8-2017

A Characterization of a Value Added Model and a New Multi-Stage Model For Estimating Teacher Effects Within Small School Systems

Julie M. Garai
University of Nebraska-Lincoln, garai.julie@gmail.com

Follow this and additional works at: http://digitalcommons.unl.edu/statisticsdiss

Part of the Statistical Models Commons

http://digitalcommons.unl.edu/statisticsdiss/19

This Article is brought to you for free and open access by the Statistics, Department of at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Dissertations and Theses in Statistics by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.
A CHARACTERIZATION OF A VALUE ADDED MODEL AND A NEW MULTI-STAGE MODEL FOR ESTIMATING TEACHER EFFECTS WITHIN SMALL SCHOOL SYSTEMS

by

Julie M. Garai

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfilment of Requirements

For the Degree of Doctor of Philosophy

Major: Statistics

Under the Supervision of Professors Walter W. Stroup and Erin E. Blankenship

Lincoln, Nebraska

August, 2017
At both the national and state level there is increasing pressure to develop metrics to determine if school systems are meeting educational objectives. All states mandate some form of assessment by standardized tests. One method currently used to model student test scores is Value Added Modeling (VAM) which models student scores as a product of classroom and school environments. One VAM approach is the Tennessee Value Added Assessment System (TVAAS) introduced in Knox County, Tennessee. This approach models student gains from year to year longitudinally allowing student scores in the current year to be correlated with previous years. Teacher effects are included in the layered model which estimates the teacher’s added value to a student score through best linear unbiased prediction.

Research using VAM typically occurs in school systems with a large number of students (e.g. New York City, Los Angeles, Chicago, etc.) or in statewide assessments that are combined across school districts (e.g. Tennessee). VAM performance in school systems with small numbers of students is unknown.

One common issue with estimation based on small samples is lack of precision. An area of statistics that has developed methodology for small sample sizes is small area estimation. One approach in this area is indirect estimation which links similar subjects together allowing the small groups to “borrow strength” from each other.
This dissertation introduces a multi-stage model that incorporates small area estimation techniques with the traditional TVAAS. The performance of both the multi-stage and TVAAS models are studied through data simulated for small school systems. The precision of predicted teacher value added scores is assessed for both modeling methods.
ACKNOWLEDGMENTS

First I would like to thank my advisers Walt and Erin. You have taught me that academia is more than research, that people are important and my feelings matter. You have challenged me to do more and be more. You taught me to learn my own limitations and that sometimes the pressure I put on myself is unreasonable. Thank you for your patience, for the hours of advice, and the pieces of your lives that you have shared with me.

Thank you to my committee members Kathy Hanford and Tiffany Heng-Moss. I appreciate all of your feedback and support through this process!

I would also like to thank my parents and siblings for your love and support. You have been so understanding. Mom and Dad thank you for teaching me that I am an intelligent woman and that with a lot of hard work I can achieve more than I thought possible. I wish you were here to see this Dad.

Thank you to my husband Matt for cooking food and reminding me to eat, for washing my clothes so I had something to wear, and for going through my code line by line to find every mistake. You have been the ultimate supporter and my biggest cheerleader. I would not have made it through this process without you. Also I want to thank my dog Luna for barking in the adjacent room as I wrote this dissertation, forcing me to take some breaks and enjoy life a little.

Finally, I want to thank Team Stroup and Danielle for accepting countless phone calls and encouraging me through this process. Also thank you to my pseudo-therapist Stacey Herceg for always listening.
Table of Contents

List of Figures x

List of Tables xii

1 Introduction 1

2 Literature Review 4

2.1 Value Added Model Overview 4

2.1.1 The Appeal of Value Added Modeling and The Modeling Process 4

2.1.2 The TVAAS Model and Associated Assumptions 6

2.1.3 Continued Research Into Assumptions Regarding Student Assessment 9

2.1.4 Concerns about VAMs 10

2.1.5 Small School Systems 11

2.2 Small Area Estimation 12

2.2.1 Background 13

2.2.2 Model-based Approach to Small Area Estimation 14

2.3 Merging Value Added Modeling and Small Area Estimation 15

3 Modeling Processes 16

3.1 Tennessee Value Added Assessment System 16
3.1.1 A Demonstrating Example of Value Added
3.1.2 The TVAAS and Associated Modeling Assumptions
3.2 Small Area Estimation
3.2.1 Small Areas: An Application in Nebraska
3.2.2 General Form of a Small Area Model
3.3 Small Area Multi-Stage Modeling Process
3.3.1 Stage One
3.3.2 Stage Two
3.3.3 Stage Three
3.3.4 Stage Four
3.4 Measuring Model Performance
3.5 Cases for Consideration
4 A Demonstration of TVAAS and the Multi-Stage Model
4.1 The Data
4.1.1 Student Data
4.1.2 Teacher Data
4.1.3 True Teacher VA Scores
4.2 Implementing TVAAS
4.3 Implementing the Small Area Multi-Stage Model
4.3.1 The First Iteration
4.3.1.1 Stage One
4.3.1.2 Stage Two
4.3.1.3 Stage Three
4.3.1.4 Stage Four
4.3.2 Evaluating Convergence
5 A Characterization of the TVAAS and the Multi-Stage Model Via Simulation

5.1 Data Generation ........................................... 52
   5.1.1 Generating Teacher Value Added Scores .................. 52
      5.1.1.1 Teacher Auxiliary Variables ...................... 53
      5.1.1.2 Ensuring Simulated Data Are Reasonable .......... 56
      5.1.1.3 Design Structure Variables ...................... 59
      5.1.1.4 Combining Teacher Auxiliary and Design Structure Variables .................. 62
   5.1.2 Generating Student Assessment Scores .................. 62
      5.1.2.1 Student Raw Scores ........................... 63
      5.1.2.2 Combining Student Raw Scores with Teacher VAM Scores .................. 64
      5.1.2.3 Replicating the Study and Ensuring Simulated Data is Reasonable .......... 65
   5.1.3 Defining the Z Matrix ................................ 66

5.2 Model Analysis ............................................ 68
   5.2.1 Generated Data ................................... 68
   5.2.2 Standard Model ................................... 69
   5.2.3 Small Area Multi-Stage Model ....................... 70

5.3 Results .................................................. 71
   5.3.1 Ranking of Teacher Scores .......................... 71
   5.3.2 Mean Square Prediction Error ....................... 76
5.3.3 Impact of Convergence Criterion ........................................ 79
5.3.4 Evaluation of Covariance Structure ................................. 81
5.4 Conclusions ................................................................. 83
  5.4.1 Teacher Rankings ...................................................... 83
  5.4.2 Mean Square Prediction Error ................................. 84
5.5 Additional Investigations .................................................. 85
  5.5.1 Larger Number of Students Per Grade ......................... 85
  5.5.2 Future Research .................................................... 91

6 Conclusions ........................................................................ 92

Bibliography ........................................................................ 94

Glossary of Terms .................................................................. 99

A Additional Figures of Interest .............................................. 101

B SAS Code ........................................................................ 102
  B.1 Setup Macro ............................................................ 102
    B.1.1 Simulate Variables and Scores for Case 1 ................. 102
    B.1.2 Simulate Variables and Scores for Case 2 ................. 107
  B.2 Compile Macro .......................................................... 108
    B.2.1 Compile Data Generated Under Case 1 .................. 108
    B.2.2 Compile Data Generated Under Case 2 .................. 111
  B.3 Analyze Macro .......................................................... 112
    B.3.1 Analyze Data Generated Under Case 1 ................. 112
    B.3.2 Analyze Data Generated Under Case 2 ................. 124
  B.4 Sim Macro ................................................................. 125
List of Figures

2.1 Student Improvement In a Box ............................... 5
2.2 Student Growth Above What is Expected .................. 6
3.1 Score for Student $i$ after Grade 1. ......................... 17
3.2 Score for Student $i$ after Grade 2. ......................... 17
3.3 Score for Student $i$ after Grade 3. ......................... 18
3.4 School Systems in Nebraska by Size. ....................... 27
4.1 Plot of Prediction Error for the Two Modeling Methods ... 50
4.2 Standard TVAAS Modeling Method Deciles and Heat Map ... 51
4.3 Small Area Multi-Stage Modeling Method Deciles and Heat Map ... 51
5.1 Years of Service Distribution ............................... 57
5.2 Assessing Simulated VAM Scores Under Both Conditions ... 61
5.3 Assessment Scores for 5 Randomly Selected Students ........ 65
5.4 Student Assessment Scores for 100 Simulated Trials For Each Grade. ... 66
5.5 Compiled Student Scores Across Grades for Case and Condition. ... 67
5.6 Resulting Deciles for Hypothetical Perfect Model ............ 71
5.7 Resulting Deciles for Standard Method with Case 1 and Condition 1 72
5.8 Results for Both Modeling Methods Under Case 1 and Condition 1 ... 74
5.9 Results for Both Modeling Methods Under Case 2 and Condition 1 ... 74
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.10</td>
<td>Results for Both Modeling Methods Under Case 1 and Condition 2</td>
<td>75</td>
</tr>
<tr>
<td>5.11</td>
<td>Results for Both Modeling Methods Under Case 2 and Condition 2</td>
<td>75</td>
</tr>
<tr>
<td>5.12</td>
<td>Comparing MSPE for Modeling Method Based on Cases and Conditions used for Generating Data</td>
<td>76</td>
</tr>
<tr>
<td>5.13</td>
<td>Confidence Interval Plot Comparing Standard and Small Area Methods</td>
<td>78</td>
</tr>
<tr>
<td>5.14</td>
<td>Number of Iterations Required for Different Convergence Criterions</td>
<td>80</td>
</tr>
<tr>
<td>5.15</td>
<td>Box plots for the MSPE for Each of the Convergence Criterions</td>
<td>81</td>
</tr>
<tr>
<td>5.16</td>
<td>Results for Both Modeling Methods Under Case 1 and Condition 1 for Design 2</td>
<td>86</td>
</tr>
<tr>
<td>5.17</td>
<td>Results for Both Modeling Methods Under Case 2 and Condition 1 for Design 2</td>
<td>86</td>
</tr>
<tr>
<td>5.18</td>
<td>Results for Both Modeling Methods Under Case 1 and Condition 2 for Design 2</td>
<td>87</td>
</tr>
<tr>
<td>5.19</td>
<td>Results for Both Modeling Methods Under Case 2 and Condition 2 for Design 2</td>
<td>87</td>
</tr>
<tr>
<td>5.20</td>
<td>MSPE for Each Case and Condition under Design 2</td>
<td>89</td>
</tr>
<tr>
<td>5.21</td>
<td>30 Students per School, Confidence Interval Plot Comparing Methods</td>
<td>90</td>
</tr>
<tr>
<td>A.1</td>
<td>Three Negative Binomial Distributions Used to Generate Years of Teaching Experience</td>
<td>101</td>
</tr>
</tbody>
</table>
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Classification System Breakdown in Nebraska [26]</td>
<td>26</td>
</tr>
<tr>
<td>4.1</td>
<td>Process Giving Rise to Student Assessment Scores</td>
<td>36</td>
</tr>
<tr>
<td>4.2</td>
<td>Portion of Student Data</td>
<td>37</td>
</tr>
<tr>
<td>4.3</td>
<td>Process Giving Rise to Teacher VAM Scores</td>
<td>39</td>
</tr>
<tr>
<td>4.4</td>
<td>Portion of Teacher Data</td>
<td>40</td>
</tr>
<tr>
<td>4.5</td>
<td>True Teacher VAM Scores</td>
<td>40</td>
</tr>
<tr>
<td>4.6</td>
<td>Covariance Parameter Estimates after Implementation of TVAAS</td>
<td>41</td>
</tr>
<tr>
<td>4.7</td>
<td>Fixed Effect Solutions and Teacher Score Estimates</td>
<td>42</td>
</tr>
<tr>
<td>4.8</td>
<td>Teacher Data Including Predicted Teacher VAM Scores</td>
<td>44</td>
</tr>
<tr>
<td>4.9</td>
<td>Covariance Parameter Estimates and Evaluation of Fixed Effect Significance</td>
<td>44</td>
</tr>
<tr>
<td>4.10</td>
<td>Solution for Fixed Effects After First Iteration</td>
<td>45</td>
</tr>
<tr>
<td>4.11</td>
<td>Portion of Solution for Random Effects After First Iteration</td>
<td>45</td>
</tr>
<tr>
<td>4.12</td>
<td>Updated Teacher VAM Scores</td>
<td>45</td>
</tr>
<tr>
<td>4.13</td>
<td>Predicted Teacher VAM Scores after the Final Stage of the Multi-Stage Model</td>
<td>48</td>
</tr>
<tr>
<td>5.1</td>
<td>Converting Years of Service to an Ordinal Variable</td>
<td>54</td>
</tr>
<tr>
<td>5.2</td>
<td>Frequency of Teachers With and Without Advanced Degrees</td>
<td>56</td>
</tr>
</tbody>
</table>
5.3 Years of Teaching Experience for Public School Teachers, 2011-2012
   Academic Year .......................................................... 57
5.4 Years of Teaching Experience for Simulated Teachers .............. 58
5.5 Percentage of Teachers that Fall Into Each Category. ............ 58
5.6 Hierarchical Structure under Consideration .......................... 59
5.7 Conditions for Variance of Teacher Random Effects ............... 60
5.8 The Z Matrix Incorporated with the Data ............................. 68
5.9 Confidence Intervals for the True Average Difference Between Modeling Methods .......................................................... 79
5.10 Confidence Intervals for the True Average Difference Between Modeling Methods .......................................................... 79
5.11 Impact of Choice of Convergence Criterion on MSPE ............ 81
5.12 Identifying Appropriate Covariance Structure for Repeated Measure in Standard Model ......................................................... 82
5.13 Confidence Intervals for the True Average Difference Between Modeling Methods .......................................................... 91
5.14 Confidence Intervals for the True Average Difference Between Modeling Methods .......................................................... 91
CHAPTER 1

INTRODUCTION

Over the past few decades, there has been a major reform movement to improve effectiveness of primary and secondary education. Specifically, there has been a focused effort on accountability assessments at the school and teacher level, particularly with the use of the value added models (VAM). One of these accountability systems was developed in the 1980s in Knox County, Tennessee: the Tennessee Value-Added Assessment System (TVAAS). The TVAAS was mandated as a component of the Education Improvement Act to provide a means of assessing the ability of teachers, schools and systems to meet goals and objectives set forth by the state of Tennessee [32].

Prior to 2001, each state was individually responsible for assessing student knowledge and material retention. Several recommendations were put in place at the national level, but not made mandatory in order to allow the states to retain autonomy over their educational processes [15]. With the passage of No Child Left Behind in 2001, local statewide initiatives became subjected to federal mandates. Specifically, statewide testing in grades 3–8 and one grade in high school were required with the goal of holding schools and districts accountable for student progress [6]. The accountability systems introduced were intended to help inform personnel decisions from low stakes evaluations, e.g. needing to participate in professional development
programs, to high stakes evaluations, e.g. bonuses, promotion, and firing of teachers [4]. In response, several methods arose as a means of linking student assessment scores to teacher evaluation. A major issue is that the use of the results from the modeling methods introduced, especially for high stakes evaluations, occurred before the models were fully understood. Consequently educational analysts continue to caution consumers (e.g. administrators) about the use of the findings from student assessment modeling for high stakes decisions. When used properly, the TVAAS and other VAMs have proven to be a valuable tool for modeling student performance.

Many facets of VAMs have been investigated, but most studies focus on urban / suburban school districts where the large number of teachers and students provide more stable value added (VA) score estimates at the teacher, school, and district level [2, 8, 16, 30, 32]. For these current models, the resulting score estimate performance in districts with smaller student / teacher populations is unknown. There is a need to provide smaller school systems with an avenue of measuring student growth and learning.

Traditionally, small area estimation is an area of study within statistics that provides an avenue in cases when small sample sizes lead to findings with inadequate precision. Indirect estimation is a technique in small area estimation that allows smaller groups of similar subjects to “borrow strength” from each other by linking the subjects together [29]. In this dissertation we propose a small area multi-stage model that builds on the findings of the TVAAS by incorporating additional information about teachers previously unaccounted for in the model. Our two main research questions are

1. How does the TVAAS model, perform in smaller school systems particularly in regards to precision of VA score estimates?
2. How does merging small area estimation techniques with TVAAS through a multi-stage model impact precision of teacher VA score estimates?

Because of the continued research into model assumptions and validity, this dissertation aims to provide a useful tool for teacher evaluation and improvement and is not intended to be used for high stakes evaluation purposes namely compensation, hiring, or firing of teachers.

Chapter 2 includes an overview of VAM. We focus on the TVAAS including ongoing research and discuss concerns surrounding VAM. We introduce basic concepts of small area estimation and present a modeling approach that incorporates small area techniques with VAM.

Chapter 3 begins by outlining the form of the TVAAS used for this dissertation including associated modeling assumptions. We proceed by presenting a general form of a small area estimation model that utilizes indirect estimation. Finally we introduce the small area multi-stage model that incorporates both TVAAS and small area estimation. We characterize the models under two cases and propose two measures for model comparison.

We include a demonstration of using both the standard TVAAS and small area multi-stage models with simulated data in Chapter 4. We discuss evaluating model convergence and close with a comparison of the findings between both modeling methods.

Chapter 5 includes an in depth characterization of both the standard TVAAS and small area multi-stage models through simulation. We discuss the process of generating reasonable data, model implementation, results and conclusions. We close this chapter with an additional investigation into the number of students per classroom and our future research plans.
CHAPTER 2

LITERATURE REVIEW

This literature review includes an overview of value added modeling (VAM), with a focus on the Tennessee Value Added Assessment System (TVAAS). We also discuss small area estimation techniques and propose a method for incorporating these techniques with the TVAAS.

2.1 Value Added Model Overview

The principal objective of VAMs is to determine the impact of individual teachers and school systems on student achievement while accounting for characteristics attributable to students’ backgrounds [1, 8, 23, 32]. With this modeling approach, school systems and teachers are measured on their ability to help students improve rather than their ability to attain a large number of high or proficient test scores.

This section discusses the appeal of VAMs along with the process of modeling student improvement. The TVAAS model is introduced along with modeling assumptions. We close with current research and concerns about VAM.

2.1.1 The Appeal of Value Added Modeling and The Modeling Process

Enthusiasm of using VAMs is that they provide an objective way of estimating teacher impact on student achievement as opposed to solely subjective classroom observation and survey based assessments. Also, teachers are compared relative to each other
rather than a pre-established threshold [13]. VAM is performed through a longitudinal study where multiple years of information are retained for each student. The model links each subsequent grade for a student together, allowing the student to serve as his or her own control and measures the student’s progress or improvement rather than the specific score [5, 23, 35]. By measuring progress, rather than proficiency, the scores of both low and high achieving students are characterized. Allowing each student to serve as his or her own control accounts for student level influences, e.g. socioeconomic status or other demographic factors, that inherently affect performance on achievement tests. Consequently the model allows for a clearer picture of the impact of school systems and teachers on student performance.

The process of modeling student improvement is well explained by Drayton 2014 and presented in Figure 2.1. A student enters a classroom at the beginning of the school year represented by a box. The student leaves at the end of the school year with academic growth implying that the box helped the student [10]. The box in this example represents the impact of classroom environment on student learning. This is the basic logic underlying VAM.

Students are expected to improve each year but any additional growth is attributed to the student’s classroom environment, namely the teacher. Figure 2.2 shows the
student growth. If the student grows as expected, we would see the gray figure, however the student grows above what is expected and has a higher actual growth represented by the black figure. The additional growth above what is expected is referred to as value added (VA) which is attributed to the teacher [10]. The expected growth is calculated based on the average growth for all students at this grade level. So approximately half of the students grow at a rate below what is expected and half grow at a rate above what is expected. Growth can be positive or negative.

Several models have been proposed for measuring student progress [8, 23]. One method is the TVAAS which was developed in Tennessee [32]. For this dissertation, we have chosen to focus on the TVAAS model which is introduced in more detail in the next section.

2.1.2 The TVAAS Model and Associated Assumptions

The TVAAS is an accountability system that utilizes a multivariate longitudinal analysis of student data and looks at student academic gain from year to year rather than the raw achievement test score [32]. Students are assessed over several grades and in multiple subjects. Along with assessment results, information regarding school
system, teacher and other demographics are recorded annually. The TVAAS allows for student scores over time and scores across subjects to be correlated.

The TVAAS utilizes a layered model to incorporate multiple years of information about each student. The layering allows the influence of previous teachers on student achievement to persist for subsequent years. Sanders, Saxton and Horn 1997 found that teachers have a significant effect on student growth [35]. The TVAAS incorporates a teacher effect for the student assessment score that persists undiminished in the model for subsequent grades.

