A Primer on Value-Added Models: Towards a Better Understanding of the Quantitative Analysis of Student Achievement

Yugo Nakamura

A dissertation to be submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Washington

2013

Reading Committee:

Margaret L. Plecki, Chair

Robert D. Abbott

Bradely S. Portin

Program Authorized to Offer Degree:

College of Education

University of Washington

Abstract

Value-added models (VAMs) have received considerable attention as a tool to transform our public education system. However, as VAMs are studied by researchers from a broad range of academic disciplines who remain divided over the best methods in analyzing the models and stakeholders without the extensive statistical background have been excluded from the technical debate, the implementation of the models have been undermined. This study conducts a comprehensive investigation of VAMs to find consensus and transparency in the understanding of the models. Unlike the majority of existing value-added research which conducts highly advanced analyses, this study takes a unique pedagogical approach. It targets a wider range of audience particularly those without an extensive background in statistics – policy makers, district officials, school principals, teachers, and parents. The study first clarifies the technical dimensions underlying the design of the models. It then applies this conceptual understanding to conduct a thorough and hands-on value-added analysis using the Washington State longitudinal data. A comprehensive diagnosis, revision, and validation of the estimated models will be conducted to ensure the most accurate, reliable, and robust estimates. The study concludes by illustrating how the conventional policy analysis which focuses only on the value-added estimates can be extended to incorporate student background equity related factors. This extended analysis enables us to identify teachers (or schools) who achieve quality and equity outcomes simultaneously by raising the performance particularly of the students who have historically underachieved. By providing a common ground for all stakeholders irrespective of their background, this study provides an essential platform for further development of VAMs and its application as a tool to improve our education system

A Primer on Value-Added Models: Towards a Better Understanding of the Quantitative Analysis of Student Achievement

Yugo Nakamura

Chair of the Supervisory Committee:

Dr. Margaret L. Plecki

College of Education, Educational Leadership and Policy Studies

I would like express my gratitude and appreciation to my Committee members for providing me the invaluable lessons, opportunities, and feedbacks during my course of study. I would also like to thank my family and friends for their unconditional support throughout my life. Special thank you goes to my grandma.

Table of Contents

CHAPTER 1: INTRODUCTION – THE CONTEXT AND NEEDS OF VALUE-ADDED MODELS

There are a number of definitions of value-added models (VAMs). Below are definitions provided by three prominent sources:

> Value-added models are the efforts to measure the effects of current teachers, schools, or educational programs on the achievement of students through taking account the differences in prior achievement and other measured characteristics that students bring with them to school…. [Specifically] they are statistical models, often complex, that attempt to attribute some fraction of student achievement growth over time to certain schools, teachers, or programs. To isolate school, teachers, or program effects, at least two years of students' test scores are taken into account, along with other student and school level variables, such as poverty and family background. (National Research Council, 2010, p.1)

> Value added estimate of a school [or teacher] is the contribution of a school to students' progress towards stated or prescribed education objectives (e.g. cognitive achievement). This contribution is the net of other factors that contribute to students' educational progress … value added models refer to the class of statistical model used in estimating this contribution using the data measured at least two points in time. (Organization of Economic Cooperation and Development, 2008, p.17)

> Value-added models seeks to measure teacher quality by estimating the impact of teachers on student achievement, holding constant other factors that affect current student performance, including contemporaneous student ability and effort, family inputs, peer influences and school quality as well as the prior contributions of family, peer, teacher and school inputs. (Sass, 2008, p.1)

As evident from these definitions, VAMs convey a powerful message regarding the quality and "effectiveness" of the teachers (and schools) in raising their students' academic performance. The models represent a vast improvement compared to prior efforts in educational reform, namely the No Child Left Behind Act, which utilized basic teacher (and school) mean scores as the basis of policy making.[1] As these mean scores do not take into account factors outside the control of teachers that also influence the student performance (such as students' prior achievement, socio-economic and contextual backgrounds), it provided unfair and unjust evaluation of teacher (and school) performance. In light of this problem, the VAMs offered the promise of "leveling the playing field" in estimating and comparing teacher effectiveness through removing and separating the effects of these external factors from the teachers. The models provide fairer, more credible, and more transparent estimates of the unique contribution the teachers (and schools) have on their students' achievement. And it therefore informs us with better, more accurate, and more reliable forms of evidence in

[1] These means scores are also referred to as the Annual Yearly Progress (AYP) Indicators or the unadjusted status indicators.

order to improve our education system. In light of this perceived potential, the VAMs received considerable attention and hope by the public at large. Policy makers embraced it as the means to improve the education accountability system, merit/performance based pay, teacher assignment and retention, program evaluation of professional development activities, targeting of resources and capacity building activities, and even school closure and restructuring. Teachers and school principals used it as a diagnostic tool to systematically monitor their students' achievement over time. Some parents also welcomed the findings as key source of evidence in choosing their child's school. And more students themselves considered it as a tool to self-monitor their academic progress. But along these hopes and aspirations came challenges and a sense of despair.

