state perspective, research is a by-product of the SLDS, not its raison d'être.*

Early Warning Systems

As SLDSs were built and extended, state and federal graduation rates were low. The research interest of SEAs, districts, and schools converged on a goal of gathering better information about who was likely fail or dropout (Balfanz and Byrnes, 2019; Conaway et al., 2015). Massachusetts was among the first states to develop an early warning system (EWS) to monitor graduation and failure (Conaway et al., 2015). Massachusetts EWS was extensive, and their SEAs used a variety of indicators to predict proficiency on statewide tests in grade 3, grade 6, and grade 9, as well as directly predicting high school graduation and dropout (Conaway et al., 2015). Immediately, reports from Massachusetts' EWS were "*highly valued by districts. The [EWS] reports [were] among the most heavily used in Massachusetts's SLDS reporting tool with over 10,000 views between February and June of 2013*" (Conaway et al., 2015, p.17S). An EWS like Massachusetts', however, was not good a candidate for wide dissemination. Massachusetts' EWS relied on more extensive data collection than was common in other states at the time, such as pre-high school measures, which were not widely collected (Balfanz & Byrnes, 2019). Over time, a literature of dropout and graduation expanded to explore a litany of early warning indicators (EWIs) with hope of finding a disseminatable EWS (Balfanz, Herzog, & MacIver, 2007; Allensworth & Easton, 2007; Balfanz & Byrnes, 2019; Bowers et al., 2012).

Optimizing EWSs: The Value of Parsimonious, Practice-Based EWIs

In the research pursuing an EWI, as well as earlier dropout literature, researchers have correlated seemingly innumerable factors with changes in national, statewide, and local dropout rates (Bowers et al., 2012; Allensworth & Easton, 2007). In a systematic review, Bowers et al. (2012) identified 36 studies with 110 total dropout indicators which ranged widely across dimensions of statistical complexity and data availability. Counts and transformations of

attendance, course failures, and other (extant) face-valid metrics were a large subset of early warning indicators. Some required novel data to be gathered, others required complex abstractions which limit the ability for these metrics to be calculated or interpreted without specialists (e.g., growth mixture modeling, hierarchical clustering analysis, etc.), and a few required both (Bowers et al., 2012).

Some of the more intuitive, but not readily available metrics used as EWIs include: student satisfaction with school (Ekstrom et al., 1986), perception of friends' interest in education (Ekstrom et al., 1986), perceptions of parents' value on completing high school (Gilbert & Devereaux, 2006), and working over 20 hours per week (Lee & Staff, 2007). Barriers to gather such variables exist at multiple levels, though. Conaway et al. (2015) explain that *"it is complicated and expensive to add data elements to district data collections, and states often experience significant pushback from districts on new data collection requirements"* (Conaway et al., 2015, p. 24S). Conaway et al. (2015) continue by saying,

> *we cannot understate the importance of the privacy and data security concerns about the use of individual-level student data in the conduct of this work. These are administrative data collected by the state for federal and state reporting and accountability purposes, not for research. There are deep tensions in the public right now over what we are collecting, how we are collecting it, and how it is used. Violations of this trust—perceived or real—will be detrimental to our ability to continue to collect, report, and analyze these data.*

Rapidly, SEAs extended the dropout literature to include the burgeoning field of EWIs, and sought metrics which were face-valid, extant, and easily conceptualized to build their EWSs (Balfanz & Byrnes, 2019). In large part, EWSs drew on cross-sectional and longitudinal dropout research from Chicago and Philadelphia schools to choose an EWI (Allensworth & Easton, 2007; Balfanz et al., 2007). Both cities valued face-valid and available metrics—attendance and course failures—which were used to predict graduation with great accuracy (Allensworth & Easton,

2007; Balfanz et al., 2007). Retrospectively commenting on choosing an EWI of dropout, Balfanz and Byrnes (2019, p. 45) write

> *Although it may seem like common sense that a student who is not attending school, failing courses, or whose behavioral issues are leading to disciplinary actions has lower odds of graduating, the leveraging of readily available data on the part of educational organizations to systematically identify at-risk students based on these variables and flag them for preemptive intervention was an innovation.*

Why More is Less: Theory

In Barry Schwartz's book, *The Paradox of Choice: Why More is Less*, Schwartz outlines a psychological phenomenon as follows "*Learning to choose is hard. Learning to choose well is harder. And learning to choose well in a world of unlimited possibilities is harder still, perhaps too hard"* (2004; p. 144). Understanding the choice paradox aids in contextualizing the original argument of Allensworth & Easton (2005; 2007), the seminal works on 9th grade on-track to graduation (9G-OTG). Without using the language, Allensworth & Easton (2007) begin with an understanding of the Paradox of Choice and used it to enable practice-driven research, demonstrating a deep understanding of the field of implementation science which is still calling researchers to use externally valid and practitioner-benefiting metrics (e.g., Brownson et al., 2022).

Though Allensworth and Easton (2007) only touch on practice-focused research, its messaging is scattered throughout these and other important papers on 9G-OTG (Allensworth, 2013; Allensworth & Easton, 2005). Others have focused on practice-driven research explicitly in field-specific (e.g., Green, 2008) or Implementation Science journals (e.g., Brownson et al., 2022). In a paper foundational to the field of Implementation Science, Green (2008) explores the research-to-practice to identify how and why the gap between theory and practice is so extensive.

The circumstances described by Allensworth & Easton (2007) are a case-study in the second of Green's two (2008) identified problems with the research-practice pipeline, the "fallacy of the empty vessel" (with the other being the "fallacy of the pipeline"). The "fallacy of the empty vessel" refers to a false belief implicit in the behavior of many researchers, wherein the researchers (a) overemphasize their own understanding of a situation, (b) underemphasize the knowledge of practitioners, and (c) fail to focus on the practical restraints observed on-the-ground by practitioners (Green, 2008). Green (2008, p. i23) explains:

> *This final phase of the transfer of original research into practice is […] a 'drop in the bucket' of what might have been of interest to the individual practitioner or of varying interest to different practitioners in different settings. The expectations that accompany its delivery often imply that the practitioner is an empty vessel into which the information can be poured and once full will spill over into action. […] [T]he recipient is far from empty with respect to the demands on the practitioner's time and resources […] and the credence he or she places on the evidence with its inevitable misfit with many of these contextual considerations.*

In summary, Allensworth & Easton (2007) understood that the Paradox of Choice manifested in an oversaturated field which was disconnected from practice. And, as a solution, the authors chose to investigate a face-valid, tangible, and even extant metric by which schools could monitor students. In this way, Allensworth & Easton (2007) positioned 9G-OTG as a practice-based early warning indicator which is easily deployed and maintained in a school, rather than the less practical research-based indicators.

The paper which—by Allensworth & Easton's (2007, p.2) admission—officially "*described and defined the 'freshman on-track indicator'*" was Allensworth & Easton (2005), though the authors explain that the metric had been used internally since the 1990s (Allensworth & Easton 2005). In Allensworth & Easton (2005), the authors outline how the metric was developed in the 1990s with a clear description of practice-based research one to two decades before Green (2008) and the growth of "Implementation Science" evident on Google Engram

(i.e., more than an 87,000% growth in the 20 years from 1999 to 2019; Google Corporation, 2024). A short quote describing the process of narrowing a four-component measure of on-track status to a simpler component elucidates the practice-based perspective:

> *Even though any one of the variables above would have worked relatively well individually as an indicator of freshman performance that would predict graduation, we combined these two variables because we believed that each contained important information relevant to Chicago Public Schools (CPS) policy about grade promotion* (Allensworth & Easton, 2005, p. 2).

Green's (2008) second fallacy—the "fallacy of the pipeline"—is evident in some of the highly effective, yet highly complex EWIs previously outlined from Bowers et al. (2012). Green (2008, p. i10) writes "*research itself is rendered increasingly irrelevant to the circumstances of practice*" as researchers implicitly emphasize other forms of validity over external validity. Green (2008) continues to explain how such behaviors are implicit when researchers unidirectionally focus on a pipeline *from* researchers *to* practitioners instead of a back-and-forth.

Allensworth and Easton's (2007) 68-page research report begins by acknowledging the complexity of a deep field—at the time mostly dropout research—before selecting a brilliantly parsimonious solution that enables schools to intervene and support students earlier. On the one-page introduction, Allensworth & Easton (2007, p. 1) write,

> *Research on dropping out has shown that the decision to persist in or leave school is affected by multiple contextual factors— family, school, neighborhood, peers-interacting in a cumulative way over the life course of a student. This suggests a daunting task for dealing with the problem of dropout—if so many factors are involved in the decision to drop out of school, including experiences outside of school and in early grades, how can any high school effort substantially address the problem?*

Allensworth & Easton (2007) advocate for on-the-ground stakeholders by pointing out the impracticality of a field built upon disparate theoretical mechanisms of action, without actionable metrics a school could reasonably monitor. In summary, they imply that dogged pursuit of a mechanism of action can push theory beyond any means of practical implementation,

degrading utility to practice (Allensworth & Easton, 2007). In response, Allensworth and Easton (2007) argue "*[w]hat is often lost in discussions about dropping out is the one factor that is most directly related to graduation—students' performance in their courses" (p. 2)*. Rather than failing to act due in the presence of excess choices, Allensworth and Easton provide a straightforward way to look at student performance in 9th grade, specifically tracking if students (a) complete 10 semester credits and (b) do not receive more than one F for a semester credit in a core subject (English, math, science, and social science). Allensworth and Easton (2007, p. 2) explain how, at the time, conversations of dropout failed to address a proximal, extant, and actionable factor for dropout—in-course performance.

> *In Chicago, we have shown that inadequate credit accumulation in the freshman year, which usually results from course failures, is highly predictive of failing to graduate four years later. Research in New York City has shown very similar connections between inadequate credit accumulation and eventual dropping out, and national data confirms this; almost all students who drop out leave school far behind in course credits.*

Therefore, 9G-OTG was proposed because it is statistically valid, easily interpretable, and calculatable from available data, all of the requirements of SEA's outlined by Conaway et al. (2015). Allensworth's research group demonstrated its utility empirically multiple times privately since the 1990s (as discussed in Allensworth & Easton, 2005) and several times in the public light (e.g., 2005, 2007; Allensworth, 2013). Hence, 9G-OTG stands out as a face-valid and easily interpretable EWI, assessable with available data. Motivated by eliminating the Paradox of Choice and maximizing stakeholder utility, Allensworth and Easton positioned 9G-OTG as a practice-driven field of research.

Why More is Less: Evidence

Returning to the systematic review by Bowers et al. (2012), examination of the presented table of precision, sensitivity, specificity, and false-positive proportion by EWI makes clear the

validity and value of predicting on-track status with extant and face-valid metrics. Face-valid metrics often had very high specificity (i.e., true-negative proportion), and—in many cases—the face-valid metrics (e.g., failing English, failing math, attending less than 80% of the time) had higher specificity than complex metrics (Bowers & Sprott, 2012). Two of three EWIs with the greatest sensitivity (i.e., proportion of true positives) were among the most complex to calculate metrics (i.e., growth mixture models of non-cumulative semester grade point average, and hierarchical clustering analysis of non-cumulative course grades), but both metrics performed worse with true negatives and false negatives, compared to face-valid metrics. Bowers et al. (2012) is a vast trove of knowledge, which in its totality, is beyond the scope of the current work. Here is how Bowers et al. (2012, p. 97) summarize their own work:

> *Other than the growth mixture models, […] some of the most accurate dropout indicators […] focus on low or failing grades. While this is important given that grades are collected regularly in schools for all students and provide an accessible data point with high face validity for teachers and administrators […], grades have historically been viewed as a subjective and "hodgepodge" assessment of student ability and academic knowledge, including academic achievement as well as class participation and behavior […] We posit that low or failing grades may constitute teacher assessment of a student's ability at both the academic components of their courses and social and behavioral components, as represented by their low and failing grades, indicators highly predictive of whether a student will persist in the system. We encourage more work in this area, since our results here suggest that low and failing grades, especially when coupled with a low number of credits in high school, are some of the most accurate indicators of students at risk of dropping out.*

Hence, Bowers et al.'s (2012) systematic review of EWIs of dropout highlights (a) low / failing grades, (b) low number of high school credits, or (c) the combination of the two as some of the most accurate indicators of students risking dropout. Such recommendations are the constituents of 9G-OTG, as defined by Allensworth & Easton (2007; i.e., accumulating 10 semester credits by 9th grade with at most 1 semester F in a core subject).

9G-OTG as an Early Warning Indicator in Oregon

For the reasons outlined above—the statistical evidence outlined in Bowers et al. (2012), its availability, and its face validity—9G-OTG stands out as a metric that is useful to all members of the research-practice pipeline. Thus, its unsurprising that the Oregon Department of Education (ODE) decided to use a variation of 9G-OTG as an EWI (ODE, 2018). As defined by ODE, 9G-OTG is a dichotomous indicator which indicates completion of at least 25% of coursework requirements by the end of 9th grade (ODE, 2018).

The State of Oregon (and ODE in particular) have performed exemplary work to engage in and support practice-focused research, adapt to local needs, and implement data driven solutions to support their students. Historically, Oregon has had graduation rates well below the national average (Greene, 2001; Heckman & LaFontaine, 2010; Atwell, Balfanz, Bridgeland, & Ingram, 2019). While still true, the difference has shrunk over time (NCES, 2022). By 2018-19, Oregon students graduated only 6% lower than the national average (80% vs. 86%; NCES, 2022). The gap has been a diminished thanks to diligent work from schools / districts (Miller, 2023; Lafollette, 2024; Bourgeois, 2024), ODE (NCES, 2007-a; ODE, n.d.-d), and even the Oregon State Legislature (Marsh et al., 2022; Oregon Senate Committee on Education, 2021; ODE, n.d.-a).

Tracking & Improving High School Graduation in Oregon

ODE began systematically tracking student data over time in 2001-02. In the 2001-02 academic year, a unique secure student identifier was assigned to every student in the state enabling granular and longitudinal tracking of educational metrics (NCES, 2007-a). At first, data was mostly compiled for compliance, and research-based decision making below a federal-level

was minimal compared to more recent years (NCES, 2007-a), paralleling national trends (Conaway et al., 2015).

In 2007, Oregon applied for federal funding to create its SLDS which streamlined and expanded data access for decision-makers (NCES, 2007-a). According to the Educational Resource Information Center (ERIC), Oregon has received three lines of federal funds for their SLDS between 2007 and 2009, totaling over $18.8 million (NCES, n.d.-b). Following in the footsteps of other states (Conaway et al., 2015), Oregon tailored the system to their own needs, acknowledged a need to improve graduation, and added an EWI to their SLDS. Specifically, the ODE began tracking a variation on Allenworth's 9G-OTG metric in the 2013-14 academic year (ODE, 2018). Since 2013-14, the State of Oregon and the US Federal agencies have demonstrated a dedication to Oregon students by increasing tracking, research, and support of on-time graduation in many ways (ODE, 2018-a; 2018-b; ODE, n.d.-c; Farley et al., 2021).

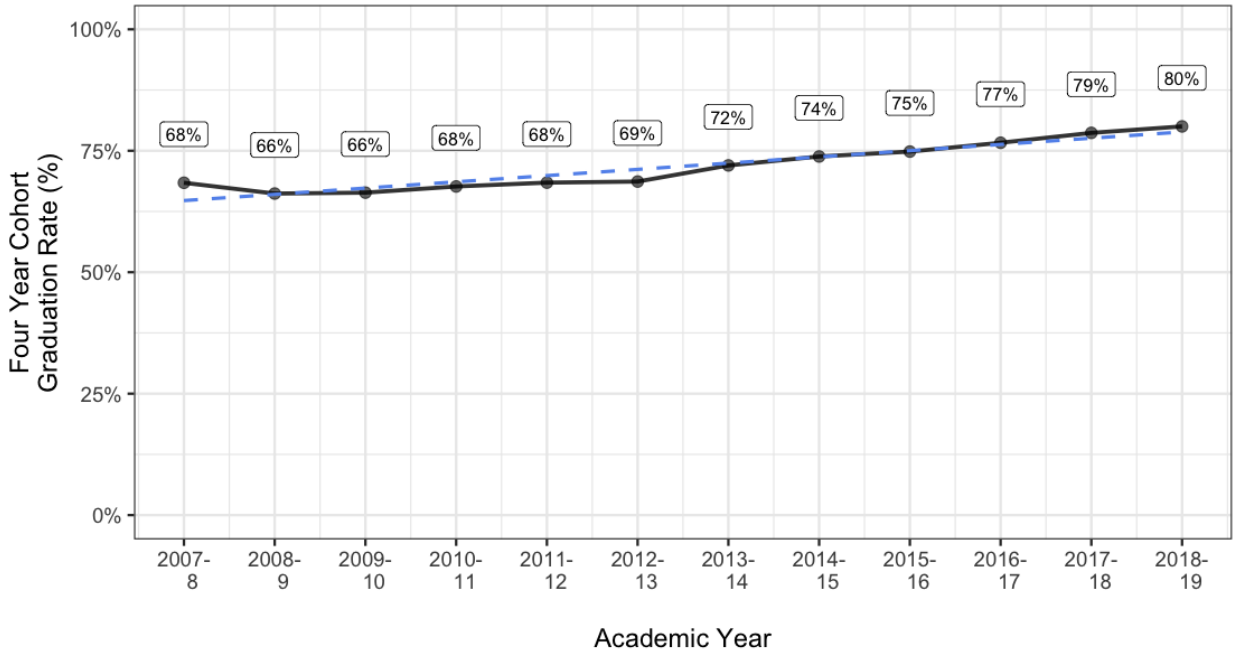
Passing ballot Measure 98 showed a continued desire to increase high school success by increasing school-level spending on a per-student basis (ODE, n.d.-a). To receive funding, schools reported the ways in which they planned to implement high school success funds. Such efforts included establishing teacher collaboration time around data, reducing chronic absenteeism, ensuring equitable assignment to advanced courses, and implementing systems to ensure on-time graduation (ODE, n.d.-a). Evaluation of the program with comparative interrupted time series provides evidence that the implementation of the program diminished school-level gaps in 9G-OTG rates (Farley et al., 2021).

Thanks to SLDS infrastructure, ODE publicly hosts graduation data in the years since funding began, allowing a comparison to the national trend (ODE, 2024). According to Balfanz and Byrnes (2019), the national average graduation rate has increased by roughly 7% since 2010-

2011. By contrast, Oregon increased graduation rates by approximately 12%, a rate 1.5 times greater than the national average (Figure 1).

Figure 1

Four Year Cohort Graduation Rates for Oregon overlaid with linear best fit (dotted line; data from ODE, 2024-b).



The aforementioned efforts—paired with essential yet unquantifiable contributions by students, teachers, and administrators—undoubtably contributed to such rapid improvements. And, by joining the SLDS with public data, further insights can be gleaned from past cohorts. Such insights are essential for Oregon in particular because on-time graduation still lags behind the national average despite rapid improvements (NCES, 2022).

Despite Growth, Room for Improvement

In Oregon, White students—a student group with nationally high graduation rates and the least history of systematic marginalization in the US—had graduation rates lower than the national average at 81% in 2018-19 (versus 89% nationally; NCES, 2022). Only New Mexico had lower graduation rates among White students at 79% (NCES, 2022). In 2018-19, Black

students in Oregon graduated at a rate of 70%, compared to a national average of 80% (NCES, 2022). For reference, 70% is lower than the racially disaggregated national average in *every year since NCES began publishing national graduation rates* in the 1969 Digest of Educational Statistics (NCES, 2007-b). The disparity between graduation rates among Black students in Oregon and nationally is substantial. In fact, despite Oregonian White students graduating at ~8% lower than average national average, the magnitude of disparity between graduation rates of Oregonian White and Black students is still in the upper half of all states (NCES, 2022). Native Americans / Alaskan Natives in Oregon had the lowest graduation rates in 2018-19 at 74%, matching the national graduation rate among this demographic group (NCES, 2022).

Along with several negative aspects of heterogeneity, some subgroup effects in Oregon are clearly an improvement over the national average. Specifically, only five states—Hawaii, West Virginia, Alabama, Nevada, and Florida—have lower disparities in graduation rates between White and Hispanic students (Oregon: 5% vs. national: 8%). These findings demonstrate the need for further investigation of on-time graduation correlates for all Oregon students, but especially for students that are Black or belonging to a historically marginalized group.