Using mixed model theory for this approach allows for the inclusion of additional covariates in the model known to be correlated with student performance on standardized tests. One criticism of TVAAS is that typically only teachers and school system are identified in the model and covariates known to be correlated with student achievement are omitted, e.g. socioeconomic status, which could lead to biased findings in regards to teachers [16, 20, 21]. Ballou, Springer and Wright 2004 found that “the introduction of controls at the student level has a negligible impact on estimated teacher effects in the TVAAS” [5]. By measuring student gains or incorporating a baseline measurement, the inclusion of additional covariates provides little additional information and often introduces serious modeling issues namely confounding and multicollinearity. Including additional covariates may seem necessary but end up being redundant if a baseline is included in the model [5].

The layered model incorporates multiple years of information for each student. Fellers 2014 presents an adaptation of the TVAAS model for one tested subject

\[
\begin{align*}
 y_{1j1} & = \mu_1 + \theta_{j1} + \epsilon_{i1j1} \\
 y_{2j1j2} & = \mu_2 + \theta_{j1} + \theta_{j2} + \epsilon_{i2j1j2} \\
 y_{3j1j2j3} & = \mu_3 + \theta_{j1} + \theta_{j2} + \theta_{j3} + \epsilon_{i3j1j2j3}
\end{align*}
\] (2.1.1)
where $y_{igj} \ldots jg$ represents student $i$'s test score in year $g$ which includes the overall mean for year $g$, $\mu_g$, as well as the random teacher effects, $\theta_{jg}$, for the current year and previous years if applicable [11]. The model assumes the following

$$
teacher \text{ effects: } \theta \sim \text{Gaussian}(0, \sigma^2_{\theta})$$
$$residuals : \epsilon \sim \text{Gaussian}(0, \Sigma)$$

with $\theta$ and $\epsilon$ assumed to be uncorrelated. Teacher effects persist in the layered model undiminished for subsequent years of assessment [5, 8, 11, 14, 23, 34]. Student residuals have a covariance structure $\Sigma$ which allows for a repeated measure structure. Possible structures for $\Sigma$ are described below.

In the literature both an unstructured covariance structure [5, 14, 23] and first order auto-regressive covariance structure[11] are used to model the repeated measure structure. Model 2.1.1 incorporates 3 years worth of information for student $i$, but additional years of information can certainly be incorporated. Also, this model assumes that student $i$ has the same teacher for the entirety of grade $g$. This assumption can be relaxed by incorporating a coefficient that allows for multiple teachers per grade [8, 23, 34]. Teacher effect estimates are found using best linear unbiased prediction (BLUP) or “shrinkage estimators” [5, 8, 11, 14, 22, 23, 34]. Similarly scores can be obtained at the district or state level, depending on the inclusion of this information in the model. Model 2.1.1 is for one tested subject (e.g. Mathematics) which is a simplification of the full TVAAS which typically models multiple tested subjects for student $i$ simultaneously.

In the next section, we discuss continued research being performed into assumptions made by VAM.
2.1.3 Continued Research Into Assumptions Regarding Student Assessment

In addition to model based assumptions (e.g. normality, independence) appropriate use of VAM relies on several other assumptions which are under continuing research. Paufler and Amrein-Berdsley 2014, Rothstein 2009 and Fellers 2014 all explore the assumption that students are randomly assigned to teachers. Paufler and Amrein-Berdsley 2014 investigated the process that administrators follow to assign students to classrooms in Arizona elementary schools. They found that overwhelmingly students are not assigned at random to classrooms. Rothstein 2009 found that bias is present when estimating teacher effects in cases where student assignment to teacher is not random. He determined that the amount of bias present is highly dependent on the quantity of information used to assign students to classrooms. Fellers 2014 found that estimates were not biased under a non-random assignment scheme but did find that the standard errors were underestimated which could lead to issues with estimate precision.

Another assumption is that the assessment is capable of capturing student growth, specifically that students are not subjected to ceiling effects which occur when a number of students score at or near the maximum possible value (it is impossible to score over 100%) which results in minimal cumulative gain. This topic is explored by Koedel and Betts 2010 and Fellers 2014. Koedel and Betts found that typically teacher effect estimates are not severely impacted by the presence of ceiling effects. They did find an issue with estimation on minimum-competency testing. In these cases, states are solely interested in determining if students have attained a minimum amount of required knowledge from state-wide education objectives. The impact on the results of VAM was more severe for these environments, because a large percentage
of students (e.g., 90%) attained the maximum score. Fellers 2014 determined that “The magnitude of bias and effect of ceiling level is not consistent across teachers” [11]. Issues arose in cases where a small number of students were at the ceiling and when a large number of students were at the ceiling. In both instances some teacher estimates were unbiased, but at least one teacher estimate showed severe bias.

Koretz 2008 explores the assumption that student test scores are a valid measure of student achievement. One example that he provides is when a mathematics assessment involves extensive reading or writing. While students may be proficient mathematically, the test penalizes immigrant students learning a new language. He presents several other examples of how testing alone may not be an adequate measure of student learning.

Many VAMs assume that any missing student records are missing solely at random. Karl, Yank and Lohr 2013 investigate the impact that this failed assumption has for estimates of teacher impact on student improvement. They introduce a correlated random effects model that allows for exploration into the sensitivity of teacher estimates for different cases of missing observations. Both patterns of data missing at random and of data not missing at random are investigated. They found that the assumption that the data is missing at random may be violated but that the impact of the violated assumption was minimal when estimating teacher effects in a VAM.

We discuss concerns and criticisms regarding VAM in the next section.

2.1.4 Concerns about VAMs

The American Statistical Association released a statement explaining that “VAMs are increasingly promoted or mandated as a component in high-stakes decisions such as determining compensation, evaluating and ranking teachers, hiring or dismissing teachers, awarding tenure, and closing schools” [1]. One of the largest concerns re-
garding VAM is not the modeling process but rather the political and micro-political uses of the results [39]. Several articles discuss the implications of using VAM for high-stakes decisions [1, 3, 4, 13, 16, 24]. It is important to note that teacher effects estimated through VAM are estimates with associated standard errors. Consequently it is not advised to use the findings for high-stakes decisions.

Additionally, several articles discuss the unintended consequences of high-stakes testing. Baker et al. 2010 explain “Surveys have found that teacher attrition and demoralization have been associated with test-based accountability efforts, particularly in high-need schools” [3]. Jiang, Sporte and Luppescu 2015 found that teachers had higher levels of stress and anxiety and believed the findings from the testing process were not worth the added stress [16]. Moore Johnson 2015 explains that “heavy reliance on VAM may lead effective teachers in high-need subjects and schools to seek safer assignments, where they can avoid the risk of low VAM scores. Meanwhile, some of the most challenging teaching assignments would remain difficult to fill and likely be subject to repeated turnover, bringing steep costs for students” [24]. High stakes use of VAM results may have more consequences than originally intended.

Instead, many educational analytic researchers suggest using VAM as a tool to help inform program evaluations, to identify teaching practices that result in higher outcomes and potentially use the knowledge for collective education improvement [24, 32]. It is intended that VAM be one means of characterizing academic progress, not the only means.

2.1.5 Small School Systems

The majority of research on VAM is conducted in large school systems. Chetty, Friendman and Rockoff 2014 study VAM in New York City School Systems, and Bacher-Hicks, Kane and Staiger 2014 extend the findings to the Los Angeles Unified

Baker et al. 2010 argue that “individual classroom results are based on small numbers of students leading to much more dramatic year-to-year fluctuations. Even the most sophisticated analyses of student test score gains generate estimates of teacher quality that vary considerably from one year to the next” [3]. Ballou and Springer 2015 echo this sentiment stating that “the measures used in... accountability systems are noisy and that the amount of noise is greater the fewer students a teacher has” [4]. Ballou and Springer present a brief investigation into the impact of the number of students on estimated teacher effectiveness. However, both articles lack an in depth analysis into the impact of small numbers of students on identification of teacher contribution to student learning. The performance of value added methodology in school systems with small numbers of students has not been investigated.

Additional knowledge about estimation for small sample sizes is needed. One area of statistics specifically focuses on developing methodology for cases with small numbers of subjects. Small area estimation is presented in the next section and ties between this area and education are introduced.

2.2 Small Area Estimation

This section includes background information regarding small area estimation and presents a standard form of a small area model.
2.2.1 Background

Sample surveys have long used small area estimation as a means to determine more precise estimates for small groups of interest. The methodology was developed for cases when using traditional estimation techniques on small samples resulted in estimates with inadequate precision [29]. Often small samples are referred to as small domains which in the education realm may encompass rural school systems as well as school systems in urban / suburban environments where the student and or teacher populations are relatively small. Additionally small area estimation techniques could be useful in larger school systems to determine more precise estimates for smaller subpopulations (e.g. demographics with small representation) within the school.

Rao 2003 introduces indirect estimation as an alternative to traditional estimation. The three types indirect estimation are: domain indirect, time indirect, and domain and time indirect. Domain indirect estimators link related groups of subjects based on similar characteristics at a given time; time indirect estimators link the present and past characteristics for the specific area; domain and time indirect estimators link related areas based on similar characteristics both in the past and the present [29]. Linking similar groups of subjects together allows for “borrowing of strength” between subjects.

Utilization of indirect estimators introduces sample design bias. If the indirect estimators link similar subjects, the bias will be small leading to a smaller variance in comparison to traditional estimators thus providing an overall reduction in mean square error (MSE). If the linking of subjects is unreasonable, the bias introduced will be large and the overall MSE will be larger in comparison to traditional estimators. It is important to realize that the bias does not go away as sample size increases. A smaller MSE leads to more precise prediction. For the small area model used in this
dissertation the bias arises because we are using shrinkage estimators which are not unbiased in the classical sense but lead to an overall reduction in prediction error.

There are several approaches to small area estimation. This dissertation focuses on model-based estimation which Ghosh and Rao 1994 explain is a special case of “a general mixed linear model involving fixed and random effects” [12]. This approach offers several advantages including the ability to derive optimal estimators, finding measures of variability that can be linked to the individual domains or groups of interest, and the flexibility to incorporate spatial or time series structures [29]. Several forms of model-based estimation exist in small area estimation. This dissertation focuses on empirical best linear unbiased prediction (EBLUP) estimators because they utilize linear mixed models (LMMs) and a frequentist perspective, a natural pairing with the TVAAS which is also a LMM considered from the frequentist framework. Additional model based approaches discussed in Rao 2003 include parametric empirical Bayes estimators and parametric hierarchical Bayes estimators.

2.2.2 Model-based Approach to Small Area Estimation

Rao 2003 introduces a generic form of an exponential family model that utilizes fixed and random effects to provide an estimate of small area numerical summaries (e.g. means, totals, proficiencies, etc.) which is shown in Model 2.2.1.

\[ z_{ij} = X_{ij}\beta + v_i + u_{ij} \]  

where \( j \) is the number of observational units in subgroup \( i \), \( v_i \) represents the error related to linking subjects and \( u_{ij} \) represents the individual error related to observational unit \( j \) in subgroup \( i \). It is assumed that the error terms \( v_i \) and \( u_{ij} \) are independent with \( v_i \overset{iid}{\sim} \text{Gaussian}(0, \sigma_v^2) \) and \( u_{ij} \overset{ind}{\sim} \text{Gaussian}(0, \sigma_u^2) \).
This model follows similarly from a designed experiment that incorporates random effects with fixed effects. Consider $z_{ij}$ to be a teacher VAM score, $X_{ij}$ to include auxiliary variables associated with a teacher which are fixed effects, and $v_i$ to be a linking structure that connects teachers possibly based on a sample of locations which is a random effect. Estimation in this case proceeds as with traditional designed experiments as presented in Stroup 2012.

The final section of this chapter introduces our proposed method of incorporating small area estimation with traditional VAM.

### 2.3 Merging Value Added Modeling and Small Area Estimation

We aim to incorporate small area estimation techniques with VAM specifically TVAAS to incorporate more information about the teachers into our estimates of teacher value added. A linking mechanism can be utilized to link teachers in similar school systems, e.g. small rural environments.

We propose using predicted teacher effects obtained from the TVAAS model using BLUP as our response variable for a small area model. This results in a multi-stage modeling process. We characterize the performance of both the TVAAS and multi-stage model in regards to precision of prediction for teacher value added scores.
CHAPTER 3

MODELING PROCESSES

The chapter begins with the Tennessee Value Added Assessment System (TVAAS), then introduces basic concepts of small area estimation. The final portion of this chapter introduces the small area multi-stage modeling process and the cases under consideration for the simulation in Chapter 4.

3.1 Tennessee Value Added Assessment System

This section includes a motivating example that introduces terminology and a visualization of the value added process. We then introduce the TVAAS along with associated modeling assumptions.

3.1.1 A Demonstrating Example of Value Added

Consider an example where students begin at 55% on average for a standardized assessment and are expected to gain 2% after each grade, \( g \).

Student \( i \) begins with a baseline score of 50% which is 5% below average. Suppose after grade \( g = 1 \) we have the situation depicted in Figure 3.1 The dotted line represents the average overall growth for all students which begins at 55%. After grade 1 students grow to 57%, a 2% gain, on average. The expected score for student \( i \) is shown by the dashed line. If the student scores as expected we anticipate a score after grade 1 of 52%, a 2% gain. However, student \( i \)'s actual growth is shown by the
solid line. Student $i$ scores a 60% which is a 10% gain, 8% above what is expected. This additional 8% is considered the value added ($VA$) by teacher $j$ in grade 1.

Figure 3.2 shows the progress of student $i$ after grade $g = 2$. If student $i$ grows at the rate expected, we anticipate a standardized score of 62%, following the average gain of 2% annually. The dashed line shows the expected growth of student $i$ to
62%. The solid line shows the actual growth of student $i$ to 75%, a gain of 15%. The additional 13% of growth is considered the VA by teacher $j$ in grade 2.

Figure 3.3 shows the results for student $i$ after grade $g = 3$. If student $i$ grows at the expected rate, represented by a dashed line, we would expect to see a score after grade 3 of 77%, a 2% gain. In grade 3 the actual growth for student $i$, represented by the solid line, is 78%, a 3% gain. The additional 1% of growth is the VA by teacher $j$ in grade 3.

Each of the grades presented in Figures 3.1-3.3 shows positive VA for each teacher: $j^1$, $j^2$, and $j^3$. It is reasonable for a student to grow at a rate below what is expected. The expected growth for each student has a central tendency and this growth is compared to the average growth trajectory overall. As a result of assuming normality, 50% of students gain above what is expected and 50% of students gain below what is expected each year. For example if a student improves between grades 1 and 2 at a rate below what is expected (e.g., a 1% gain rather than 2%) this results in negative VA for teacher $j$ for grade $g = 2$. The TVAAS introduced in Section 3.1.2 has the flexibility to allow each $t_{jg}$ for grades $g = 1, 2, 3$ to be attributed to a
different teacher, to be the same teacher, or some combination.

### 3.1.2 The TVAAS and Associated Modeling Assumptions

The Tennessee Value Added Assessment System (TVAAS) utilizes a layered model that incorporates all available years of information for each student. A unique characteristic of this model is that teacher effects are considered to be persistent without diminishing over time. Additionally, the flexibility of the TVAAS allows for all tested subjects for a student to be modeled simultaneously. A baseline measurement for each student is included which allows each student to serve as his or her own control or “block factor” [32]; because the model is blocked by each individual student, it is redundant to include covariates known to be highly correlated with student success (e.g., socioeconomic factors) in the model [33].

Typically several grades worth of data for each student are stored, including scores on standardized assessments and the teacher the student has for each grade. We denote the first measured grade of information for student $i$ as $g = 1$. If the student is first assessed in grade 6, in our model $g = 1$ will correspond to grade 6. We denote subsequent years of assessment as $g = 1, 2, \ldots, Y$ where $Y$ is the most recent year of assessment. We use the terms grade and year interchangeably. Model 3.1.1 represents the first layer for the TVAAS model which corresponds to grade $g = 1$:

$$s_{i,1,j^1} = \mu + X_i \beta + t_{j^1} + e_{i,1,j^1}. \quad (3.1)$$

For this model, $s_{i,1,j^1}$ denotes the standardized test score of student $i$ in grade $g = 1$ with teacher $j$ in grade $g = 1$ for a specific subject (e.g., Mathematics); $\mu$ denotes the common intercept for all students; $X_i$ is the baseline score for student $i$; $\beta$ is the regression coefficient; $t_{j^1}$ is the teacher effect for teacher $j$ in grade $g = 1$; $e_{i,1,j^1}$
represents the individual residual for student $i$ in grade $g = 1$ with teacher $j$. Sanders and Horn 1994 discuss the benefits of considering teacher effects to be random, such as the flexibility to accommodate reality where teachers may teach several different grades or switch grades [32]. Thus we assume teacher effects are random with an associated distribution where

$$\begin{bmatrix} t_{j1} \\ e_{i,1,j1} \end{bmatrix} \sim \text{Gaussian}\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix} \right).$$

The structure of $R$ and $G$ are discussed below.

Model 3.1.2 shows a general form of the layered model which is an adaptation of models presented in the literature [11, 23, 32]. The total number of measured grades is represented by $Y$. The general layered model is

$$s_{i,g,J} = \mu + X_i\beta + \sum_{g=1}^{Y} t_{jg} + \sum_{g=1}^{Y} e_{i,g,J}$$

(3.1.2)

where $s_{i,g,J}$ denotes the standardized score for student $i$ in grade $g$ with teachers $J$; $\mu$ is the common intercept for all students; $X_i$ is the baseline score for student $i$ which remains constant for each measured grade $g$; $\beta$ denotes the common regression coefficient; $\sum_{g=1}^{Y} t_{jg}$ is the sum of the random teacher effects for all previous teachers and the current teacher, which remain persistent and undiminished in the model for subsequent grades; $\sum_{g=1}^{Y} e_{i,g,J}$ represents the random residual for student $i$ after grade $g$ with teachers $J$. Let $w$ be the total number of teachers. The random teacher effects
from 3.1.2 are assumed to have the following distribution

\[
\begin{bmatrix}
  t_1 \\
  t_2 \\
  \vdots \\
  t_{w-1} \\
  t_w
\end{bmatrix}
\sim \text{Gaussian}
\begin{bmatrix}
  \begin{bmatrix}
    0 \\
    0 \\
    \vdots \\
    0
  \end{bmatrix}
  \begin{bmatrix}
    1 & 0 & \ldots & \ldots & 0 \\
    0 & 1 & \ldots & \ldots & 0 \\
    \ldots & \ldots & \ddots & \ldots & \vdots \\
    0 & \ldots & \ldots & 1 & 0
  \end{bmatrix}
\end{bmatrix}.
\]

(3.1.3)

Defining the teacher effects in this manner allows for teacher \( t_1 \) to possibly teach grade \( g = 1 \) which we denote \( t_{1,1} \) or grade \( g = 2 \) which we denote \( t_{1,2} \). Regardless, teacher \( t_1 \) is the same in both instances. We assume a constant variance across teachers. It is possible to relax this restriction and allow the teachers to have unique variances.

The student residuals from 3.1.2 are assumed to have the following distribution

\[
e_{i,g,J} \sim \begin{bmatrix}
  0 \\
  0 \\
  \vdots \\
  0
\end{bmatrix},
\]

\[
R = \begin{bmatrix}
  \sigma_1^2 & \sigma_{12} & \sigma_{13} & \ldots & \sigma_{1Y} \\
  \sigma_{21} & \sigma_2^2 & \sigma_{23} & \ldots & \sigma_{2Y} \\
  \sigma_{31} & \sigma_{32} & \sigma_3^2 & \ldots & \sigma_{3Y} \\
  \vdots & \vdots & \vdots & \ddots & \vdots \\
  \sigma_{1Y} & \sigma_{2Y} & \sigma_{3Y} & \ldots & \sigma_Y^2
\end{bmatrix}.
\]

(3.1.4)

For grade \( g = 1 \) we define \( J = j^1 \), for grade \( g = 2 \) we define \( J = j^1j^2 \), and for grade \( g = Y \) we define \( J = j^1j^2\cdots j^Y \). Thus \( J \) is a compilation of teachers past and present.