As the models were considered for its actual implementation, a variety of technical and analytical challenges surfaced. In a way, a Pandora's box of VAMs was opened. The challenges consisted of psychometric issues over the appropriate test designs in measuring student progress (e.g. difficulty of vertically linked and scaled tests); statistical and analytical problems over choosing which modeling options to use and how to address the students' non-random selection of teachers (and schools) which plagues the model estimates; and data capacity and collection challenges over maintaining a reliable and comprehensive data infrastructure overtime. Over the last decade, the research community warned the public of these technical challenges noting that the models were still in its infancy. A series of public forums, research articles, specialized journal editions, think tank, and government reports highlighted the implementation challenges. More importantly, the researchers continuously highlighted the danger of using the VAM estimates as the only indicator in informing high stakes policies. As addressed by Michael Kolen and Henry Braun two leading and pioneering researchers of VAMs in a national report titled "Getting Value Out of Value-Added" published in 2010 by the National Research Council (NRC) and the National Academy of Education (NAE),

> Are estimated teacher and school effects largely due to idiosyncrasies of statistical methods, measurement error, the particular test examined, and the scales used? Or are the estimated teacher and school effects due at least in part to educationally relevant factors? These questions need to be answered clearly before a value-added model is used as the sole indicator to make important educational decisions (Kolen, p.39)

> There is not one dominant VAM. Each major class of models has shortcomings. There is no consensus on the best approaches, and little work has been done on synthesizing the best aspects of each approach. There are questions about the accuracy and stability of value-added estimates of schools, teachers, or program effects. More needs to be learned about how these properties differ, using different value-added techniques and under different conditions. Most of the workshop participants argued that steps need to be taken to improve accuracy if the estimates are to be used as a primary indicator for high-stakes decisions; rather, value-added estimates should best be used in combination

> with other indicators. But most thought that the degree of precision and stability does seem sufficient to justify low-stakes uses of value-added results for research, evaluation, or improvement when there are no serious consequences for individual teachers, administrators, or students. (Braun, p.54)

But these voices were unheard by the general public including policy makers. As the political climate and the public consensus continue to press for reforms especially due to frustration over the NCLB Act, VAMs continued on the road to adoption and implementation. But it was only a matter of time for this ticking bomb would explode. As the detailed ranking of the teacher VA estimates were made public (e.g. in the Los Angeles Times); bonuses and hiring decisions were decided based on the estimates; educators were accused of engaging in cheating to raise the students' scores; teachers narrowed the curriculum and taught to the tests; and teachers sacrificed team teaching and cooperation among their colleagues, public outrage intensified. But as the general public tried to question the validity, reliability, and accuracy of the models, they were left refuge to the methodological complexity constructed by the elite technocrats. The mathematical formulas and the statistical jargons were incomprehensible to the common eye. As Harris (2011) addresses, this resulted in a grand divide between the economists (one of the technocrats) and the educators. For example, Ms. Isaacson, a teacher in New York who received low value-added scores despite being highly praised by her principal and colleagues, noted the following:

> The [value-added] calculation is daunting. It is based on 32 variables … These 32 variables are plugged into statistical model that looks like one of the those equations that in "Good Will Hunting" only Matt Damon was capable of solving. The process appears transparent, but it is clear as mud, even for smart people like teachers, principals – and I hesitate to say this – journalists. I may have two Ivy League degrees but I am lost. I find it impossible to understand in plain English. (New York Times, March 6, 2011)

And Diane Ravitch a noted historian of education, noted:

> We should make do with fewer economists in education. These practitioners of dismal science have flocked to education reform, though most know little about teaching and learning…. There is certain kind of madness in thinking that economists who never set foot in a classroom can create a statistical measure to tell us how to best educate children. It seems some will never be satisfied until they have a technical process to override the judgments of those who work in schools and are in daily contact with teachers and children. I don't know of any other nation in the world that is so devoted to this effort to turn education into a statistical problem that can be solved by a computer. (EdWeek, January 18, 2011)

And as the general public also learned that the technicians themselves are divided over the stability, accuracy, and reliability of the estimates, it further ignited their concern. For example, studies have found that 20 to 30% of the teachers in the top quartile (25%) of the ranking ended up in the bottom quartile in a course of just one year.[2] Another study which reanalyzed the LA Times ranking found that the choice over which variables to include in the model significantly changed the ranking of roughly half of the 6,000 teachers.[3]

In light of the public protest, policy makers have finally caught up in recent years with the flux of VA research. Many have now acknowledged that the models are not "magic bullet" policy tools. In an effort to resolve the conflict and provide responses, the report "Getting Value Out of Value-Added" summarized areas of emerging consensus and areas of disagreement regarding the uses of VAMs. It also asked the research community to work on a number of unresolved areas concerning the development and implementation of VAMs.[4] Furthermore, in order to spread the awareness and understanding of VAMs for education practitioners, various training modules and capacity building activities have been developed. William Sanders, another pioneer of VAMs and founder of the Tennessee Value-Added Assessment System (TVAAS) has designed various graphical tools to effectively report on the VA findings in order to help teachers' instruction and policy makers' decisions. The Value-Added Research Center (VARC) at the University of Wisconsin Madison also provides online modules to help interpret the value-added findings using various matrices and tables. Battelle for Kids (BFK), an education consulting company in partnership with the Ohio Department of Education, built an interactive tool kit and an online training session to assist educators in using VAMs as a medium for improving instruction. Yet, all of these efforts fall short in unraveling the complicated mechanics of VAMs. Users are not aware of how the results are derived or the assumptions and caveats underlying the results. In a majority of cases, only the simplest model is introduced without any consideration of differences and similarities in comparison to other prominent models. More importantly, the statistical jargon and building blocks essential in the understanding of the models – both simple and complex – are never clarified.