9G-OTG & Graduation in Oregon: Associations and Differential Effects

One of the first evaluations of 9G-OTG reported students who were not on-track dropped out at a rate more than 16 times higher than those who were on track (ODE, 2018-a). A follow-up analysis of dropout in 2015-16 cohort reported those off-track in 9th grade to be at more than 10 times higher risk of dropout by the end of their junior year, compared to those on-track (ODE, 2017). In 2016-17, evaluation of this metric's ability to predict on-time graduation was feasible.

Roughly 91% of students who were 9G-OTG graduated within four years, by comparison this number was approximately 40% for those who were not 9G-OTG (ODE, 2019).

A Poisson regression estimating "risk" of graduating on time explored main effects and interactions of several covariates with on-track status among Oregonian students (ODE, n.d.-c). All included main effects other than 9G-OTG decreased the risk ratio of a student graduating. Experiencing houselessness and having transitory enrollment or "high mobility" within a school year were the largest main effects after 9G-OTG (ODE, n.d.-c). Other variables which decreased the risk ratio of graduating were students with a disability flag, a poverty flag, and (one or more) disciplinary incidents (ODE, n.d.-c). Some protective factors were identified which could increase risk ratio of graduating. Regular attendance in 9th and 10th grade were significantly more impactful for those off-track than on-track, being associated with greater chances of graduation after 9th grade. Similarly, enrollment in career and technical education (CTE) programs was associated with a boost in risk ratio of graduating for those on-track but increased the odds significantly more for those off-track in 9th grade. Math course enrollment in 9th grade was also highly predictive of differential impact of 9G-OTG on the risk ratio of graduating. Enrollment in some pre-requisites to Algebra 1 or no math decreased risk ratio among all students, but particularly those off-track (ODE, n.d.-c). Few demographic factors were included in this Poisson regression, but 9G-OTG was significantly less predictive for students who identify as male, compared to female-identifying students (ODE, n.d.-c).

ODE has explored subgroup effects for on-time to graduation, mostly focusing on student-level factors (e.g., attendance, demographics, disciplinary incidents). The ability of 9G-OTG to predict on-time graduation is not constant across subgroups. For example, regular attenders—defined as attending 90% or more of possible days— who are 9G-OTG have even

larger differences in dropout rate (ODE, n.d.-c). Those who were neither on-track nor regularly attending classes in 9th grade represented 60% of dropouts in 10th grade, despite being less than 9% of the cohort (corresponding to a dropout risk ratio of 36.5, compared to on-track regular attenders; ODE, n.d.-c).

Just as with dropout, differences in regular attendance status and demographics were observed in graduation outcomes (ODE, 2018-b). Native American students showed the lowest 4-year graduation rates for both on-track status and off-track status, with those on-track graduating at 8% lower rate than average (i.e., 83% vs. 91%) and those off-track graduating at a 9% lower rate (i.e., 31% vs. 40%). Asian and Hispanic students who were off-track graduated at higher rates (59.7% and 43.4%, respectively) than the disaggregated average for off-track students (40.1%; ODE, 2018-a). Other racial/ethnic groups differed from the disaggregated average, but to a lesser extent. Male students also displayed lower graduation rates whether on- or off-track, compared with female students. English learners, economically disadvantaged students, and students with disabilities also display systematic differences in graduation by on-track status (ODE, 2018-b).

Among Oregon students, the risk ratio of graduating decreases with a count of disciplinary incidents per year (ODE, n.d.-c). ODE (n.d.-c) reported disciplinary incidents are differentially predictive by 9G-OTG status, with variable having disproportionate influence on off-track students, compared to on track students. Despite being highly predictive, disciplinary incidents were infrequent in the data (only 9% of 9th graders had one disciplinary incident and only 4% had more than 1; ODE, n.d.-c).

Heterogenous graduation rates have been documented in Oregon across town size or "rurality," with students in small towns graduating at 5% higher rates than disaggregated

Oregonian students, on average (Clinton & Reeder, 2017). In some small towns, pronounced differences appear among some subgroups, compared to all Oregonian students (Clinton & Reeder, 2017). Native American / Alaskan Native students, for example, graduated at nearly 20% higher when in small towns, compared to other students (Clinton & Reeder, 2017). By contrast, Asian students graduated at nearly 11% lower rates in small towns, compared to all students (Clinton & Reeder, 2017).

In the studies of 9G-OTG above, ODE uses an appraisal or confirmatory data analysis (CDA) lens to evaluate 9G-OTG and associated factors and differential effects in Oregon. Notably, ODE has leveraged at least one publicly available application of EDA involving graduation and 9G-OTG (ODE, n.d.-c). They apply a classification tree to the entire sample, as well as across subsamples as a means of discovering heterogeneity in graduation rates among demographic groups and probe exploratory data analysis (EDA) results with a CDA method (ODE, n.d.-c).

Prior Applications of Discovery at ODE

A binary decision tree classified on-time graduation based on mobility, attendance, current math course, 9G-OTG, and demographics (ODE, n.d.-c). The authors report 9G-OTG status being the most predictive factor, with 89% of those on-track graduating and only 41% of those off-track graduating. Among those off-track in 9th grade, the next most influential factor was 10th grade attendance, followed by CTE participation. At only 19% graduating, those off-track in 9th grade who had less than 81% attendance in 10th grade had the lowest graduating rate in the sample (ODE, n.d.-c). In fact, aforementioned Poisson regression was conducted to follow-up EDA findings from the classification tree.

The results of ODE (n.d.-c) support the validity of 9G-OTG as Oregon's EWI. However, methodological limitations prevent much more from being gained from this paper. Of large concern, is the use of a classification tree without cross-validation, hyperparameter tuning, pruning, or any other procedure to increase the generalizability of these results and decrease the likelihood of overfitting the training data (Yang & Shami, 2022; Strobl, 2009; Hothorn et al., 2006). With extensive theoretical foundation for 9G-OTG, its documented validity in Oregon specifically, and the magnitude of differences between the group, the classification tree likely identifies the correct first split—9G-OTG—but subsequent splits are likely to be less accurate and generalizable. Using a classification tree also fails to account for dependencies due to nesting, which would be better handled with mixed effects extensions of the model (Hajjem et al., 2011; Sela and Simonoff, 2012).

ODE-c hoped to use classification trees to support invariance in the effectiveness of 9G-OTG on graduation. Specifically, ODE (n.d.-c) re-estimates the classification tree separately by race/ethnicity groups and compares the results to the whole sample. By using smaller subsamples, however, this approach exacerbates the issues with classification trees outlined above, limiting confidence in the generalizability of these findings. Furthermore, uneven geographic distribution of, e.g., racial/ethnic groups (Whitson, 2017; Clinton & Reeder, 2017) increase the likelihood of pooling some school-, district-, or zip-code-level effects into estimates of student-level differences with such an approach.

From an epistemological framework, the use of a decision tree to undergo discovery focuses on an exploration of how included factors are *associated with an outcome* (i.e., on-time graduation). In other words, classification trees return a breakdown of the factors which group observed data (graduates versus non-graduates) based on observed variables (e.g., 9G-OTG,

attendance percentage, etc.) and how likely a member of a given group is to have graduated. This is because the objective function of a classification tree is some metric of classification accuracy (Strobl et al., 2009). Being the first split in the classification tree, 9G-OTG is the variable which minimized misclassification of on-time graduation the most, among variables in the model (Strobl et al., 2009). Such a finding identifies 9G-OTG as the most predictive factor of graduation (among this sample and set of variables), validating this as a stronger EWI relative to other variables in the model.

New Avenues for Discovery

Other methods of discovery, notably extensions of MOB like the GLM tree and the GLMM tree are able to provide different and more suitable results to the secondary research question of subgroup effects (Zeileis et al., 2008; Fokkema et al., 2018). Rather than focusing on accuracy of predicting graduation—like in classification trees—MOB focuses on consistency in the entire logistic regression, asking "*does 9G-OTG work as an EWI just as well for everyone?*" (Zeileis et al., 2008; Fokkema et al., 2018). MOB does not return a dichotomous "*yes it fits universally*" or "*no it does not fit universally*," but a series of subgroups based on both the baseline probability of the group graduating when off track (i.e., the intercept) and the additive influence being on track (i.e., the slope; Zeileis et al., 2008).

Restating the Parametric Model

The specific parametric model to explore heterogeneity in 9G-OTG and on-time graduation is equivalent to a very face-valid, practitioner focused contingency table used to document differential incidence of a metric within a sample by subgroup. By applying GLMM trees to the multilevel logistic regression of interest, subgroups are returned which have the

greatest difference along the sum of all cells in the 2 x 2 contingency table made by (a) on-/off-track and (b) graduating/not graduating (Table 1).

Table 1

Restating parametric model explored for heterogeneity with multilevel logistic regression of graduation on 9G-OTG as contingency table.

	9G-OTG	Not 9G-OTG
Graduated	$\frac{EWI_{true\ positive}}{Total}$	$\frac{EWI_{false\ negative}}{Total} = \frac{Back\ on-Track}{Total}$
Did Not Graduate	$\frac{EWI_{false\ positive}}{Total} = \frac{fell\ off-track}{Total}$	$\frac{EWI_{true\ negative}}{Total}$

Considering the metric in context, though, identifying off-track students and getting them to graduate within 4 years is the intended feature of an EWI. Therefore, referring to these cases as "false" versus "true" positives or negatives is a carryover of the epistemology of statistically assessing model predictions, and are used for ease of communicating. The model returns circumstances and factors which maximize group differences among the sample as measured by a pooled aggregate of all cells of the contingency table (Table 1), which can be expressed in a number of mathematically equivalent ways (discussed below; Alexander et al., 2015; Nahhas, 2023; Ben-Shachar et al. 2020).

Research Questions

Explicitly, two successive research questions were asked:

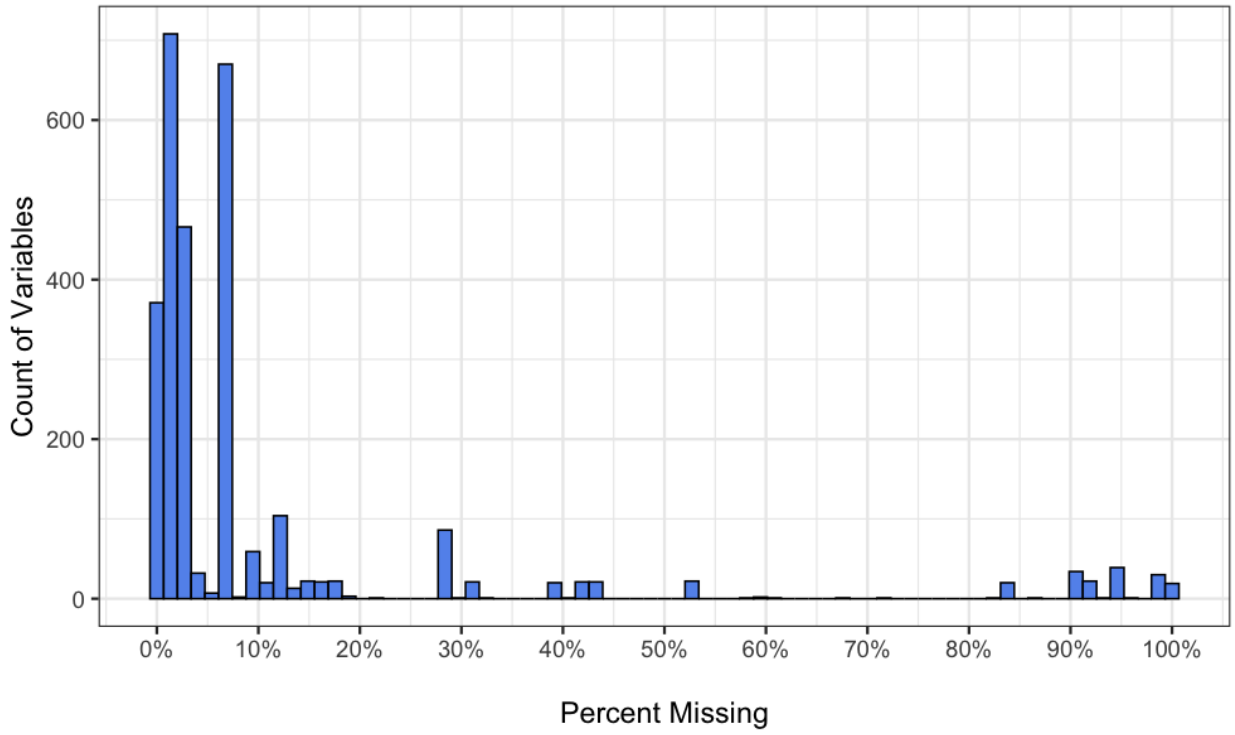
1. Is the relationship between graduation in 4 years and 9G-OTG consistent across student, school, district, and zip-code level factors in the SLDS and other publicly available data sources?

2. If not, what will emerge as most influential variable(s) to heterogenous proportions of students across in the 2 x 2 contingency matrix made by 9G-OTG and on-time graduation across student, school, district, and zip-code level factors?

Method

Extensive processing of the data was required to compile these data and organize them for analysis with GLMM trees. In total, roughly 1000 lines of R code was required to clean and join these data even relying on loops, iterations, and custom functions to decrease repetitious code. A total of 3,154 variables joined to each student's data. Figure 2 shows a histogram which counts the number of variables with a given percent missing.

Figure 2
Histogram showing percent of a variable which was missing data before deletion.



Compiling such a large dataset required joining across multiple sources. Most sources offered extensive manuals on how to process data, the level of measurement, assumptions of the

data, and more. These manuals guided use of the data. The project amassed complete data for 1,570 of variables (1566 splitting variables, 2 random effects, 1 outcome, and 1 covariate), meaning exhaustive discussion of theory and (coding-based) decision points was not reasonable. Instead, the method section documented the names of surveys, the levels of measurement, and the topics measured. Additionally, when transformations and normalizations—referred to as "feature engineering"—were applied, they were documented. Readers interested in nuances of the data other than those given here are redirected to the respective citation for the user manual.

Data

All data joins began with student data and used "left joins" which filtered new data based on matching an ID variable to prior data. Hence, variables were only added to pre-existing data where ID variables (e.g., school ID, district ID, zip code) matched across two datasets (Wickham et al., 2019). Notably, school- and district-level identifiers frequently differed across sources and relatively minor differences in exact characters used as names of clustering units prevented using clustering unit names directly (e.g., Eugene 4J versus Eugene 4 J versus Eugene4j). In such cases, regular expressions (R Core Team, 2024) were leveraged in conjunction with ODE's institution lookup table (ODE, 2024-b) to match the irregular names to school IDs in the SLDS. Regular expression was also used to fix errata which would have otherwise prevented the ability of data to be matched (e.g., NCES' file labeled "CDP02_105_04000US41_91143652328.txt" line 32 to 150 have Oregon schools listed in Georgia, Florida, Delaware, and other US States; NCES, n.d.-d).

At all levels of data, some level of feature engineering was done to decrease runtime. First, variables were removed if they could not have an influence on the outcome. For continuous

variables, variances were calculated and any which had exactly zero variance were removed from the analysis. If categorical variables had only one value, they were similarly removed.

Another decision made to decrease runtime was categorizing variables said to be continuous, but with exactly two or three observed values in the data. To do so, the number of unique observed values were counted in continuous variables. If exactly two or three values existed in the data, the variable were converted to categorical variables. Such a re-coding changes the instability test as described by Hjort and Koning (2002; Zeileis et al., 2008) instead of the extension of Andrews (1993) described for continuous variables (Zeileis et al., 2008; Zeileis & Hornik, 2007). With only one or two observed thresholds in continuous variables in the data, those (observed) thresholds are the only values which can be returned by the threshold selection done on continuous variables. In other words, Hjort and Koning (2002) can converge on the same (observed) threshold of instability faster than method which uses fluctuation tests extended from Andrews (1993; Zeileis & Hornik, 2007; Zeileis et al., 2008).

Missing Data

Presently, GLMM trees lack an internal means to impute missing data or pool estimates from multiple samples. To prevent bias from single imputation, data were limited to cases which were complete, but only after variables with excessive missingness (> 5%) were removed from the final dataset (Table 2). Variables were removed from the data if more than 5% were missing. Once the dataset was comprised of cases with 95% or more complete observations, row-wise deletion of cases was conducted, eliminating about 9% of students per year. Table 2 shows the decrease in sample size with progressive joins. The largest decrease in number of students were those with missingness in student-level data (Table 2 Step 0 to Step 1). The most missing

variable in this sample was 9G-OTG status with a total of 3,568 and 3,985 students missing a 9G-OTG identifier in the 2016-17 and 2017-18 cohorts, respectively.

Student-Level Data

The sample includes two cohorts of Oregon 9th graders obtained from the state's SLDS, specifically the 2013-14 and 2014-15 freshman cohorts who graduated on-time in 2016-17 and 2017-18. Later cohorts were not included because of the disruption due to the COVID-19 pandemic. Student-level characteristics were obtained through an IES Grant (award number: R305S210005, Farley et al., 2021). Student-level disciplinary incidents, test score, and attendance data were provided to UO for offshoot collaborations of the original Farley et al. (2021) research grant. The SLDS data housed 9G-OTG status for all students. Students were on-track to graduate if they had completed 25% of the course credits required to graduate by the end of their 9th grade year (ODE, 2018-a).

On-Time Graduation. Graduation was defined with ODE's average daily membership (ADM) end date codes, with 6 possible codes being included as on-time graduation, as long as the student reached the milestone 3 school years after assessing 9G-OTG status (i.e., completing high school in 4 years). Specifically, students were coded as graduating on time if they:

1. met requirements for and was awarded a high school diploma,
2. completed Non-Diploma-Track Program and Received Certificate (i.e., special education students in Individualized Educational Programs [IEP]),
3. completed a vocational program recognized by state or district and Received Certificate,
4. received High School Diploma Equivalency Certificate / General Educational Development (GED) credential,

5. met requirements for a High School Diploma, but it is not yet awarded (e.g., those who have met requirements to graduate but choose to remain continue their enrollment), or
6. met requirements for and was awarded a High School Diploma, continuing to post graduate scholars program (under SB 1537).

Transitory and Multiple Enrollment. The Oregon Student Membership Manual outlines broad strokes of data handling student-level data with essential information regarding enrollment data (Gill, 2023). Students with multiple attendance records were reported in SLDS data. In total, n = 2,708 (2016-17) and n = 2,321 (2017-18) students were reported to be a part of more than 1 ADM program in the same year. Gill (2023, p. 9) provided a brief explanation for such cases: "*Students in multiple programs, who have withdrawn and re-enrolled or who have transferred schools, may have multiple records reflecting enrollment on the Accountability dates in October and/or May,*" before explaining the order of precedence used in reporting. However, the student's dual or transitory enrollment in multiple ADM programs may reasonably influence the student, so a dichotomous indicator of dual ADM enrollment within a single year was created before retaining the case with higher precedence.

Standardized Test Scores. These cohorts progressed through Oregon schools at a time where most students took the same standardized tests, but a few took older "legacy" tests measured on a different scale (ODE, 2019). To prevent deletion among student with the legacy test where possible, a uniform scale was made. First, tests were mean centered and scaled with standard deviations by test type (legacy versus new), subject, and school year. Some students did not have all test scores reported and for others, one or two test scores were coded as missing. Specifically, legacy math and English tests were converted to missing values for the 7 students which took each test. Such a sparse distribution of legacy math and English scores limited

confidence in scaling these scores and pooling with other tests. By contrast, nearly 1000 students had taken the legacy science test, allowing scores to be scaled within the subgroup and cases to be retained. Once centered-and-scaled scores were made by test type, subject, and year, they were averaged with missing data omitted. In other words, if students had two or three subject tests completed, the score was an average. However, if only one test was available, the individual test score was used.

High School Attendance. Attendance data also was available inconsistently across students' grades, making it difficult to include each year independently. Average attendance was calculated across 9th to 11th grade as a proportion of total school days attended, allowing omission of missing data. Like the aggregated and standardized testing metric, this prevents deletion when a student did not have three years of attendance data.

Disciplinary Incidents. All student-level disciplinary incidents in the student's senior year were included in the analysis as count variables (i.e., counts of incident types for each student), as well as a variable representing their cumulative number of incidents until the end of their 11th grade, and a cumulative metric which included senior year incidents. Disciplinary incidents were provided as counts of approximately thirty action categories.

Language of Origin Codes. Originally, the processes above constituted all student-level feature engineering. However, assessment of instability along language of origin codes was not feasible and the algorithm crashed on all cores. Warnings showed the number of tests required was larger than the computer could process, so the variable was collapsed into fewer categories before re-estimation. The top 10 most common language of origin codes were retained as categories and all others were converted to a category labeled "other."