Consider an example with a total number of measured grades \( Y = 3 \). Model 3.1.2 simplifies to the layered model presented in 3.1.5.
Here the subscripts 1, 2 and 3 represent the 3 measured grades for student $i$. Teacher effects, $t_{jg}$, remain in the model for subsequent grades. In year 2, the teacher effect $t_{j1}$ from grade $g = 1$ remains in the model and persists without diminishing. Here the random teacher effects follow the same distribution presented in 3.1.3. The distribution for the student random residual $e_{i,g,j}$ presented in 3.1.4 simplifies to the distribution presented in 3.1.6

\[
\begin{bmatrix}
e_{i,1,j1} \\
e_{i,2,j1,j2} \\
e_{i,3,j1,j2,j3}
\end{bmatrix} \sim \text{Gaussian} \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix}. \tag{3.1.6}
\]

Assuming an unstructured covariance matrix for student residuals allows for the student score to vary from year to year and the correlation between scores to vary. An unstructured covariance matrix makes the fewest assumptions about the relationship between current and previous scores [23, 34].

The layered models presented in 3.1.2 and 3.1.5 are a simplification of the full TVAAS model introduced by Sanders, Saxton and Horn 1997. Our models focus on only one subject for each student while the full model incorporates information for all tested subjects and allows for correlation between subjects for each student. Our models also represent a situation where student $i$ has only 1 teacher for the duration of grade $g$. It is possible to include a coefficient along with $t_{jg}$ that incorporates shared
classroom teaching or students that have different teachers throughout a school year due to transferred classrooms [23].

The matrix form of Model 3.1.2 is $s = X\xi + Zt + e$

where $s = \begin{bmatrix} s_{1,1,J} \\ s_{1,2,J} \\ \vdots \\ s_{1,Y,J} \\ \vdots \\ s_{n,1,J} \\ s_{n,2,J} \\ \vdots \\ s_{n,Y,J} \end{bmatrix}_{(n\times Y) \times 1}$ for student $i = 1, \ldots, n$

for student $i = 1, \ldots, n$

in grade $g = 1, 2, \ldots, Y$

in grade $g = 1, 2, \ldots, Y$

with $J = j^1$

with $J = j^1$

when $g = 1$

when $g = 2$

\[ J = j^1 j^2 \ldots j^Y \] when $g = Y$

\[ J = j^1 j^2 \ldots j^Y \] when $g = Y$

where $X = \begin{bmatrix} 1 & X_1 \\ 1 & X_1 \\ \vdots & \vdots \\ 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \\ 1 & X_n \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}_{(n\times Y) \times 2}$

and $\xi = \begin{bmatrix} \mu \\ \beta \end{bmatrix}_{2 \times 1}$
with \( Z = \)
\[
\begin{vmatrix}
1 & 0 & \ldots & \ldots & 0 & 0 & 0 & \ldots & 0 \\
1 & 1 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 1 & \ldots & \ldots & 1 & 0 & 0 & \ldots & 0 \\
0 & 0 & 0 & \ldots & 0 & 1 & 0 & \ldots & 0 \\
0 & 0 & 0 & \ldots & 0 & 1 & 1 & 0 & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \ldots & 0 & 1 & 1 & \ldots & 1
\end{vmatrix}
\]

\( t = \)
\[
\begin{bmatrix}
t_1 \\
t_2 \\
\vdots \\
t_w
\end{bmatrix}_{w \times 1}
\]

for teacher \( t = 1, \ldots, w \), and \( e = \)
\[
\begin{bmatrix}
e_{1,1,J} \\
e_{1,2,J} \\
\vdots \\
e_{1,Y,J} \\
\vdots \\
e_{n,1,J} \\
e_{n,2,J} \\
\vdots \\
e_{n,Y,J}
\end{bmatrix}_{(n \times Y) \times 1}
\]

We assume that \( t \sim \text{Gaussian}(0, G) \) and \( e \sim \text{Gaussian}(0, R) \) where \( G \) is defined as in 3.1.3 and \( R \) is defined as in 3.1.4. The resulting mixed model equations are
\[
\begin{bmatrix}
X' R^{-1} X & X' R^{-1} Z \\
Z' R^{-1} X & Z' R^{-1} Z + G^{-1}
\end{bmatrix}
\begin{bmatrix}
\xi \\
t
\end{bmatrix}
= \begin{bmatrix}
X' R^{-1} s \\
Z' R^{-1} s
\end{bmatrix}.
\]

The main goal of implementing TVAAS is to obtain an estimate of the VA by
teachers to student scores. Solving 3.1.7 we obtain teacher VA estimates through best linear unbiased prediction (BLUP). Because $G$ and $R$ are unknown, we utilize restricted maximum likelihood to obtain their estimates. If we define $K$ as a linear combination of fixed effects of interest and $M$ as a linear combination of random effects of interest, we obtain the predictable function $K'\tilde{\xi} + M't$ by using the estimated values of $\xi$, $\tilde{\xi}$, and $t$, $\tilde{t}$, obtained by solving the mixed model equations [22]. The resulting predictor for teacher $j$ is referred to as the teacher value added model score or VAM score for teacher $j$. Ballou, Sanders and Wright 2004, show this process in detail [5].

Our two objectives with this dissertation are first to determine how TVAAS performs with small sample sizes (e.g. small number of students per classroom and small number of teachers per school) and second to determine how performance is affected by the incorporation of small area estimation techniques. To assess model performance, we are focusing on how accurate the resulting teacher rankings are and how precisely the VAM scores are estimated. This section provides a sufficient starting place for the first objective. The next section introduces small area estimation concepts necessary to assess our second objective.

3.2 Small Area Estimation

This section includes an introduction to small area estimation including an example based on the state of Nebraska as well as an applicable form of a small area model.

3.2.1 Small Areas: An Application in Nebraska

Small area estimation is an area of statistics developed for cases where small sample sizes lead to issues with estimation and prediction. One method of small area estimation, called indirect estimation, involves the linking of subjects in a study based
on similar characteristics [29]. Linking subjects increases the effective sample size and allows for borrowing of strength between subjects. If the linking is reasonable the bias introduced by linking subjects together will be small, resulting estimates will be more precise, and there will be a reduction in mean square error (MSE). While linking increases precision, unreasonable linking increases bias even more, offsetting increased precision and therefore increasing MSE.

For an example, consider the state of Nebraska. School districts in Nebraska are divided into four classification sizes 2 through 5. Table 3.1 shows the breakdown for the different classifications. In Nebraska, there is only one school district in Class 4

<table>
<thead>
<tr>
<th>Classification</th>
<th>Number of Inhabitants in the District</th>
<th>Total Number of Districts in this Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>&lt;1,000</td>
<td>18</td>
</tr>
<tr>
<td>3</td>
<td>1,001–99,999</td>
<td>225</td>
</tr>
<tr>
<td>4</td>
<td>100,000–199,999</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>200,000+</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3.1: Classification System Breakdown in Nebraska[26]

(Lincoln Public Schools) and one school district in Class 5 (Omaha Public Schools). The largest number of school districts are Class 3.

In the interest of readability, consider a subset of Nebraska public schools. A visual depiction of a possible linking process for 20 public school districts out of 245 total districts is shown in Figure 3.4. The two locations in Nebraska that fall into the largest category are shown with the darkest circle (Class 4 and 5 districts). We have divided Class 3 school districts into two categories. The middle category (identified with gray circles) represents the 85 school districts in Nebraska with at least 500 total students in grades PK-12. The smallest category (identified with white circles) represents Class 2 and small Class 3 school districts in Nebraska, which have less than
500 total students. 158 school districts in Nebraska fall into this category.

For Figure 3.4, a small area model utilizing indirect estimation may link teachers in similar size school districts (e.g. the white circles) together. This is considered linking based on location [29]. Effectively, teachers deemed to be in similar areas are grouped together. This is especially helpful in smaller school districts where both the number of teachers and number of students per class are small. Other forms of indirect estimation link teachers based on time, e.g. linking the same teacher to previous years of information, or link based on both time and location, e.g. linking teachers based on similar locations and including previous time points.

### 3.2.2 General Form of a Small Area Model

Once we have established a grouping mechanism (e.g. location) we are able to incorporate additional information about the teachers not included in the TVAAS model into the estimation process. Several variables are known about teachers (e.g. years of teaching experience, advanced degree, professional development program participation, etc.); in small area estimation these variables are referred to as auxiliary
variables. A general form of a small area estimation model is

$$z_{kl} = Y_k \gamma + v_k + u_{kl}$$

(3.2.1)

where \(z_{kl}\) is the response variable for teacher \(l\) within small area \(k\); \(Y_k\) contains the auxiliary variables of interest for small area \(k\); \(\gamma\) are the associated parameters of interest for the auxiliary variables identified in \(Y_k\); \(v_k\) represents the error among the small areas and \(u_{kl}\) represents the individual error for teacher \(l\) within each small area \(k\). It is assumed that

$$\begin{bmatrix} v_k \\ u_{kl} \end{bmatrix} \overset{ind}{\sim} \text{Gaussian} \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_V^2 & 0 \\ 0 & \sigma_U^2 \end{bmatrix} \right).$$

The flexibility of this model allows for linking across teachers (e.g. linking teachers from separate schools together), linking within teacher (e.g. linking current year information for teacher \(l\) with previous years of information) or a combination of the two. For our study, we have chosen to focus on two auxiliary variables, level of experience (1-4) and whether or not the teacher has an advanced degree. However, any auxiliary variable deemed to be reasonable can be included. The auxiliary variables chosen for our study are not meant to be exhaustive.

The small area estimation modeling structure can also be viewed as a pseudo-blocking structure. Grouping teachers based on similar characteristics leads to a cluster of similar teachers where the variance within the cluster is small relative to the variance between clusters. The following section discusses the integration of the two models (3.1.2, 3.2.1) presented above.
3.3 **Small Area Multi-Stage Modeling Process**

The small area multi-stage modeling process begins with the TVAAS then incorporates additional information about the teachers by using small area estimation techniques. The iterative process combines variables related to the teachers with student assessment scores in order to obtain more information. Assuming that teacher VAM scores are meaningful, this process aims to address known issues with student assessment data namely imprecise estimates of teacher VA. We seek to determine how incorporating small area estimation affects the resulting teacher estimates and associated precision.

This section introduces the four stages of the small area multi-stage model.

- **Stage One**- Implement TVAAS to obtain estimates of our parameters and predicted teacher VAM scores, $\hat{t}_j$

- **Stage Two**- Use the predicted teacher VAM scores as the response variable for our small area estimation model and obtain new teacher estimates, $\tilde{t}_j$

- **Stage Three**- Combine the new teacher estimates, $\tilde{t}_j$, from Stage Two with the parameter estimates from Stage One to obtain new estimated student scores.

- **Stage Four**- Assess model convergence

### 3.3.1 **Stage One**

The initial phase of this process involves extracting the teacher VAM scores from equation 3.1.2 using BLUP consistent with the Mixed Model Equations 3.1.7. The obtained teacher VAM scores are denoted $\hat{t}_j$ and are stored along with parameter estimates $\hat{\mu}$ and $\hat{\beta}$. During the first iteration of the process, ending with Stage One would result in the TVAAS VAM scores for the teachers.
3.3.2 Stage Two

The teacher VAM scores, \( \hat{t}_j \), obtained in Stage One now act as the response variable. The two auxiliary variables we are considering for analysis are whether or not the teacher has an advanced degree and level of experience (e.g. teachers with 0-1 years of experience are grouped together). Other or additional auxiliary variables considered to be reasonable could be included in the small area model.

The small area model presented in 3.2.1 is adapted to incorporate our auxiliary variables of interest. Teacher predictions \( (\tilde{t}_{jklmn}) \) based on the model are stored. The model for this stage is

\[
\hat{t}_{jklm} = \kappa + \alpha_m + \tau_o + \alpha\tau_{mo} + d_k + s(d)_{kl} + u_{jklm} \implies \tilde{t}_j \tag{3.3.1}
\]

where \( \kappa \) represents the overall mean, \( \alpha_m \) represents the main effect of advanced degree, \( \tau_o \) represents the main effect of level of experience, \( \alpha\tau_{mo} \) represents the interaction between advanced degree and level of experience, \( d_k \) and \( s(d)_{kl} \) are the random effects due to linking teachers based on location (i.e. district and school within district), and \( u_{jklm} \) is the residual for teacher \( j \).

3.3.3 Stage Three

Student scores are updated by incorporating parameter estimates obtained in Stage One and teacher predicted scores obtained in Stage Two with baseline scores for the students. The new teacher predicted values \( \hat{t}_j \) are substituted into the equation from Stage One for \( t_j \) along with the predicted values of \( \hat{\mu} \) and \( \hat{\beta} \). The updated model is presented in 3.3.2 where \( Y \) is used to represent the most recent grade for the
assessment score of student $i$.

\[
\begin{align*}
\hat{s}_{i,1,j^1} & = \mu + X_i \hat{\beta} + \tilde{t}_{j^1} \\
\hat{s}_{i,2,j^1,j^2} & = \mu + X_i \hat{\beta} + \tilde{t}_{j^1} + \tilde{t}_{j^2} \\
& \vdots \\
\hat{s}_{i,Y,j^1,j^2,\ldots,j^Y} & = \mu + X_i \hat{\beta} + \tilde{t}_{j^1} + \tilde{t}_{j^2} + \cdots + \tilde{t}_{j^Y}.
\end{align*}
\]

(3.3.2)

The result is a predicted value for the assessment score of student $i$ in grade $g$ with teachers $J$ as defined for Model 3.1.2.

### 3.3.4 Stage Four

We evaluate model convergence after each iteration. Consider $\bar{s}_{i,g,J}$ to be the student score for the previous iteration and $\hat{s}_{i,g,J}$ to be the score for the current iteration. To determine convergence, we have chosen to use the average relative change in student scores. If the change falls above the convergence criterion, namely

\[
change = \frac{1}{n} \times \frac{1}{Y} \times \sum_{i=1}^{n} \sum_{g=1}^{Y} \left( \frac{\bar{s}_{i,g,J} - \hat{s}_{i,g,J}}{\bar{s}_{i,g,J}} \right) > C
\]

(3.3.3)

where student $i = 1, \ldots, n$ and grade $g = 1, \ldots, Y$, the multi-stage model continues into another iteration. If the convergence criterion falls below $C$, the process terminates and model performance is measured. A discussion of choosing a convergence criterion follows in Section 5.3.3. While we considered the average relative change in student scores as our measure of model convergence, a different metric could be utilized.
3.4 Measuring Model Performance

Often a primary goal with VAM is to rank teachers and take appropriate action. For example high ranking teachers are deemed to have the greatest impact on students and school districts may want to determine what can be learned from their methods. The benefit of studying these models through simulation is that the true teacher VA is known. If an objective is to identify teachers in the top 10% and teachers in the bottom 10%, we know if the modeling method is correctly identifying these teachers. As one measure of model performance, we rank the teacher VAM scores for both the standard TVAAS model and small area multi-stage model. We then compare the predicted teacher scores for both methods with the true known rankings.

Additionally we can determine how precisely the standard TVAAS model and the small area multi-stage model estimate the true value of each teacher VA score. Consider the true teacher VA score for teacher $j$ to be $T_j$ and $\tilde{t}_j$ to be the predicted teacher VAM score found from one of the modeling methods. The mean square prediction error (MSPE) is calculated as

$$MSPE = \frac{1}{w} \times \frac{1}{Y} \times \sum_{j=1}^{w} \sum_{g=1}^{Y} (T_{jg} - \tilde{t}_{jg})^2. \quad (3.4.1)$$

The MSPE is our second measure of model performance and is found for both the TVAAS and small area multi-stage models. In order to obtain the estimated teacher VAM scores, $\tilde{t}_j$, for each model we

1. **TVAAS**: store the teacher estimates after Stage One of the first iteration of the multi-stage model. These are referred to as the predicted scores from the *standard method*, because TVAAS is currently an approach used in practice.

2. **Multi-stage**: store the teacher estimates after Stage Two of the most recent
iteration once the multi-stage model has converged. These are referred to as the predicted scores from the small area method, because the multi-stage model incorporates small area estimation.

Defining the MSPE as in formula 3.4.1 is flexible and allows for cases when the same teacher has taught multiple grades.

For each of the two modeling methods, the teacher effect estimates, \( \tilde{t}_{ij} \), are ranked and organized into deciles. The resulting teacher deciles are compared with the original decile to see if the teachers are properly identified. Then all teacher effect estimates are compared to their true original value and the MSPE resulting from Equation 3.4.1 is recorded.

3.5 Cases for Consideration

Because the linking structure for model 3.2.1 is manipulable through data generation and simulation, the impact of the linking structure is assessed. Specifically, two cases are considered for analysis:

- **Case 1:** Teachers linked by auxiliary variables perform similarly, so linking the teachers together is reasonable.

- **Case 2:** Teachers linked by auxiliary variables perform differently, and so linking teachers together is unreasonable.

Data generated for Case 1 follow the form of Equation 3.3.1.

For Case 2, while the values of teacher auxiliary variables are still known, those variables do not correlate with teacher VA. The data generated for this case utilize the following model

\[ t_{jklmo} = \kappa + d_k + s(d)_{kl} + u_{jklmo}. \] (3.5.1)
Here $\kappa$ represents the overall mean, $d_k$ and $s(d)_{kl}$ are the random effects due to location, and $u_{jklmo}$ is the residual for teacher $j$. In context, teachers generated under Case 2 are grouped by level of experience, but level of experience has no bearing on teacher VA. Considering this case allows us to determine the consequences when the auxiliary variables chosen to model teacher VA are independent of teacher VA. Specifically we want to assess how the small area method performs when key assumptions made about the linking structure are violated.

In the next chapter, we provide a demonstration of the modeling processes for both the TVAAS and small area multi-stage models. We provide an in depth investigation into the two modeling process through simulation in Chapter 5.
CHAPTER 4

A DEMONSTRATION OF TVAAS AND THE MULTI-STAGE MODEL

This demonstration begins by outlining the scenario that gives rise to the data and the data structure. Discussion includes implementation of the TVAAS along with implementation of the multi-stage small area models including the first and final iterations. Relevant results and findings are presented for both models. The demonstration ends with a comparison of the two modeling methods. We implement these models in SAS/STAT software Version 9.4 of the SAS system for Windows [36]. However the models could easily be implemented using different software.

The example that we introduce involves 5 districts each with 6 schools. There are 3 grades at each school each taught by a separate teacher, resulting in 90 total teachers. There are 15 students in each class, thus 450 total students in the sample that are measured for 3 grades, resulting in 1350 scores. The data for this example are simulated by linking teachers based on level of experience and whether or not the teacher has an advanced degree (i.e. under Case 1).

4.1 The Data

There are two sets of data that we are interested in: student data and teacher data. Both are introduced in subsequent subsections.
### 4.1.1 Student Data

Students are assessed for 3 consecutive years and their scores are recorded. We also have information about where the student began (i.e. a baseline) and the teachers that the student has for each grade. Table 4.1 provides an outline of the process that gives rise to the data following the form of *What Would Fisher Do* (WWFD) introduced in Stroup 2012 where SV represents the sources of variation [38].

![Table 4.1: Process Giving Rise to Student Assessment Scores](image)

**What Would Fisher Do**

<table>
<thead>
<tr>
<th>Design Structure SV</th>
<th>df</th>
<th>“Treatment” SV</th>
<th>df</th>
<th>Combined SV</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>teacher</td>
<td>90 - 1</td>
<td>overall mean</td>
<td>1</td>
<td>overall mean</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>baseline</td>
<td>1</td>
<td>baseline</td>
<td>1</td>
</tr>
<tr>
<td>student(teacher)</td>
<td>(15 - 1) * 90</td>
<td>“parallels”</td>
<td>90 * 15 - 3</td>
<td>student(teacher)</td>
<td>1260 - 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>mean, baseline</td>
<td></td>
<td></td>
<td>= 1258</td>
</tr>
<tr>
<td>Total</td>
<td>1349</td>
<td>Total</td>
<td>1349</td>
<td>Total</td>
<td>1349</td>
</tr>
</tbody>
</table>

WWFD process leads to the first layer of our TVAAS model presented in Model 4.1.1.

\[
s_{i1j1} = \mu + X_i \beta + t_{j1} + e_{i1j1}. \tag{4.1.1}
\]

We have 3 years of information for each student, which leads to a model with 3 layers. The full model is presented in 4.1.2

\[
s_{i1j1} = \mu + X_i \beta + t_{j1} + e_{i1j1};
\]

\[
s_{i2j1j2} = \mu + X_i \beta + t_{j1} + t_{j2} + e_{i2j1j2}; \tag{4.1.2}
\]

\[
s_{i3j1j2j3} = \mu + X_i \beta + t_{j1} + t_{j2} + t_{j3} + e_{i3j1j2j3}.
\]
Model 4.1.2 is the same as Model 3.1.5 which is introduced in Section 3.1.2.