[2] Rothstein (2011) who reanalyzed the VA analyses of the Measures of Effective Teaching (MET) project administered by the Gates Foundation.
[3] Briggs (2011)
[4] Some of these areas are as follow.

- How might the econometric and statistical models incorporate features from the other paradigms that are missing in their own approaches?
- What are the effects of violating the model assumptions on the accuracy of value-added estimates? For example, what are the effects on accuracy of not meeting assumptions about the assignment of students to classrooms, the characteristics of the missing data, as well as needed sample sizes?
- How could the precision of value-added estimates be improved? Instability declines when multiple years of data are combined, but some research shows that there is true variability in teacher performance across years, suggesting that simply pooling data across years might introduce bias and not allow for true deviation in performance.
- How do policy makers, educators, and the public use value-added information? What is the appropriate balance between the complex methods necessary for accurate measures and the need for measures to be transparent?
- How might value-added estimates of effectiveness be validated? One approach would be to link the estimates of school, teacher, or program effects derived from the models with other measures of effectiveness to examine the extent that the various measures concur. Some past studies have looked at whether value-added modeling can distinguish certified and noncertified teachers in an effort to validate the National Board for Professional Teaching Standards certification. In other words, value-added estimates are treated as the criterion. Another approach would be to turn that on its head and ask: How well do the value-added estimates agree with other approaches to evaluate the relative effectiveness of teachers?
- What are the implications of Rothstein's results about causality or bias for both the economics and the statistical approaches?

Consequently, the models remain incomprehensible to those without an extensive background in statistics. And the grand divide between the technocrats and educators continues.

In light of these challenges, the main objective of this dissertation is to end this interdisciplinary divide and provide an explanation in a way that can be understood by parties without an extensive background in statistics. These include policy makers, district officials, school principals, teachers, and parents who all play a pivotal role in the education system but have been excluded from the technical debates regarding VAMs. This study takes the view that the heated debate, confusion, frustration, and anxiety over the design and usage of VAMs have been going on unnecessarily too long. Now is the time to arrive at a consensus of VA research. This study assumes that the challenges and questions raised regarding VAMs must be and can be met in order to realize an educational system which is founded on sound evidence where all parties involved are equally informed. The findings can and should benefit all the constituents who support the success of this system. Moreover, as tests scores and various forms of data continue to pour into our daily lives together with numerous empirical analyses, it is ever more important to understand the statistical methods to manage this vast amount of information. These methods will become the critical tool for making informed decisions in this information based era.

To complete this objective, unlike the most of the current research of VAMs that conduct highly advanced statistical analyses, this study takes a unique pedagogical approach. This approach consists of two parts. Part I re-examines the foundation and purpose of statistics with a focus on the linear regression models which lie at the heart of VAMs. Once the linear regression models are examined, its framework will be applied to the construction of VAMs. This will consist of two major tasks: First, defining the VA parameter using the fixed and/or random effects. The second task focuses on how to take factors outside the control of teachers into account. The linear regression model assumptions which ensure the accuracy (unbiasedness) and reliability (efficiency) of the estimates will then be highlighted. These assumptions will become the guiding framework for conducting VA analysis in practice. Part II applies the conceptual understanding attained in Part I to a teacher VA analysis using Washington State longitudinal data. Unlike the majority of existing VA research which often hastily jump to the policy analysis, this study conducts a more rigorous and comprehensive analysis of the models. It undertakes a thorough diagnosis, revision, and validation of all the estimated models. In this study, only the VAMs in which the accuracy and reliability are justified are used for policy analysis. This study asserts that only through the rigorous assessment of the models can we achieve the utmost confidence in utilizing the VA results for the betterment of our education system. The outline of the chapters underlying the two parts is provided below.

Part I: CONCEPTUALIZATION OF VALUE-ADDED MODELS

Chapter 2 Statistics as a Science of Parsimony
Chapter 3 Statistics as a Science of Assumptions

CHAPTER 2: STATISTICS AS A SCIENCE OF PARSIMONY

Statistics is a science of parsimony. The word parsimony is generally defined as closeness or efficiency driven by a desire to avoid excess, unnecessary complexity or redundancy. In the discipline of statistics, the word means getting the task done in as few terms as possible.[5] Statistics, in fact, play an important part in our daily lives as it helps us make informed, rational, and valid decisions in the face of ever accumulating amounts of information. In this section, the two major forms of "parsimonious" tasks underlying statistics are introduced: descriptive statistics and inferential statistics. An explanation of these two concepts and how they play an important part in our lives is provided. Special attention is given to linear regression models, a popular statistical tool which studies the relationship between variables and which lies at the heart of VAMs.