Table 2

Counts of observations per year as additional data were incorporated to SLDS.

Step 0: Student-Level ODE Data						
Year	12th Graders	Schools	Districts	Zip Codes	Variables	Percent Missing Cells[1]
2016-17	n = 38,399	j = 431	k = 190	l = 422	p = 41	5.5%
2017-18	n = 40,275	j = 429	k = 191	l = 429	p = 41	5.5%
Step 1: Total Included Student-Level Data[2]						
Year	12th Graders	Schools	Districts	Zip Codes	Variables	Percent Missing Cells[1]
2016-17	n = 33,480	j = 362	k = 185	l = 407	p = 70	0.13%
2017-18	n = 33,733	j = 360	k = 181	l = 408	p = 70	0.16%
Step 2: All Joined Student & School-Level Variables						
Year	12th Graders	Schools	Districts	Zip Codes	Variables	Percent Missing Cells[1]
2016-17	n = 33,229	j = 350	k = 175	l = 403	p = 1,589	28.82%
2017-18	n = 33,250	j = 350	k = 174	l = 404	p = 1,589	28.85%
Step 3: All Joined Student-, School-, and District-Level Variables						
Year	12th Graders	Schools	Districts	Zip Codes	Variables	Percent Missing Cells[1]
2016-17	n = 33,229	j = 350	k = 175	l = 403	p = 2,105	22.82%
2017-18	n = 33,250	j = 350	k = 174	l = 404	p = 2,105	22.82%
Step 4: All Joined Student-, School-, District, and Zip-Code-Level Variables						
Year	12th Graders	Schools	Districts	Zip Codes	Variables	Percent Missing Cells[1]
2016-17	n = 33,229	j = 350	k = 175	l = 403	p = 3,154	15.99%
2017-18	n = 33,250	j = 350	k = 174	l = 404	p = 3,154	15.97%
Step 5: After Removal Criteria						
Year	12th Graders	Schools	Districts	Zip Codes	Variables	Percent Missing Cells[1]
2016-17	n = 28,949	j = 265	k = 114	l = 177	p = 1,576	0%
2017-18	n = 29,065	j = 263	k = 114	l = 178	p = 1,576	0%

[1] Percent missing cells always calculated at student-level = (total missing cells) / (number of rows x number of columns)

[2]The greatest missingness were in the 9G-OTG variable, with a total of 3,568 and 3,985 missing in 2016-17 and 2017-18 respectively.

School-Level Data Sources

Civil Rights Data Collection. Civil Rights Data Collection (CRDC) data were available for school districts in Oregon biennially and the 2017-18 survey was used for this sample, as it was the most recent survey which contained both cohorts (CRDC, 2020). The CRDC (2024) self-describes as a

> *longstanding and important aspect of the U.S. Department of Education Office for Civil Rights' overall strategy for administering and enforcing civil rights laws that prohibit discrimination based on race, color, national origin, sex, disability, and age by schools, school districts and other entities that received Federal financial assistance from the Department.*

The 2017-18 CRDC school-level data contains 16 topical "modules" (e.g., school characteristics, course offerings, placement exam scores, absenteeism, discipline, school faculty/staff, etc.) spread across 30 data files (CRDC, 2020). Each contains between 1 and 140 variables about the school. See (CRDC, 2020) for all variables.

Planned Implementation of High School Success Funds. Measure 98 funding was primarily given to districts, but in a few cases were disbursed to individual schools (Farley et al., 2021). Measure 98 data was thus reported on the school-level based on the funding recipients' intention to fully (data tracking + high school success coaches), partially (data tracking alone), or not implement the components of Measure 98 when applying for funding. When joined with student-level data, implementation data reported a total of 40 cases for which implementation did not match across datasets. Of those 40 cases, 28 were able to be matched manually using the ODE's institution lookup tool (ODE, 2024-b), but 2 districts (district ID = 3477 and 3476, representing 16 unique schools and 325 students) were coded as missing. The ODE institution lookup tool also provided a metric of school type (standard public, charter, or alternative; ODE, 2024-b).

District-Level Data Sources

NCES hosts demographic, social, economic, and housing data gathered from the Census American Community Survey (ACS) which have been converted to calculate district-level estimates about the entire population of the district, as well as tabulations calculated with restricted samples of only school-age children and their parents (NCES, n.d.-d). Tabulated data are referred to as the ACS-ED collectively and are reported as averages over a five-year span. All tabulations of school-age children and their parents were included in the analysis. The 2013-17 and 2014-18 data were averaged with missingness omitted, extracting information from cases

which were not missing in at least one of the two years. In total 219 parent level variables were extracted from demographic, economic, and social variables (NCES, n.d.-d). For school-age children, the ACS-ED offered district level tabulations of 300 variables.

Zip-Code-Level Data Sources

Three data sources were compiled at the zip-code level. The Oregon Health Authority (OHA) reports a dichotomous measure of rurality (rural vs. urban) reported at the zip-code-level (Office of Rural Health [ORH], n.d.). The Agency for Healthcare Research and Quality (AHRQ) is housed within the Department of Health and Human Services, and report zip-code-level data measuring more than 352 social determinants of health (SDOH; AHRQ, 2020). AHRQ data is reported yearly, and 2013 and 2014 data were averaged by zip code. Finally, zip-code-level Census data were taken from American Community Survey 5-Year (ACS-5), specifically the 2013-17 ACS-5 was used for these zip codes (US Census Bureau, 2018). The ACS-5 contains 698 variables for each zip-code in Oregon, and reports over more restricted ranges (e.g., 1 year) were not available at the zip code level.

Software

All analyses were conducted using R version 4.3.1 (R Core Team, 2024). All data manipulation and visualization were done with the *tidyverse* suite of packages (Wickham et al., 2019). The *glmertree* package was used for fitting GLMM trees (Fokkema et al., 2018), and the *gardenr* package (Loan, 2023) was used for tuning hyperparameters on GLMM trees. The *rsample* package (Frick, et al., 2023) was used to split training/testing data, create cross-validation datasets from training data, and create the maximum entropy grid used in grid search. The Census application programming interface (API) was queried with the *tidycensus* package (Walker & Herman, 2024).

Model Estimation

Fixed-Effects Structure

Graduation status was regressed on 9G-OTG status with a binomial link function to handle the dichotomous outcome. Those who did not graduate within 4 years were coded as 0 and any of the 6 conditions mentioned above (see *On-Time Graduation*) were coded as 1. Student-level 9G-OTG was also coded dichotomously with 0 indicating a student off-track to graduate and 1 indicating a student on-track to graduate. No other fixed effects were included.

Random-Effects Structure

A series of logistic regressions of on-time graduation on 9G-OTG were conducted using varied random effects specifications and the preferred random-effects structure was chosen with a combination of likelihood ratio tests (LRTs), Akaike information criterion (AIC), and Bayesian information criterion (BIC). Accounting for school-level differences with a random intercept was preferred by all metrics to a single-level regression without random effects (Table 3). Similarly, accounting for students with a district-level intercept was preferred to the model with no effect (Table 3). Having the same degrees of freedom, LRTs were unable to compare models with a random intercept for schools versus for districts, but information criteria (ΔAIC = 2296; ΔBIC = 2297) preferred school-level random intercepts to district-level random intercepts. The school-level random intercept was then compared to a model which explicitly nested schools within districts, which found no added benefit of the nested effect (Table 3). A cross-classified model added a random intercept for zip code and was compared to the model with a random intercept for schools alone. The cross-classified model being the preferred of all tested conditions (see Table 3).

Table 3

Tested clustering conditions (preferred models bolded).

Nesting						
Random Effects	*lme4* Syntax	ΔBIC	ΔAIC	Δχ2	Δ df	p
None	—	—	—	—	—	—
School	**(1 | school)**	**8672**	**8690**	**8695**	**2**	**<0.001**
School Nested in District	(1 | district/school)	-17.68	-8.7	0	1	1.00
Cross-Classified						
	Model Comparison	ΔBIC	ΔAIC	Δχ2	Δ df	p
School	(1 | school)	—	—	—	—	—
School & Zip Code	**(1 | school) + (1 | zip)**	**17.00**	**25.97**	**27.97**	**1**	**<0.001**

Anchoring Parametric Model in GLMM Tree

With fixed- and random-effects documented, the entire parametric model used to optimize the GLMM tree is thus defined: cross-classified random intercepts account for systematic deviations seen at the school- and zip-code-level, enabling more accurate estimation of the association between on-time graduation and 9G-OTG. By applying MOB to the fit of the cross-classified logistic regression, this GLMM tree reports subgroups in the intercept and slope of the relationship between 9G-OTG and on-time graduation. The model will report fixed effects coefficients from the logistic regression by subgroup recursively until further subgrouping does not improve model fit.

Manually Specified Hyperparameters

In accordance with *Paper 2*, default settings were intentionally chosen for the *ranefstart* hyperparameter, initializing the algorithm with the GLM tree. The *cluster* hyperparameter was also set to its default value, meaning the model only used random effects to account for heterogeneity (without secondary corrections to standard errors).

Hyperparameter Optimization

Based on the results of *Paper 1,* hyperparameter optimization (HPO) was conducted to identify a more parsimonious tree. A maximum entropy grid was used to cover a hyperparameter space with 10 values (boundaries defined in Table 4).

Table 4

Boundaries of hyperparameter space by variable.

Name	Meaning	Boundaries of Tested Hyperparameter Space
alpha	Minimum threshold of significance in parameter instability tests required for a split to be taken.	[0.01, 0.33]
minsize	Minimum number of observations in a node.	[0.1%, 10%]
maxdepth	Maximum depth of the tree (number of total splits the model is allowed to find).	[2, 20]
trim	The proportion of outliers removed in assessing instability of a given split (i..e., in calculating instability statistic and p-value)	[0.01, 0.3]
prune	Should models be pruned by AIC, BIC, or do not prune?	AIC, BIC, no pruning
bonferroni	Use Bonferroni corrections to adjust threshold of instability statistic?	TRUE/FALSE

The hyperparameter value which minimized the mean classification accuracy across 5 folds of cross-validation (CV) was used to fit the final GLMM tree. Table 5 outlines the process of HPO with maximum entropy grid search (MEGS).

Assessing & Interpreting GLMM Tree

The performance of the model was assessed with the ability of trained model—tuned with cross-validation across 75% of the data—to predict new cases (25% of the data), assessing the accuracy of the model. Next, all of the data (training and testing) were used with optimal hyperparameters to fit the final model for interpretation; specifically, model-identified splits and regression parameters were used to reflect on possible means of follow-up research and student support.

Table 5

Hyperparameter Optimization Procedure with Maximum Entropy Grid Search.

Required Inputs	***Traversing Hyperparameter Space***	***Assessing Final Model***
Data • *Train with 75%* • *Retain 25% to Test Accuracy* *Cross-Validation Procedure (5-fold)* *Model* *Graduation ~ OTG + (1| school) + (1|zip)* *A Defined Hyperparameter Space* *Maximum Entropy Grid Search to Select Testing States*	*Initialize a maximum entropy grid of 10 hyperparameter states from hyper parameter space.* *Create 5-fold Cross-Validation Object from training data.* *For each fold of the cross-validated object {* *Fit a model with the hyperparameter grid and k-fold training data.* *Assess classification accuracy with k-fold testing data.* *}* *Take mean of classification accuracy for each hyperparameter state across the 5 folds* *Select hyperparameter state with best mean classification accuracy across 5 folds.*	*Using hyperparameters selected after MEGS:* 1. *Fit model with training data, extract AIC, BIC, and number of terminal nodes* 2. *Predict outcomes of withheld final testing data (25% of original sample)* 3. *Calculate final classification accuracy with simulated values vs. predictions.*

Parallel Processing

To minimize computation time, parallel processing was used to distribute the task across multiple cores on a 2021 Apple M1 Pro Chip with 32 GB random-access memory (RAM) and 10 computer processing unit (CPU) cores. In parallel, three cores processed one hyperparameter state across each of the k-folds using the *foreach* (Microsoft Corporation & Weston, 2022) and *doparallel* (Microsoft Corporation & Weston, 2022) packages. This was repeated with three hyperparameter states (one per core), assessing a total of ten hyperparameter states over four batches (i.e., indices 1-3, 4-6, 7-9, and 10). Between batches, the computer was restarted to free

memory swaps and caches resultant from overloading onboard RAM. Using more than 3 hyperparameter indices in parallel was not possible with available RAM.

Results

The fastest hyperparameter index completed 5-fold CV in 10.80 hours, and the slowest index took 24.81 hours, totaling 158.5 hours (or 6.6 days) of computation time. After training, the final model correctly predicted graduation of 93.03% of unseen students from testing data. In direct comparison to the 9G-OTG metric, the model had a greater percent of true-positives (i.e., observed on-time graduation). 9G-OTG outperformed the model on true negatives (Table 5). In comparison to model-based predictions, 9G-OTG could be phrased as having a higher probability of false negatives.

Table 6

Accuracy of 9G-OTG versus GLMM Tree Estimate on Graduation among (unseen) Testing Data.

	9G-OTG (87% accurate)		Model Estimate (93.3% accurate)	
	9G-OTG	Not 9G-OTG	Predicts Graduation	Predicts No Graduation
Graduated	82.0%	3.33%	90.5%	1.17%
Did Not Graduate	9.64%	4.99%	5.81%	2.52%

Hyperparameter Optimization

Table 7 shows the mean fit statistics by hyperparameter state for the 10 states tested, ordered by mean classification accuracy over cross-validated datasets. Several combinations of hyperparameters resulted in identical performance, size, and fit to the data (i.e., Table 7 Rows 2 through 7; Rows 8 to 10). Identical estimates and standard errors of these fit metrics over 5-fold cross-validation are unlikely without identical terminal node predictions, however. Without

further analysis, it is unclear if the models differed to an extent which did not affect final predictions. Regardless of if split points were identical, the findings supported a convergence of model performance, size, and fit to the data across a range of hyperparameter states (Table 7). For these data, three states emerged, all very accurate in making predictions (~93%), with only small differences among their predictive accuracies (range 0.08%).

Table 7

Results from hyperparameter tuning across the k-folds of training data (final hyperparameters bolded).

Mean Class. Acc. (SE)	Mean Number Terminal Nodes (SE)	Mean AIC (SE)	Mean BIC (SE)	Max Tree Depth	Alpha	Trim Proportion	Correction to Multiple Testing	Minimum Node Size	Pruning Method
93.15 (0.09)	**9.6 (0.4)**	**13341 (58)**	**13593 (67)**	**5**	**0.13**	**0.23**	**Bonferroni**	**56**	**None**
93.12 (0.12)	25.6 (0.81)	12831 (41)	13489 (33)	13	0.03	0.3	None	3,340	BIC
93.12 (0.12)	25.6 (0.81)	12831 (41)	13489 (33)	4	0.13	0.29	None	1,210	BIC
93.12 (0.12)	25.6 (0.81)	12831 (41)	13489 (33)	6	0.04	0.02	None	1,000	AIC
93.12 (0.12)	25.6 (0.81)	12831 (41)	13489 (33)	16	0.25	0.23	None	902	AIC
93.12 (0.12)	25.6 (0.81)	12831 (41)	13489 (33)	6	0.23	0.26	None	4,069	BIC
93.12 (0.12)	25.6 (0.81)	12831 (41)	13489 (33)	3	0.04	0.18	None	3,436	AIC
93.07 (0.09)	2.8 (0.2)	13680 (55)	13760 (52)	9	0.24	0.03	FALSE	1,247	None
93.07 (0.09)	2.8 (0.2)	13680 (55)	13760 (52)	10	0.06	0.18	FALSE	2,477	None
93.07 (0.09)	2.8 (0.2)	13680 (55)	13760 (52)	20	0.24	0.04	FALSE	2,645	None

The average number of terminal nodes across CV datasets varied substantially as hyperparameter states were explored (range of CV mean terminal node size = [2.8, 25.6]; Table 7). Across the CV datasets, the preferred hyperparameter state displayed a mean of 9.6 terminal nodes (SE = 0.4). Hyperparameter states with worse performance had either many more (M = 25.6 terminal nodes; SE = 0.81) or many fewer (M = 2.8 terminal nodes; SE = 0.2), indicating over- and underfitting the training data relative to the optimal state, respectively.

When the GLMM tree was pruned using either information criterion—AIC or BIC—estimated models (unsurprisingly) displayed better information criteria than models which did not prune. However, at least with these data, preferrable information criteria came at the cost of parsimony, with AIC and BIC models containing the most terminal nodes regardless of other hyperparameters. Information-criteria pruning improved AIC, relative to the preferred model (Table 7). BIC—which penalizes model size relative to AIC—differed to a much lesser extent considering the standard errors of BIC across CV datasets.

Optimal Hyperparameter State

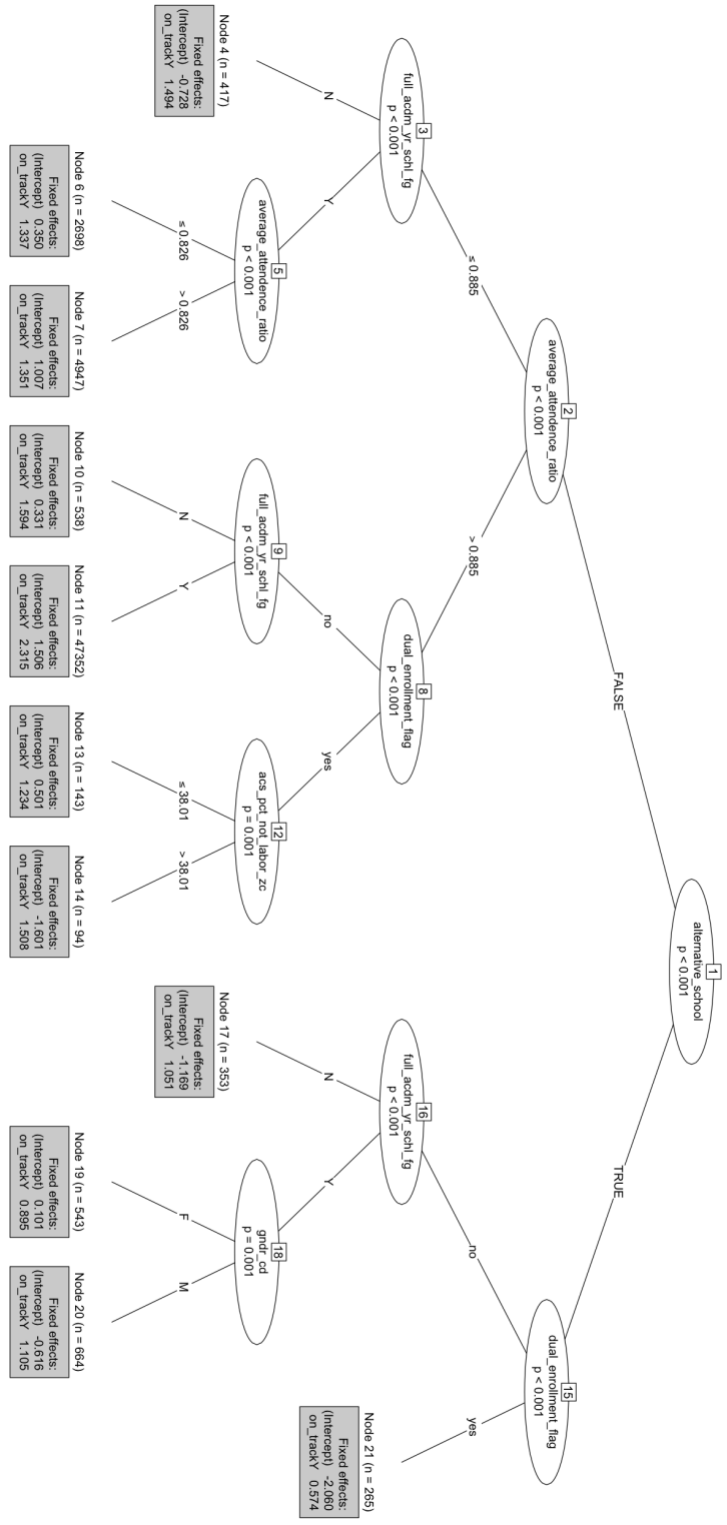
The final model used a combination of the six hyperparameters which optimized classification accuracy on unseen cases, averaged over CV datasets (Table 6). Two of the tested hyperparameters were optimized with their default values, specifically (a) applying Bonferroni corrections for multiple testing within parameter stability tests and (b) not pruning the model with information criteria. One hyperparameter was more restrictive than defaults: limiting depth of the tree to be less than 5, which prevents subgroups of the parametric model from being defined by ≥ 5 successive splits.