Table 4.2 shows a portion of the student data for our example. Information regarding the district and school attended by the student is known and included in the data set. Student scores are separated so that one line of the data set represents information about the student for the grade when the score is observed. We have

<table>
<thead>
<tr>
<th>Obs</th>
<th>district</th>
<th>school</th>
<th>grade</th>
<th>teacher</th>
<th>student</th>
<th>score</th>
<th>baseline</th>
<th>_Z1</th>
<th>_Z2</th>
<th>_Z31</th>
<th>_Z32</th>
<th>_Z61</th>
<th>_Z62</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>98.5245</td>
<td>100.117</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>31</td>
<td>1</td>
<td>99.8781</td>
<td>100.117</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>61</td>
<td>1</td>
<td>96.6832</td>
<td>100.117</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>98.2857</td>
<td>100.117</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>32</td>
<td>1</td>
<td>98.4688</td>
<td>99.453</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>62</td>
<td>1</td>
<td>95.5265</td>
<td>99.453</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 4.2: Portion of Student Data

information about students from two separate schools. For student \( i \) the baseline score remains the same for each grade, because it relates to where that student began. The student data set also includes the columns of the \( Z \) matrix (reduced to save space). Each column represents a specific teacher. Because the TVAAS is a layered model with persistent teacher effect, all previous year teachers are denoted with a 1 as well as the current year. For example, consider student 1 from school 1. This student
was taught by teacher 1 in grade 1. The column in the \( Z \) matrix related to teacher 1, \( Z_1 \), is denoted with a 1 for grades 1, 2 and 3 for this student. In grade 2, the student was taught by teacher 31, and the column related to teacher 31, \( Z_{31} \), is denoted with a 1 for grades 2 and 3. In the final year, the student was taught by teacher 61, and the column related to this teacher, \( Z_{61} \), is denoted with a 1.

Model 4.1.3 is the layered model for student 1 incorporating the data presented in Figure 4.2.

\[
\begin{align*}
 s_{1,1,1^1} &= 98.5245 = \mu + 100.117 \beta + t_{1^1} + e_{1,1,1^1} \\
 s_{1,2,1^1,31^2} &= 99.8781 = \mu + 100.117 \beta + t_{1^1} + t_{31^2} + e_{1,2,1^1,31^2} \\
 s_{1,3,1^1,31^2,61^3} &= 96.6832 = \mu + 100.117 \beta + t_{1^1} + t_{31^2} + t_{61^3} + e_{1,3,1^1,31^2,61^3}
\end{align*}
\] (4.1.3)

We continue with the analysis of this model in Section 4.2.

### 4.1.2 Teacher Data

We have 5 districts, \( d \), with 6 schools per district, \( s \). Each school has 3 teachers one for each grade. Several potential variables relating to teachers are reasonable to include in our analysis. For this example we have chosen to focus on two: whether or not the teacher has an advanced degree, \( \alpha \), and years of teaching experience, \( \tau \), which we have categorized into 4 levels as in Table 5.1.

<table>
<thead>
<tr>
<th>Level ( o )</th>
<th>Years of Experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>( o = 1 )</td>
<td>0-1</td>
</tr>
<tr>
<td>( o = 2 )</td>
<td>2-5</td>
</tr>
<tr>
<td>( o = 3 )</td>
<td>6-10</td>
</tr>
<tr>
<td>( o = 4 )</td>
<td>11+</td>
</tr>
</tbody>
</table>
Table 4.3 provides an outline of the process that gives rise to the teacher data following the WWFD process [38]. The response variable at this stage is obtained utilizing predictable functions when solving the TVAAS model. The predicted values are represented as $\hat{t}_{jklmo}$ in Model 3.2.1 which arises following the WWFD process.

$$\hat{t}_{jklmo} = \kappa + \alpha_m + \tau_o + \alpha \tau_{mo} + d_k + s(d)_{kl} + u_{jklmo} \quad (4.1.4)$$

Here, $\kappa$ represents the overall mean teacher score; $\alpha_m$ represents whether or not the teacher has an advanced degree; $\tau_o$ represents the level of experience; $\alpha \tau_{mo}$ represents the interaction between advanced degree and years of experience; $d_k$ represents the random effect of district $k$; $s(d)_{kl}$ represents the random effect of school $l$ within district $k$; $u_{jklmo}$ is the individual teacher error. Figure 4.4 shows a portion of the teacher data for our example. Notice that our data set does not include the response variable for teacher $j$. We will obtain this value as a result of implementing the TVAAS model. The data set is organized so each line represents a separate teacher.
Teacher $j = 1$ works in district $k = 1$ and school $l = 1$ and teaches grade 1. The teacher has 5 years of teaching experience which corresponds to level $\tau_2$ and does not have an advance degree $\alpha_0$. We implement this model in Section 4.3.

### 4.1.3 True Teacher VA Scores

The benefit of studying these models via simulation is that the true teacher VA scores, $T_j$, are known. Table 4.5 includes a few of the true teacher scores for reference. Knowing the target values allows us to see which modeling method, the standard or small area, estimates the true teacher VA scores more closely.

<table>
<thead>
<tr>
<th>Obs</th>
<th>district</th>
<th>school</th>
<th>grade</th>
<th>teacher</th>
<th>yr</th>
<th>ad</th>
<th>level</th>
<th>true_vam</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>-0.39989</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>31</td>
<td>11</td>
<td>1</td>
<td>4</td>
<td>0.47620</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>61</td>
<td>7</td>
<td>1</td>
<td>3</td>
<td>0.06279</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>-0.75870</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>32</td>
<td>12</td>
<td>1</td>
<td>4</td>
<td>-0.00098</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>62</td>
<td>18</td>
<td>1</td>
<td>4</td>
<td>0.29419</td>
</tr>
</tbody>
</table>

Table 4.5: True Teacher VAM Scores
4.2 Implementing TVAAS

For the student data, we implement the model identifying school, district, student and grade as classification variables. Because the $Z$ matrix is defined within the data set, we can identify the teacher effects as a collection which results in a single effect with multiple degrees of freedom [37]. The fixed effect components of our model are the intercept ($\mu$) and baseline ($X_i$) and the teacher effect ($t_{jg}$) is random. We also define a repeated measure structure implemented using the Cholesky root which is a parametrization of an unstructured covariance matrix that is computationally more efficient. To improve estimation, we have set starting values for each of the random effects.

Our goal after implementing this model is to obtain estimates for the teacher VA scores. First, we investigate the values of the covariance parameters to ensure that our teacher effect estimate is non-trivial. Covariance parameter estimates are provided in Table 4.6. If the effect is estimated to be zero, additional investigation into the modeling process and the data is necessary. One possible solution is to set starting values for covariance parameters which aids in the estimation process. For subsequent

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Subject</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td></td>
<td>0.2276</td>
</tr>
<tr>
<td>CHOL(1,1)</td>
<td>distri<em>school</em>studen</td>
<td>1.4987</td>
</tr>
<tr>
<td>CHOL(2,1)</td>
<td>distri<em>school</em>studen</td>
<td>0.9511</td>
</tr>
<tr>
<td>CHOL(2,2)</td>
<td>distri<em>school</em>studen</td>
<td>1.5421</td>
</tr>
<tr>
<td>CHOL(3,1)</td>
<td>distri<em>school</em>studen</td>
<td>0.7017</td>
</tr>
<tr>
<td>CHOL(3,2)</td>
<td>distri<em>school</em>studen</td>
<td>1.1197</td>
</tr>
<tr>
<td>CHOL(3,3)</td>
<td>distri<em>school</em>studen</td>
<td>1.5008</td>
</tr>
</tbody>
</table>

Table 4.6: Covariance Parameter Estimates after Implementation of TVAAS

analysis we need both the fixed effect parameter estimates for the student model, $\hat{\mu}$
and $\hat{\beta}$, and the teacher VAM scores, $\hat{t}_j$ obtained using BLUP. Table 4.7 shows this relevant information; we selected a few teacher BLUPs for reference. Incorporating

<table>
<thead>
<tr>
<th>Solution for Fixed Effects</th>
<th>Solution for Random Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effect</td>
<td>Estimate</td>
</tr>
<tr>
<td>Intercept</td>
<td>-5.7495</td>
</tr>
<tr>
<td>baseline</td>
<td>1.0590</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.7: Fixed Effect Solutions and Teacher Score Estimates

our findings with Model 4.1.3 for student 1, we have

$$\hat{\mu} = -5.7495, \quad \hat{\beta} = 1.0590, \quad \hat{t}_{11} = -1.0729, \quad \hat{t}_{312} = 0.2767, \quad \hat{t}_{613} = 0.1181.$$  

The values $\hat{t}_1$, $\hat{t}_{31}$ and $\hat{t}_{61}$ are the teacher VAM scores. For the standard modeling method, the modeling process would stop here. The teacher VAM scores $\hat{t}_j$ are saved and for our example will be referred to as the standard teacher VAM scores.

For the multi-stage modeling process the predicted teacher scores, $\hat{t}_j$, serve as our response variable for the teacher data during the first iteration of the model.

### 4.3 Implementing the Small Area Multi-Stage Model

One iteration of the multi-stage modeling process includes the following stages

1. Stage One- Implementing the TVAAS model to obtain predicted teacher VAM scores and estimated parameter values.

2. Stage Two- Incorporating predicted teacher VAM scores with additional information about the teachers to obtain updated teacher VAM scores.

3. Stage Three- Combining parameter estimates from Stage One with updated teacher VAM scores from Stage Two to obtain new predicted student scores.
4. Stage Four- Determining the average relative change in student scores between Stage One and Stage Three to decide if model convergence has occurred.

We will discuss the multi-stage modeling process for the first iteration and the final iteration.

4.3.1 The First Iteration

We will discuss the first iteration of the small area multi-stage model with relevant results, from Stage One to Stage Four.

4.3.1.1 Stage One

For the first iteration, we can use the results of implementing the standard method discussed in Section 4.2. We record the parameter estimates, \( \hat{\mu} \) and \( \hat{\beta} \), along with the predicted teacher VAM scores, \( \hat{t}_j \). The predicted scores, \( \hat{t}_j \) are the response variable for Stage Two of the model.

4.3.1.2 Stage Two

We incorporate the teacher predicted scores \( \hat{t}_j \) with the teacher data discussed in Section 4.1.2. Table 4.8 shows the revised data incorporating the teacher predictions, \( \hat{t}_j \), found in Stage One which are referred to as \( vam \). Notice that the \( vam \) variable in Table 4.8 has the same values that we found for the teacher effect estimates in Table 4.7. We implement Model 4.1.4 with \( vam \) as a function of the fixed effects advanced degree, \( \alpha_m \), and level of experience \( \tau_o \). We identify district, \( d_k \), and school within district, \( s(d)_{kl} \), as random effects.

After model implementation, we have the findings given in Table 4.9. It is important to ensure that we carried out the analysis correctly. Table 4.9 includes the Type 3 test for fixed effects. We can see that the residual degrees of freedom (53) matches
Table 4.8: Teacher Data Including Predicted Teacher VAM Scores

<table>
<thead>
<tr>
<th>Obs</th>
<th>district</th>
<th>school</th>
<th>grade</th>
<th>teacher</th>
<th>yr</th>
<th>ad</th>
<th>level</th>
<th>vam</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>-1.07302</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>-0.59875</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>31</td>
<td>11</td>
<td>1</td>
<td>4</td>
<td>0.27669</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>32</td>
<td>12</td>
<td>1</td>
<td>4</td>
<td>-0.15051</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>61</td>
<td>7</td>
<td>1</td>
<td>3</td>
<td>0.11813</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>62</td>
<td>18</td>
<td>1</td>
<td>4</td>
<td>-0.00496</td>
</tr>
</tbody>
</table>

Table 4.8: Teacher Data Including Predicted Teacher VAM Scores

the degrees of freedom that we calculated through the WWFD process presented in Table 4.3. There are two possible uses for our findings: to determine if the model

Table 4.9: Covariance Parameter Estimates and Evaluation of Fixed Effect Significance

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Subject</th>
<th>Estimate</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>district</td>
<td>0.01814</td>
<td>0.01752</td>
</tr>
<tr>
<td>school</td>
<td>district</td>
<td>0.01126</td>
<td>0.01290</td>
</tr>
<tr>
<td>Residual</td>
<td></td>
<td>0.07941</td>
<td>0.01555</td>
</tr>
</tbody>
</table>

Table 4.9: Covariance Parameter Estimates and Evaluation of Fixed Effect Significance

effects are significant and/or to use estimated values to update our information about the teachers. For this dissertation, we are focusing on the latter because our objective is to obtain updated teacher VAM scores which we need for future analysis.

To calculate the predicted teacher scores, we can use the predicted values (or solution) for each of the components of our model. Table 4.10 includes the fixed effects solution and Table 4.11 includes a portion of the random effects solution. Table 4.8 shows that teacher 1 does not have an advanced degree (\(ad = 0\)), has level 2 experience, and teaches in school 1 of district1. Using the solutions from Tables
Table 4.10: Solution for Fixed Effects

| Effect  | ad  | level | Estimate | Standard Error | DF | t Value | Pr > |t| |
|---------|-----|-------|----------|----------------|----|---------|------|---|
| Intercept | 0.2713 | 0.08705 | 4 | 3.36 | 0.0283 |
| ad | 0 | -0.3099 | 0.06927 | 53 | -3.47 | 0.0010 |
| ad | 1 | 0 | . | . | . | . |
| level | 1 | -0.3901 | 0.1802 | 53 | -2.17 | 0.0349 |
| level | 2 | -0.1392 | 0.1179 | 53 | -1.18 | 0.2430 |
| level | 3 | -0.1097 | 0.1359 | 53 | -0.81 | 0.4200 |
| level | 4 | 0 | . | . | . | . |
| ad*level | 0 | 0.3473 | 0.2477 | 53 | 1.40 | 0.1667 |
| ad*level | 0 | -0.05823 | 0.1659 | 53 | -0.35 | 0.7270 |
| ad*level | 0 | 0.01440 | 0.1918 | 53 | 0.08 | 0.9404 |
| ad*level | 1 | 0 | . | . | . | . |
| ad*level | 1 | 2 | 0 | . | . | . |
| ad*level | 1 | 3 | 0 | . | . | . |
| ad*level | 1 | 4 | 0 | . | . | . |

Table 4.11: Portion of Solution for Random Effects After First Iteration

| Effect  | school | Subject | Estimate | Std Error | DF | t Value | Pr > |t| |
|---------|--------|---------|----------|------------|----|---------|------|---|
| Intercept | 0.1194 | 0.0659 | 53 | -1.38 | 0.1738 |
| school | 1 | district 1 | -0.05142 | 0.0233 | 53 | -0.56 | 0.5799 |
| school | 2 | district 1 | -0.06990 | 0.0171 | 53 | -0.76 | 0.4493 |
| school | 3 | district 1 | 0.08838 | 0.0168 | 53 | 0.96 | 0.3394 |
| school | 4 | district 1 | 0.03888 | 0.0168 | 53 | 0.42 | 0.6739 |
| school | 5 | district 1 | -0.02802 | 0.0295 | 53 | -0.30 | 0.7643 |
| school | 6 | district 1 | -0.05202 | 0.0193 | 53 | -0.57 | 0.5739 |
| school | 1 | district 2 | 0.06364 | 0.0875 | 53 | 0.73 | 0.4705 |
| school | 2 | district 2 | 0.04027 | 0.0168 | 53 | 0.44 | 0.6623 |
| school | 2 | district 2 | 0.03035 | 0.0345 | 53 | 0.32 | 0.7466 |
| school | 3 | district 2 | -0.06614 | 0.0238 | 53 | -0.74 | 0.4640 |
| school | 4 | district 2 | 0.02102 | 0.0200 | 53 | 0.02 | 0.9819 |
| school | 5 | district 2 | 0.04574 | 0.0266 | 53 | 0.49 | 0.6236 |
| school | 6 | district 2 | -0.01081 | 0.0207 | 53 | -0.12 | 0.9080 |
| Intercept | 0.1661 | 0.0666 | 53 | 1.92 | 0.0607 |
| school | 1 | district 3 | 0.1403 | 0.0304 | 53 | 1.51 | 0.1375 |

4.10 and 4.11, we find the predicted value of teacher 1 to be

\[
\tilde{t}_1 = \hat{\kappa} + \hat{\alpha}_0 + \hat{\tau}_2 + \hat{\alpha}_0 \hat{\tau}_0 + \hat{d}_1 + s(d)_{11}
\]

\[
= 0.2713 + -0.3099 + -0.1392 + -0.05823 + -0.1194 + -0.05142
\]

\[
= -0.40685
\]

(4.3.1)

We continue this process for each of the teachers. Table 4.12 provides the predicted VAM scores for a subset of the teachers referred to as \textit{vam\_new} obtained following the process in 4.3.1. The answers vary slightly due to rounding. The relevant information

<table>
<thead>
<tr>
<th>Obs</th>
<th>district</th>
<th>school</th>
<th>grade</th>
<th>teacher</th>
<th>yr</th>
<th>ad</th>
<th>level</th>
<th>vam</th>
<th>vam_new</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>-0.07283</td>
<td>-0.40686</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>31</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>0.27665</td>
<td>0.10051</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>61</td>
<td>7</td>
<td>1</td>
<td>3</td>
<td>0.11813</td>
<td>-0.00621</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>-0.59872</td>
<td>-0.42335</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>32</td>
<td>12</td>
<td>1</td>
<td>4</td>
<td>-0.15053</td>
<td>0.08202</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>62</td>
<td>18</td>
<td>1</td>
<td>4</td>
<td>-0.00496</td>
<td>0.08202</td>
</tr>
</tbody>
</table>

Table 4.12: Updated Teacher VAM Scores
for student 1 in school 1 is

\[ \tilde{t}_1 = -0.40686, \quad \tilde{t}_{31} = 0.10051, \quad \tilde{t}_{61} = -0.00921. \]

We will use the new predicted teacher VAM scores, \( \tilde{t}_j \) to update the student scores in subsequent stages of the multi-stage model.

4.3.1.3 Stage Three

We incorporate parameter estimates obtained from the TVAAS model with new predicted teacher VAM scores to get an updated student score. From Stage One, we found

\[ \hat{\mu} = -5.7495, \quad \hat{\beta} = 1.0590, \]

and from Stage Two, we found

\[ \tilde{t}_1 = -0.40686, \quad \tilde{t}_{31} = 0.10051, \quad \tilde{t}_{61} = -0.00921. \]

This process of updating Model 4.1.3 for student 1 is

\[
\hat{s}_{1,1,1} = \hat{\mu} + X_1 \hat{\beta} + \tilde{t}_1, \\
\hat{s}_{1,2,1,312} = \hat{\mu} + X_1 \hat{\beta} + \tilde{t}_{11} + \tilde{t}_{312}, \\
\hat{s}_{1,3,1,312,613} = \hat{\mu} + X_1 \hat{\beta} + \tilde{t}_{11} + \tilde{t}_{312} + \tilde{t}_{613}. \\
\]

This process is repeated for all students.

4.3.1.4 Stage Four

As we begin the next iteration, the new predicted student scores from Stage Three are used as our response variable for the TVAAS model in Stage One. We define, \( \hat{s}_{i,j} = \hat{s}_{i,1,j}, \hat{s}_{i,2,j}, \hat{s}_{i,3,j} \), and \( \hat{s}_{i,3,1,312,613} = \hat{s}_{i,3,j} \). For the next iteration,
we analyze the following model

\[
\begin{align*}
\bar{s}_{i,1,j^1} &= \mu + X_i \beta + t_{j^1} + e_{i,1,j^1} \\
\bar{s}_{i,2,j^1,j^2} &= \mu + X_i \beta + t_{j^1} + t_{j^2} + e_{i,2,j^1,j^2} \\
\bar{s}_{i,3,j^1,j^2,j^3} &= \mu + X_i \beta + t_{j^1} + t_{j^2} + t_{j^3} + e_{i,3,j^1,j^2,j^3}
\end{align*}
\] (4.3.2)

with the same assumptions from teacher effects and residual as in Model 4.1.1. The baseline, \( X_i \), for student \( i \) remains the same for each iteration and our objective is to estimate the model parameters \( \mu \) and \( \beta \) and obtain estimates of the teacher VAM scores, \( \hat{t}_j \).