Descriptive Statistics

Descriptive statistics pertains to the summarization of the salient features of the outcome variable(s) of interest. It is concerned with the simplification of a large amount of information (data) about the variable(s) into manageable and more readily understandable forms. It attempts to "parsimoniously" reveal and extract the essence and main messages of the underlying data patterns using selective summary indicators. As Bartholomew et al. (2008) describe, "the aim of descriptive statistics is to find ways of summarizing a large amount of data in a form which our eye and minds are able to grasp the underlying pattern immediately. It is a judicious way of conveying information and involves a substantial reduction of the volume of data." In practice, we carry out this process by selecting elements which seem significant or representative of the variable of interest such as the mean. In other cases when multiple variables are involved, we can calculate the mean by grouping similar things together or use slightly more advanced tools such as correlation to describe the extent of association between the variables. Descriptive statistics, in fact, plays an important part in our daily lives. For the most part, it is done unconsciously and we are hardly aware of the summarizing process that is taking place in our brains. But in an age when technological advances continually increase the amount of information, the role of descriptive analysis has clearly become ever more important. The only way we can make sense of this vast amount of data and to make informed decisions based on it is to extract the salient features underlying the data. To better illustrate the close ties between descriptive statistics and our daily lives, the following examples are provided.

In education policy, policy makers such as the Secretary of Education must constantly manage massive amounts of data such as student test scores from a national assessment in order to monitor the health and progress of the nation's education system. But in doing so, the Secretary cannot examine the performance of each and every student as this creates massive collection of confusing short stories. Instead, the Secretary must obtain the key features and messages underlying the massive data in order to see the overall picture of the

[5] Klockars (2005)

education system. The most important descriptor for this purpose is the average test score which provides a measure of the typical and central value of the entire data set. It provides an indication of the overall quality (performance level) of the education system. Tracking the average scores across time will also enable the Secretary to get a sense of the overall direction the nation is headed. To obtain a finer picture of student performance, the Secretary can also calculate average test scores by different group of students e.g. by demographic backgrounds, gender, geographical region, etc. Together with variance or standard deviation of test scores (which measures the spread of the student test scores from the average), these findings can provide an indication of the equity (dispersion of performance) of the education system. These descriptors provide key "snapshots" and messages that help the Secretary make decisions and form policy interventions. Other policy makers (such as the Superintendent, state and district officials, school leaders, and teachers) can similarly engage and benefit from the descriptive analysis at their respective level of operations. The importance of descriptive statistics is clearly not confined to the education scene. The chairman of the Federal Reserve must rely on a number of descriptors such as inflation rate, unemployment rate, production growth rate, etc., to monitor economic systems. The sports world is also filled with "stats" to evaluate a players' performance. For example, in baseball, descriptors such as the batting average (BA) and runs batted in (RBI) are widely used. Sports agents use these "stats" to help decide on the contracts for the players and coaches also use these summaries to decide on the best lineup to win a game. Furthermore, at the grocery store, parents utilize a number of descriptive statistics shown on the food label such as the total number of calories, average amount of sugar, parentage of vitamins, etc., to cook a balanced meal. In all these examples, the descriptors simplify the complexity of the outcome variable(s) into tractable and manageable parts. In essence, it provides us with efficient "snapshots". And based on these summaries, we are able to make informed decisions.

In formal statistical language, the mean and the variance are examples of univariate descriptive statistics. Univariate descriptive statistics are summaries of the data of one ("uni") variable at a time.[6] The mean belongs to the type of univariate descriptors called central tendency descriptors which describe the typical values of the data/distribution or where the distribution tends to fall. In addition to the mean it includes median and mode.[7] On the other hand, the variance belongs a group of univariate descriptors called dispersion descriptors which describe how spread out the distribution is around the mean. In addition to the variance, it includes standard deviation, range, and percentile.[8] Other useful descriptors include the skew (symmetry) and kurtosis (peakedness). Skew describes how concentrated the distributions are at the low or high end of the scale. It

[6] It is contrast to bivariate or multivariate descriptive statistics which describes the extent of relationship and association inherent between multiple variables. Specific examples of these descriptors (i.e. regression) are explained later in this section.

[7] Median – the value in the middle of the data set when the measurements are arranged in order of magnitude. Also referred to as the 50th percentile. Mode – the value occurring the most often in the data.