Some hyperparameters were less restrictive, most notably allowing the model to report subgroups as small as 54. At each parameter stability test, the model calculated instability

statistics with only the middle 77% of the data for each subgrouping. This specification chooses splits based on the influence of the 23% of data comprising the tails of the distribution by omitting these "outliers" from calculation of the *Instability Statistic* and *p* value. Additionally, the minimum threshold required for a split to be considered was set to 0.13 from 0.05. This term, *alpha*, is the threshold that the *Instability Statistic*'s simulated *p* value must be smaller than for GLMM trees to report a split. Importantly, these hyperparameters are not necessarily reached, as they are imposed in combination. For example, limiting tree depth in a large sample may prevent the minimum node size or other hyperparameters from being close to their threshold. In fact, this was seen with all instability statistics having $p \leq 0.001$ at all partitions (Figure 3). Additionally, the was allowed to report groups as small as 55 students, but the smallest subgroup identified was $n = 94$ (Figure 3).

Figure 3.

Tuned GLMM tree structure regressing on-time graduation on 9G-OTG.



Maximizing Interpretability of Results

As a variation of a 2 x 2 design, results can be presented in numerous ways via mathematical transformations (Alexander et al., 2015; Nahhas, 2023). Based on ODE reports which provide effect sizes, risk ratios were calculated as the "risk" of graduating (ODE, n.d-c; ODE, 2019; ODE, 2018-a; ODE-2018-b). Unlike ODE's report which has a single intercept—and therefore a single baseline probability—relative risk of an on-track group is in reference to a node-specific baseline probability. Hence, absolute risk was expressed as a probability and additional caution is advised in interpreting other estimates.

First, results are presented as absolute "risk" of graduating for those on-and off-track by subgroup (Figure 4), additionally the odds ratio of graduating for those on-track is included (Figure 5). To simplify interpretation further, absolute "risk" of graduating was calculated for those on and off track by accounting for the baseline probability and the risk ratio. Absolute risk for off-track students was the baseline probability, specifically by exponentiating the intercept and calculating baseline probability as:

$$probability = \frac{e^{\beta 0}}{1 + e^{\beta 0}}$$

(Nahhas, 2023). Risk ratios were calculated from the model with the *effectsize* package and were used to convert from relative to absolute risk (Ben-Shachar et al. 2020). In a 2 x 2 design, the risk ratio is the

$$Risk\ Ratio = \frac{\text{Risk}_{Exposed}}{\text{Risk}_{Unexposed}} = \frac{\text{Risk}_{On-Track}}{\text{Risk}_{Off-Track}}$$

Thus, the risk of graduating on time for 9G-OTG is,

$$\text{Risk}_{On-Track} = Risk\ Ratio * \text{Risk}_{Off-Track}$$

where $Risk_{off\text{-}track}$ is equivalent to the baseline probability when off track, defined above (Nahhas, 2023). Odds ratios are also provided, as well, which provide node-specific effect sizes of the odds of graduating among on track students divided by the odds in off-track students.

Splitting Variables

The branching structure of the GLMM tree is shown in Figure 3. Terminal nodes report the size of the subgroup, the *p* value from tests of parameter instability, and node-specific fixed effects. Coefficients from terminal nodes are included in Table 8 and are shown in Figure 3 ("Intercept" = node specific intercept; onTrackY = difference in log-odds when on-track). Additionally, these results were transformed and presented visually in two ways absolute risk (Figure 4) and odds ratio for 9G-OTG (Figure 5) are reported by node. Absolute risk is highly interpretable, but odds ratios give a singular metric of the effectiveness of the intervention within terminal node (i.e., within the context).

Figure 4.

Absolute risk of graduating and 95% confidence intervals (expressed as probability) by node and subgroup.

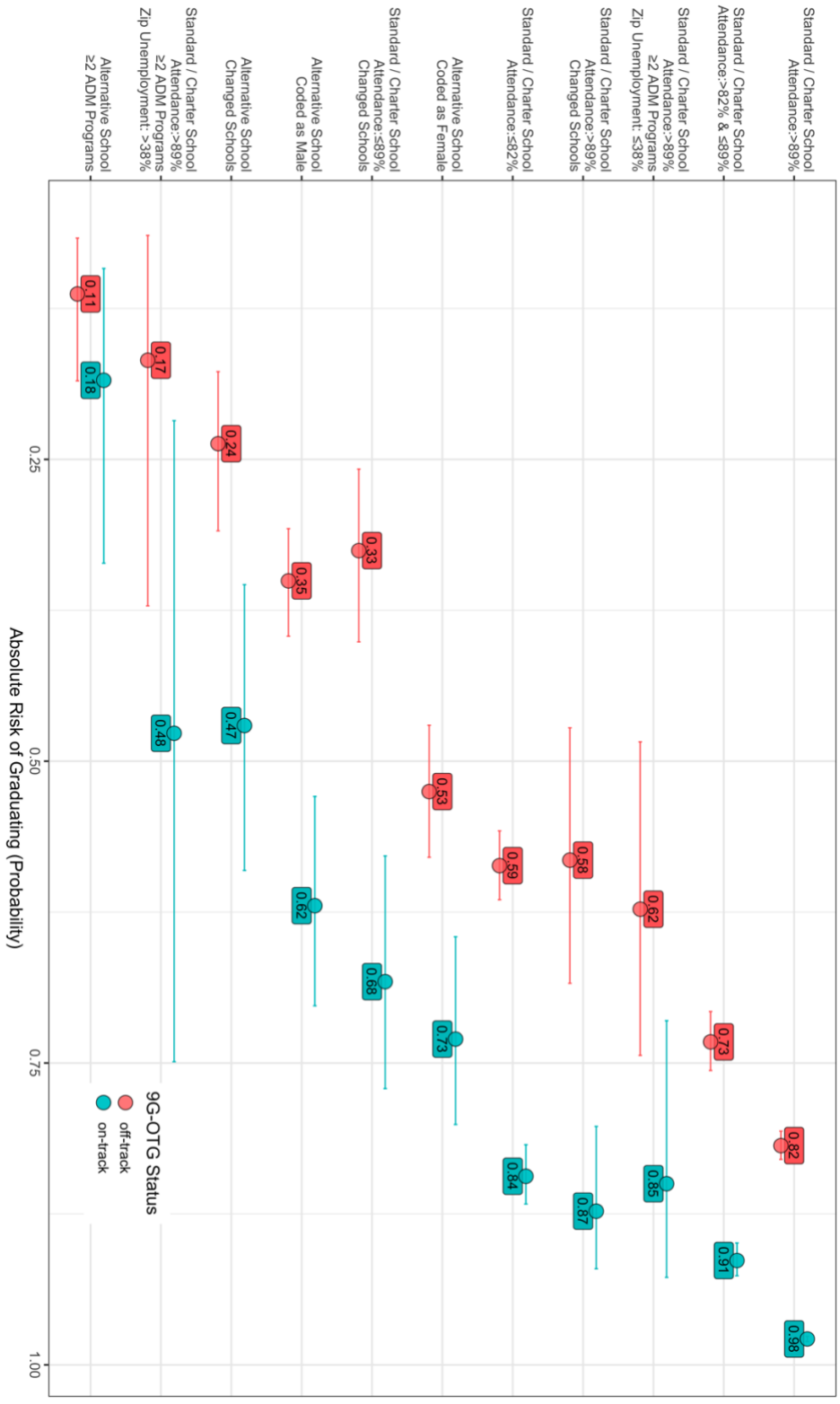


Figure 5.

Odds ratios and 95% confidence intervals for those 9G-OTG (dotted line = null odds).

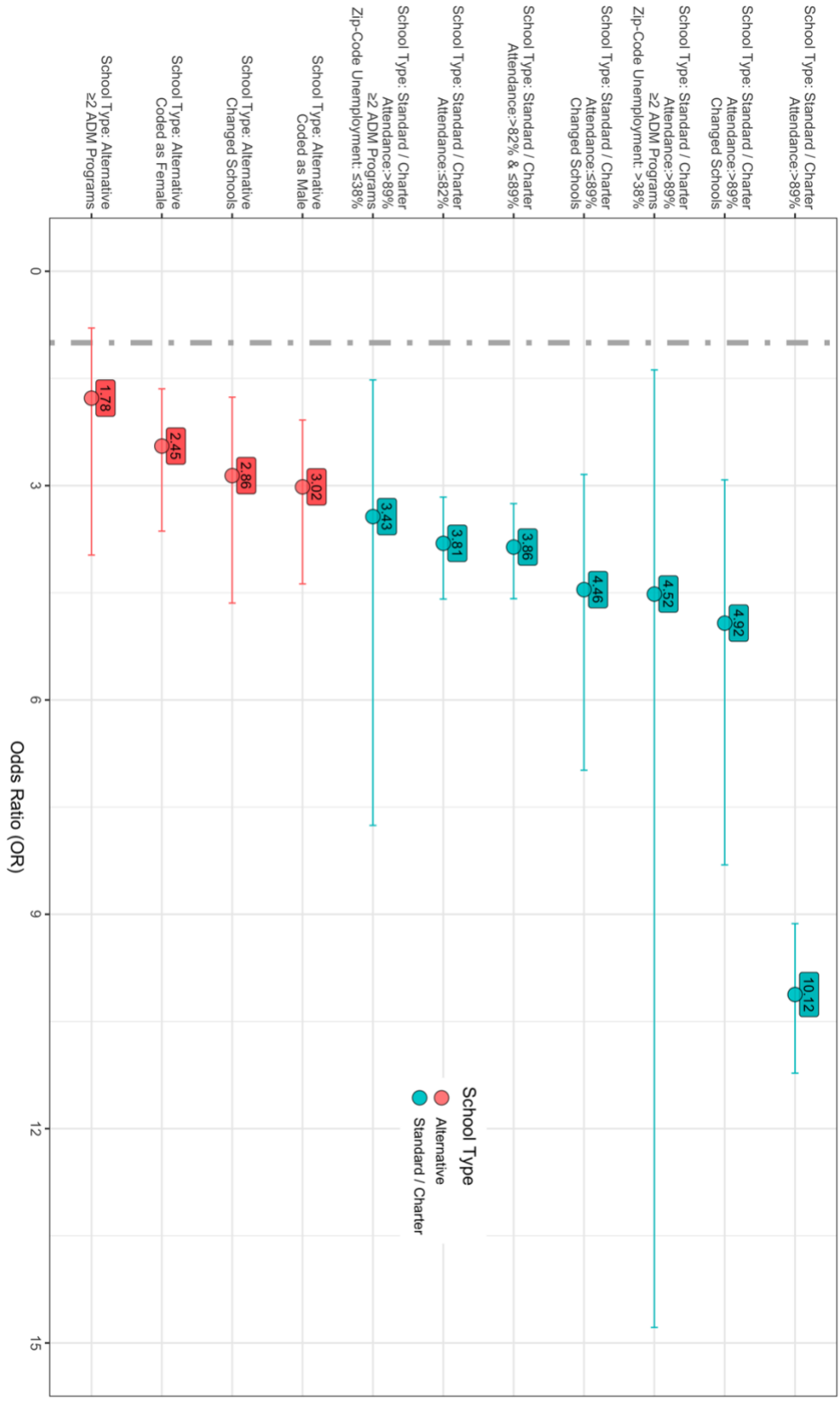


Table 8

Fixed-effects coefficients estimated from GLMM tree terminal nodes.

Coefficient	Estimate	Standard Error	*z*	*p*	$e^{(Estimate)}$
Standard / Charter School Attendance: ≤89% Changed Schools					
Intercept	−0.728	0.167	−4.356	<0.001	0.483
On Track	1.494	0.229	6.52	<0.001	4.456
Standard / Charter School Attendance: ≤82%					
Intercept	0.35	0.06	5.822	<0.001	1.419
On Track	1.337	0.095	14.07	<0.001	3.808
Standard / Charter School Attendance:>82% & ≤89%					
Intercept	1.007	0.064	15.841	<0.001	2.738
On Track	1.351	0.088	15.432	<0.001	3.86
Standard / Charter School Attendance:>89% Changed Schools					
Intercept	0.331	0.225	1.469	0.142	1.392
On Track	1.594	0.267	5.976	<0.001	4.925
Standard / Charter School Attendance:>89%					
Intercept	1.506	0.04	37.359	<0.001	4.507
On Track	2.315	0.053	43.954	<0.001	10.123
Standard / Charter School Attendance:>89% ≥2 ADM Programs Zip Unemployment: ≤38%					
Intercept	0.501	0.288	1.737	0.082	1.65
On Track	1.234	0.416	2.968	0.003	3.434
Standard / Charter School Attendance:>89% ≥2 ADM Programs Zip Unemployment: >38%					
Intercept	−1.601	0.548	−2.921	0.003	0.202
On Track	1.508	0.605	2.493	0.013	4.518
Alternative School Changed Schools					
Intercept	−1.169	0.186	−6.273	<0.001	0.311
On Track	1.051	0.247	4.251	<0.001	2.86
Alternative School Coded as Female					
Intercept	0.101	0.112	0.898	0.369	1.106
On Track	0.895	0.202	4.424	<0.001	2.447
Alternative School Coded as Male					
Intercept	−0.616	0.1	−6.161	<0.001	0.54
On Track	1.105	0.189	5.83	<0.001	3.018

The variable which most contributed to instability in the logistic regression was alternative school status (Figure 3). Students which enrolled in alternative schools had the lowest log odds of graduating when off-track and the increase in probability of graduating on time when on-track was less than the estimate seen in other subgroups. Among alternative schools, differences were seen in subgroupings of three variables. Students in alternative schools who were enrolled in more than one ADM program in a single year were very unlikely to graduate when off- or on-track in 9th grade (Figure 4). Cross-tabulation of on-track status and on-time graduation shows this clearly when grouped by alternative school status (Table 9).

Table 9.

Cross Tabulation of 9G-OTG and observed graduation for alternative vs. other (standard public and charter) schools.

	Standard Public / Charter Schools		Alternative Schools	
	9G-OTG	Not 9G-OTG	9G-OTG	Not 9G-OTG
Graduated	84.3%	9.22%	19.9%	21.8%
Did Not Graduate	2.91%	3.58%	16.3%	41.9%

Two variables appeared on both sides of the first split, the dual ADM enrollment flag, and the full academic year flag (i.e., within-year student mobility; Gill, 2023). The dual ADM enrollment flag was calculated based on student records existing for more than one ADM program type in a given year. Some students with more than one ADM program type are “shared time” students, in both public and private institutions, others transfer schools, and others may be a member of one school while being enrolled in a program that is shared across multiple schools/districts (e.g., an alternate program; Gill, 2023).

Students in Alternative Schools

Those who were in alternative schools and were enrolled in more than one ADM program graduated at the lowest rates regardless of on- or off-track status. Alternative schools comprised 48 of schools in the sample (j = 45 of 268 in 2016-17; j = 43 of 263 in 2017-18). Confidence intervals for the absolute probabilities of graduating by subgroup demonstrated substantial overlap in on- and off-track status (Figure 4), which is only true of 3 of the 11 terminal node groups. When students were enrolled in only one ADM program type in a given year, their stability at a single school was associated with logistic regression coefficients. Specifically, those who changed schools had lower absolute risk of graduating when off-track or on-track, compared to those who stayed at one school all year. Among those who attended the same school all year, those coded as female had higher absolute risk of graduating when off-track (probability graduating = 0.53) or on-track (0.72) compared to those coded as male who were off-track (0.35) and on-track (0.62) in this sample. Comparing those coded as female to those coded as male in this subgroup showed overlap between the confidence intervals of their probability to graduate when on-track, however, the confidence intervals do not overlap for those off-track. The baseline probability of graduating is roughly a coin-flip for those coded as female in this sample when off-track, but the model has relative confidence that most observations coded as male will not graduate when off-track among this subgroup.

Students in Standard Public & Charter Schools

GLMM trees segmented alternative schools from other school types in the sample, which included both standard public and charter schools. After removing missingness, the final sample retained 39 charter schools (j = 37 in 2016-17; j = 36 in 2017-18) and 192 standard public schools (j = 186 in 2016-17; j = 184 in 2017-18). Among students not in alternative schools,

average high school attendance was the most influential variable to the parametric model (Figure 3). Table 10 shows cross tabulations of 9G-OTG and graduation students who attend above and below the threshold of ~88.5%.

Table 10

Cross Tabulation of 9G-OTG and observed graduation based on first attendance threshold (≤88.49% vs > 88.49%) for students not at alternative schools.

	Attendance Rate ≤88.49%		Attendance Rate > 88.49%	
	9G-OTG	Not 9G-OTG	9G-OTG	Not 9G-OTG
Graduated	57.5%	21.0%	88.8%	7.25%
Did Not Graduate	7.78%	13.7%	2.90%	1.89%

Lower Attendance. When attendance was less than ~88.5% three subgroups were identified. Among students who remained in the same school for a full academic year, another split in attendance improved fit of the parametric model at 82.6%. Those in the lower-attending terminal node had lower absolute probability of graduating when off track (probability = 0.59), compared to those with greater rates of attendance (i.e., 82.6% and 88.5%; probability = 0.73). When on-track, though, students across these terminal nodes displayed smaller disparities in absolute probability than their off-track counterparts (attendance ≤82.6% node on-track probability = 0.82; attendance [82.6%, 88.5%] node on-track probability = 0.91), however, neither confidence intervals overlapped (Figure 4).

Higher Attendance & Singular ADM Program Enrollment. Among students with attendance rates above 88.9%, dual enrollment in an ADM program within a year was the most influential variable to stability of the parametric model. The impact of ADM program status (single vs. ≥2) was further partitioned with a different variable among both groups. First,

examining those in only 1 ADM program led to a partition in the sample based on if they transfer schools in their senior year or not.

The overwhelming majority of Oregon students are contained in the nodes described by >88.9% attendance, enrollment in a single ADM program type, and retention in the same school all year (n = 47,352). Notably, students in this—the most prevalent—group are shown to have high baseline odds of graduating (absolute probability = 0.82) with near complete confidence (probability = 0.98) that an on-track student in this sample will graduate on time. Even when students in a single ADM program—as well as not attending alternative schools and having high attendance rates—changed schools their senior year, their absolute risk of graduating was still 0.58 when off-track and 0.87 when on-track.

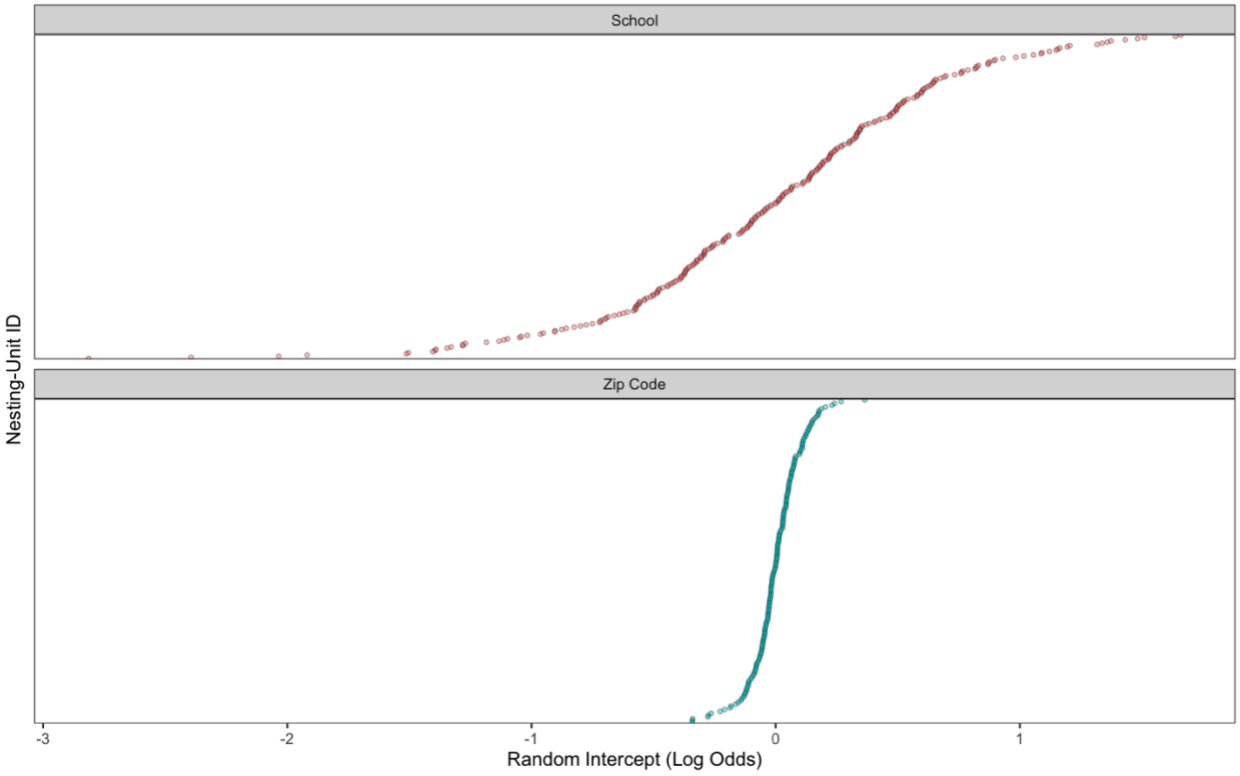
Higher Attendance & Multiple ADM Program Enrollment. Among students not attending an alternative school with high-attendance and enrollment in multiple ADM programs, a zip-code-level variable from the ACS-5 was most influential to the parametric model. Specifically, students from zip codes with ≤ 38% of adults over 16 employed have much higher absolute probabilities of graduating when off-track or on-track, compared to those in zip codes with > 38% (Figure 4). Among the node with high unemployment, the model has the second lowest absolute odds of graduating (0.17). Considering the size of this subgroup (n = 94), the standard errors for estimate of on-track and off-track were much larger than other groups, and when these students were on-tracking their absolute probability of graduating was only 0.48 with confidence intervals that overlap heavily with the off-track estimate. Corresponding students in higher-employed zip codes displayed the third highest absolute risk of graduating (probability = 0.85), with confidence intervals that overlapped with the second highest estimate of risk of graduating (Figure 4).