4.3.2 Evaluating Convergence

We have chosen to measure convergence by evaluating the average relative change for the predicted student scores between the current iteration of the multi-stage model and the previous iteration of the multi-stage model. Consider \( \bar{s}_{i,g,J} \) to be the predicted student score obtained during Stage Three of the previous iteration of our model. For the first iteration of the model, \( \bar{s}_{i,g,J} \) is the original student score. Let \( \hat{s}_{i,g,J} \) be the predicted student score obtained during Stage Three of the current iteration. Equation 3.3.3, referenced below, is used to calculate the average relative change in predicted student score.

\[
\text{change} = \frac{1}{450} \times \frac{1}{3} \times \sum_{i=1}^{450} \sum_{g=1}^{3} \left( \frac{\bar{s}_{i,g,J} - \hat{s}_{i,g,J}}{\bar{s}_{i,g,J}} \right)
\]

where \( J \) is a compilation of teachers so that \( J = j^1 \) when \( g = 1 \), \( J = j^1j^2 \) when \( g = 2 \), and \( J = j^1j^2j^3 \) when \( g = 3 \). If the average relative change falls above \( C = 0.0001 \), where \( C \) is our convergence criterion, we determine that the model has not converged and we move to the next iteration of the model. We repeat the multi-stage process.
until model convergence is obtained.

### 4.3.3 The Final Iteration

After the model is iterated three times, we determine the average relative change to be

\[
change = \frac{1}{1350} \times \sum_{i=1}^{450} \sum_{g=1}^{3} \left( \frac{\hat{s}_{i,g,J} - \hat{s}_{i,g,J}}{\hat{s}_{i,g,J}} \right) = 0.0000674 < 0.0001
\]

Thus our model has converged.

Selected predicted teacher scores from iteration 3 are presented in Table 4.13 where

<table>
<thead>
<tr>
<th>Obs</th>
<th>district</th>
<th>school</th>
<th>grade</th>
<th>yr</th>
<th>ad</th>
<th>level</th>
<th>teacher</th>
<th>vam</th>
<th>vam_new</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>-0.4450</td>
<td>-0.4345</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>11</td>
<td>1</td>
<td>4</td>
<td>31</td>
<td>0.09386</td>
<td>0.08957</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>7</td>
<td>1</td>
<td>3</td>
<td>-0.01676</td>
<td>-0.02109</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>-0.46219</td>
<td>-0.45219</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>2</td>
<td>12</td>
<td>1</td>
<td>4</td>
<td>32</td>
<td>0.07596</td>
<td>0.07193</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>14</td>
<td>1</td>
<td>4</td>
<td>0.07596</td>
<td>0.07193</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.13: Predicted Teacher VAM Scores after the Final Stage of the Multi-Stage Model

\(vam\) represents the teacher predicted score from iteration 2 and \(vam\_new\) represents the teacher predicted score from iteration 3. The teacher VAM scores from iteration 3, the final iteration, are referred to as the small area teacher VAM scores.

Consider the findings in our example for teacher \(j = 1\). From Table 4.5, we know that the true VAM score for this teacher is \(-0.39989\). The standard TVAAS model predicted that the VAM score for this teacher is \(-1.0729\) (Table 4.7) and the small area multi-stage model predicted that the VAM score for this teacher is \(-0.43455\) (Table 4.13). These are the findings for one teacher. In the next section we will discuss how to evaluate and compare both the standard and small area methods for all teachers in our sample.
4.4 Comparing Results for the Two Modeling Methods

We are using two measures to evaluate the models including

1. Evaluating prediction error between the true teacher VA score and the predicted VAM score for each of the different modeling methods

2. Dividing the VAM scores into deciles and creating frequency tables and heat maps to identify where the teacher falls in reality and where the different modeling methods predict the teacher to be.

These are the two measures that we have considered; other measures of interest could be utilized to compare the methods.

To find the prediction error, we take the true teacher VA score, $T_j$, and subtract the predicted VAM score. For our methods, standard and small area, we define the prediction error as

$$
\text{standard : } T_j - \hat{t}_j, \quad \text{small area : } T_j - \tilde{t}_j.
$$

Figure 4.1 shows box plots of the prediction errors for each of the modeling methods. Each method has one outlier denoted with a circle. We determined that the large outlier for the standard method is teacher $j = 17$ with a prediction error of 0.984 and the large outlier for the small area method is for teacher $j = 43$ which has a prediction error of 1.042.

The mean square prediction error (MSPE) for our example is calculated for each modeling method as

$$
\text{standard : } \frac{\sum_j (T_j - \hat{t}_j)^2}{90}, \quad \text{small area : } \frac{\sum_j (T_j - \tilde{t}_j)^2}{90}.
$$
Figure 4.1 shows that the small area method has a lower MSPE on average than the standard method.

Another commonly used evaluation measure involves the ranking of teachers. We begin by dividing the true teacher VA scores into deciles (identified as *Original*) and then separating the predicted scores for both the standard method and small area method into deciles. Figure 4.2 presents the findings for the standard method and Figure 4.3 presents the findings for the small area method. Evaluating the contingency tables we see that the small area method is better able to identify teachers in the bottom 10% (Decile 1) and top 10% (Decile 10) than the standard method. Analyzing the heat maps, we can see that the small area method is more concentrated than the standard method, implying that the small area method more precisely predicts the true decile for the teachers than the standard method.

In Figure 4.1, we see that the standard method has a large outlier which we identify as teacher \( j = 17 \). Teacher \( j = 17 \) originally falls into Decile 10, and the standard method predicts that the teacher is in Decile 4 which can be seen in the
Figure 4.2: Standard TVAAS Modeling Method Deciles and Heat Map

<table>
<thead>
<tr>
<th>Original Decile</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

Figure 4.3: Small Area Multi-Stage Modeling Method Deciles and Heat Map

<table>
<thead>
<tr>
<th>Original Decile</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

heat map of Figure 4.2. We also see in Figure 4.1, that the small area method has a large outlier which we identify as teacher \( j = 43 \). The original decile for this teacher is 10 and the small area method predicts that the teacher is in Decile 10. While this prediction error leads to a larger MSPE for the small area method, the small area method does correctly identify the teacher when we consider the ranking.

In conclusion, for this example it appears that the two modeling methods perform similarly in regards to MSPE, but considering teacher rankings reveals that the small area multi-stage model more closely predicts the true decile for teacher VA scores. We also acknowledge that while the small area method may perform better in regards to teacher rankings, it only correctly identifies 56% of teachers in the bottom 10% and 78% of teachers in the top 10%. Both methods are struggling to correctly identify the teachers.
CHAPTER 5

A CHARACTERIZATION OF THE TVAAS AND THE MULTI-STAGE MODEL VIA SIMULATION

This simulation study involves generating data consistent with realistic student achievement scores and teacher VA scores, implementing the TVAAS and multi-stage models, compiling findings and analyzing the results. This chapter covers each of these models for both cases discussed in Chapter 3.

5.1 Data Generation

This section discusses the components related to data generation. The process of generating teacher VA scores are discussed first followed by the generating of student assessment scores. The process of defining the layered Z-matrix from 3.1.7 is also discussed.

5.1.1 Generating Teacher Value Added Scores

The initial step in data generation for teacher scores involves defining auxiliary variables and VA scores for the teachers. For Case 1 where linking is reasonable Model 3.3.1 includes fixed effects which need to be specified and random effects which need to be generated. This section discusses both the specification of the fixed components of the teacher VA scores (e.g. the auxiliary variables) and generation of associated mean responses as well as the generation of random effects related to location. We
include brief investigations to ensure that values generated are reasonable.

5.1.1.1 Teacher Auxiliary Variables

The fixed components we are focusing on for our simulation are advanced degree \((ad)\) and level of experience which is related to years of teaching experience \((yr)\). To define reasonable values for our variables, we choose to reference the Nebraska Education Profile which states that in the 2014-2015 school year, 52.16\% of teachers in Nebraska had an advanced degree [27]. In the simulation, to determine if a teacher has an advanced degree, each teacher is assigned a random number, \(deg \sim \text{Uniform}(0, 1)\). The value of \(deg\) then assigns a value to \(ad\)

\[
ad = \begin{cases} 
  deg \leq 0.52, & ad = 1 \\
  \text{else}, & ad = 0 
\end{cases}
\]

If the value of \(deg\) is less than or equal to 0.52, the teacher was deemed to have an advanced degree. We define \(\alpha_m\) to be the fixed effect associated with advanced degree, and it has two levels.

Because years of teaching experience is a count variable that must be greater than or equal to 0, several possible distributions exist to generate plausible values of this variable, including Poisson, generalized Poisson, and Negative Binomial. After much trial and error, we determined that a mixture distribution leads to the most realistic results. First each teacher is assigned a random number, \(cutoff \sim \text{Uniform}(0, 1)\) then based on the value of \(cutoff\), the teacher years of experience variable, \(yr\), uses
the following mixture distribution

\[
yr = \begin{cases} 
  cutoff < 0.05, & yr \sim \text{NegBin}(2, 0.5) \\
  0.05 \leq cutoff < 0.25, & yr \sim \text{NegBin}(3, 0.4) \\
  \text{else}, & yr \sim \text{NegBin}(4, 0.2)
\end{cases} \tag{5.1.1}
\]

Pictorial representation of these three distributions are available in Figure A.1 of the Appendix, and associated SAS code is available in Appendix B.1. Note that the mixture distribution presented in Model 5.1.1 without correction possibly leads to extreme values (e.g. a teacher with 60 years of teaching experience). Thus a correction was applied so that if a teacher is assigned more than 50 years of experience, the year is reduced by 30.

We have chosen to reduce years of service from a continuous random variable to an ordinal one. We found that it is reasonable to consider teachers with similar years of teaching experience to be grouped together. Let \(\tau_o\) be a new variable, level, that splits teachers into 4 groups as defined in Table 5.1. We are considering teachers

<table>
<thead>
<tr>
<th>Level (o)</th>
<th>Years of Experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>(o = 1)</td>
<td>0-1</td>
</tr>
<tr>
<td>(o = 2)</td>
<td>2-5</td>
</tr>
<tr>
<td>(o = 3)</td>
<td>6-10</td>
</tr>
<tr>
<td>(o = 4)</td>
<td>11+</td>
</tr>
</tbody>
</table>

Table 5.1: Converting Years of Service to an Ordinal Variable

with 0–1 year of experience to be new teachers, teachers with 2–5 years to have some experience, teachers with 6–10 years to have moderate experience, and teachers with 11 or more years to be highly experienced. We assume that teachers within each level
will perform similarly, and so we have variable $\tau_o$ which has four levels.

For Case 1, teachers are linked based on whether or not they have an advanced degree and based on level of experience. Essentially the auxiliary variables we generate are meaningful when measuring teacher VA. For this case, the means for both auxiliary variables was set to be

$$
\alpha_1 = 0.2 \ , \ \alpha_0 = -0.2 \ ,
\tau_1 = -0.35 \ , \ \tau_2 = -0.2 \ ,
\tau_3 = -0.05 \ , \ \tau_4 = 0.1 \ .
$$

We have defined these means so that a teacher with an advanced degree will have higher VA than a teacher without an advanced degree. We have set the means for level of experience so that as experience increases teacher VA also increases. The mean values chosen for our simulation are arbitrary and not important. What is important is if the multi-stage model can accurately measure the values. Because advanced degree and level of experience will have a varied impact for each teacher, we allow the mean value for each auxiliary variable to change. This variation results in a residual, $u_{jklmo}$, which follows

$$
\quad u_{jklmo} \sim \text{Gaussian}(0, 0.0625).
$$

A variance of $\sigma^2 = 0.0625$ leads to a standard deviation of $\sigma = 0.25$. Using the empirical rule this implies that the individual teacher variability typically fluctuate $\pm 0.5$. We have chosen this level of variation so it is possible for teachers without an advanced degree to have a lower VA than teachers with an advanced degree. Similarly, it is possible for a teacher with a lower level of experience to have a higher VA than a teacher with more experience.
For Case 2, we have the information provided by the auxiliary variables, but teacher VA is independent of the auxiliary variables. Thus the only portion of the teacher VA score present at this portion of the data generation is the individual unit level variation, \( u_{jklmn} \), which can be seen in Model 3.5.1. We define the residual for this case just as we did for Case 1

\[
 u_{jklmo} \sim \text{Gaussian}(0, 0.0625).
\] (5.1.3)

In the next subsection we proceed with a brief investigation to ensure that the teacher auxiliary variables, associated mean responses, and residual terms lead to data that is realistic.

5.1.1.2 Ensuring Simulated Data Are Reasonable

We generate 90,000 teachers following the process presented in Section 5.1.1.1. Table 5.2 shows the frequency for \( ad \). The percentage of teachers with an advanced degree is 51.98% which is approximately 52%, as defined in the simulation.

<table>
<thead>
<tr>
<th>Advanced Degree</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>( ad = 1 )</td>
<td>46783</td>
<td>51.98</td>
</tr>
<tr>
<td>( ad = 0 )</td>
<td>43217</td>
<td>48.02</td>
</tr>
</tbody>
</table>

Table 5.2: Frequency of Teachers With and Without Advanced Degrees

is 51.98% which is approximately 52%, as defined in the simulation.

Figure 5.1 shows the simulated distribution of years of teaching experience. For our teachers, the simulated mean is 12.89. The Nebraska Education Profile states that the average years of teaching experience in the 2014-2015 school year is 14.34 [27]. However, additional metrics regarding the variability of teaching experience are not readily available. To verify the validity of the distribution presented in Figure 5.1
additional information is obtained from the National Center for Education Statistics (NCES) [25]. The NCES provides information regarding the number of years of full-time teaching experience for public school teachers. The most recent information provided is for the 2011-2012 academic year is Considering the distribution for years

<table>
<thead>
<tr>
<th>Years</th>
<th>Percentage*</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;3</td>
<td>9.0</td>
</tr>
<tr>
<td>3-9</td>
<td>33.3</td>
</tr>
<tr>
<td>10-20</td>
<td>36.4</td>
</tr>
<tr>
<td>&gt;20</td>
<td>21.3</td>
</tr>
</tbody>
</table>

Table 5.3: Years of Teaching Experience for Public School Teachers, 2011-2012 Academic Year

*Percentage of all public school teachers in the U.S.

of service presented in Figure 5.1, our simulation shows the following The findings between our simulation and the national values are similar.
Table 5.4: Years of Teaching Experience for Simulated Teachers

<table>
<thead>
<tr>
<th>Years</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;3</td>
<td>11.28</td>
</tr>
<tr>
<td>3-9</td>
<td>31.17</td>
</tr>
<tr>
<td>10-20</td>
<td>35.49</td>
</tr>
<tr>
<td>&gt;20</td>
<td>22.06</td>
</tr>
</tbody>
</table>

After reducing years of teaching experience to level of experience, we want to ensure that simulated teachers fall into each of the \( ad \times level \) combinations. Table 5.5 shows the percentage of the 90,000 teachers that fall into each category combination. Teachers are falling into each combination, but there are fewer teachers falling into level 1, \( \tau_1 \), which may need to be monitored through the simulation process if estimation issues arise.

This process verifies that the values we choose as our parameter values lead to reasonable data. However, other values could be chosen and justified. The important things is that our model is able to capture the relationship between the variables that we have defined.
5.1.1.3 Design Structure Variables

Both the linking and non-linking models (3.3.1, 3.5.1) presented in Chapter 3 incorporate random effects due to location. We choose to focus on grades 6 through 8 for our simulation. The hierarchical structure we consider for this simulation study mimics the 75 smallest Class 3 middle schools in Nebraska for the 2015-2016 school year[26]. On average these schools have 14.94 students per grade ($SD = 5.4$), which leads to a cohort where all students will have the same sixth, seventh and eighth grade teachers. The design structure for this simulation study is shown in Table 5.6. Five districts are assumed to be linked together; each district includes six schools; each school has one sixth grade, one seventh grade, and one eighth grade class. Each school within a district is assumed to have 15 students.

In Figure 5.6, we are considering the five school districts and 6 schools within

<table>
<thead>
<tr>
<th>District</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>...</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>6th Grade</td>
<td>6th Grade</td>
<td>6th Grade</td>
<td>6th Grade</td>
<td></td>
<td>6th Grade</td>
</tr>
<tr>
<td>7th Grade</td>
<td>7th Grade</td>
<td>7th Grade</td>
<td>7th Grade</td>
<td>...</td>
<td>7th Grade</td>
</tr>
<tr>
<td>8th Grade</td>
<td>8th Grade</td>
<td>8th Grade</td>
<td>8th Grade</td>
<td></td>
<td>8th Grade</td>
</tr>
</tbody>
</table>

Table 5.6: Hierarchical Structure under Consideration
each district to be a sample of all similar districts and school systems. Thus both district \((d_k)\) and school system within district \((s(d)_{kl})\) are considered random effects and have associated distributions. Table 5.7 presents the two conditions we consider for simulating the random effects. For the first condition, referred to as Condition 1,

\[
d_k \sim \text{Gaussian}(0, 0.04)
\]

\[
s(d)_{kl} \sim \text{Gaussian}(0, 0.0625)
\]

Table 5.7: Conditions for Variance of Teacher Random Effects

Condition 1: Low Variances

\[
d_k \sim \text{Gaussian}(0, 0.16)
\]

\[
s(d)_{kl} \sim \text{Gaussian}(0, 0.25)
\]

Condition 2: High Variances

districts are similar as are the schools within each district. Teacher scores generated under Condition 1 vary from about \((-1.5, 1.5)\) once the random effect is incorporated with the fixed effects generated in Section 5.1.1.1. For the second condition, referred to as Condition 2, there is more variability between districts and between schools within each district. Teacher scores generated under this condition vary from about \((-2.5, 2.5)\). Considering the scale of student scores we generate in Section 5.1.2 the ranges for both conditions are reasonable. Teacher VA scores above 2.5 (or below -2.5) are considered extreme. Thus Condition 2 represents the maximum reasonable VA for our study. We have chosen to consider two different variance conditions for our random effects in order to determine their impact on the standard and small area methods in both Cases 1 and 2.

One key assumption for teacher scores with VAM is that the teacher effects are assumed to be approximately normal and centered at zero. We also want to be sure that the teacher effects for Cases 1 and 2 appear to follow the same distribution. This allows for fair comparison of the modeling methods without extraneous variability arising from the data generation process. For the Cases and Conditions under consideration, we have found the distributions presented in Figure 5.2. We can see
Condition 1: Low Variances

Condition 2: High Variances

Figure 5.2: Assessing Simulated VAM Scores Under Both Conditions

that for Condition 1, the teacher effects for both Case 1 and Case 2 appear to be similarly distributed and assuming normality in both cases appears to be reasonable. For Condition 2, the scores are more spread out than for Condition 1. The added variability between school systems has led to a wider distribution. Under Condition 2, the teacher effects for both Cases 1 and 2 appear to be similarly distributed and assuming normality seems reasonable. Assuming all distributions are centered at zero also seems reasonable. For implementing the VAM, the teacher effect assumptions are met.
5.1.1.4 Combining Teacher Auxiliary and Design Structure Variables

Once we have quantities for the auxiliary and design structure variables, we define the teacher VA score, $t_{jklmo}$, for Case 1 as

$$t_{jklmo} = \alpha_m + \tau_o + 2 \times \alpha_m \times \tau_o + d_k + s(d)_{kl}. \tag{5.1.4}$$

We have chosen to place a higher weight on the interaction because assuming that the relationship between advanced degree and level of teaching experience changes over time may be reasonable. We want to determine if the multi-stage model captures this relationship.

For Case 2 we define the teacher VA score, $t_{jklmo}$, as

$$t_{jklmo} = d_k + s(d)_{kl} + u_{jklmo} \tag{5.1.5}$$

where $u_{jklmo}$ is defined as in Equation 5.1.3.

For each simulated data set, the teacher VA scores at this stage are the true scores for the teachers, denoted $T_j$. These scores are ranked and saved for comparison after the implementation of the standard and small area methods.

5.1.2 Generating Student Assessment Scores

Considering three grades of information for each student allows us to use the reduced 3 layer TVAAS model presented in 3.1.5. There are two pieces to the student assessment scores presented in this model: the piece related to the individual student and the piece related to the student’s learning environment, namely the teacher. The following sections discuss the data generation of individual student contribution and the process of combining the student contribution, or raw score, with teacher VA from Section
5.1.1.