[8] Standard deviation – the square root of the variance. It is the average difference between observed values and the mean. It expresses the spread in the same units as the original measurements unlike the variance. Range – the distance between the lowest (minimum) and highest (maximum) value. Percentile – the percentage of data points which lie below a certain value when the values are ordered e.g. if a person's score lies in the 90th percentile, then 90% of the individuals/cases lie lower than this score. The 25th percentile is also referred to as the first quartile, 50th percentile as the median, 75th percentile as the third quartile.

indicates the degree of symmetry in the data.[9] Kurtosis describes how concentrated the distribution is around a single value, usually the mean. It assesses the extent to which the distribution is peaked or flat.[10] In all, these descriptors provide a good illustration of the underlying patterns of the object being studied. Now, when two or more variables are present, it is our natural interest to describe the extent of association inherent between these variables. The descriptors used for this purpose are called bivariate (two variables) and multivariate (more than two variables) descriptive statistics. And the key descriptor extensively used today for this purpose is regression.[11] In understanding regression, it is important to note that the features of univariate descriptors are not forgone. Regression is in fact a simple extension of the univariate indicator - the mean. It is a summary of a series of means of a variable for (conditioned on) different values of another variable. The former variable is referred to as the outcome or dependent variable which is explained by the latter variable(s) referred to as the explanatory or independent variable(s). As formally described by Miller (2008), "the basic problem of regression is that of determining the conditional mean – the average values of the outcome variable Y for given values of the explanatory variables X." Graphically, regression is essentially a function which describes, joins, or connects the conditional means of the outcome variable for different fixed, given, and known values of the explanatory variables. Through providing a numerical description of the function connecting the conditional mean values, regression models (more specifically linear regression models) allow us to interpret this description as "effect", "cause", or "prediction" of the explanatory variable(s) on the outcome variable. This sends a powerful message that has real life implications. It is also the reason for its popularity in the research world. .

Inferential Statistics

In the above section, a variety of descriptors to parsimoniously summarize the variable(s) of interest was described. However, in practice, we cannot stop with these descriptors alone. The analyst (e.g. the Secretary of Education) does not just want to know well the particular students collected in the dataset perform but rather want to know how well the entire population of students across the nation performs. Depending on the sampling scheme, the available data is only a particular group of subjects (observations) who just happen to be collected at a specific time and does not comprise the entire population of observations we would like to investigate. Thus, we need to be able to somehow use this information obtained from this sample of observations to make an informed guess about what the entire group of observations (outside and beyond the observed data) might look like. This "parsimonious" use of information from a limited sample of observations to make informed guesses and generalizations about the entire population from which they came is the second

[9] It is calculated by summing the cubes (power three) of the differences between each observation and the mean and then dividing by the cubes of the standard deviation. In a perfectly symmetric distribution e.g. normal distribution, the skew value is zero. The mean, median and mode is also equal. A positive skew is indicated by a tailed distribution (long and flat) toward the upper end of the scale (and vice versa for negative skew).

[10] It is calculated by raising the sum of the squares of the differences between each observation and the mean to the fourth power and then dividing by the fourth power of the standard deviation. The basis value use for comparison is three, which is the kurtosis of a normal distribution. Higher the value (above three), the more peaked is the distribution and lower the value (below three) more flat is the distribution.

[11] Other descriptors include correlation coefficient, partial correlation coefficient, semi-partial correlation coefficient, principal components analysis, etc.

major task of statistics known as inferential statistics. Inferential statistics goes beyond summarization by aiming to discover something about the process (population, general truths, data generating process) which has generated the data, and hence to discover something about a wider class of data of which the present set is only a sample.[12] It is a cost effective, efficient, and "parsimonious" use of information as we try to describe the population without having to actually measure all of it. In practice, the population may be finite and rather small. In such instance, we might conceivably measure all of its subjects,[13] but in most instances, the population is very large and too expensive to measure. Thus instead of measuring the entire population we select a "random" sample of the population. This is a group of observations selected in such a way that each member of the population has an equal chance of being included in the sample and each observation is independent of the others. A random sample provides an unbiased representation of the population. We can then use this information to estimate and infer the features of the population.

Statistical inference is also a familiar and everyday happening as we often base our expectations on sample experiences. Together with descriptive statistics, it plays an important role in our daily lives and decision making processes. An anecdotal example is a child trying to convince his/her parents to purchase a trendy item (e.g. PlayStation). The child overzealously generalizes that "everybody" in the school has the item based on the selective sample of friends who obtain the product. In more carefully planned experiments and investigations, examples of statistical inference include health experts trying to detect a possible outbreak of an epidemic via a collection of sample of observations (e.g. chicken, cows, etc.). The collection of data from the entire population of chickens is virtually impossible and even if not extremely expensive. We must therefore rely on a sample of chickens to infer and generalize the features of the entire population. Other examples include an engineer trying to decide on the basis of sample data whether the true average lifetime of a machine (e.g. a personal computer) is at least 5 years; an agronomist trying to decide on the basis of experiments whether one kind of fertilizer produces higher yield of a certain crop than another; a pharmaceutical company trying to evaluate on the basis of sample of patients whether a new possible anti-cancer drug effectively works on all patients; etc. All these examples illustrate how we must "parsimoniously" utilize a limited sample of data to make an informed generalization and decision regarding the much bigger population.