Random Effects

GLMM trees estimate global random intercepts, such that the coefficient is not dependent upon terminal node, as is the case with fixed effects. The variance across school-level log-odds of graduation (random intercept v*ariance* = 0.79, *SD* = 0.84) was larger than that for zip code (random intercept *variance* = 0.049; *SD* = 0.22). Figure 6 shows the distribution of random intercepts for both clustering units (in log-odds).

Figure 6.

Random Intercepts estimated by final GLMM tree.



Discussion

Among students in the Oregon SLDS, a confluence of student-, school-, and zip-code-level data were shown to create instability in the association between completion of 25% of course work by the end of 9th grade (9G-OTG) and the probability of graduating in 4 years. In most cases, previous research has documented the same or similar variables being associated with graduation directly or in combination with another variable in the same sample (ODE, n.d.-c; 2019; 2018-b; 2018-a) as well as in meta-analysis (Zheng et al., 2023).

Using more than 1,500 variables, GLMM trees explored subgroups which deviated most from other observations along the sum of a 2 x 2 contingency table made by (a) on-/off-track and (b) graduating/not graduating. In doing so, this paper used EDA to discover circumstances and factors which—when pooled—maximized group differences among the sample, focused within the context of the driving research question. With this parametric model, group differences were reported as they cumulatively influenced (a) probability of graduating when off-track, (b) probability of graduating when on track, and (c) the variation around these probabilities. Therefore, some subgroups had very different probabilities across groups and others had very similar probabilities of graduating across groups. Some subgroups displayed wide ranges of probabilities to graduate when on- or off-track and others had minimal differences. Each subgroup allows insight into factors which may decrease the chance of graduating on-time for on- and/or off-track students. For example, changing schools appears to be detrimental to probability of on-time graduation among several student groups. Conversely, the odds of graduation differed between those coded as male versus female in some groups, but not others.

Despite evidence of heterogeneity in the association, students who were 9G-OTG consistently had significantly higher probability of graduating than their off-track counterparts

across 8 of 11 subgroups. On the whole, 9G-OTG was shown to be a powerful EWI of on-time graduation in Oregon, despite being a binary, common-sense, metric assessed three years before graduation. The model was robust to invariance across over 1500 student-, school-, district-, and zip-code-level factors, except for six variables: alternative school (school-level), average attendance ratio (student-level), enrollment in 2+ ADM programs within a year (student-level), student mobility (student-level), percent of workforce unemployed ≥ 16 years old (zip-code-level), and sex (student-level).

Contextualized in Past Research

At least in practice as implemented, the results of the GLMM tree housed within the statistical theory of null and alternative hypotheses (of course bounded by the inferential-derived hyperparameters). In other words, a lack of an identified effect does not mean the effect does not exist outside or even inside the sample. With the recursive structure, instability in the parametric model was reduced to the greatest extent by the final structure. The optimal model specification found a confluence of influences which performed better over cross validation than any other recursively partitioned solution. Therefore, divergences from past research should be taken with a grain of salt and pursued if not in alignment with theory, as these effects may exist as well. As such, all models are wrong, but their relative utility depends on the design, data, and statistical validity. The model here used one algorithm to reproducibly explore invariance in an EWI's predictivity across effects and returned results which can help generate theory.

More than any other factor in the model, alternative school status was associated with a differential set of graduation probabilities for on- and off-track students. The metric of on-track to graduate appears to function differently for students in alternative schools than those in standard public / charter schools, and on-track students in alternate schools had smaller odd of

graduating ratios than students in other subgroups. Some factors were found to further partition subgroups among both alternative and non-alternative (i.e., standard public and charter) schools, such as being flagged as a mobile student (one not in the same school district for an entire year).

School Type. Alternative school status was most influential on heterogeneity in the relationship between 9G-OTG and on-time graduation. Alternative schools vary drastically across the state, with each school having a self-defined mission and all students having their own individualized education plan (US Department of Education [DOE], 2017). Oregon Laws (2021) define placement of alternative schools as *"those whose educational needs and interests are best served by participation in such programs and will include:*

A. *Students identified pursuant to Oregon Revised Statute (ORS) 339.250:*
    i. *Who are being considered for suspension or expulsion pursuant to ORS 339.250*
    ii. *Who have been suspended or expelled pursuant to ORS 339.250*
    iii. *Whose attendance patterns have been found to be so erratic that the students are not benefiting from the regular educational program; or*
    iv. *Who have had a second or subsequent occurrence within any three-year period of a severe disciplinary problem;*
B. *Students identified pursuant to ORS 329.485 (Statewide assessment system) and OAR 581-022-1110(5) who do not meet the standards or who exceed all of the standards at any benchmark level;*
C. *Students admitted to the district pursuant to ORS 339.115 (Admission of students) who have not yet turned 21 prior to the start of the school year and who need additional instruction to earn a diploma in compliance with Oregon Administrative Rule (OAR) 581-022-2000 (Diploma Requirements);*

D. *Students whose parents or legal guardians apply for the student's exemption from compulsory attendance on a semiannual basis as provided in ORS 339.030 (Exemptions from compulsory school attendance) and OAR 581-021-0076 (Exemption from Compulsory Attendance); and*

E. *Others who are individually approved for placement consistent with the district's board policies regarding the placement"*

Using aggregated school-level data from the SLDS, Zvoch et al. (2023) reported invariance in a GLMM parameterized as an interrupted time series (C-ITS) model evaluating effectiveness of Measure 98 by implementation status, though the GLMM tree was used to modify final parameterization (i.e., as a diagnostic tool, ensuring distribution of residuals were uncorrelated with observed variables). Zvoch et al. (2023) found invariance in their parametric model between (a) standard public schools and (b) alternative schools + charter schools, which inspired the C-ITS to control for intercept differences by school types. Numerous methodological and design choices could have led to divergent groupings of charter schools, including differences in the parametric model in GLMM trees, the level-of-analysis, or the sample restriction from missing data. Regardless, differences in these and associated processes are not fully understood across school type, particularly for charter schools. Furthermore, that conference presentation did not tune hyperparameters or use nearly as many covariates.

Attendance. ODE's classification tree of on-time graduation places reports a threshold in attendance of 81% being the second most important predictor of graduation, putting their threshold relatively close to the threshold between node 6 and 7 (ODE, n.d.-c). In the same report, ODE shows that the number of years as a regular attender—defined as 90% or more attendance—drastically increases odds of graduating for those on- and off-track. The 90%

attendance threshold reported by ODE is very close to the threshold of 88.5% in average attendance rate reported for all students which are in standard public and charter schools. In addition to explorations with Oregon students, a large body of research has demonstrated the value of attendance on graduation (e.g., Allensworth and Easton, 2005; Allensworth and Easton 2007; Bowers et al., 2012).

Mobility. ODE reports have also reported high school mobility as a significant direct predictor of on-time graduation among Oregon Students, and a predictor with differential impact based on 9G-OTG status (ODE, n.d.-c). Quantitative differences between the Poisson regression and the GLMM tree led to differential representations of variance and covariance. For example, the Poisson regression imposes a constant interaction effect of student mobility. The GLMM tree modeled the relationship flexibly by terminal node, but the effect arose independently in three branches of the final GLMM tree. Furthermore, the GLMM tree identified a variable relatively similar to mobility—dual/transitory enrollment in an ADM program within a year. Partial dependency and/or marginal effects plots should compare models with these two specifications and assess differences in practice, as results appear very harmonious.

Coded Sex. In ODE's Poisson regression, lower risk of graduation was reported for on-track students coded as male, compared to those coded as female. The GLMM tree returned this effect, but only within a high-risk subgroup of students: alternative school students. Among subgroup members coded as male (specifically those enrolled in one ADM program and who did not transfer schools), the probability of graduating when off-track was lower than among their female coded counterparts, and substantially so (difference of 0.18). Among on-track students, the difference between students coded as male and female was buffered (difference of 0.09).

Career and Technical Education ODE (n.d.-c) reports CTE enrollment significantly increases the risk of on-time graduation. Though CTE participation is not measured in student-level data, successful completion of these programs (which resulted in a completion certificate) counted as on-time graduation among a subset of students. The presence of CTE programs is available at the school-level, but the factor did not emerge as influential to the relationship between 9G-OTG and graduation.

***Economic Indicators*.** The GLMM tree did not identify student-level free and reduced-price (FRL) status or student disability status, both of which were shown to significantly decrease relative risk of graduating in ODE's Poisson regression. The variable most directly related to income from the GLMM tree—percent of adults unemployed—was measured at the zip-code-level. Though significant, the relative risk ratio for students coded as FRL indicated only 3% difference from the intercept (relative risk = 0.97). Although not directly comparable from the GLMM tree (i.e., with node-specific intercepts), large differences can be seen among similar students living in different zip codes. In low employment zip codes, absolute risk of graduating for off-track (probability = 0.17) and on-track (0.48) students were much lower than their counterparts in zip codes with greater employment (off-track: 0.62; on-track: 0.85). These findings are not necessarily at odds with ODE's use of student-level factors, as recent meta-analyses report significant effects of aggregated and disaggregated economic factors on on-time graduation (Zheng et al., 2023). Student disability status was not available for inclusion in the analysis.

These data contain several factors outlined by Zheng et al. (2023) which were shown to influence on-time graduation directly through meta-analysis. Student-level demographic factors (e.g., racial/ethnic status, special education status, economic factors, English Learner status)

were not identified as influential to instability in the association of 9G-OTG and on-time graduation. Our data do not include GPA but do incorporate standardized test scores. Zheng et al. (2023) discuss school-level improvement strategies, which would be measurable in a later sample (i.e., implementation of Measure 98 funds influencing a later cohort).

Improving 9G-OTG: For Whom Does 9G-OTG Not Differentiate?

Three terminal nodes displayed overlapping predictions of absolute risk for on- and off-track students. In such cases, excessive variation in the intercept, slope, or both, limit the effectiveness as 9G-OTG as an EWI in this sample. Despite predicting overlapping probabilities within the node, the point estimate of on-time graduation probability was always estimated to be higher among 9G-OTG students.

The absolute probability of these three nodes' 9G-OTG students differed qualitatively, however. For the lowest-performing node which did not statistically differentiate between on- and off-track students—i.e., in alternative schools enrolled in more than 1 ADM program in a year—the upper bound of the confidence interval for on-track did not overlap with a null absolute probability (i.e., 0.50). Hence, the model reported students would not graduate in this group even when on-track to graduate. The middle of the three nodes which did not differentiate between 9G-OTG status was the second lowest performing node. Students in this node had a confluence of risk and protective factors, and predicted probability was nearly null probability (0.48) with confidence intervals well above and below 0.50. Specifically, these students have very high average attendance rates, but are enrolled in more than 1 ADM program and live in neighborhoods with high unemployment. Following the decision tree by ODE (n.d.-c), students in this node would all be predicted to graduate whether on-or off-track.

The third node whose absolute risk of graduating cannot be distinguished is the other half of the split made by zip-code-level unemployment. Specifically, this node was comprised of students with more than 1 ADM program from non-alternative schools that attend regularly. However, these zip codes had lower unemployment than their counterparts. The lower unemployment node boasted the third best risk graduating when off-track (62%) and fourth best risk of graduating on-track (85%). Being close estimates and a relatively small node, the confidence intervals of these estimates overlapped. However, unlike their high-unemployment counterparts, students on-track in low-unemployment neighborhoods had lower bounds of probability estimates above null probability (0.50).

Limitations & Future Research

Student-Focused

Perhaps the greatest limitation of the study involves a loss of data from deletion of missing observations. Compared to the percentage of students lost to deletion, the percentage of nesting units lost were much larger. Using random effects to account for the unmodeled influence of missing clustering units mitigates but does not eliminate the problem (Bates et al., 2015). Zip codes are a notable example, which decrease from over 400 observed units to under 120 when variables were deleted. Uneven deletion can bias the model systematically, particularly if the process relating to missingness was also related to either 9G-OTG status or on-time graduation.

As a project of discovery, there was no attempt to causally link any covariates with 9G-OTG, graduation, or their interrelation. Correlational analyses may identify unexpected findings, encouraging limited resources to be allocated to exploring (atheoretical) effects which may not generalize. In this analysis, identified variables previously have been linked to 9G-OTG, on-time

graduation, or both. In concordance with theory, these findings support future exploration of the combination of factors which lead to differential outcomes among students (and differential predictivity of 9G-OTG). Furthermore, model comparisons can explore the adequacy of representing data as GLMM trees in comparison to parametric models with interactions (instead of segmented parametric models).

Differences in the parametric model between alternative and other schools is a major cause of concern. Students can be placed in alternative schools for numerous reasons (Sagor, 1999). Despite regulation, these programs differ to an extent that makes quantitative exploration of these data difficult (Oregon Secretary of State [SOS], 2024; Oregon Laws, 2021; Sagor, 1999). In many ways, GLMM trees are well-suited for such complex circumstances (Sagor, 1999). However, observational studies, teacher panels, or other qualitative methods of knowledge generation may be more impactful to improving outcomes for these students in the short term. In the meantime, ODE, IES, and academic partners should work to make 9G-OTG an equitable metric for students at all schools.

Students, teachers, and administrators should be made aware of some differential impacts across conditions, with great importance on supporting students who change schools within a year. Greater exploration should pair qualitative follow-up and CDA approaches to assess if the effect of dual/transitory enrollment in ADM program types is a unique effect or a collinear measure of a student transferring schools.

Analyst-Focused

As publicly available data become increasingly common, SEAs and other organizations may leverage GLMM trees as means of discovery, focused through the theoretical lens of an analyst-specified parametric model. However, this project was an example of how computational

resources limit the pace of research in the implementation of GLMM trees on large datasets. On the same data estimated without random effects, the GLM tree algorithm is orders of magnitude faster than a similar GLMM tree, suggesting the estimation of random effects may be a bottleneck in this process. Nonetheless, increasing computational efficiency of the algorithm may facilitate uptake of the model.

In addition to efficiency in finding an optimal solution, large datasets require specialized solutions in software like R which retains most data in RAM (R Core Team, 2024). In this project, large differences in computation existed across the number of variables in the data. Rather than failing, the algorithm ran for more than 48 hours, only returning three hyperparameter states (equating to 15 successful models or 3 indexes across 5-fold CV). Although SLDS and joined data sets are large compared to other social science projects, the dataset itself just more than half a gigabyte. Regardless, 100% of RAM (32 GB) was used processing the data in parallel, forcing partially processed data to be written to physical memory swaps (max ~49 GB) and caches (~max 7 GB). The process was terminated after reading and writing 2.75 *terabytes* of data to swaps and caches (i.e., filling the swap/cache roughly 50 times) without returning the fourth hyperparameter index. This highlights the computational complexity of GLMM trees, and the extensive number of tests conducted to converge on a final model, and underscores the value add of increasing computational efficiency of the algorithm as dissemination increases as well as increasing available resources for computation even in social science.

Summative Conclusion

As individual components, *Paper 1* and *Paper 2* made small methodological contributions to subfields of machine learning and inferential statistics surrounding the relatively new generalized linear mixed-effects model (GLMM) tree (Fokkema et al., 2018). With methodological validity demonstrated, *Paper 3* applied GLMM trees to make (similarly small) contributions to the literatures of early warning indicators (EWIs), on-time graduation, and state longitudinal data systems (SLDS).

Primary Contributions

Paper 1 began with theoretical knowledge about model optimization and hyperparameter tuning before testing if those findings generalize to GLMM trees. Through simulation, a means of appraisal, best practice in applying GLMM trees, a method of discovery, was clarified. *Paper 1* elucidated the role of hyperparameter optimization (HPO) in GLMM trees. HPO-tuned trees performed equivalently with untuned trees but were able to do so with fewer terminal nodes. The magnitude of difference varied, but some tuned models made equivalent predictions on testing data with a single subgroup, instead of five, ten, or even fifteen subgroups made from the same data. *Paper 1* leaned into the ML aspects of GLMM trees and compared HPO approaches to optimize predictions of unseen data.

Paper 2, on the other hand, focused on the GLMM trees' inferential roots. *Paper 2* critically engaged with the literature of clustered effects in social science and asserted GLMM trees' cluster robust corrections to instability statistics might be vestigial. *Paper 2* simulates data to demonstrate the effect and provides the first confirmation that the method can perform well in cross-classified spaces. Comparing the results of *Paper 2* to the highly similar Fokkema and Zeileies (preprint), models with random intercepts only were in harmony but differences were

found in conditions with a random slope—something not tested in *Paper 2*. Finally, in hopes of increasing use of GLMM trees in extremely high dimensional spaces, *Paper 2* demonstrated the baseline preference of influential-variable-selection across levels when the magnitudes were equivalent. By simulating data with equivalent underlying effects, *Paper 2* was the first to probe the performance of GLMM trees in spaces where variables may exert similar magnitude influence on the parametric model across the levels whether real or a byproduct of sampling variation. The default specification of algorithm initialization and clustered adjustments performed the closest to chance over the simulations, whereas other specifications showed more bias. Results suggest selection between identical effects is somewhat sample size dependent and the default settings for algorithm initialization and cluster-robust corrections is least likely to over-select based on sample size alone (i.e., preferring variables at lower-level to higher-level units).

Paper 1 established a protocol to optimize predictive accuracy and identify parsimonious ways to express variation in the parametric model. *Paper 2* clarified the best specification for dual- versus singular-correction with SLDS data. Both of which were needed to ensure valid results were extracted from a high-stakes and novel implementation of GLMM trees. Regressing on-time graduation, a dichotomous predictor, on 9G-OTG (a binary EWI), *Paper 3* used GLMM trees to identify circumstances which were associated with the largest total disruption to the parametric model, controlling for school- and zip-code-level variation in intercepts. A school-level factor, alternative school status, was associated with the largest total variation in the aforementioned 2 x 2 contingency table (on-/off-track x graduation/no graduation). Alternative school students displayed the lowest odds ratio of graduating when on-track across all conditions. As of now, it is unclear if the differences Fokkema and Zeileis (2023) report are from

differences in random effects (random slopes vs. fixed slopes), fixed effects (growth models with only few waves and an identity link function vs. students in schools), sample size, or something else. However, harmony between their random intercept models and *Paper 2* despite differences in all past listed factors suggests the differences may be in complexity of random effects.

By interweaving these disparate literatures, though, the entirety of the dissertation provided a greater impact than the sum of the components independently. The dissertation exposes educational researchers to a novel EDA method which is unique in its ability to explore within a targeted theoretical framework. At the same time, the dissertation provides new evidence of how to improve the model (i.e., tuning procedures and specification), enabling those learning about the model through this avenue to see (still emerging) best practice.