5.1.2.1 Student Raw Scores

For data generation purposes it is assumed that each student has a unique baseline score and that assessment scores in subsequent years for each student are correlated. Following Fellers 2014, student baseline scores are generated assuming that student $i$’s baseline score follows the distribution $X_i \sim \text{Gaussian}(100, 4)$ [11]. In practice, this means that students begin at different levels in a classroom. We assume that the assessment is able to capture student’s true knowledge at the time of evaluation and that the assessment allows for students to improve. Specifically, we assume that students are not affected by ceiling effects. For our simulation we are assuming that we have scores for 3 grades for each student. From Model 3.1.5, the portion of the score attributable to an individual student is

$$
\begin{align*}
 r_{i1} &= X_i + e_{i1} \rightarrow \text{Grade 1} \\
 r_{i2} &= X_i + e_{i2} \rightarrow \text{Grade 2} \\
 r_{i3} &= X_i + e_{i3} \rightarrow \text{Grade 3}.
\end{align*}
$$

(5.1.6)

The baseline score $X_i$ is the same for each grade presented in Model 5.1.6. Following Fellers(2014) student errors, $e_{ig}$ are simulated assuming a first order auto-regressive process within each student, specifically

$$
\begin{align*}
 e_{i1} &= \epsilon_{i1} \\
 e_{i2} &= \rho \times e_{i1} + \epsilon_{i2} \\
 e_{i3} &= \rho \times e_{i2} + \epsilon_{i3}.
\end{align*}
$$
where $\epsilon_{ig} \sim \text{Gaussian}(0, 2.25)$ and $\rho = 0.7$ [11]. For our simulation, grades 1, 2 and 3 correspond to the grades 6, 7 and 8.

As mentioned in Section 5.1.1.3, 15 students are generated per school within each district. The 15 students are assumed to have had the same teacher for grade 6, the same teacher for grade 7, and the same teacher for grade 8. Each school within a district has only 15 students, i.e. one cohort, included in the simulation study.

5.1.2.2 Combining Student Raw Scores with Teacher VAM Scores

After data manipulation to ensure that each student has a separate line for each grade (6-8), student raw scores are combined with teacher VA scores generated in Section 5.1.1 using the following process

$$s_{i,g,j} = r_{i1} + t_{j1}$$
$$s_{i,g,j} = r_{i2} + t_{j1} + t_{j2}$$
$$s_{i,g,j} = r_{i3} + t_{j1} + t_{j2} + t_{j3}$$

where $s_{i,g,j}$ is referred to as the student assessment score. For Case 1, the teacher scores generated using Equation 5.1.4 are used for $t_j$, and for Case 2 the teacher scores generated using Equation 5.1.5 are used for $t_j$.

After defining the student assessment scores using Model 5.1.7, we have selected 5 students at random from Case 1 and Condition 1. Figure 5.3 plots the scores for the 5 students and shows the student progress from grades 6 through 8. An increase between grades implies that teacher $j$ for student $i$ in grade $g$ added positive value to the student assessment score. The student grew at a rate above what is expected. A decrease between grades implies that teacher $j$ for student $i$ in grade $g$ added negative value to the student assessment score. The student grew at a rate below
Figure 5.3: Assessment Scores for 5 Randomly Selected Students

what is expected.

5.1.2.3 Replicating the Study and Ensuring Simulated Data is Reasonable

After one trial, we generate information for 5 districts each with 6 schools. The school has 15 students measured in grades 6, 7 and 8 with all 15 students having the same teacher for each grade. 1000 trials are simulated for each Case and Condition. Figure 5.4 shows the generated scores for all students under Case 1 and Condition 1 for each grade. We can see that assessment scores for each grade are similarly distributed around 100, which is expected based on our starting parameters from Section 5.1.2.1. The variability appears to increase slightly as grade increases. Normality of student scores seems reasonable for all grades.

Combining student scores across grades for the 1000 simulated trials is shown in Figure 5.5 for each Case and Condition. Within Condition 1, we can see that means
and standard deviations are very similar for the data generated under the linking and non-linking models. For Condition 2, we can see that means and standard deviations are also similar. The standard deviations are slightly higher for Condition 2 which is expected because districts and school systems are generated to vary more than with Condition 1. Verifying that the means and standard deviations are similar within each condition allows us to infer that any differences related to the modeling method are due to the modeling procedure not violation in model assumptions.

5.1.3 Defining the Z Matrix

Teacher effects are assumed to be persistent without diminishing as demonstrated in Model 5.1.7. The random effect matrix, $Z$, needs to include a separate column.
Condition 1: Low Variances

Condition 2: High Variances

Figure 5.5: Compiled Student Scores Across Grades for Case and Condition.

for each teacher in the trial and include the coefficients associated with the layered effects.

For example, Table 5.8 shows an example where Student 1 has Teacher 1 in grade 6, Teacher 2 in grade 7 and Teacher 3 in grade 8. Linking this example with Equation 5.1.7, for Student $i = 1$, $t_{j1}$ represents the VA score for Teacher 1, $t_{j2}$ represents the VA score for Teacher 2, and $t_{j3}$ represents the VA score for Teacher 3. Student 2 was in different classrooms with teachers 4, 5, and 6. Thus for Student $i = 2$, $t_{j1}$ represents the VA score for Teacher 4, $t_{j2}$ represents the VA score for Teacher 5, and $t_{j3}$ represents the VA score for Teacher 6.
<table>
<thead>
<tr>
<th>Student</th>
<th>Grade</th>
<th>Teacher</th>
<th>t1</th>
<th>t2</th>
<th>t3</th>
<th>t4</th>
<th>t5</th>
<th>t6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 5.8: The \( \mathbf{Z} \) Matrix Incorporated with the Data

Because our simulation includes 5 school districts, 6 schools within district, and 3 grades, we have 90 total teachers per trial, and thus our \( \mathbf{Z} \) matrix has 90 columns. Once the \( \mathbf{Z} \) matrix has been defined, it remains the same for each iteration of the small area multi-stage model.

5.2 Model Analysis

Two modeling methods are utilized for analysis, the standard method and small area method. Model 3.1.5 is the the standard method and the model is analyzed using PROC HPMIXED from SAS/STAT software Version 9.4 of the SAS system for Windows. BLUP is used to obtain the predicted teacher VAM scores.

The first stage of the small area method also follows Model 3.1.5. Subsequent stages of the multi-stage model incorporate small area estimation techniques by analyzing Model 3.3.1 and which is performed using PROC GLIMMIX from SAS/STAT software Version 9.4 of the SAS system for Windows [36]. New teacher predicted scores are obtained based on the parameter estimates for the auxiliary and design structure variables from the small area model.

5.2.1 Generated Data

With each simulation, the number of teachers falling into each \( ad \times level \) combination are saved along with summary statistics for the teacher’s true VAM score. Addition-

ally the raw student scores are stored for each simulation. Full code for this portion of the analysis is provided in Appendix B.2.

5.2.2 Standard Model

Student scores are analyzed using PROC HP MIXED from SAS/STAT software Version 9.4 of the SAS system for Windows [36]. The scores are assumed to be normally distributed. Student baseline is modeled as a fixed effect and teacher scores are assumed be random effects. Because each student is measured for three grades, the scores are repeated measures. In practice, the covariance structure chosen for student scores is unstructured [5, 23, 34]. Assuming this structure allows for minimal assumptions to be made about the relationship between student scores. To improve computation, in our simulation, we have chosen to use the Cholesky structure because it is a special parametrization of the unstructured model where diagonal elements of the matrix are forced to be positive. This reduces computational errors by forcing the covariance structure to be at least positive definite. We investigate the appropriateness of the Cholesky structure in comparison to other possible structures and present the findings in Section 5.3.4. To improve optimization, starting values for the covariance parameters are selected when we estimate model parameters.

BLUPs obtained from this model for the teacher scores are the standard VAM scores used by educational analysts. Teacher BLUPs and parameter estimates for the overall mean and baseline are stored and used for subsequent modeling and analysis. The teacher scores obtained at this point in the modeling process are referred to as the standard teacher scores, since they were obtained using the standard method.

To assess the standard model performance, the standard teacher scores are ordered and ranked. The rankings are then compared to the known values from data generation for the teacher VA scores. Additionally, the MSPE is calculated for each
simulation as introduced in Equation 3.4.1.

5.2.3 Small Area Multi-Stage Model

Teacher BLUPS are merged with original information about the teachers (e.g. level, ad, district, school). Then, using PROC GLIMMIX from SAS/STAT software Version 9.4 of the SAS system for Windows [36], we use the teacher BLUPs as our response variable. Advanced degree and level of experience are our fixed effects. Because we are considering district and school within district to be a sample of all similar entities, we continue to treat district and school within district as random effects. After model estimation has completed, we store the new teacher predicted scores.

We monitor the change in student scores between iterations. If the average change falls below our convergence criterion $C$, defined in Equation 3.3.3, we have extracted all information due to the linking of teachers. If the average change falls above the convergence criterion, we iterate the multi-stage model.

To move to the next iteration, we use the updated student scores as our response variable for the first stage of the multi-stage model. Because we have used estimates to produce new student scores, a covariance structure is no longer needed for the repeated measure; there is no additional parameter $\rho$ correlating the student assessment scores. As with the standard model, starting values are selected for covariance parameters to improve optimization.

Once convergence has been obtained, the teacher scores found in the most recent run are referred to as the small area teacher scores. The MSPE is found for the small area teacher scores and the true teacher scores we generated. Additionally, the small area teacher VAM scores and standard teacher VAM scores are stored along with the true teacher VA scores for each simulation. For this process we have chosen to use $C = 0.0001$. We evaluate the selection of convergence criterion in Section 5.3.3.
Full code for the standard and small area methods can be found in Appendix B.3.

5.3 Results

In this section both the results in regards to teacher rankings and MSPE are presented. In addition the selection of convergence criterion and chosen covariance parameter structure are discussed.

5.3.1 Ranking of Teacher Scores

True teacher rankings are found at the beginning of each simulation (VA scores). Then teachers are ranked at both the end of the standard method and the small area method (VAM scores). Teacher rankings are aggregated into deciles. A lower decile represents a teacher with a lower VA or VAM score. If a modeling method performs perfectly, we would find the results presented in Figure 5.6 which includes a frequency table comparing the original decile for the teacher VA score and the final decile for the teacher VAM score. Additionally a heat map for the frequency table is included in the Figure.

<table>
<thead>
<tr>
<th>Original Decile</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>100%</td>
</tr>
<tr>
<td>9</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>100%</td>
</tr>
<tr>
<td>8</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>100%</td>
<td>0.0%</td>
</tr>
<tr>
<td>7</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>100%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>6</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>100%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>5</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>100%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>4</td>
<td>0.0%</td>
<td>0.0%</td>
<td>100%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>3</td>
<td>0.0%</td>
<td>100%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>2</td>
<td>0.0%</td>
<td>100%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>1</td>
<td>100%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

Figure 5.6: Resulting Deciles for Hypothetical Perfect Model

Of the teachers known to be in the bottom 10% (Original Decile 1), 100% of them are correctly identified as being in the bottom 10% by the perfect model. For the heat map, higher percentages are represented with darker colors and the lower percentages are represented with the lightest colors. The heat map is completely dark
along the diagonal with no shading, because the model is perfectly identifying the teachers. While the perfect model may not be reasonable in practice, it does provide a point of comparison for all subsequent cases and conditions.

Considering Case 1, where teacher VA scores are generating using linking, and Condition 1, low variability across districts and schools, Figure 5.7 shows the percentage of teacher VA scores that began in each decile and where the standard method places the teacher VAM score. The Figure includes the results for 1000 simulations. In Table 5.7 of teachers known to be in the bottom 10% (Original Decile 1), 48.9% of them are correctly identified by the standard method as being in the bottom 10%; the standard method also incorrectly identifies 23.7% of the teachers known to be in the bottom 10% as being between the bottom 10% and 20%. We can see more clearly from the heat map that the standard modeling method best identifies teachers in the bottom 10% (Original Decile 1) and in the top 10% (Original Decile 10). However, there is a lot of variability in prediction over teachers in the middle (Deciles 4-6).

Figures 5.8-5.11 show the frequency tables and heat maps for each modeling method under the cases and conditions considered. We can see for the low variance condition in Figures 5.8 and 5.9 that both methods are struggling to correctly identify the true teacher decile. This is especially apparent at the middle deciles where the shading in the heat maps is wide. For Case 1 presented in Figure 5.8, the small
area method is performing slightly better than the standard method at the extremes (Deciles 1 and 10). The small area method is also more focused over the middle deciles (3–8) than the standard method, however both are struggling with prediction accuracy. For Case 2, presented in Figure 5.9, the two methods appear to be performing similarly. The standard method is slightly better than the small area method for both the top 10% and bottom 10% of teachers (Deciles 1 and 10). As with Figure 5.8, both methods are still struggling over the middle deciles (3–8) to correctly identify the teachers.

Figures 5.8 and 5.9 are for Condition 2, where there is higher variability between districts and schools. Considering Case 1 which is presented in Figure 5.8, the small area method better identifies teachers in the bottom 10% and top 10% than the standard method. Also, the small area method more precisely estimates the teacher scores over the middle deciles (3-8) which is more apparent in the heat map, because the gray shading is more narrow than for the standard method. For Case 2 which is presented in Figure 5.11, the two modeling methods are performing similarly when identifying the top 10% of teachers. However, the small area method appears to be slightly better than the standard method at identifying teachers at all deciles. There is more variability in prediction for both methods over middle deciles (3–8).

Comparing the two conditions, Figures 5.10 and 5.11 under the condition with higher variability between districts and schools show that both the standard and small area modeling methods perform better than under the low variance condition shown in Figures 5.8 and 5.9. For example, considering the teachers correctly identified in the top 10% (Decile 10) is higher for both cases and methods under Condition 2 than under Condition 1. Increased variability between districts and schools has improved the correct identification of teachers in the top 10%, bottom 10% and across the middle deciles.
Figure 5.8: Results for Both Modeling Methods Under Case 1 and Condition 1

Figure 5.9: Results for Both Modeling Methods Under Case 2 and Condition 1
Figure 5.10: Results for Both Modeling Methods Under Case 1 and Condition 2

Figure 5.11: Results for Both Modeling Methods Under Case 2 and Condition 2
5.3.2 Mean Square Prediction Error

The MSPE is calculated for each modeling method at the end of each of 1000 simulations. The resulting MSPEs from the simulations are provided for each case and condition in Figure 5.12. For Case 1, the average MSPE is lower for the small area modeling method than for the standard method. For Case 2, the average MSPE is lower for the standard modeling method than for the small area method for Condition 1. For Case 2 and Condition 2, there is not a clear difference between the two methods on average. For Condition 1 we defined the variance between districts and schools within each district to be smaller. We can see that the variability within Condition 1 is smaller for both modeling methods than for Condition 2, which is expected. Also, we can see that the small area modeling method is more likely to have outliers than the standard method pulling the mean far above the median.

Figure 5.12: Comparing MSPE for Modeling Method Based on Cases and Conditions used for Generating Data
Because Figure 5.12 shows that MSPE is highly skewed, assuming normality is unreasonable. Thus we choose to model MSPE with the Gamma distribution which is continuous, bounded by zero and commonly skewed. All three of these conditions are necessary to model MSPE. To formally assess if the two methods are performing significantly different, we analyze Model 5.3.1 which follows the notation presented in Stroup [38]

\[ m_{pqrs} \sim \text{Gamma}(\psi_{pqrs}, \varphi) \]

\[ \eta_{pqrs} = \log(\psi_{pqrs}) \]

\[ \eta_{pqrs} = \eta + \lambda_p + \delta_q + \nu_r + \lambda\delta_{pq} + \lambda\nu_{pr} + \delta\nu_{qr} + \lambda\nu\delta_{pqr} \]

where \( m_{pqrs} \) represents the MSPE for case \( p \), condition \( q \), modeling method \( r \), and simulation \( s \); \( \psi_{pqrs} \) represents the mean and \( \varphi \) the variance; \( \eta_{pqrs} \) is the link function and \( \eta \) is the overall mean on the model scale; \( \lambda_p \) represents the main effect of condition \( p \) which has two levels, low variance and high variance; \( \delta_q \) represents the main effect of case, which has two levels, linking and non-linking; \( \nu_r \) represents the main effect of modeling method \( r \) which has two levels, standard and small area. The two and three-way interaction terms are also included in the model. We choose to use restricted subject specific pseudo-likelihood (RSPL) as our method of estimation due to the findings in Couton and Stroup 2013 that Type I error is adequately controlled and coverage probabilities are reasonable [9].

We utilize PROC GLIMMIX from SAS/STAT software Version 9.4 of the SAS system for Windows [36] to carry out our analysis. Figure 5.13 is a mean plot on the data scale that shows the confidence intervals for each method under the different cases and conditions. A smaller MSPE is preferred because it implies we have a lower error in our predicted teacher scores in comparison to the true scores.
both cases with linking, the small area modeling method performs better than the standard modeling method. For the low variance condition, the standard modeling method performs better with non-linking than the small area modeling method. A difference between the two methods at the high variance condition with non-linking is not apparent from this plot alone.

Table 5.9 shows the confidence intervals for the true average difference between modeling methods for each of the cases and conditions. We can see that for Case 1 and Condition 1, the standard method has an MSPE that is 0.1106 to 0.1615 higher than the small area method. The standard method has a higher MSPE on average for Case 1 regardless of condition. For Case 2 and Conditions 1, the small area method has an MSPE that is 0.1392 to 0.1901 higher than the standard method on average. The difference between the two methods for Case 2 and Condition 2 is much smaller. The small area method has a higher MSPE by 0.0102 to 0.0612 on average than the
Simple Effect Comparison of condition*case*method
Least Squares Means by condition*case

<table>
<thead>
<tr>
<th>Condition</th>
<th>Case</th>
<th>Method1</th>
<th>Method2</th>
<th>Estimate</th>
<th>SE</th>
<th>Alpha</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>standard</td>
<td>sm_area</td>
<td>0.1361</td>
<td>0.01298</td>
<td>0.05</td>
<td>0.1106</td>
<td>0.1615</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>standard</td>
<td>sm_area</td>
<td>-0.1647</td>
<td>0.01298</td>
<td>0.05</td>
<td>-0.1901</td>
<td>-0.1392</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>standard</td>
<td>sm_area</td>
<td>0.1738</td>
<td>0.01298</td>
<td>0.05</td>
<td>0.1484</td>
<td>0.1993</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>standard</td>
<td>sm_area</td>
<td>-0.0357</td>
<td>0.01298</td>
<td>0.05</td>
<td>-0.0612</td>
<td>-0.0102</td>
</tr>
</tbody>
</table>

Table 5.9: Confidence Intervals for the True Average Difference Between Modeling Methods

standard method.

Table 5.10 shows the confidence intervals for the true average difference between the cases for each of the modeling methods. We can see that for the standard method,

Simple Effect Comparison of condition*case*method
Least Squares Means by condition*case

<table>
<thead>
<tr>
<th>Method</th>
<th>Case</th>
<th>Case</th>
<th>Estimate</th>
<th>SE</th>
<th>Alpha</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>standard</td>
<td>Case1</td>
<td>Case2</td>
<td>0.0655</td>
<td>0.009179</td>
<td>0.05</td>
<td>0.0475</td>
<td>0.0835</td>
</tr>
<tr>
<td>sm_area</td>
<td>Case1</td>
<td>Case2</td>
<td>-0.1896</td>
<td>0.009179</td>
<td>0.05</td>
<td>-0.2076</td>
<td>-0.1716</td>
</tr>
</tbody>
</table>

Table 5.10: Confidence Intervals for the True Average Difference Between Modeling Methods

the MSPE is lower on average when teacher VA scores are generated without using linking (Case 2). The small area method has a significant reduction in MSPE on average when linking is used to generate teacher VA scores (Case 1).

5.3.3 Impact of Convergence Criterion

For our analysis, we chose to use a convergence criterion of $C = 0.0001$. We also assessed the impact of choosing a more lenient criterion, $C = 0.001$, and a more strict criterion $C = 0.00001$. For all three criteria, the results of the standard model are the same; only the multi-stage model is impacted.
We focus on Condition 1 and Case 1. Figure 5.14 shows the number iterations required for the multi-stage model under each of the three criteria. As the strictness of our criterion increases, the number of iterations increases.

Figure 5.15 provides a box plot for the MSPE for each of the convergence criterions. As the convergence criterion increases, it appears that the average MSPE increases as does the number of outliers.

We further investigate the impact of the additional iterations on MSPE that we discussed in Section 5.3.2. Table 5.11 presents the median and mean estimates of MSPE as well as 95% confidence intervals for average MSPE under each of the convergence criterions. We find that as we increase the strictness of criterion, our average MSPE increases. Thus we choose a moderate criterion, 0.0001 to benefit from teacher
Figure 5.15: Box plots for the MSPE for Each of the Convergence Criteria

<table>
<thead>
<tr>
<th>Criterion Used</th>
<th>Median</th>
<th>Mean</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.001</td>
<td>0.0648</td>
<td>0.0672</td>
<td>0.0661</td>
<td>0.0683</td>
</tr>
<tr>
<td>0.0001</td>
<td>0.0664</td>
<td>0.0702</td>
<td>0.0691</td>
<td>0.0714</td>
</tr>
<tr>
<td>0.00001</td>
<td>0.0672</td>
<td>0.0717</td>
<td>0.0706</td>
<td>0.0729</td>
</tr>
</tbody>
</table>

Table 5.11: Impact of Choice of Convergence Criterion on MSPE

information while not over-modeling the teacher data.