In formal statistical language, the numerical descriptions and features of the population we wish to infer are known as the parameters. Parameters represent the true characteristics of the population under study. Examples include the usual univariate descriptors such as the average and the variance and when multiple variables are involved, the correlation and regression coefficients. But these parameters are never going to be

[12] Bartholomew et al. (2008)

[13] In such case, it can be argued that inferential statistics is not necessary and we can conclude with only the descriptive statistics as our statistical analysis. Yet, we can also always argue for a hypothetical population e.g. in case of time series. This latter argument is often highly debated especially as the original statistical theory (i.e. random sampling) is not properly implemented. Please refer to Bartholomew et al. (2008), and Klockars (2005) for further explanation.

known or realized. We must estimate them using the sampled data. The corresponding numerical descriptions using the sampled data are known as estimates. And the manner or the mathematical formula in obtaining the estimates is known as the estimator. As you may have noticed, parameters and estimates are both examples of descriptive statistics described earlier. Statistical inference basically extends the descriptive statistics across the worlds of samples and population. More specifically, it bridges the two worlds by making informed probabilistic guesses of the unknown parameters using the parallel descriptor from the sampled data. By "informed probabilistic guesses", statistical inference first provides a specific value or guess of the parameter derived from the estimator and sampled data values. This is known as point estimation. The point estimate is then supplemented (or coated and shielded) with a margin of error and degree of uncertainty. This defines a range of plausible values around the point estimate that the parameter is likely to take. And this is also known as confidence interval estimation or hypothesis testing.[14] The probabilistic link between sample estimates and population parameter is established through the way the sample is drawn (i.e. randomization). The "randomness" of the sample is an important condition in conducting statistical inference. It is this condition which enables us to find the sampling distribution of the estimator which lies at the base of constructing the confidence interval and the probabilistic statements regarding the parameter values.[15] More specifically, since the estimator (the mathematical formula that translates the random sample to the estimate value) is a direct function of a set of random variables (the random sample), it also becomes a random variable.[16] That is, the estimator is also an unknown and stochastic variable with a probabilistic distribution. Now, sometimes the estimators are simple and straightforwardly derived. For example, the mathematical formula for calculating averages is a good estimator for the population mean.[17] But in other occasions (e.g. linear regression models), finding an estimator is not so obvious with the presence of competing candidates. In such a case, the various statistical properties of random variables (or the features of a probabilistic distribution) provide us with the effective criteria to judge the appropriateness and viability of the different estimators. The two most important properties for this purpose are unbiasedness and efficiency. Other properties include consistency, sufficiency, and robustness as described below.

[14] As Miller (2004) describes, traditionally, the problem of statistical inference is divided into problems of estimation and tests of hypotheses. The main difference between the two kinds of problems is that in problems of estimation we must determine the specific point value of the parameter, where as in tests of hypotheses we must decide whether to accept or reject a specific value or range of specific value of a parameter.

[15] Statistical theory is based upon the assumption of randomly drawn samples from a population. As described earlier, for a sample to be considered random, it must have been selected in such a way that each observation is independent of the others and has an equal probability of being included in the sample. This enables us to treat each observation as a random variable. It is not fixed or known with specific values but rather a stochastic variable with a probabilistic distribution. Being a random variable, the observed sample values essentially (probabilistically) represent any observation that could have been selected in the sample. This condition allows us to treat the sampled observations as if a repeated sampling has been conducted. That is, we can imagine as if the random sampling continued over and indefinite number of occasions. And for each sample we obtain an estimate for the parameter. The collection all estimates across the repeated samples then formulates the sampling distribution of the estimator which is used for the statistical inference (making probabilistic statements) of the population parameter.

[16] Random variables have a property where any direct function of random variables is also a random variable. Please refer to any probability theory textbooks such as Larsen (2005), Casella and Berger (2002), and Lindsey (1995).

[17] The proof that the mathematical formula in calculating averages is an unbiased, consistent, and efficient estimator of the population mean is provided in any introductory mathematical statistics textbooks.

Unbiasedness

Unbiasedness pertains to the accuracy of the estimator, that is, whether the estimator actually provides the approximation of the parameter of interest. An unbiased estimator is on average (across the repeated samples) equal to the population parameter it is supposed to estimate. It only differs from the actual population parameter due to random error and not to systemic error.[18] Formally, unbiasedness is represented by the expectation of the estimator equal to the population parameter of interest.[19]

$$Unbiasedness\colon\ E(estimator) = parameter$$

This implies that the sampling distribution of the unbiased estimator has the parameter as its center or the long run average. But now, if the estimator is biased, then the estimator is on average systemically and consistently different from the true parameter. The expectation of the estimator is no longer equal to the parameter of interest as shown below.

$$Bias\colon\ E(estimator) = parameter + bias$$

If we calculate the estimate for each sample and repeat this process infinitely many times, the average of all these estimates will be "off target" and does not tend towards its true parameter value. There is something inherently wrong with the estimator and the size of this problem (bias) is represented by $E(estimator) - parameter$. And this bias will not disappear by simply increasing more observations (in the sample) into the estimator. Graphically, the mean or center of the sampling distribution is no longer equal to the parameter. To summarize, Klockars (2005) intuitively explain the difference between unbiased and biased estimators as follow.