Furthermore, the paper is the first verification of GLMM trees successfully identifying most influential splitting variables across multiple levels at once, whether nested or cross-classified. The application demonstrates a robustness of the state's EWI to more than 1500 covariates and finds invariance only along variables either previously reported by ODE (ODE, n.d.-c, 2019; 2018-b, the UO (Scalise et al., 2023; Zvoch et al., 2023; Farley et al., 2021), or others (Zheng et al., 2023) to be associated with 9G-OTG and/or graduation. By identifying theoretically driven covariates, the model reaffirms the grasp the state has on their students and their needs. However, the findings shed new light on how factors might compound to the detriment of student success, thanks to the branching structure of GLMM trees.

Using predicted probabilities by subgroup, I reported segmented variations in the entire process underlying 9G-OTG, graduation, and their interrelation among two cohorts of the Oregon SLDS with over 1500 variables measured across four clustering levels (student, school, district, and zip code). Such findings differ from, for example, a linear interaction by partitioning

observations based on the entire parametric model. As a result, all model-identified information ensures a targeted theoretical focus whereby both on- and off-track students influence which variable the model deems most influential. Furthermore, these findings allow a unique approach that allows differential impact to be assessed across multiple clustering levels at once (e.g., student-, school-, and zip-code-level).

Recursive Emergence

Unfortunately, two variables—student mobility and ≥2 ADM codes—which emerged were relatively crude metrics of multiple similar but unique circumstances. Complicating circumstances, these crude metrics emerged recursively on both sides of the primary split. As the statistical and ML adage goes "*garbage in, garbage out*," and recommendations should avoid such an issue via a granular, iterative investigation of the relation (i.e., measured circumstances changing the parametric model) and this statistical approximation of the underlying processes. Perhaps these results will be interpretable directly by an expert of the ADM system and student mobility at ODE, but from a data-analytic perspective, data available in codebooks and the gap between codebooks and practice limit recommendations about these subgroups without greater understanding of data and measured contexts.

First, quality of the ≥2 ADM code measure needs to be verified, completely understood, and evaluated in more depth. For example, data handbooks do not provide exhaustive lists by which students can have multiple ADM programs (Gill, 2023). With multiple ADM programs and some students enrolled in more than two enrollment codes, a more nuanced exploration of these students' circumstances is required. Dummy coding ADM program type could reveal additional patterns if explored with additional MOB approaches or theory-driven means of appraisal. Such a problem is inevitable in a big data approach, but SEAs should prioritize

verifying and improving veracity and granularity in measurement of the most heterogenous effects.

Similarly, convergent circumstances can lead a student to be coded as "no" on the full academic year flag including entering a school more that 10 days after the first day, transitioning to/from another district, and extended gaps in enrollment (Gill, 2023). In ODE publications, this variable is often described as "student mobility" (e.g., ODE, n.d.-c). The ways in which identified students are getting knocked off track—even across broadly different school circumstances—needs to be evaluated with *post-hoc* tests, theory, and methods of appraisal. Regardless of the findings, these results suggest that non-traditional students require additional supports, roughly corresponding to students which transfer schools (i.e., student mobility) and students enrolled in multiple ADM programs. In most cases, previous research has documented the same or similar variables being associated with graduation directly or in combination with another variable in the same sample (ODE, n.d.-c; 2019; 2018-b; 2018-a) as well as in meta-analysis (Zheng et al., 2023).

Standard Public and Charter Schools

Of effects which were not recursively emergent, effects in standard public-school students echo past research with students outside of Oregon, with the SDLS directly, or both (ODE, n.d.-c; 2019; 2018-b; 2018-a; Zheng et al., 2023; Balfanz & Byrnes, 2019; Bowers et al., 2012; Scalise et al., 2023; Zvoch et al., 2023). ODE's efforts to optimize student success are clear with how extensively past reports have covered these and adjacent variables in the past (ODE, n.d.-c; 2019; 2018-b; 2018-a), despite *Paper 3* recursively assessing invariance along more than 1,500 variables. In addition to granularization of two broad metrics identified

recursively across school type—student mobility and ADM enrollment program status (1 vs ≥2)—parametric invariance was found in attendance and zip-code-level employment.

Zip-Code Unemployment

Neighborhood-level factors have been associated with high school graduation, including with economic and employment measures (Vartanian and Gleason 1999; Wodtke et al., 2011; Ensminger et al., 1996). Many factors, including the one outlined in *Paper 3*—percentage of zip code unemployed (16 years and older)—are not within the control of ODE directly. In *Paper 3*, the influence of zip-code unemployment on the parametric model emerged as influential within subgroups defined by several splits. Zip code unemployment became important among a relatively small subgroup (n = 237) which had very good attendance (≥88.5%) but were enrolled in more than one ADM program in a given year. When neighborhood unemployment was below ~38%, students in this subgroup displayed third and fourth highest probability of graduating when off- and on-track, respectively. Analogous student which lived in neighborhoods with unemployment >38%, by contrast, had the second lowest probability of graduating when off-track, the third lowest probability of graduating when on-track, and the least precision in estimate of graduating when off-track. Neighborhood unemployment is not an outcome within the control of ODE. Therefore, ODE should attempt to parse a mechanism of action by which neighborhood unemployment knocks students off track. Only in doing so will ODE be able to clarify what malleable targets are within their control. Considering the high attendance rate among these students (>88.5%), their engagement appears to be high, meaning practical circumstances may take their focus off on-time graduation. One possibility that could be tested with qualitative forms of appraisal is that these students may leave work to get jobs that support their immediate or indirect family. Prior to embarking on processes of extensive data-collection, 5-year

graduation rates should be assessed for this group to explore if these students still graduate but at a slower pace.

Attendance Data

Based on the findings of Allensworth and Easton (2005; 2007) and others (Bowers et al., 2012)—including Oregon SLDS analyses (ODE, n.d.-c; ODE, 2017)—invariance by attendance is among the most expected findings which could emerge from the model. Allensworth and Easton explain that individually or in combination attendance, failures, credits earned, and grade point average (GPA) are strong predictors of graduation, and their selection was because they "*believed that each contained important information relevant to Chicago Public Schools (CPS) policy about grade promotion"* (2005, p.2).

Attendance data has been explored among SLDS students by ODE, with data-driven thresholds identified at 81% attendance (ODE, n.d.-c) and theoretically coded models setting the threshold at 90% (ODE, 2017). Tuned GLMM trees returned nearly identical thresholds which segmented students at 82.6% and 88.9%. Considering how unexpected the findings, the consistency of findings across approach, and the pre-existing body of literature, ODE should seriously consider utilizing some aspect of attendance in combination with course failure to enhance the 9G-OTG metric.

Extending Oregon's Early Warning System

The recommendations in this section are to students outside of the alternative school system (i.e., standard public and charter school students), and recommendations of modifying EWIs in alternative schools are not considered without further discussion with stakeholders and research (discussed below). The most direct extension is a model-based prediction of graduation by subgroup, but such a solution is problematic for operational consistency and longitudinal

research. Furthermore, model-based predictions are not consistent with past research on EWIs supporting the validity of face-valid metrics in this context discussed throughout the *Context* and *Paper 3* (e.g., Allensworth and Easton 2005; Bowers et al., 2012). A better solution is to make a still face-valid and interpretable metric that is ordinal instead of dichotomous. Risk profiles can be established whereby a student is not dichotomously on-track or off-track but is assigned as a face-valid metric with differential recommendations by group. For example, past findings have returned interaction of the current metric (9G-OTG = completion of 25% of courses by 9th grade) and binned attendance scored as a grade ("below B- [0%, 83%")", "B [83%, 90%)", and "A [90, 100%]). These risk profiles are shown in Table 1 with relative risks ordered based on the results of *Paper 3*. Such a system could be investigated retrospectively and be retained in longitudinal research.

Table 1

Possible ordinal EWI based on percentage course completion and binned attendance; recommendations are given by cell in parenthesis.

		Binned "Attendance Grade"		
		A [90%, 100%]	B [83%, 90%)	B- or Worse [0%, 83%)
Completed 25% of Core Courses at End of 9th Grade	Yes	Graduation Highly Probable (No recommendations)	Graduation Probable (Monitor Attendance)	Graduation Probable (Monitor Attendance)
	No	Graduation Probable (Provide Academic Supports)	Concerning Graduation Probability (Dual Supports)	Extremely Concerning Graduation Probability (Dual Supports)

The new system suggests a 4-level ordinal metric with tiered risk categories by color, with greater distinction among those least probable to graduate (Table 1). In fact, such a metric could be re-evaluated by term or semester, making the metric not a 9G-OTG, but an OTG metric continuously assessing probability of graduating throughout high school. Conaway et al. (2015 p.

17S) writes "*[w]hat gets state education agencies (SEAs) excited about SLDSs, and the focus for the bulk of their investment and development efforts, is the opportunity to provide much more useful, timely information back to district personnel and the public.*" They continue "*[m]ost states have focused their SLDS programs on making data available to districts and the public in real time, in a format that provides insight into critical district issues*" (Conaway et al. 2015, p. 18S). Conaway et al. (2015) emphasize the value and timeliness of *yearly* reports from the SLDS. With relatively small investments in data engineering, student attendance data and course failures could be updated such that a term-by-term metric of OTG was available.

Automated notifications (e.g., via email) could be sent to the individual monitoring each student's success. In some cases, this could be as simple as emailing parents and the school counselor to increase awareness. If schools have high school success teams, such teams are the logical target to receive these notifications. Any time a student changes risk category (i.e., by failing classes or passing one of the two attendance thresholds), an email would be sent. Customized warning messages should not ignore or undervalue the psychological importance of presentation and interpretable, face-valid, and minimal explanations should be constructed with focus groups, computer scientists, and user-interface/user-experience (UI/UX) experts. Warnings need to be adapted to the stakeholder which receives the custom warning. Customized warnings should provide the following at minimum:

- The prior value of the ordinal risk variable,
- The new value of the ordinal risk variable,
- The reason for the students changes in status (i.e., recent increase in absences vs. a course failure),
- The action recommended to support the student,

- Short, expandable/minimizable explanations of why the metric of interest is important which has been vetted by stakeholder groups,
- Links to additional resources, and
- Additional stakeholder-specific information.

Policy could be put in place at state-level as well whereby one of several approved stakeholder-types must independently acknowledge receipt of the recommendation and then document conversations with the student. Exact implementation of the plan requires focus groups comprised of all stakeholder types, Oregon DOE stakeholders, researchers, and UI/UX designers. Other strategies are possible, but are of lower priority, mostly dependent upon the desire of ODE and other stakeholders.

Alternative Schools

Oregon State Law designates alternative schools as those *"designed to best serve students' educational needs and interests and assist students in achieving the academic standards of ... the state*" (US DOE, 2017) To receive public funds, schools must adhere to Private Alternative Education Standards established by the Oregon State Board of Education, meaning:

> *Each school must have a mission statement that clearly identifies the student population to be served and articulates the unique education option provided by the program; an education plan, which outlines program goals and strategies in the following areas: student equity, academic achievement, social skills development, and successful transitions to further education and training; a requirement for all staff of reference and criminal background checks and regular evaluations; and an instructional program that is aligned with state content*

Complexity in the Literature

With such broad variation in the schooling context, it is no surprise that alternative school status was the greatest contributor to instability in the parametric model. Furthermore, Oregon

Laws (2021) outline that students in alternative schools are among the highest risk population, but in extremely variable ways. Engagement with the alternative school literature highlights the complexity of the space in Oregon as well as nationally (e.g., Hadderman, 2002; Sanders, 2009; Aron, 2006; Saber, 1999; Dillow, 2003). In a US Department of Labor report, Aron (2006) explains the diversity in educational problems that a given student faces is a large reason for so much complexity. Curriculum also varies drastically among these schools with any "accountable activity" allowed, which are defined as (US DOE, 2017):

> *tutorial instruction; small group instruction; large group instruction; personal growth and development instruction; counseling and guidance; computer assisted instruction; vocational training; cooperative and/or supervised work experience; instructional activities provided by institutions accredited by the Northwest Association of Schools and Colleges; supervised community service activities performed as part of the instructional program; and supervised independent study in accordance with a student's educational goals, including classroom or equivalent work supervised by school district officials that serve as one component of the student's educational plan and profile and not the entire part. Examples of this include required and elective courses, supervised independent study, career-related learning experiences, and project-based learning. Or. Admin. Rules§581-023-0008(2).*

Therefore, alternative schools vary from other schools in at least three major school-level ways mission, curriculum, and type(s) of instructional (i.e., accountable) activity. Furthermore, as explained in *Paper 3*, students arrive for one of seven distinct reasons and an eighth formalized catch-all equating to students being recommended for placement in alternative school (Oregon Laws, 2021). Considering this variation in the sample, context, and treatment (i.e., type of material and instruction), little value is gained by evaluating these programs in aggregate without more granular data. In other words, generalizing these findings is nearly if not outright impossible with current SLDS and external data. When tested systematically, the most robust studies have found null effects between groups. The *What Works Clearinghouse* contains results from three randomized studies of alternative school effectiveness in high-risk youth from

Cincinnati, Wichita, and Stockton (Dynarski and Wood, 1997). Despite sample sizes adequate to detect group differences (*N*s = 358, 375, and 902) and randomization, none of these studies report effectiveness of the program showing null differences between groups.

Experts in alternative school education also endorse varied opinions regarding the value of alternative schools, with some suggesting the segregation of the most at-risk students being more harmful than beneficial (Sagor, 1999). Therefore, stakeholders must define optimal outcomes within and across the alternative school system to optimize an EWS for their students. Next, ODE can gather more granular and frequent data on these schools and their students. Until these are both done, quantitative work will be much less impactful than qualitative work on this sample (and should precede quantitative work to shape what is gathered).

Recommending Framework for Alternative School Students

I happened to teach at a high school for two years nearly a decade ago, but I am first and foremost a methodologist. In the years since being a high school teacher, I have shifted focus to research various adolescent risk and protective factors but have consistently worked to create better outcomes for adolescents as they transition into adulthood in a way that aligns with my skills. Many great academics have said that [*impact*] = [*depth*] x [*breadth*] of a contribution, with John Seeley being the first I heard say this phrase during my Master's training. The *depth* of contributions I can give to alternative school and other students pales in comparison to that of a well-trained, highly experienced educator or member of a SEA. Thus, I contribute to improving the lives of students by playing support along a *breadth* of possibilities and across all student groups—making sure everyone is as informed as possible, using all available information, and maximizing validity of research. Iteratively, I can help extract additional knowledge, but there are those who understand much of this already.

Despite training in prevention science and education, I do not understand this set of schools, students, and best practices adequately to make recommendations on how to change, but I can aid in directing research and resources. As *Paper 3* emphasized, the goal was *theory generation*. It is not only against the project's goal for me to tell teachers, parents, administrators, and Oregon Department of Education employees *how* to best work with these students or make recommendations but undermines their lived experience. I emphasize the inutility of falling prey to Green's (2008) fallacy of the empty vessel, meaning stakeholders' input is required. With on-site work, collaborative meetings, in-class observations, and other means of knowledge generation not available quantitatively, I could make other recommendations. However, such recommendations are outside the scope of this dissertation and cannot be made responsibly with available data. Concretely, this means telling stakeholders *where* the data suggests a problem, rather than *how* problems should be fixed.

First and foremost, stakeholders need to critically examine the purpose of 9G-OTG and an EWS broadly before choosing a future direction in general, but especially for alternative school students. From a prevention standpoint, an EWS should be early enough to intervene and accurate enough that it is predictive, yet the EWI needs to be malleable to change. However, if a metric is predictive, early, and malleable, the metric can appear to perform poorly to an evaluator, making the metric into the "boy who cried 'wolf!'" Without clarification of the state's exact desire for an EWI in aggregate and at the individual-level, recommendations on *how* to change the metric cannot be made effectively.

Therefore, my recommendations for ODE, administrators, and other alternative-school stakeholders are as follows. First, ask if four-year graduation is the goal for this population or if something else is more important (e.g., five-year graduation). If other goals are measured and not

used, re-run this analysis with the outcome of interest. If four-year graduation results are the goal—or if five-year results parallel four-year—a few effects should be explored in addition to processes identified by qualitative research.

As stated in the *Repeated Recursive Emergence* section, dual enrollment in ADM programs and student mobility need to be explored, though this exploration should not be restricted to alternative schools. In alternative schools, a unique phenomenon emerged whereby those coded as female had better odds of getting back on track than those coded as males, but on-track males had much lesser difference in the probability of graduating when on-track. Such an effect should be investigated to understand what ways schools successfully support off-track students coded as female, relative to off-track students coded as male.

Impact to GLMM trees

Paper 1

Few publications have validated GLMM trees, especially with simulated data (Jorink, 2018; Fokkema et al., 2018; Fokkema and Zeileis, preprint). *Paper 1* is the first systematic investigation of the comparative influence of untuned GLMM trees against those optimized with maximum-entropy grid search and a genetic algorithm. The findings of *Paper 1* suggest that tuning is important to GLMM trees. However, the lack of added benefit for genetic algorithms and the ability of genetic algorithms to find the optimal solution in only a few iterations demonstrates the relative stability of the model performance across many hyperparameters. Though the essential conditions to test for *Paper 3*, this finding should be verified across a range of parametric models and simulated data before being stated as a general truth about GLMM trees across varied data and models.

If the finding does not generalize, the stability in GLMM tree performance and fitness function observed in *Paper 1* is due to the type of data: a highly probable binary outcome regressed on a highly predictive (also highly probable) binary covariate. However, subgroupings of a theoretically driven parametric model may have less variance in estimated fitness functions, compared to an atheoretical, data-driven models of the same data (e.g., neural networks, random forests, etc.). Regardless, future research could assess this by estimating, visualizing, and modeling the fitness function of GLMMs across varied simulated data and hyperparameter states. Fully atheoretical frameworks could be applied to the same subgroupings and differential variance in their fitness functions across the same data could be explored. Finding that similar data are modeled with more stability (i.e., less variance in the fitness function across hyperparameter states) through GLMM trees compared to atheoretical models, for example, would increase support that minimal coverage of the hyperparameter space may be sufficient in GLMM trees. Regardless, this paper demonstrates how researchers can explore the hyperparameter space and that the optimal state may be identified with relatively few tested states.

Another fascinating insight of the model is the way tuning influences the GLMM tree, and the evidence is uniquely fitting for a hybrid of inferential and ML models. Instead of simply improving predictive accuracy—the dogma in ML (Yang & Shami, 2020; Belkin et al., 2019)—smaller models with equivalent predictive accuracy were identified after tuning, even though hyperparameters were optimized directly with classification accuracy (not size). At least for data like these, tuned GLMM trees are preferred to untuned trees. Finding that smaller models tend to predict the outcome with greater accuracy is based upon the "classical" paradigm of bias-variance tradeoff (Belkin et al., 2019).

Though seemingly common-sense that data adhere to the classical paradigm of the bias-variance tradeoff, such findings are important in an era of available data, increased tracking, and efficient computation. Emerging evidence from applied research at Google suggest the bias-variance tradeoff may not hold in scenarios complex enough to undergo interpolation (Belkin et al., 2019; Zhang et al., 2016; Zhang et al., 2021-a), but such scenarios are beyond the scope of discussion here. Briefly stated, interpolation is a concept gaining rapid popularity in ML whereby models perform equivalently when predicting (unseen) testing data and (seen) training data through very high parameterization (Belkin et al., 2019; Zhang et al., 2021-a). Interpolation is more common among extremely large, complex models such as back-propagating convolutional neural networks (Zhang et al., 2021-a; Zhang et al., 2016; Belkin et al., 2019), but can be seen in some extensions of more familiar methods, including boosted random forests and boosted penalized regression (Belkin et al., 2019). For predicting continuous or dichotomous outcomes, Belkin et al. (2019) suggests as many parameters as there are observations are required, making interpolation not possible in this dataset (Belkin et al., 2019). Even if interpolation is unlikely, such findings highlight the need for validating methods with new levels of dimensionality (cases by variables) and new modeling frameworks, instead of applying them based on our past knowledge. Regardless of being in a "classical" or the "modern" regime (as Belkin et al., 2019 call them), the optimal model would be selected through HPO. See Belkin et al. (2019), Zhang et al. (2021), Zhang et al., (2016), and others working in spaces where the number of observed variables outweighs number of observations for greater discussion of interpolation.

The existence of such new phenomena demonstrates how statisticians' expectations can break down as dimensionality of data increase. This is why studies of methodological

falsification must keep pace with applications of methodological discovery (Belkin et al., 2019). As ML models rapidly advance our ability to make discoveries, we must use simulation or other tools of appraisal to establish proper working conditions for these models, as *Paper 1* does. Furthermore, the existence of interpolation provides a new—albeit extremely high—bar for how effective a data-driven EWI could be with enough data and an adequate modeling framework.