5.3.4 Evaluation of Covariance Structure

The TVAAS typically uses an unstructured covariance matrix (UN) for the student repeated measure [5, 23, 34], however a first order autoregressive AR(1) structure is also found in the literature as a possible covariance structure [11]. Our data is simulated using an AR(1) structure so it seems likely that this would be the best covariance structure for our data.

We we have chosen to use a Cholesky (Chol) covariance structure. PROC HP-
MIXED from SAS/STAT software Version 9.4 of the SAS system for Windows offers four options for repeated measures covariance structures: UN, Chol, AR(1), and variance component only (VC) [36]. We investigate the selection of covariance structure for Condition 1 and Case 1. One method of covariance structure selection is to identify the structure with the lowest corrected Akaike’s Information Criterion (AICC). We perform a brief investigation with 100 simulated implementations of the standard TVAAS model. Table 5.12 shows the AICC averaged over the 100 simulations. Both

<table>
<thead>
<tr>
<th>Covariance Structure</th>
<th>AICC</th>
</tr>
</thead>
<tbody>
<tr>
<td>UN</td>
<td>5014.65</td>
</tr>
<tr>
<td>Chol</td>
<td>5014.65</td>
</tr>
<tr>
<td>AR(1)</td>
<td>5050.93</td>
</tr>
<tr>
<td>VC</td>
<td>5454.84</td>
</tr>
</tbody>
</table>

Table 5.12: Identifying Appropriate Covariance Structure for Repeated Measure in Standard Model

UN and Chol have the lowest AICC of all structures and the AICC for UN and Chol is identical. This is not unexpected, because Chol is a re-parameterization of UN. Because UN can have issues defining covariance parameters, e.g. issues with the covariance matrix not being positive definite, Chol is a great alternative that ensures the variance-covariance matrix will be at least positive semi-definite [37]. Additionally, Chol and UN make the fewest assumptions about the covariance parameters, allowing each variance and covariance parameter to be different. Thus our choice of modeling the repeated measure structure for student scores with Chol is reasonable.
5.4 Conclusions

This section discusses our conclusions for both the teacher rankings and MSPE approaches to model comparison.

5.4.1 Teacher Rankings

The methods more precisely identify teachers in the bottom 10% and the top 10% (Deciles 1 and 10) while there is less precision in identifying teachers in the middle (Deciles 3-8).

In cases where teachers are generated using linking, the small area multi-stage model performs better than the standard method. The percentages are higher in the tables and for the heat map the extremes are darker for the small area modeling method. Teachers in the top and bottom are more frequently identified with this method. The gradient is also more focused over middle levels with the small area modeling method, this is notable in both the low variance and high variance conditions.

The standard method appears to be slightly better than the small area method for the non-linking case with low variance condition. The table shows that teachers in the extremes are better identified with the standard method and the heat map shows that the extremes are darker. However the difference between the standard and small area modeling methods for the non-linking case with high variance condition is less apparent. In fact the small area method appears to be slightly better at identifying teachers at all deciles than the standard method.

In conclusion, if a linking structure is reasonable to assume, the small area multi-stage model appears to improve the accuracy of the standard TVAAS model. The standard method performs better than the small area method when variability be-
tween districts and schools is lower and linking is not reasonable to assume; using the multi-stage model after the standard TVAAS model leads to less accuracy. The best method is not clear for the high variability condition when linking is not reasonable to assume. Regardless, both models perform better under conditions where there is higher variability between districts and schools within districts.

While one method may be performing better than the other, there are still issues with correctly identifying the teachers. This is especially noticeable for the low variance condition, where the methods across both cases are still misidentifying 40–55% of teachers in the top 10% and 40–55% of teachers in the bottom 10%.

### 5.4.2 Mean Square Prediction Error

In all cases and conditions, the small area method is more likely to have extreme outliers than the standard method. This causes the average MSPE to be shifted especially in the high variance condition. Consequently modeling MSPE as a Gaussian random variable is unreasonable.

Further investigation into the true average difference between the methods revealed that the small area method has a significantly lower MSPE on average than the standard method under cases where teachers VA scores are generated using linking. The standard method has a significantly lower MSPE on average than the small area method under the low variance condition when teacher VA scores are generated without using linking. However, there is not a relatively large difference between the two methods under the high variance condition when teacher VA scores are generated without using linking in comparison to each of the other cases and conditions. For the standard method, the non-linking structure leads to a lower MSPE on average and for the small area method, the average MSPE is significantly lower when the teacher scores are generated using the linking structure.
In conclusion, if assuming a linking structure is reasonable, the small area method is preferable to the standard method. If the linking is unreasonable, the standard method is preferred when districts and schools are similar. As the variability between schools increases, the difference between the methods is less apparent.

## 5.5 Additional Investigations

In this section we conduct an investigation into the impact of simulating more students per school on both the standard and small area modeling methods. We also discuss our plan for future research.

### 5.5.1 Larger Number of Students Per Grade

Revisiting information about school systems in Nebraska reveals that of Class 3 schools, the smallest 200 have 32.46 students on average ($SD = 1.63$) [26]. We investigate the performance of the standard TVAAS and small area multi-stage modeling methods with 30 students per class (referred to as Design 2) rather than 15 as in Section 5.1.1.3 (referred to as Design 1).

We assess the correct identification of teacher rankings as analyzed in Section 5.3.1. Figures 5.16–5.19 summarize our findings for each of the cases and conditions under consideration. Both methods better identify teachers under Condition 2 in Figures 5.18 and 5.19 where 68%–76% of teachers in the bottom 10% (Decile 1) and 68%–76% of teachers in the top 10% (Decile 10) are correctly identified. Teachers over middle levels (Deciles 3–8) are more precisely estimated under this condition which is visible in the heat map with darker shading along the diagonal. For Condition 1 in Figures 5.16 and 5.17 only 51%–64% of teachers in the bottom 10% and 52%–66% of teachers in the top 10% are correctly identified. There is a lot of variability in identifying teachers over middle levels seen in the heat map where the middle diagonal is less
Figure 5.16: Results for Both Modeling Methods Under Case 1 and Condition 1 for Design 2

<table>
<thead>
<tr>
<th></th>
<th>Original Decile</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.2%</td>
<td>0.7%</td>
<td>2.4%</td>
<td>7.4%</td>
<td>23.1%</td>
<td>66.1%</td>
</tr>
<tr>
<td>2</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.2%</td>
<td>0.7%</td>
<td>2.4%</td>
<td>7.4%</td>
<td>23.1%</td>
<td>66.1%</td>
</tr>
<tr>
<td>3</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.2%</td>
<td>0.7%</td>
<td>2.4%</td>
<td>7.4%</td>
<td>23.1%</td>
<td>66.1%</td>
</tr>
<tr>
<td>4</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.2%</td>
<td>0.7%</td>
<td>2.4%</td>
<td>7.4%</td>
<td>23.1%</td>
<td>66.1%</td>
</tr>
<tr>
<td>5</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.2%</td>
<td>0.7%</td>
<td>2.4%</td>
<td>7.4%</td>
<td>23.1%</td>
<td>66.1%</td>
</tr>
<tr>
<td>6</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.2%</td>
<td>0.7%</td>
<td>2.4%</td>
<td>7.4%</td>
<td>23.1%</td>
<td>66.1%</td>
</tr>
<tr>
<td>7</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.2%</td>
<td>0.7%</td>
<td>2.4%</td>
<td>7.4%</td>
<td>23.1%</td>
<td>66.1%</td>
</tr>
<tr>
<td>8</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.2%</td>
<td>0.7%</td>
<td>2.4%</td>
<td>7.4%</td>
<td>23.1%</td>
<td>66.1%</td>
</tr>
<tr>
<td>9</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.2%</td>
<td>0.7%</td>
<td>2.4%</td>
<td>7.4%</td>
<td>23.1%</td>
<td>66.1%</td>
</tr>
<tr>
<td>10</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.2%</td>
<td>0.7%</td>
<td>2.4%</td>
<td>7.4%</td>
<td>23.1%</td>
<td>66.1%</td>
</tr>
</tbody>
</table>

Figure 5.17: Results for Both Modeling Methods Under Case 2 and Condition 1 for Design 2

<table>
<thead>
<tr>
<th></th>
<th>Original Decile</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>2</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>3</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>4</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>5</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>6</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>7</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>8</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>9</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>10</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Original Decile</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>2</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>3</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>4</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>5</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>6</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>7</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>8</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>9</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>10</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Original Decile</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>2</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>3</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>4</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>5</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>6</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>7</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>8</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>9</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>10</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>
Figure 5.18: Results for Both Modeling Methods Under Case 1 and Condition 2 for Design 2

<table>
<thead>
<tr>
<th>Original Decile</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Standard Decile</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Small Area Decile</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

Figure 5.19: Results for Both Modeling Methods Under Case 2 and Condition 2 for Design 2
focused and shading has a wider spread.

For Case 1 in Figures 5.16 and 5.18, the small area method better identifies teachers in the bottom 10% and top 10% than the standard method. The heat map shows that the teacher scores are more focused along the diagonal for the small area method than the standard method for this case.

For Case 2 in Figures 5.17 and 5.19, the standard method better identifies teachers in the bottom 10% and top 10% than the small area method. The improvement for Condition 2 is slight. For Condition 1, the standard method is slightly more precise when estimating middle level teachers than the small area method. For Condition 2, the variability in prediction over middle levels is similar between the two methods.

Both methods improve for all cases and conditions under Design 2 in comparison to Design 1 given in Figures 5.8–5.11. The methods are better able to identify teachers in the bottom 10% and top 10%. The standard method improves by 6%-10% when identifying teachers in the bottom 10% or top 10% while the small area method improves by 3%-6% when identifying teachers in the same categories. While one method may perform better under certain cases or conditions, the methods are still misidentifying 24%-49% of teachers in the bottom 10% or top 10%.

We assess the impact of the larger number of students on MSPE as analyzed in Section 5.3.2. Figure 5.20 shows the resulting MSPE for each of the cases and conditions we investigate for Design 2. For comparison we have included Figure 5.12 from Section 5.3.2 which shows the findings for Design 1.

For Design 2, we find that for each case and condition the average MSPE is lower for both methods in comparison to Design 1. The standard method for Design 2 appears to be similar to the small area method on average for Case 1 and significantly lower than the small area method on average for Case 2. For the standard method the MSPEs are more precise under Design 2 than Design 1. The small area method
under Design 2 is more likely to have extreme outliers than the standard method. The median MSPE for the small area method is lower than the standard method for Case 1. The variability for the small area method is similar between Design 2 and Design 1.

We proceed with analysis of Model 5.3.1. Figure 5.21 presents the confidence intervals for the different modeling methods in each of the cases and conditions considered under Design 2. For cases with linking, a significant difference between the two modeling methods is not apparent; however in cases with non-linking the standard method notably reduces MSPE on average.
The confidence intervals for the true average difference between methods are presented for each combination of Case and Condition in Table 5.13. Further investigation into performance of the methods for the different linking cases yields the confidence intervals presented in Table 5.14. For the standard method, the average difference between the two cases is much smaller relative to the average difference between the two cases for the small area method. The linking structure used to generate teacher VA has a much larger impact on the small area method than the standard method.

In conclusion, for Design 2, the method of utilizing the teacher VAM scores impacts the decision as to which method is preferred. If the objective is to identify teachers using ranks, the small area method shows promise for cases where linking is reasonable to assume. If the objective is to identify the true teacher VA score, the standard method may be preferable.
Simple Effect Comparison of condition*case*method
Least Squares Means by condition*case

<table>
<thead>
<tr>
<th>Condition</th>
<th>Case</th>
<th>Method1</th>
<th>Method2</th>
<th>Estimate</th>
<th>SE</th>
<th>Alpha</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>standard</td>
<td>sm_area</td>
<td>0.0633</td>
<td>0.01538</td>
<td>0.05</td>
<td>0.0331</td>
<td>0.0934</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>standard</td>
<td>sm_area</td>
<td>-0.3534</td>
<td>0.01538</td>
<td>0.05</td>
<td>-0.3836</td>
<td>-0.3233</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>standard</td>
<td>sm_area</td>
<td>0.0224</td>
<td>0.01538</td>
<td>0.05</td>
<td>-0.0078</td>
<td>0.0525</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>standard</td>
<td>sm_area</td>
<td>-0.2519</td>
<td>0.01538</td>
<td>0.05</td>
<td>-0.2820</td>
<td>-0.2217</td>
</tr>
</tbody>
</table>

Table 5.13: Confidence Intervals for the True Average Difference Between Modeling Methods

Simple Effect Comparison of condition*case*method
Least Squares Means by condition*case

<table>
<thead>
<tr>
<th>Method</th>
<th>Case</th>
<th>Case</th>
<th>Estimate</th>
<th>SE</th>
<th>Alpha</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>standard</td>
<td>1</td>
<td>2</td>
<td>0.0565</td>
<td>0.01088</td>
<td>0.05</td>
<td>0.0352</td>
<td>0.0778</td>
</tr>
<tr>
<td>sm_area</td>
<td>1</td>
<td>2</td>
<td>-0.2890</td>
<td>0.01088</td>
<td>0.05</td>
<td>-0.3103</td>
<td>-0.2677</td>
</tr>
</tbody>
</table>

Table 5.14: Confidence Intervals for the True Average Difference Between Modeling Methods

5.5.2 Future Research

For future research, we are interested in determining how well the modeling methods perform as the linking structure becomes more informative. Also, we want to investigate how the modeling methods perform under known issues with VAM, e.g. in the presence of ceiling effects. Finally we plan to investigate how incorporating small area estimation impacts the ability of TVAAS to predict future teacher VA scores.
CHAPTER 6

CONCLUSIONS

VAMs in the literature currently include examples from major metropolitan areas (e.g. Los Angeles and New York City) or statewide initiatives (e.g. Tennessee). There is no research available that specifically focuses on smaller schools within metropolitan areas or schools in rural areas. Through our research, we have assessed how traditional VAMs perform in smaller school systems. We also propose a new multi-stage modeling methodology that incorporates small area estimation techniques with a traditional VAM.

While VAMs have been shown to provide stable estimates for schools with a large number of students, we have found issues with these models correctly identifying teachers in cases where schools have a small number of students. Our research provides an avenue for administrators in small school systems to assess teacher impact on student growth. The multi-stage model shows promise especially when determining teachers in the extremes (e.g. top 10% and bottom 10%).

However, additional research is needed before this model should be implemented in school systems. Investigation is needed in regards to model convergence. Currently the changes between iterations (especially an increase in MSPE as the convergence criterion becomes more strict) are troubling. Additionally, we need to partner with small school districts to ensure that generated data accurately represents student
achievement test scores and teacher VA scores for these schools.

Finally, additional analyses into cases where modeling assumptions are violated (e.g., the presence of ceiling effects) is needed to assess the sensitivity of parameter estimates obtained from our small area multi-stage model.
BIBLIOGRAPHY


GLOSSARY OF TERMS

**Auxiliary Variables:** Variables known about subjects in a study that provide additional information after the subjects have been linked together and are not accounted for in the TVAAS model

**Best Linear Unbiased Prediction (BLUP):** A method of estimating random effects included in mixed models and obtaining predicted values for the random variables

**Case 1:** The case where linking is used to generate teacher value added scores

**Case 2:** The case where no-linking structure is used to generate teacher value added scores

**Condition 1:** Variability between districts and schools within each district is small

**Condition 2:** Variability between districts and schools within each district is large

**Design 1:** Data is generated with 15 students per school

**Design 2:** Data is generated with 30 students per school

**Mean Square Error (MSE):** A measure of model performance

**Mean Square Prediction Error (MSPE):** The squared error in prediction. In this dissertation, MSPE is found by taking the true teacher value added and
subtracting the teacher value added model score obtained through one of the modeling methods. This difference is squared and averaged over all teachers.

**Original Decile:** For the simulated data, the decile where the true teacher value added score falls

**Small Area Estimation:** An area of Statistics that has developed methodology when samples are small and typical estimation methods lead to imprecise estimates

**Small Area Method:** In this dissertation, the small area multi-stage model approach to modeling student assessment data which includes four stages

**Standard Method:** In this dissertation, the Tennessee Value Added Assessment System approach to modeling student assessment data

**Tennessee Value Added Assessment System (TVAAS):** A value added modeling approach introduced in Tennessee in which teacher effects persist in subsequent grades without diminishing, and student scores are analyzed using repeated measures with an unstructured covariance matrix

**Value Added (VA):** The true value added by a teacher. This is a parameter to be estimated through Value Added Modeling

**Value Added Model (VAM):** A process of modeling student assessment scores over time by treating teacher contribution to a student score as added value

**Value Added Model Score (VAM score):** The estimated value added by a teacher. This is a statistic obtained after a value added model is analyzed and best linear unbiased prediction is used
APPENDIX A

ADDITIONAL FIGURES OF INTEREST

Figure A.1: Three Negative Binomial Distributions Used to Generate Years of Teaching Experience

Section 5.1.1.1: These distributions were considered as a mixture distribution which is presented in 5.1.1. A different combination or alternate parametrizations could certainly be considered.
APPENDIX B

SAS CODE

B.1 Setup Macro

This section includes the macro used to simulate all variables for Case 1 and the changes that were made for Case 2.