> A biased estimator is one which, on the average, tends to overestimate (or underestimate) the parameter. It is not necessarily that the estimator will always be too large or too small but that on the average, it is too large or too small. With an unbiased estimate we can make the statements such as "my estimate is 53 and I don't know whether this is too big or too little" while with biased estimate the statement would be "my estimate is 53 and that is probably too small (or too large depending on the direction of the bias)." (p.28)

Efficiency

Unbiasedness is generally one of the most important properties we look for in an estimator. However, unbiasedness alone does not mean that estimator is particularly good. There can be many unbiased estimators

[18] Random error unlike systemic error represents slight or minor differences or irregularity occurring from many natural imperfections of human decision e.g. imprecise recording, sampling, accidents in experiment setting, etc. It occurs by pure change and do not have systematic or consistent pattern and on average it has a value of zero. It is something we must live with, as no human endeavor can be made perfect in applications.

[19] Expectation or the expected value is a probability theory language used to denote the average value of random variable. It is the average value if the sampling continued over and indefinite number of observations. It is interchangeably used as the long run average of the random variable. It is what we expect in theory to happen in the long run.

but with different degree of dispersion in the estimates. For example, under the assumption that the true parameter has the value of 50, one unbiased estimator can have estimates of 40, 60, 30, 70 while another with 51, 49, 52, 48. But it is clear from this example that the latter estimator is preferable as estimates deviate less from the true parameter. Any specific estimate will be closer to the parameter. This "spread" or "variability" property of an estimator is known as efficiency. Efficiency pertains to the reliability and precision of the estimator. It is represented by the variance (or standard deviation) of the sampling distribution of the estimator. The so called "best" estimator would be the one with the least variability away from the parameter. Such estimator has the most tightly packed and peaked sampling distribution centered around the parameter of interest.

Unbiasedness and efficiency are the two most important statistical properties of an estimator as they both directly affect the statistical inference process. Specifically, unbiasedness affects the point value of the estimates (the numerator of the test statistic in statistical inference) and efficiency affects the standard error of the estimates that determine the confidence interval (the denominator of the test statistic). This is illustrated in the *t*-test statistic that is used for the statistical inference of the mean estimator ($\bar{X}$) and the linear regression estimator ($\hat{\beta}$) as shown below.

$$\begin{array}{lll} unbiasedness & \rightarrow & \dfrac{\bar{X}-\mu}{std(\bar{X})} \; or \; \dfrac{\hat{\beta}-\beta}{std(\hat{\beta})} \sim t_{df} \\ efficiency & \rightarrow & \end{array}$$

Conducting the most accurate and reliable statistical inference is the ultimate goal of statistical analysis. For this reason, it is crucial to obtain the most unbiased and efficient estimators. The estimator which is both unbiased and efficient has come to be known in the statistical literature as the "minimum variance unbiased estimator" or shortly as the "best unbiased estimator" (BUE).

Consistency:

Consistency pertains to the accuracy (unbiasedness) of the estimator when the sample size is increased. When the sample size is increased towards infinity and the estimator takes on values that are closer and closer (converges) to the parameter, then the estimator is consistent. It is also referred to as asymptotic unbiasedness.[20] Mathematically, consistency is described as follows.

An estimator is a consistent estimator of the parameter if and only if for each constant c > 0

$$\lim_{n\to\infty} P(\mid estimator - parameter \mid < c) = 1$$

This equation implies that when the sample size is sufficiently large, we can be practically certain that any form of consistent and systemic error will be less than any pre-assigned positive constant. Consistency of an

[20] Asymptotic property is a probability theory concept which studies the convergence in probability/distribution of a random variable. That is, the limiting property of an estimator when the sample size is approached towards infinity.

estimator can be evaluated by examining the role of sample size in the mathematical formula of the estimator. If the estimator mathematically converges to the parameter of interest as the sample size approaches infinity, then the estimator is consistent.

Sufficiency:

Sufficiency pertains to whether or not the estimator utilizes all the relevant information in the sample relevant to the estimation of the parameter. If an estimator is sufficient, then all the knowledge about the parameter that can be gained from the individual sample values (including the order of the parameter) can be obtained from the value of the estimate alone.[21] In essence, it is like the degree of representation of the data and population parameter by the estimator/estimate. A sufficient estimator is said to "nest" (or encompass) any other estimator as it better entails the information underlying the data and parameter. Mathematically, an estimator is a sufficient if and only if for each value of the estimator the conditional probability distribution of the random sample given the estimate value is independent of the parameter. That is, the estimate provides a good and pure representation of the parameter and the population distribution. In practice, sufficiency of the estimator can be derived by considering whether the mathematical formula of the estimator and the assumptions under which it derived satisfies the condition above.[22]

Robustness:

Robustness or insensitivity pertains to the extent to which the estimator is adversely affected by its underlying factors or assumptions. An estimator is robust if its sampling distribution is not seriously affected by these factors. This implies that the statistical properties such as the unbiasedness or efficiency are insensitive and not severely altered by the values of the underlying factors. These factors often include sample sizes, outlying observations, measurement and recording errors, missing data, etc. Robustness or sensitivity analysis can best be conducted by means of simulations (i.e. using the advanced computing techniques) which intentionally alter the numerical values of the underlying factors to assess the subsequent performance of the estimator. When the estimates are stern and do not show huge discrepancy, then this provides evidence of robustness of the estimator/estimates.