Paper 2

Fokkema and Zeileis' preprint of their article in-preparation for publication emphasizes the timeliness and value of answering *Paper 2*'s research question in the field of model-based recursive partitioning, as well as extending their work to cross-classified designs. Despite working with very different parametric models and data structures, findings from Fokkema and Zeileis (2023) and *Paper 2* are mostly harmonious. Across the two papers, robustness to GLMM tree specification differed by circumstance, with both papers reporting relative invariance to specification if data were at level-1 and modeled random intercepts (without random slopes). *Paper 2* shows robustness of effect for cross-classified structures and nested structures, neither of which have been tested before. Subsequent studies can continue to probe the performance of GLMM trees with additional parametric models. *Paper 2* also verified the capability of GLMM trees to correctly identify the most influential splitting variables measured across nested and cross-classified levels simultaneously. This finding supports the ability of the model to be applied flexibly to diverse circumstances.

Two more robust simulations are suggested to reinforce *Paper 2* directly. First, no research—including that presented in *Paper 2*—have demonstrated the ability to recover the only global effects from GLMM trees. Random effects, the only global parameters, can be estimated from simulated data across conditions to determine if the same or other specifications of

algorithm initialization and cluster-robust corrections are preferred. Based on the iterative nature of the model, it is possible that researchers will have to preference minimizing bias in one aspect of the model over the other depending on if the random-effects structure or the splitting structure are of greatest importance. Two other improvements to *Paper 2* can either be done in the continuous space (i.e., using identity link functions) or by independently saving the simulated error terms used within the simulating function. With binary outcomes, residual variance is not directly assessable from the binomial link function. If observation-level residuals from simulating functions were saved, these can be subtracted from predicted probabilities and a pseudo metric of residual variance could be calculated for the model. This added evaluation of simulated versus recovered random effects and residual variance would drastically increase confidence in the reported findings of *Paper 2*.

Paper 3

Paper 3 offers substantially fewer methodological contributions of GLMM trees but does offer a few. In *Paper 3*, GLMM trees identified the most influential split to be alternative school status. As defined by Oregon law, alternative schools by definition have individualized education plans for each student, meaning the underlying process differs (i.e., navigating high school) for this group relative to other students. This validates the application of GLMM tree in practice with applied educational data. Tuned GLMM trees parsed a subgroup of students which are given education plans and goals that vary across the subgroup (i.e., compared to standard public) and within the subgroup (i.e., compared to other alternative schools). Such a finding lends confidence to applied researchers in the applied validity of the method, hopefully contributing to the method through dissemination. High uptake is especially likely among those evaluating educational data due to the organization of high schools in the US. Clustering can be found

across multiple levels of educational research from students in classrooms to schools in districts and even districts in states. GLMM trees can be used more frequently, be validated across additional designs, and be extended methodologically by disseminating the utility of GLMM trees for educational research.

In addition to demonstrating real-world validity, *Paper 3* had interesting results which deserve to be explored in depth. Of course, *Paper 3* results represent 1 instance of the 500 simulations used in *Paper 1* and *Paper 2*, and generalization should be taken with caution before explored explicitly with simulation or more datasets. In *Paper 3*, "islands of stability" were found whereby unique hyperparameter states appeared to make identical terminal node predictions in aggregate. More work needs to explore if these islands of stability are a byproduct of these data (e.g., binomial link function, highly probable outcome, highly predictive and probable EWI etc.), or occur in models with more variation in the sample. Furthermore, there are ways that classification accuracy, number of terminal nodes, and information criteria could be identical but models to differed slightly and were not explored. One obvious difference could be cluster-by-cluster random effects. Future work could explore if these (or models with more variant outcomes like LMMs) estimate random effects to be identical by clustering unit or only in aggregate. Besides the only global parameters—random effects—differences may exist in the threshold between two groups on a continuous variable (e.g., average attendance ratio split threshold) or may differ without a difference in terminal node predictions in aggregate on testing data (i.e., predicted dichotomous on-time graduation). Assessing a parametric model with a binomial link function may have "stabilized" terminal node predictions even if the underlying predicted probabilities varied to some extent. As of now, this suggests some models can have

islands of stability on which models converge based on relatively invariant fitness across the hyperparameter space.

A notable difference between *Paper 1* and *Paper 3* was in the relationship of model parsimony and classification accuracy. In *Paper 1*, hundreds of cases were simulated allowing a generalized additive mixed model (GAMM) to show classification accuracy decreasing slowly but continuously as model size increased over simulated data. These results partially support *Paper 3*'s finding. However, with real data, solutions with a middle-ground of complexity slightly outperformed their more and less parsimonious counterparts. In fact, the middle-ground of complexity being optimal is a demonstration of optimizing (or at least approximating) the variance explained in by model and the training bias in predictions within the "classic paradigm" of the bias-variance tradeoff (Belkin et al., 2019).

Impact to Other Evaluations

In addition to direct uptake by SEAs, MOB and GLMM trees can be incorporated into evaluations more broadly. The following sections aim to bridge the gap between applying these findings in SLDSs and other evaluative contexts in and beyond social science.

Targeted Discovery & Reporting

Incorporating exploratory methods as a step of the evaluation cycle allows more complete understanding of relationship between the variable and covariate of interest, making it a targeted tool of discovery. In a high-dimensional space, GLMM trees can explore invariance in the parametric model without losing power (Fokkema et al., 2018). Exploring invariance in the parametric model recursively allows straightforward identification of interactions which are unspecified, higher order (i.e., interactions of more than two variables), or of unknown functional form (Zeileis et al., 2008). GLMM trees collapse to the specified GLMM when there is no

evidence of systematic invariance (Fokkema et al., 2018). Therefore, evaluative contexts which are properly modeled by the GLMM can be shown empirically and reproducibly in one test, which informs the reader about the stability of the process being modeled. Specifically, evaluators and researchers can report the variable(s) tested for splitting, instability statistic(s), and associated p-value(s) from the GLMM tree.

The Cycle of Evaluation & Model Parameterization

When differences are identified in an evaluation context, the results can be classified as expected or unexpected. Expected results are those which were likely based on prior research, evaluator logic, or another form of knowledge generation. When instability in the parametric model returns an expected result, the evaluator may return to the literature and/or stakeholders to understand the correct way to specify the effect, i.e., in a way that minimizes bias in the evaluative parametric model. In some cases—like the home language/test type example in the *Context*—the underlying process suggests the GLMM tree is the best way to model the data. The process in the subgroup was fundamentally different than other groups, and retention of the subgroup decreases accuracy and precision in the primary covariate of interest (time studying on a final exam).

In other cases, for example, a GLMM tree may separate subgroups based on a large disparity in the intercept of the model alone, despite having consistent slopes with the other portion of the sample. In those cases, controlling for the intercept differences in the subgroups (e.g., with a dummy code for a categorical variable) may enable valid estimate of the effect of interest with all available cases. In this way, GLMM trees are powerful tools for evaluators, but are insufficient to determine the "optimal" modeling framework without a grasp of the theory underlying the evaluative context. Best practice dictates documentation of this process is

essential, including documentation as to why such an “expected” result was not included in the original parametric model.

Recovering unexpected findings with GLMM trees should similarly be followed by consulting the theoretical base underlying the evaluation and the stakeholder on the ground. Unexpected effects can be followed up with any additional means of knowledge generation, including qualitative, correlative, causal, and logical (i.e., extending past theory with new observation directly). In high dimensional spaces, correlative approaches should be tested to determine if the model-identified (and unexpected) association is highly predicted by a confluence of multiple circumstances (e.g., demographic differences emerging as “most important” though causal differences are from multiple discriminatory correlates). Interviews of on-the-ground stakeholders and other qualitative methods may return similar understandings. With this knowledge, theory and causal designs can identify and validate possible mechanisms of actions and build confidence in the findings.

Appendix A

Glossary of Abbreviations

Abbreviation	Full Term
9G-OTG	9th Grade on Track to Graduate
AGQ	Adaptive Gauss-Hermite Quadrature
AHRQ	Agency for Healthcare and Research Quality
AIC	Akaike Information Criteria
ACS	American Community Survey
AutoML	Automated Machine Learning
ADM	Average Daily Membership
BIC	Bayesian Information Criteria
BOBYQA	Bounded by Quadratic Approximation
CTE	Career and Technical Education
CRDC	Civil Rights Data Collection
C-ITS	Comparative Interrupted Time Series
CPU	Computer Processing Unit
ctree	Conditional Inference Tree
CDA	Confirmatory Data Analysis
CV	Cross-Validation
DOE	Department of Education
ECLS-K	Early Childhood Longitudinal Study-Kindergarten
EWI	Early warning indicator
EWS	Early Warning System
ERIC	Educational Resource Information Center
EDA	Exploratory Data Analysis
FRL	Free and Reduced-Price Lunch
GED	General Educational Development Credential
GAMM	Generalized Additive Mixed Effects Model
GLMM	Generalized Linear Mixed-Effect Model
GLMM Tree	Generalized Linear Mixed-Effects Model Tree
GLM	Generalized Linear Model
GA	Genetic Algorithm
GB	Gigabyte
GS	Grid Search
GCM	Growth Curve Model
IEP	Individualized Educational Programs
IES	Institute of Education Sciences
LRT	Likelihood Ratio Test
ML	Machine Learning

Abbreviation	Full Term
MEGS	Maximum Entropy Grid Search
MAE	Mean Absolute Error
MOB	Model-Based Recursive Partitioning
NBER	National Bureau of Economic Research
NCES	National Center for Educational Statistics
ORH	Office of Rural Health
OAR	Oregon Administrative Rule
ODE	Oregon Department of Education
OHA	Oregon Health Authority
ORS	Oregon Revised Statute
SOS	Oregon Secretary of State
PDP	Partial Dependency Plots
RAM	Random Access Memory
RE	Random Effects
RMSE	Root Mean Square Error
SDOH	Social Determinants of Health
SB	State Bill
SEA	State Educational Agency
SLDS	State Longitudinal Data System
UX	User Experience
UI	User Interface

Appendix B

Glossary of Selected Terms

Term	Simple Definition
Appraisal	Falsifiable approaches to knowledge generation.
Cluster (MOB hyperparameter)	Name of the GLMM tree hyperparameter used to determine which level (if any) should influence corrections to the
Clustered Observations	Grouped observations whereby observations within a unit or cluster are more similar to other observations within the unit, compared to across (e.g., students in school, students in a zip code).
Discovery	Approaches which can offer novel insights.
Dummy-coding	A method used in statistical modeling where categorical variables are represented as binary variables (0 or 1).
Fitness function	Function against which optimization occurs.
Fixed Effects	An effect which is constant across levels of the model; differentiated from a random effect which is not constant across levels.
Heterogeneity	Variation
Hyperparameter	Modifiable configuration which determines how an algorithm functions; values are not estimated from the model.
Hyperparameter Optimization	Systematic process of calibrating a machine learning algorithm to observed data.
Hyperparameter Space	All possible combination of hyperparameter states for a given model.
Hyperparameter State	The configuration of all hyperparameters in the model.
Hyperparameter Tuning	Process of calibrating a machine learning algorithm to observed data. A general term including HPO.
Interaction	Product of two main effects.
Knot	Curves in the relationship between the predictor and outcome.
Level-A / Level-B	Distinction used for non-nested clusters.
Level-1	The observation-level of analysis in multilevel modeling
Level-N (when N is a number greater than 1)	Level of nesting which are [n-1] levels abstracted from the observations (e.g., schools can be level-2 and districts can be level-3).
Link Function	A function used by generalized linear models (and mixed-effects extensions) to allow estimation of non-linear outcomes.
Meta-heuristic optimization	Approaches to finding the best model using a set of decision rules (a component of AutoML)
Modular	A system comprised of independent subunits which communicate.
Nested Observations	A type of clustering with only hierarchical structure and observations only belong to one group (e.g., students in schools, schools in districts).
Overfitting	Generalizing sample-specific information from training data.
p-hacking	Manipulating statistical models and/or data to achieve a statistically significant p-value.
Parameter	The underlying true value in the population measured by a statistic; an "estimated parameter" is an estimate of the true value from a sample.
Power	The ability to detect an effect when the effect is present
Ranefstart	Name of GLMM tree hyperparameter used to determine algorithm initialization.
Recursive Partitioning	Repeatedly splitting data based on criteria.
Spline	Smoothed functions used by GAMMs

Term	Simple Definition
Split Variable	A variable selected to partition the data in recursive partitioning models.
Terminal Node	A subgroup of a recursive partitioning model for which no further splits are made.
Testing Data	Data withheld from model calibration, allowing model performance to be assessed with unseen data.
Training Data	Data with predictors and outcomes used to calibrate a machine learning model.
Type I Error	A false positive.
Type II Error	A false negative.

References

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2017). When should you adjust standard errors for clustering? *The Quarterly Journal of Economics, 138*(1), 1-35.

Agency for Healthcare Research and Quality [AHRQ]. (2020). Social Determinants of Health Database. *Department of Health and Human Services*.

Aitkin, M., & Longford, N. (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society: Series A (General), 149*(1), 1-26.

Albert, J. (2022). *A course in exploratory data analysis (based on Tukey's EDA book)*. Department of Mathematics and Statistics, Bowling Green State University. Retrieved from https://bayesball.github.io/EDA/

Alexander, L. K., Lopes, B., Ricchetti-Masterson, K., and Yeatts, K. B. (2015). Common Measures in Statistics and Epidemiological Literature. *ERIC notebook.* University North Carolina Chapel Hill Department of Epidemiology. Retrieved from https://sph.unc.edu/wp-content/uploads/sites/112/2015/07/nciph_ERIC3.pdf

Alibrahim, H., & Ludwig, S. A. (2021). Hyperparameter optimization: Comparing genetic algorithm against grid search and bayesian optimization. *IEEE Congress on Evolutionary Computation* (pp. 1551-1559).

Allensworth, E. (2013). The use of ninth-grade early warning indicators to improve Chicago schools. *Journal of Education for Students Placed at Risk (JESPAR), 18*(1), 68-83.

Allensworth, E., & Easton, J. Q. (2005). *The on-track indicator as a predictor of high school graduation*. Consortium on Chicago School Research, University of Chicago.

Allensworth, E. M., & Easton, J. Q. (2007). What Matters for Staying On-Track and Graduating in Chicago Public High Schools: A Close Look at Course Grades, Failures, and Attendance in the Freshman Year (Research Report). *Consortium on Chicago School Research*.

Andrews, D. W. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica: Journal of the Econometric Society*, 821-856.

Andrews, M. (2021). *Doing data science in R: An introduction for social scientists.* SAGE Publishing. Retrieved from https://www.mjandrews.org/book/ddsr/

Arend, M. G., & Schäfer, T. (2019). Statistical power in two-level models: A tutorial based on Monte Carlo simulation. *Psychological methods, 24*(1).

Arnold, M., Voelkle, M. C., & Brandmaier, A. M. (2021). Score-guided structural equation model trees. *Frontiers in psychology*, *11*, 564403.

Atwell, M. N., Balfanz, R., Bridgeland, J., & Ingram, E. (2019). Building a Grad Nation: Progress and Challenge in Raising High School Graduation Rates. *Educational Resource and Information Center Annual Update*. Retrieved from https://files.eric.ed.gov/fulltext/ED597661.pdf

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390-412.

Balfanz, R., & Byrnes, V. (2019). Early warning indicators and intervention systems: State of the field. *Handbook of student engagement interventions*, 45-55.

Balfanz, R., Herzog, L., & MacIver, D. J. (2007). Preventing student disengagement and keeping students on the graduation path in urban middle-grades schools: Early identification and effective interventions. *Educational Psychologist, 42*(4), 223-235.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48. https://doi.org/10.18637/jss.v067.i01

Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences, 116*(32), 15849-15854.

Ben-Shachar, M. S., Lüdecke, D., & Makowski, D. (2020). effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software*, *5*(56), 2815.

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, *24*.

Bourgeois, M. (2024). 2023 graduation rates in Oregon tie for second-highest: ODE. *Koin6.* Retrieved from https://www.koin.com/news/education/oregon-2023-graduation-rates/

Bowers, A. J., & Sprott, R. (2012). Examining the multiple trajectories associated with dropping out of high school: A growth mixture model analysis. *The Journal of educational research, 105*(3), 176-195.

Bowers, A. J., Sprott, R., & Taff, S. A. (2012). Do we know who will drop out? A review of the predictors of dropping out of high school: Precision, sensitivity, and specificity. *The High School Journal*, 77-100.

Brandmaier, A. M., von Oertzen, T., McArdle, J. J., & Lindenberger, U. (2013-a). Structural equation model trees. *Psychological methods*, *18*(1), 71.

Brandmaier, A. M., Prindle, J. J., McArdle, J. J., & Lindenberger, U. (2016). Theory-guided exploration with structural equation model forests (Vol. 21, No. 4, p. 566). *American Psychological Association*.

Brandmaier, A. M., von Oertzen, T., McArdle, J. J., & Lindenberger, U. (2013-b). Exploratory data mining with structural equation model trees. *Contemporary issues in exploratory data mining in the behavioral sciences* (pp. 96-127). Routledge.

Brannen, J., & Coram, T. (Eds.). (1992). *Mixing methods: Qualitative and quantitative research (Vol. 5)*. Aldershot: Avebury.

Brown, D., de Bruin, S., de Sousa, K., Aguilar, A., Barrios, M., Chaves, N., Gómez, M., Hernández, J. C., Machida, L., Madriz, B., Mejía, P., Mercado, L., Pavón, M., Rosas, J. C., Steinke, J., Suchini, J. G., Zelaya, V., & van Etten, J. (2022). Rank-based data synthesis of common bean on-farm trials across four Central American countries. *Crop Science*.

Brownson, R. C., Shelton, R. C., Geng, E. H., & Glasgow, R. E. (2022). Revisiting concepts of evidence in implementation science. *Implementation Science, 17*(1), 26.

Box, G. E. & Luceño, A. (1997). *Statistical Control: By Monitoring and Feedback Adjustment*. John Wiley & Sons.

Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of human resources, 50*(2), 317-372.

Civil Rights Data Collection [CRDC]. (2020) *Civil Rights in Education*. US Department of Education. https://www2.ed.gov/about/offices/list/ocr/data.html

Civil Rights Data Collection (CRDC). (2024) Civil Rights in Education. *US Department of Education*. https://civilrightsdata.ed.gov/about/crdc

Clinton, C., & Reeder, B. (2017). Graduation rates in small towns exceed the statewide rate, while those in medium-sized towns catch up. *Oregon Department of Education Internal Reports*. Retrieved from https://www.oregon.gov/ode/reports-and-data/researchbriefs/pages/internalresearchbriefs.aspx

Conaway, C., Keesler, V., & Schwartz, N. (2015). What research do state education agencies really need? The promise and limitations of state longitudinal data systems. *Educational Evaluation and Policy Analysis, 37*(1_suppl), 16S-28S.

Darwin, C. (1859). On the origin of species: facsimile of the first edition. Retrieved from: https://libarch.nmu.org.ua/bitstream/handle/GenofondUA/17782/a895bf8fc586f4f5ca56cacfa00bd63f.pdf?sequence=1

Dynarski, M., & Wood., R. (1997). Helping high-risk youth: Results from the alternative school demonstration program. *Mathematica Policy Research, Inc.*

Education Resources Information Center [ERIC] (2019). 50-State Comparison. U.S. Department of Education. https://files.eric.ed.gov/fulltext/ED607481.pdf

Ekstrom, R. B., Goertz, M. E., Pollack, J. M., & Rock, D. A. (1986). Who drops out of high school and why? Findings from a national study. *Teachers college record*, *87*(3), 356-373.

Elshawi, R., Maher, M., & Sakr, S. (2019). Automated machine learning: State-of-the-art and open challenges. *arXiv*.

Elman, C., Gerring, J., & Mahoney, J. (Eds.). (2020). *The production of knowledge: Enhancing progress in social science*. Cambridge University Press.

Ensminger, M. E., Lamkin, R. P., & Jacobson, N. (1996). School leaving: A longitudinal perspective including neighborhood effects. *Child development*, *67*(5), 2400-2416.

Falkner, S., Klein, A., & Hutter, F. (2018). BOHB: Robust and efficient hyperparameter optimization at scale. *International conference on machine learning* (pp. 1437-1446). PMLR.

Farley, D., Messer, J., Scalies, K., Tate, B., and Zvoch, K. (2021). A Study in Equity: Oregon's 9th Grade Transition. *Institute of Education Sciences Funding Opportunities*. https://ies.ed.gov/funding/grantsearch/details.asp?ID=4533

Fielding, A., & Goldstein, H. (2006). Cross-classified and multiple membership structures in multilevel models: An introduction and review.