B.1.1 Simulate Variables and Scores for Case 1

%MACRO setup1(sim, seed, ndist, nscho, nteach, dege, yre, degvar, yrvar,
    dist_var, scho_var, ngrade, nstud, stuvar, arrho);
/*Define Auxiliary Variable Values for Teachers*/
data var1;
    call streaminit(&seed-1);
    do district = 1 to &ndist;
        do school = 1 to &nscho;
            do grade = 1 to &ngrade;
                cutoff = ranuni(&seed);
                deg = ranuni(&seed+1);
                if cutoff < 0.05 then do;
                    yr = rand('NEGBINO', 0.5, 2);
                    if deg < 0.52 then ad = 1;
                    else ad = 0;
                    output;
                end;
                else if cutoff < 0.25 then do;
                    yr = rand('NEGBINO', 0.4, 3);
                    if deg < 0.52 then ad = 1;
                    else ad = 0;
                    output;
                end;
                else do;
                    yr = rand('NEGBINO', 0.2, 4);
                    if deg < 0.52 then ad = 1;
                    else ad = 0;
                    output;
                end;
        end;
    end;
end;
proc sort data=var1;
  by district school grade;
run;

/*Assign Values to the Variables*/
data mean1;
  set var1;
  if yr < 2 then level = 1;
  else if yr < 6 then level = 2;
  else if yr < 11 then level = 3;
  else level = 4;
  exp = rannor(&seed + 10) * &yrvar + level * &yc - 0.5;
  if ad = 1 then do;
    dg = rannor(&seed + 10) * &degvar + &degc;
    group = 4 + level;
    output;
  end;
  else do;
    dg = rannor(&seed + 10) * &degvar - &degc;
    group = level;
    output;
  end;
run;

proc sort data = mean1;
  by district school grade;
run;

/*Define Design Structure and Associated Variances*/;
data topol;
  do district = 1 to &ndist;
    dist_var = &dist_var * rannor(&seed + 100);
  do school = 1 to &nscho;
    scho_var = &scho_var * rannor(&seed + 100);
  do grade = 1 to &ngrade;
    bk = dist_var + scho_var;
    output;
  end;
end;
keep district school grade bk;
run;

proc sort data = topol;
  by district school grade;
run;

/*Define VA Scores for Each Teacher*/
data score1;
  merge mean1 topol;
  by district school grade;
  VAM = dg + exp + 2 * dg * exp + bk;
  drop dg exp bk group;
run;

/*Number the Teachers*/
data order1;
do count - 1 to &ngrade\&nteach;
if count<\&nteach then do;
   teacher= count;
   grade = 1;
   output;
end;
else if count<2\&nteach then do;
   teacher= count;
   grade = 2;
   output;
end;
else do;
   teacher= count;
   grade = 3;
   output;
end;
end;
drop count;
run;

proc sort data = order1;
   by grade;
run;

proc sort data = score1;
   by grade district school;
run;

data VAM1;
   merge score1 order1;
   by grade;
run;

proc sort data = VAM1;
   by district school teacher;
run;

data teacher_info1;
   set VAM1;
   drop VAM;
run;

proc sort data = teacher_info1;
   by teacher;
run;

/*/Define Starting Scores for Each Student;*/

/data students1:
   do district = 1 to &ndist;
   do school = 1 to &nscho;
   do student = 1 to &nstud;
      baseline = 100 + 2*ranor(&seed + 1000);
      score0a = baseline;
      w1 = &stuv*a * ranor(&seed + 1000);
      score1a = baseline + w1;
      w2 = &arrho * w1 + &stuv*a * ranor(&seed + 1000);
      score2a = baseline + w2;
      w3 = &arrho * w2 + &stuv*a * ranor(&seed + 1000);
      score3a = baseline + w3;
      output;
   end:
end;
end;
drop baseline w1-w3;
run;

proc sort data = students1;
  by district school student;
run;

/*Format Teacher Information for Merge*/
data teachA1;
  set VAM1;
  drop ad level yr grade teacher;
run;

proc transpose data = teachA1 out = teachersA1 prefix = vam;
  by district school;
run;

data teachB1;
  set VAM1;
  drop ad level yr grade VAM;
run;

proc transpose data = teachB1 out = teachersB1 prefix = teach;
  by district school;
run;

/*Define Values for Teacher and Organize for Each Student */
data teachersFinal1;
  merge teachersA1 teachersB1;
  by district school;
  te1 = vam1;
  te2 = vam2;
  te3 = vam3;
  t1 = teach1;
  t2 = teach2;
  t3 = teach3;
  do student= 1 to &nstud;
    output;
  end;
  drop _NAME_ vam1-vam3 teach1-teach3;
run;

proc datasets noprint;
  delete teachal teachbl teachersal teachersbl order1 topol var1;
quit;
run;

/*Merge Student Starting Scores with Teacher VA Scores*/
data merged1;
  merge students1 teachersFinal1;
  by district school student;
  score0 = score0a;
  score1 = score1a + te1;
  score2 = score2a + te1 + te2;
  score3 = score3a + te1 + te2 + te3;
  drop score0a score1a score2a score3a te1-te3;
run;

/*Add the Student Grade */
data final1;
  set merged1;
  do grade = 1 to 3;
    if grade = 1 then do;
      teacher = t1;
      score = score1;
      baseline = score0;
      output;
    end;
    else if grade = 2 then do;
      teacher = t2;
      score = score2;
      baseline = score0;
      output;
    end;
    else do;
      teacher = t3;
      score = score3;
      baseline = score0;
      output;
    end;
  end;
drop score0-score3 t1-t3;
run;

/*Define the Z-Matrices*/
data lags1;
  set final1;
  tchr=put(teacher,z2.);
  l1=lag1(tchr);
  l2=lag2(tchr);
  lag0tchr=tchr;
  if lag1(student)=student then lag1tchr=l1;
  if lag2(student)=student then lag2tchr=l2;
  drop l1 l2 tchr;
r

/*Utilize Glimmix to Output the Z matrix*/
proc glimmix data = lags1 outdesign=zmatrixL0;
  class lag0tchr lag1tchr lag2tchr;
  effect T-MM(lag0tchr lag1tchr lag2tchr / noeffect);
  model score = baseline/s;
  random t/s;
r
proc datasets noprınt;
  delete merged1 teachersfinal1;
r

/* Create a starting point to calculate MSPE later for teacher VAM scores */
data compare1;
  set vaml;
  vam_orig=vam;
  keep teacher vam_orig;
r
proc sort data = compare1;
  by teacher;
r
%END setup1;
run;
quit;

B.1.2 Simulate Variables and Scores for Case 2

For this case, teacher value added scores are generated without linking. The following two data sets, topol and score1 listed below for the Setup1 macro are changed for Case 2.

```plaintext
/*Define Design Structure and Associated Variances*/;
data topol;
do district=1 to &ndist;
dist_var = &dist_var*ranor(&seed+100);
do school = 1 to &nscho;
scho_var = &scho_var*ranor(&seed+100);
do grade = 1 to &ngrade;
   bk=dist_var+scho_var;
output;
end;
end;
end;
keep district school grade bk;
run;

/*Define VA Scores for Each Teacher*/
data score1;
merge mean1 topol;
by district school grade;
VAM = dg+exp+2*dg*exp+bk;
drop dg exp bk group;
run;
```

Data is generated for Case 2 in the Setup2 macro. Data sets in the Setup2 macro are named differently, e.g. data sets are identified with a 2 rather than a 1. Otherwise, the macros are identical except for the changes given below.

```plaintext
%MATERIAL setup2(sim, seed, ndist, nscho, nteach, dgec, yrc, degvar, yrvar,
   dist_var, scho_var, ngrade, nstul, stuvar, arrho);
/*Define Design Structure and Associated Variances*/;
data topol2;
do district=1 to &ndist;
dist_var = &dist_var*ranor(&seed+100);
do school = 1 to &nscho;
scho_var = &scho_var*ranor(&seed+100);
do grade = 1 to &ngrade;
   eijk=ranor(&seed+200)*0.25;
   bk=dist_var+scho_var;
output;
end;
end;
```
end;
keep district school grade bk eijk;
run;

/* Define VAM Scores for Each Teacher*/
data score2;
merge mean2 topo2;
by district school grade;
VAM = bk+eijk;
drop dg exp bk group eijk;
run;

%END setup2;
run:
quit;

The parameter values for the Setup1 and Setup2 macros are provided in Appendix B.4.

B.2 Compile Macro

This section includes macros used to save generated scores for future analysis. The macros for both Cases 1 and 2 are included.

B.2.1 Compile Data Generated Under Case 1

After the data is simulated, certain attributes are compiled and stored for future evaluation. The compilation was done in the same manner for both the Compile1 and Compile2 macros which correspond to Case 1 and Case 2.

%MACRO compile1(sim);
/* Ensure that each ad*level combination has at least 1 teacher */
proc freq data = mean1;
tables group/ out=count1;
run;

proc transpose data = count1 out=compA1 prefix=group;
id group;
run;

data compA1;
retain sim;
set compA1;
if _name_ = "COUNT";
    sim = &sim;
    drop _name_ _label_;
run;

/* Obtain information about teacher VA scores */
data teacher_val;
    set score1;
    va=vam;
run;

proc means data = teacher_val;
    var va;
    output out=tsuml;
run;

proc transpose data = tsum1 out=compB1;
    id _STAT_;
run;

data compB1;
    retain sim;
    set compB1;
    sim = &sim;
    if _name_ = "va";
    drop _name_;
run;

/* Combine information about the teachers for each simulation */
data teacher_details1;
    merge compA1 compB1;
    by sim;
run;

/* Save teacher VA Scores */
data VamFull1;
    retain sim district school grade teacher;
set vam1;
sim - &sim;
vam_orig - vam;
drop vam;
run;

proc sort data = VamFull1;
  by teacher;
run;

/* Save Student Simulated Scores */
data StuFull1;
  retain sim district school grade teacher;
  set Finall1;
  sim - &sim;
run;

/* Save information after each simulation */
%if &sim = 1 %then %do;
  data teachsetL;
    set teacher_details1;
  run;

data stuscoreL;
  set StuFull1;
run;
%end;

%if &sim > 1 %then %do;
  proc append base = teachsetL data = teacher_details1;
  run;

  proc append base = stuscoreL data = StuFull1;
  run;
%end;

%mend compile1;
run;
quit;

B.2.2 Compile Data Generated Under Case 2

In the Compile2 macro, data sets are named differently, e.g. data sets are identified with a 2 rather than a 1, and N is used to identify that the case is non-linking where as in Compile1 L is used to identify that the case in linking. A subset of the changes are included in the program below.

```
%MACRO compile2(sim);

/* Save teacher VAM Scores */
data VamFull2;
    retain sim district school grade teacher;
    set vam2;
    sim = &sim;
    vam_orig=vam;
    drop vam;
run;

/* Save information after each simulation */
%if &sim - 1 %then %do;
    data teachsetN;
    set teacher_details2;
    run;

    data stuscoreN;
    set StuFull2;
    run;
%end;

%if &sim > 1 %then %do;
    proc append base = teachsetN data = teacher_details2;
    run;

    proc append base = stuscoreN data = StuFull2;
```
run;
%
end;
%
end compile2;
run;
quit;

The parameter value, \&sim for the Compile1 and Compile2 macros are provided in AppendixB.4.

B.3 Analyze Macro

This section includes the macros used to analyze data under Cases 1 and 2.

B.3.1 Analyze Data Generated Under Case 1

In this macro, both the standard TVAAS and multi-stage small area modeling methods are implemented. The Analyze1 macro is for data generated under Case 1.

/* The First Iteration Evaluates the Standard TVAAS Method */
proc hpmixed data=zmatrixL0;
   class district school student grade;
   effect T~collection(_z1-_z90);
   model score = baseline / s;
   random T / s;
   repeated grade / type=chol subject=district*school*student;
   parms (0.1) (0.25) (0.25) (0.25) (0.25) (0.25) (0.25);
ods output SolutionR~VAM_scoresL&i ParameterEstimates=stu_infoL&i.;
run;
/* Get the teacher VAM scores into workable format for small area estimation and identify rankings */
data Vam_Est1;
set VAM_scoresL&i.;
step = substr(T, 3, 2);
teacher = input(step, 2.);
vam = estimate;
keep teacher vam;
run;

/* Verify that the teacher estimates produced are non-zero */
There are several ways to do this */
proc means data = Vam_Est1 max min;
var vam;
output out = verify1 max = max;
run;

data _null_; 
set verify1;
call symput('max', max);
run;

/* If the estimates are zero, then the process ends here */
%if &max = 0 %then %do;

data std_error1;
sim = symgetn('sim');
method = "standard";
i = 1;
mse = .;
run;

data tch_error1;
set VamFull1;
vam_stnd = .;
run;
%if &sim = 1 %then %do;
data basesetL;
   set std_error1;
run;
data teachvamL;
   set tch_error1;
run;
%end;

%if &sim > 1 %then %do;
   proc append base = basesetL data = std_error1;
run;
   proc append base = teachvamL data = tch_error1;
run;
%end;
%goto DONE;
%end;

/* If estimates are non-zero, the process continues */
data teacherL&i.;
   merge teacher_info1 vam_est1;
   by teacher;
r
proc sort data = teacherL&i.;
   by teacher;
r
PROCUT DIFF1;
   retain teacher vam_orig vam;
   merge compareL&i.:;
   by teacher;
e = vam_orig - vam;
e2 = e**2;
/*Find the MSPE for the standard method */
data StndDiff1;
   retain teacher vam_orig vam;
   merge compareL&i.:;
   by teacher;
e = vam_orig - vam;
e2 = e**2;
keep teacher vam_orig vame e2;
run;

proc means data = StndDiff1 mean;
var e2;
output out = StndMse1 mean - MSE;
run;

data null;
set StndMse1;
call symput('StndMSE', MSE);
run;

/* Add the teacher VAM scores for standard method to the VA scores —
Used to identify rankings later */
data VamFull1;
merge VamFull1 teacherL&i ;
by teacher;
vam_stnd=vam;
drop vam;
run;

/* Incorporate additional information about the teachers to improve VAM
estimates*/
proc glimmix data = teacherL&i ;
class district school ad level;
model vam=ad|level;
random intercept school / subject = district;
output out=vam_newL&i . pred=p;
run;

proc sort data = vam_newL&i .;
by district school teacher;
run;

/* Get updated teacher VAM estimates in a form to be combined with student
information*/
data teach_lagL&i.;
set van_newL&i.;
t1=lag1(p);
t2=lag2(p);
tlag0=p;
if lag1(school)=school then tlag1=t1;
if lag2(school)=school then tlag2=t2;
if tlag1=. then tlag1=0;
if tlag2=. then tlag2=0;
do student = 1 to &nstud;
output;
end;
keep tlag0–tlag2 district school grade teacher student;
run;

proc sort data = teach_lagL&i.;
by district school student grade;
run;

data parm_estA1;
set stu_infoL&i.;
keep effect estimate;
run;

proc transpose data = parm_estA1 out=parm_estB1;
id effect;
run;

/* Include Parameter Values obtained from HpMixed */
data parm_estC1;
set parm_estB1;
nm=intercept;
beta=baseline;
do district = 1 to &ndist;
do school = 1 to &nscho;
do student = 1 to &mstud;
output;
end;
end;
end;

drop _NAME_ intercept baseline;
run;

data stu_parml;
merge students1 parm_estC1:
by district school student;
baseline=score0a;
do grade = 1 to &ngrade;
output;
end;
drop score0a score1a score2a score3a;
run;

/ * New Student Scores */
data combine_newL&i.;
merge stu_parml teach_lagL&i.:
by district school student grade;
score_new=m+b*baseline+tlag0+tlag1+tlag2;
keep district school student grade score_new;
run;

/ * Add New Student Scores to Z–matrix and Measure the Change in Scores */
data zmatrixL&i.;
merge combine_newL&i. zmatrixL0;
by district school student grade;
diff = (abs(score–score_new))/score;
drop lag0tchr lag1tchr lag2tchr _X1 _X2;
run;

proc means data = zmatrixL&i.;
var diff;
output out=convL&i. mean=value;
run;
data convL&i.;
    set convL&i.;
    call symput('criterion', value);
    i = symgetn('i');
run;

proc datasets noprint;
    delete parm_estA1 parm_estB1 parm_estC1 stu_parm1 stnddiff1 stndmse1;
run;
%

/* Each Iteration Utilizes Obtained Student Scores from Previous Iteration*/;
%if &i > 1 %then %do;
    %let j = %eval(&i - 1);
%

/* Implement TVAAS */

proc hpmixed data=zmatrxL&j.;
    class district school student grade;
    effect T=collection(_z1-_z90);
    model score_new = baseline / s;
    random T / s;
    parms (0.05) (0.005);
    ods output SolutionR-VAM_scoresL&i. ParameterEstimates=stu_infoL&i.;
run;

data vam_estL&i.;
    set VAM_scoresL&i.;
    step = substr(T, 3, 2);
    teacher=input(step, 2.);
    vam=estimate;
    keep teacher vam;
run;

data teacherL&i.;
    merge teacher_info1 vam_estL&i.;
by teacher;
run;

/*Continue to Multi–Stage Model– Incorporate Auxiliary Variables*/

proc glimmix data = teacherL&i.;
class district school ad level;
model vam = ad | level;
random intercept school / subject = district;
output out = vam_newL&i. pred = p;
rungl

proc sort data = vam_newL&i.;
by district school teacher;
rungl

/* Format Results to Update Student Scores*/
data teach_lagL&i.;
set vam_newL&i.;
t1 = lag1(p);
t2 = lag2(p);
tlag0 = p;
if lag1(school) = school then tlag1 = t1;
if lag2(school) = school then tlag2 = t2;
if tlag1 = . then tlag1 = 0;
if tlag2 = . then tlag2 = 0;
do student = 1 to &nstud;
output;
end;
keep tlag0–tlag2 district school grade teacher student;
rungl

proc sort data = teach_lagL&i.;
by district school student grade;
rungl

data parm_estA1;
set stu_infoL&i.;
keep effect estimate;
run;

proc transpose data = parm_estA1 out=parm_estB1;
  id effect;
run;

/* Use Parameter Estimates from HpMixed */
data parm_estC1;
  set parm_estB1;
  mm-intercept;
  beta-baseline;
  do district = 1 to &mdist;
    do school = 1 to &mscho;
      do student = 1 to &mstud;
        output;
        end;
      end;
    end;
  end;
  drop _NAME_ intercept baseline;
run;

data stu_parm1;
  merge students1 parm_estC1;
  by district school student;
  baseline-score0a;
  do grade = 1 to &ngrade;
    output;
    end;
  drop score0a score1a score2a score3a;
run;

/* Updated Student Scores */
data combine_newL&i.;
  merge stu_parm1 teach_lagL&i.;
  by district school student grade;


```plaintext
score_new = beta * baseline + tlag0 + tlag1 + tlag2;
keep district school student grade score_new;
run;

data zmatrix_try1;
set zmatrixL&j . ;
score = score_new;
drop score_new diff;
run;

/* Combine Student Scores with Z-Matrix and Evaluate Convergence */
data zmatrixL&i . ;
merge combine_newL&i . zmatrix_try1;
by district school student grade;
diff = (abs(score - score_new))/score;
run;

proc means data = zmatrixL&i . ;
var diff;
output out=convL&i . mean= value;
run;

data convL&i . ;
set convL&i . ;
if value < &cutoff then check = 1;
else check = 0;
call symput('criterion', value);
call symput('check', check);
i = symgetn('i ');
run;

%put Linking;
%put Simulation &sim;
%put Iteration &i ;
%put &criterion ;
%put &check ;
```
%if &check = 1 %then %goto EXIT;

proc datasets noproct;
   delete parm_estA1 parm_estB1 parm_estC1 stu_parm1 zmatrix_trl;
run;
%end;

%if &i = &maxiter %then %goto DONE;
%end;

/* After Model Convergence */

%EXIT:

/* Save Teacher VAM Scores */
data VamFull1;
   merge VamFull1 teacherL&i.;
   by teacher;
   vam_sae=vam;
   drop vam;
run;

/* Compute MSE for Small Area Method */
data SaeDiff1;
   retain teacher vam_orig vam;
   merge compare1 teacherL&i.;
   by teacher;
   e = vam_orig - vam;
   e2 = e**2;
   keep teacher vam_orig vam e e2;
run;

proc means data = SaeDiff1 mean;
   var e2;
   output out = SaeMse1 mean = MSE;
run;
data _null_;  
  set SaeMse1;  
  call symput('SaeMSE', MSE);  
run;

data stdl1;  
  sim = symgetn('sim');  
  method = "standard";  
  i = 1;  
  mse = symgetn('stdMSE');  
run;

data sael;  
  sim = symgetn('sim');  
  method = "sm_area";  
  i = symgetn('i');  
  mse = symgetn('SaeMSE');  
  stop = symgetn('criterion');  
run;

data resultsL&sim.;  
  retain sim method i;  
  set stdl sael;  
run;

/* Save Information After Each Simulation */

%if &sim - 1 %then %do;  
  data basesetL;  
  set resultsL&sim.;  
run;

data teachvamL;  
  set VamFull1;  
run;  
%end;
%if &sim > 1 %then %do;
    proc append base = basesetL data = resultsL&sim.;
    run;
    proc append base = teachvamL data = VamFull1;
    run;
%end;

%DONE:

%mend analyze1;
run;
quit;

B.3.2 Analyze Data Generated Under Case 2

For the Analyze1 macro, L is used with data to identify that it is evaluated under Case 1. For the Analyze2 macro, data is identified with an N to specify that evaluation occurs when data is generated under Case 2. The code that saved data after the Analyze2 macro is implemented is included below.

%MACRO analyze2(sim, criterion, cutoff, ndist, nscho, nteach, ngrade, nstud, max, maxiter);
%if &sim = 1 %then %do;
    data basesetN;
    set resultsN&sim.;
    run;
    data teachvamN;
    set VamFull2;
    run;
%end;

%if &sim > 1 %then %do;
%macro simA(numSim, sim_seed, ndist, nscho, nteach, ngrade, nstud, conv);
%let datetime_start = %sysfunc(TIME());
%put START TIME:%sysfunc(datetime(),datetime14.);

%do numSim = 1 %to &numSim;
%let seed = &sim_seed + &numSim;

%setup1(&numSim, &seed, &ndist, &nscho, &nteach, 0.2, 0.15, 0.1, 0.1, 0.2,
0.25,
&ngrade, &nstud, 1.5, 0.7);
%compile1(&numSim);
B.4.2 Simulate Data Under Condition 2

To simulate data under Condition 2, the SimB macro is identical to SimA, except the parameters used for the setup macros have changed. The parameters for the macros are presented below.

%macro simB(numsim, sim_seed, ndist, nscho, nteach, ngrade, nstud, conv);
  %setup1(&nsim, &sim_seed, &ndist, &nscho, &nteach, &ngrade, &nstud, 0.2, 0.15, 0.1, 0.1, 0.4, 0.5, &ngrade, &nstud, 1.5, 0.7);
  %setup2(&nsim, &sim_seed, &ndist, &nscho, &nteach, &ngrade, &nstud, 0.2, 0.15, 0.1, 0.1, 0.4, 0.5, &ngrade, &nstud, 1.5, 0.7);
%endo simB;
&ngrade, &nstud, 1.5, 0.7);
%mend simB;

%simB (1000, 19870404, 5, 6, 30, 3, 15, 0.0001);
run;
quit;