Fokkema, M., & Zeileis, A. (2023). Subgroup detection in linear growth curve models with generalized linear mixed model (GLMM) trees . *arXiv*, *2309*(05862).

Fokkema, M. (n.d.). gamtree: Generalized additive model (GAM) trees. *GitHub*. https://github.com/marjoleinF/gamtree

Fokkema, M., Edbrooke-Childs, J., & Wolpert, M. (2021). Generalized linear mixed-model (GLMM) trees: A flexible decision-tree method for multilevel and longitudinal data. *Psychotherapy Research*, *31*(3), 329-341.

Fokkema, M., Smits, N., Zeileis, A., Hothorn, T., & Kelderman, H. (2018). Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. *Behavior Research Methods*, *50*(5), 2016-2034. https://doi.org/10.3758/s13428-017-0971-x

Fong, D. W., Kane, T. C., & Culver, D. C. (1995). Vestigialization and loss of nonfunctional characters. *Annual Review of Ecology and Systematics*, *26*(1), 249-268.

Freudenberg, N., & Ruglis, J. (2007). Reframing school dropout as a public health issue. *Preventing Chronic Disease, 4*(4), A107.

Frick, H., Chow, F., Kuhn, M., Mahoney, M., Silge, J., & Wickham, H. (2024). rsample: General Resampling Infrastructure. https://rsample.tidymodels.org

Gerring, J. (2011). Social science methodology: A unified framework (2nd ed.). Cambridge University Press.

Gilbert, S. & Devereaux, M.S. (1993). Leaving school: Results from a national survey comparing school leavers and high school graduates 18 to 20 years of age. *Statistics Canada*, *Policy Commons*, CS81-575/1993E-PDF

Gill, C. (2023). Oregon Student Membership Manual for 2022-23 School Year. *Oregon Department of Education*. https://www.oregon.gov/ode/reports-and-data/students/Documents/studentmembershipmanual2022-23.pdf

Google Corporation (2019). Google Ngram Viewer [Search: *Implementation Science*. Years: *1999 to 2019*]. Retrieved from http://books.google.com/ngrams May 2024.

Google Corporation (2024). Google Scholar [citation count]. Retrieved from https://scholar.google.com.

Goldstein, H. (1986). Multilevel Mixed Linear Model Analysis Using Iterative Generalized Least Squares. *Biometrika, 73*(1), 43-56

Greene, J. P. (2001). High School Graduation Rates in the United States. Revised. *Education Resources Information Center [ERIC]*. Retrieved from http://files.eric.ed.gov/fulltext/ED466523.pdf

Green, L. W. (2008). Making research relevant: if it is an evidence-based practice, where's the practice-based evidence? *Family practice, 25*(suppl_1).

Greenwell, B. M. (2017). pdp: An R package for constructing partial dependence plots. *R J.*, *9*(1), 421.

Heckman, J. J., & LaFontaine, P. A. (2010). The American high school graduation rate: Trends and levels. *The review of economics and statistics*, *92*(2), 244-262.

Karapetyan, S., Zeileis, A., Henriksen, A., & Hapfelmeier, A. (2023). Tree models for assessing covariate-dependent method agreemen . *arXiv*, *2306*(04456).

Hartwig, F., & Dearing, B. E. (1979). *Exploratory data analysis (No. 16).* SAGE publising.

Hajjem, A., Bellavance, F., & Larocque, D. (2011). Mixed effects regression trees for clustered data. *Statistics & probability letters, 81*(4), 451-459.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. PLoS biology, 13(3), e1002106.

Hertel, L., Baldi, P., & Gillen, D. L. (2022). Reproducible hyperparameter optimization. *Journal of Computational and Graphical Statistics, 31*(1), 84-99.

Hjort, N. L., & Koning, A. (2002). Tests for constancy of model parameters over time. *Journal of Nonparametric Statistics, 14*(1-2), 113-132.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, *15*(3), 651-674.

Hothorn, T., & Zeileis, A. (2015). partykit: A modular toolkit for recursive partitioning in R. *Journal of Machine Learning Research*, *16*, 3905-3909. https://jmlr.org/papers/v16/hothorn15a.html

Huber, C., Benda, N., & Friede, T. (2022). Subgroup identification in individual participant data meta-analysis using model-based recursive partitioning. *Advances in Data Analysis and Classification*, 16(3), 797-815.

Institute of Education Sciences [IES]. (2024). History of the SLDS Grant Program. *U.S. Department of Education, Institute of Education Sciences.* https://nces.ed.gov/programs/slds/pdf/History_of_the_SLDS_Grant_Program_Oct2023.pdf

Jee, C. (2021). *Modularization in Software Engineering*. Medium. Retrieved from https://medium.com/@caitlinjeespn/modularization-in-software-engineering-1af52807ceed

Johnson, M. D., Hsieh, W. W., Cannon, A. J., Davidson, A., & Bédard, F. (2016). Crop yield forecasting on the Canadian Prairies by remotely sensed vegetation indices and machine learning methods. *Agricultural and Forest Meteorology*, 218, 74-84.

Jones, P. J., Mair, P., Simon, T., & Zeileis, A. (2020). Network trees: A method for recursively partitioning covariance structures. *Psychometrika*, *85*(4), 926-945.

Jorink, M. (2018). Recursive partitioning of growth curve models with generalised linear mixed-effects regression trees. (Master's thesis). Leiden University.

Kern, C., Li, Y., & Wang, L. (2021). Boosted kernel weighting–using statistical learning to improve inference from nonprobability samples. *Journal of Survey Statistics and Methodology*, 9(5), 1088-1113.

Krueger, P. M., Tran, M. K., Hummer, R. A., & Chang, V. W. (2015). Mortality attributable to low levels of education in the United States. *PLoS ONE, 10*(7), e0131809. doi: 10.1371/journal.pone.0131809

Kuenzi, J. J., & Zota, R. R. (2023). The Education Sciences Reform Act (ESRA): A Primer. CRS Report R47481, Version 2. *Congressional Research Service*. https://sgp.fas.org/crs/misc/R47481.pdf

Kuhn, M., & Johnson, K. (2019). *Feature engineering and selection: A practical approach for predictive models*. Taylor & Francis. Retrieved from https://bookdown.org/max/FES/

Ker, H. W. (2014). Application of hierarchical linear models/linear mixed-effects models in school effectiveness research. *Universal Journal of Educational Research*, *2*(2), 173-180.

Kuhn, M., & Silge, J. (2022). Tidy modeling with R. "O'Reilly Media, Inc.". https://www.tmwr.org/

Lakens, D., & Caldwell, A. R. (2021). Simulation-based power analysis for factorial analysis of variance designs. *Advances in Methods and Practices in Psychological Science, 4*(1), 2515245920951503.

Lafollette, J. (2024). Oregon, along with Bend-La Pine Schools, saw increased graduation rates in 2023. *Source Weekly.* Retrieved from https://www.bendsource.com/news/oregon-along-with-bend-la-pine-schools-saw-increased-graduation-rates-in-2023-20570522

Lang, M. N., Schlosser, L., Hothorn, T., Mayr, G. J., Stauffer, R., & Zeileis, A. (2020). Circular regression trees and forests with an application to probabilistic wind direction forecasting. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *69*(5), 1357-1374.

Lin, X., & Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *61*(2), 381-400.

Lee, J. C., & Staff, J. (2007). When work matters: The varying impact of work intensity on high school dropout. *Sociology of Education*, *80*(2), 158-178.

Loan, C. M. (2023). {gardenr}: hyperparameter tuning tools for generalized linear and generalized linear mixed-effects model trees. https://chhr1s.github.io/gardenr/articles/intro-gardenr.html#compare-this-to-a-tree-with-default-hyperparameters

Luo, G. (2016). A review of automatic selection methods for machine learning algorithms and hyper-parameter values. *Network Modeling Analysis in Health Informatics and Bioinformatics*, *5*, 1-16.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). *Is psychology suffering from a replication crisis? What does "failure to replicate" really mean?* American Psychologist, 70(6), 487.

Miller, E. (2023). Oregon's graduation rate went up last year. Here's some of what's working. *Jefferson Public Radio*. Retrieved from https://www.ijpr.org/education/2023-02-06/oregons-graduation-rate-went-up-last-year-heres-some-of-whats-working

Moulton, B. R. (1986). Random group effects and the precision of regression estimates. *Journal of Econometrics, 32*(1), 385-397.

Moulton, B. (1990). An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units. *The Review of Economics and Statistics, 72*(2), 334-338.

Nahhas, R. W. (2023). Introduction to Regression Methods for Public Health Using R. Dayton: Creative Common. Retrieved from https://www.bookdown.org/rwnahhas/RMPH/blr-interp.html

National Bureau of Economic Research [NBER]. (2024). *Standards of Conduct*. Retrieved from *https://www.nber.org/about-nber/standards-conduct*

National Center for Education Statistics [NCES]. (2007-a). Application for grants under the Statewide Longitudinal Data Systems: Oregon. U.S. Department of Education. U.S. *Department of Education, Institute of Education Sciences*. https://nces.ed.gov/programs/slds/pdf/Oregon2007.pdf

National Center for Education Statistics [NCES]. (2007-b). Digest of Education Statistics: Table 100. High school graduates, by sex and control of school: Selected years, 1869-70 through 2007-08. https://nces.ed.gov/programs/digest/d07/tables/dt07_100.asp

National Center for Education Statistics [NCES]. (2020). American Community Survey (ACS). https://nces.ed.gov/programs/edge/TableViewer/acsProfile/2020

National Center for Education Statistics [NCES]. (2023). Leveraging a State-of-the-Art Statewide Longitudinal Data System to Improve Education and Workforce Outcomes in New Mexico. *U.S. Department of Education, Institute of Education Sciences*. https://nces.ed.gov/programs/slds/state.asp?stateabbr=NM

National Center for Education Statistics [NCES]. (2022). Public high school graduation rates. Condition of Education. *U.S. Department of Education, Institute of Education Sciences*. https://nces.ed.gov/programs/coe/indicator/coi.

National Center for Education Statistics [NCES]. (n.d.-a). Fast facts: Historical reports. U.S. Department of Education. U.S. Department of Education, Institute of Education Sciences. https://nces.ed.gov/fastfacts/display.asp?id=932

National Center for Education Statistics [NCES]. (n.d.-b). Statewide Longitudinal Data Systems Grant Program: About the SLDS Grant Program. *U.S. Department of Education, Institute of Education Sciences*. https://nces.ed.gov/programs/slds/about_SLDS.asp

National Center for Education Statistics [NCES]. (n.d.-c). Statewide Longitudinal Data Systems Grant Program: Grantee states. U.S. Department of Education, Institute of Education Sciences. https://nces.ed.gov/programs/slds/stateinfo.asp

National Center for Education Statistics [NCES]. (n.d.-d). American Community Survey – Education Tabulation (ACS-ED). *U.S. Department of Education, Institute of Education* https://nces.ed.gov/programs/edge/Demographic/ACS

Neal, B., Mittal, S., Baratin, A., Tantia, V., Scicluna, M., Lacoste-Julien, S., & Mitliagkas, I. (2018). A modern take on the bias-variance tradeoff in neural networks. arXiv preprint *arXiv*:1810.08591.

Nguyen, D., Gupta, S., Rana, S., Shilton, A., & Venkatesh, S. (2020). Bayesian optimization for categorical and category-specific continuous inputs. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 04, pp. 5256-5263).

Oregon Department of Education [ODE]. (n.d.-a). High School Success. *Oregon Department of Education, State of Oregon*. https://www.oregon.gov/ode/students-and-family/graduationimprovement/pages/hss.aspx

Oregon Department of Education [ODE]. (n.d.-c). Predictors of High School Graduation. State of Oregon, Oregon Department of Education. https://www.oregon.gov/ode/students-and-family/GraduationImprovement/Documents/PredictorsofHSGraduation.pdf

Oregon Department of Education [ODE]. (n.d.-d). Graduation in Oregon: Critical Elements Leading to Positive Outcome shttps://www.oregon.gov/ode/students-and-family/GraduationImprovement/Documents/Graduation%20in%20Oregon%20Report%20FINAL.pdf

Oregon Department of Education. (2018-a). Data brief: Freshman on-track and freshman attendance as predictors of sophomore dropout status. State of Oregon, Oregon Department of Education. https://www.oregon.gov/ode/reports-and-data/Documents/databrief_ontrack_dropout.pdf

Oregon Department of Education. (2018-b). Data brief: On-track status as a predictor of graduation. State of Oregon, Oregon Department of Education. https://www.oregon.gov/ode/reports-and-data/Documents/databrief_ontrack_yr4_v3.pdf

Oregon Department of Education [ODE]. (2017). Data brief: Freshman on-track as a predictor of junior year achievement and outcomes. State of Oregon, Oregon Department of Education. https://www.oregon.gov/ode/reports-and-data/Documents/databrief_FreshmanOnTrackFollowup.pdf

Oregon Department of Education [ODE]. (2019). Data Request Specification: UO Data Request: 9G-OTG Study. *State of Oregon, Oregon Department of Education*.

Oregon Department of Education [ODE]. (2024-a). Graduation Reports: Media files, summary and trend reports, policy and technical manuals. State of Oregon, Oregon Department of Education. https://www.oregon.gov/ode/reports-and-data/students/Pages/Cohort-Graduation-Rate.aspx

Oregon Department of Education [ODE]. (2024-b). Institution Lookup Table. State of Oregon, Oregon Department of Education. https://www.ode.state.or.us/instID/

Marsh, Gomberg, Anderson, Hayden, Courtney, Girod, Golden, Alsonso, Leon, Hieb, Levy, Sosa, & Wilde. (2022). House Bill 4026. 81st *Oregon Legislative Assembly*. Retrieved from https://olis.oregonlegislature.gov/liz/2022R1/Downloads/MeasureDocument/HB4026/Enrolled

Office of Rural Health (ORH), [n.d.]. ORH Urban/Rural Definition. *Oregon Health Authority (OHA).* https://www.oregon.gov/oha/HSD/AMHPAC/Documents/OR-Zip-Codes-Urban-Rural-Designations.pdf

Oregon Secretary of State [SOS]. (2024). Standards for Public Elementary and Secondary Schools. *State of Oregon. Chapter 581, Division 22.* https://secure.sos.state.or.us/oard/viewSingleRule.action?ruleVrsnRsn=290555

Oregon Senate Committee on Education. (2021). Senate Bill 744. 81st *Oregon Legislative Assembly.* Retrieved from *https://olis.oregonlegislature.gov/liz/2021R1/Downloads/MeasureDocument/SB744/Enrolled*

Oregon Laws. (2021). OAR 581-022-2505. Alternative Education Programs. https://oregon.public.law/rules/oar_581-022-2505

Petticrew, M., & Roberts, H. (2003). Evidence, hierarchies, and typologies: horses for courses. *Journal of epidemiology & community health*, *57*(7), 527-529.

Powell, M. J. (2009). The BOBYQA algorithm for bound constrained optimization without derivatives. *Cambridge NA Report NA2009/06, University of Cambridge, Cambridge*, *26*, 26-46.

Quan, Z., Wang, Z., Gan, G., & Valdez, E. A. (2020). Hybrid tree-based models for insurance claims . *arXiv*, *2006*(05617).

R Core Team (2024). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. https://www.R-project.org/.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological bulletin*, *86*(3), 638.

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *arXiv, 1706(05098)*.

Rusch, T., Lee, I., Hornik, K., Jank, W., & Zeileis, A. (2013). Influencing elections with statistics: Targeting voters with logistic regression trees. *The Annals of Applied Statistics*, 1612-1639.

Sagor, R. (1999). Equity and excellence in public schools: the role of the alternative school. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas, 73*(2), 72-75.

Scalise, K., Zvoch, K., Loan, C. M., & Guha, A. (2023). Using student- and school-level characteristics to predict on-time graduation: An investigation of the 2013-14 and 2014-15 freshman cohort. *Internal Report for IES*. Available upon request.

Schwartz, B. (2004). *The paradox of choice: Why more is less.* HarperCollins Publishers.

Scrucca, L. (2013). GA: A package for genetic algorithms in R. *Journal of Statistical Software, 53*, 1-37.

Sela, R. J., & Simonoff, J. S. (2012). RE-EM trees: a data mining approach for longitudinal and clustered data. *Machine learning, 86*, 169-207.

Shirilla, P., Solid, C., & Graham, S. E. (2022). The benefits of longitudinal data and multilevel modeling to measure change in adventure education research. *Journal of Experiential Education*, *45*(1), 88-109.

Shrout, P. E., & Rodgers, J. L. (2018). *Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis*. Annual review of psychology, 69, 487-510. https://doi.org/10.1146/annurev-psych-122216-011845

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. Psychometrika, 80(2), 289-316. https://doi.org/10.1007/s11336-013-9388-3

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods*, *14*(4), 323.

Tamborini, C. R., Kim, C., & Sakamoto, A. (2015). Education and lifetime earnings in the United States. *Demography, 52*(4), 1383-1407. https://doi.org/10.1007/s13524-015-0429-3

Therneau, T. (2015). A package for survival analysis in S. *R package version*, *2*(7), 2014.

Tiendrébéogo, S., Some, B., Kouanda, S., & Dossou-Gbété, S. (2019). Survival analysis of data of HIV infected persons receiving antiretroviral therapy using a model-based binary tree approach. *Journal of Mathematics and Statistics*, 15(1). 354-365. https://doi.org/10.3844/jmssp.2019.354.365

Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 2). Addison-Wesley.

United States [US] Census Bureau (2018). 2013-17 American Community Survey 5-Year (ACS-5). Retrieved from *tidycensus* R package by Walker & Herman (2024).

US Department of Education [DOE]. (2017). Oregon State Regulations. US Department of Education. https://www2.ed.gov/about/inits/ed/non-public-education/regulation-map/oregon.html

Vartanian, T. P., & Gleason, P. M. (1999). Do neighborhood conditions affect high school dropout and college graduation rates? *The Journal of Socio-Economics*, *28*(1), 21-41.

Vaughn, M. G., Salas-Wright, C. P., & Maynard, B. R. (2014). Dropping out of school and chronic disease in the United States. *Journal of Public Health*, *22*, 265-270. https://doi.org/10.1007/s10389-014-0615-x

Walker, K. & Herman, M. (2024). tidycensus: Load US Census Boundary and Attribute Data as 'tidyverse' and 'sf'-Ready Data Frames. R Package. https://walker-data.com/tidycensus/

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., & Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686.

Whitson, B. (2017). American Indian / Alaskan Native Students in Oregon: A Review of Key Indicators. *Oregon Department of Education*. Retrieved from https://www.oregon.gov/ode/reports-and-data/researchbriefs/Documents/Internal/American_Indian_Alaska_Native_Students_In_Oregon.pdf

Wodtke, G. T., Harding, D. J., & Elwert, F. (2011). Neighborhood effects in temporal perspective: The impact of long-term exposure to concentrated disadvantage on high school graduation. *American sociological review*, *76*(5), 713-736.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *65*(1), 95-114.

Wood, S. N. (2017). Generalized additive models: An introduction with R (2nd ed.). Chapman and Hall/CRCC. https://doi.org/10.1201/9781315370279

Wood, S. N., Pya, N., & Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, *111*(516), 1548-1563.

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, 415, 295-316.

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, *17*(2), 492-514.

Zeileis, A., & Hornik, K. (2007). Generalized M-fluctuation tests for parameter instability. *Statistica Neerlandica*, *61*(4), 488-508.

Zeileis, A., Köll, S., & Graham, N. (2020). Various versatile variances: an object-oriented implementation of clustered covariances in R. *Journal of Statistical Software*, *95*, 1-36. https://doi.org/10.18637/jss.v095.i01

Zagoruyko, S., & Komodakis, N. (2016). Wide residual networks. *arXiv*:1605.07146. https://doi.org/10.48550/arXiv.1810.08591

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021-a). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3), 107-115.

Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2021-b). Dive into deep learning. *arXiv.* https://doi.org/10.48550/arXiv.2106.11342

Zheng, Y., Gao, X., Shen, J., Johnson, M. R., & Y. Krenn, H. (2023). A Meta-Analysis of the Predictors of On-time High School Graduation in the United States. *NASSP Bulletin*, *107*(2), 130-155.

Zvoch, K., Loan, C. M., & Scalise, K. (2023, October). Identifying implementation and subgroup effects with mixed effects trees in a statewide intervention. Paper presented at the Annual Meeting of the American Educational Research Association, Indianapolis, IN.