To sum, in conducting a viable and healthy statistical analysis (both descriptive and inferential statistics), ensuring the accuracy (unbiasedness), reliability (efficiency), and other statistical properties such as robustness becomes the most crucial task. The conditions and assumptions under which these properties are ensured become the goal in guiding our analysis. This challenge directly applies to the linear regression models which lie at the heart of VAMs. As a thorough understanding of linear regression models is critical for the successful implementation of VAMs, the following section solidifies the definition and mechanics underlying these

[21] Miller (2004)
[22] This process is slightly rigorous with mathematics and probability theory. Intuitive examples are provided in Hoff (2011).

models. The subsequent chapter will then provide a clear and intuitive explanation of the different assumptions in ensuring the "best" (for efficiency) "linear" (for linear regression) "unbiased" estimates (BLUE) for the linear regression models and VAMs.

Further Explanation of Regression Models

Regression models are a popular statistical method used today in exploring the relationship between variables. They are extensively used in both the hard sciences (such as in medicine, physics, agriculture) and behavioral sciences (such as in economics, sociology, psychology, etc.). The main reason for its popularity is due to its ability to provide a numerical description/summary of the structural relation between the explanatory variables and the outcome variable.[23] And as introduced earlier, this numerical description give rise to the popular interpretation of "effects", "cause", "prediction", "determination", and "contribution" of the explanatory variables on the outcome variable. These interpretations send strong and powerful messages which may suggest change and possible interventions.

Yet, what is often overlooked today is the formal and precise definition of regression. As described earlier, regression is any function which defines the conditional expectation (mean) of the dependent variable given the values of the independent variables. It is the average values of the outcome variable for different fixed values of the explanatory variables. Mathematically, this is represented as follows for a continuous variable y (which is in fact the definition of conditional expectation of random variable Y given the fixed values of x).[24]

$$\mu_{Y/x} = E(Y/x) = \int_{-\infty}^{\infty} y \cdot f(y/x)dy$$

The regression function can take any form as long as it captures the conditional mean of the outcome variable for different values of the explanatory variables. There are infinite possibilities. For example, non-parametric models (such as local averaging models or kernel density) apply a computationally advanced algorithm to iteratively estimate the conditional mean values for successive values of the explanatory variables to capture the most accurate regression function. It is a very flexible and reliable estimation process. But in exchange for its flexibility, it cannot provide us with a numerical descriptor which summarizes the relation or "effects" inherent between the variables. Moreover, it is not a parsimonious model as it attempts to maximally capture the underlying pattern of the data (e.g. a curve constantly alternating its directions does not really give us a summative measure of the pattern). For this reason, non-parametric models are best used as a diagnostic tool in examining the data.

[23] Structural relation unlike correlation has a direction of influence in the association of variables. Namely, the outcome variable is explained, affected, and determined by the explanatory variables.

[24] y is the specific values of the outcome variable. And f(y/x) is the distribution or frequency of the y values given x. Taking the product of y and f(y/x) and summing across all its values is precisely the definition of conditional average. This is clear when y is a categorical variable with a restricted set of values it can take. When y is a continuous variable, the integration precisely does this calculation for us.

In modern statistical practice, we focus our attention on a special and simplified kind of regression function known as a linear regression function.[25] Linear regression, unlike non-parametric models, is defined (more precisely as "constrained" or "simplified") in terms of parameter(s) that describe the structural relation between the explanatory variables and the outcome variable.[26] For this reason they are also referred to as parametric models. But moreover, the parameters are only allowed to be raised to the first power which gives rise to the term "linear". Linear regression models therefore impose some constraints and structure of the form of regression function. This is best illustrated by means of examples where $E(Y/x) = \alpha + \beta_1 x_1 + \beta_2 x_2$ is a linear regression while $E(Y/x) = \alpha + \beta_1^2 x_1 + \log(\beta_2) x_2$ is not a linear regression. The former is linear in the parameters β_s while the latter is not linear in terms of the parameters. But linear regression does not necessarily have to be linear in the explanatory variables. For example, $E(Y/x) = \alpha + \beta_1 x_1^2 + \beta_2 x_1 x_2 + \beta_3 \log(x_3)$ is an example of linear regression. The key advantage of the linear regression models as already described lies in its simplicity and interpretability. The parameters provide us with the numerical description of the magnitude of the structural relation inherent between the explanatory variables and the outcome variable. For example, in the case of simple linear regression with only one explanatory variable (raised only to its first power), the parameter represents the change on the average value of the outcome variable for a unit increase in the explanatory variables. This is essentially the magnitude of the slope coefficient of a straight line.[27] And in the case of multiple regression with multiple explanatory variables, the parameter (for a particular variable) represent the change in the average value of the dependent variable for unit increase in the explanatory variable after removing the effects (common dependencies) of the other variables.[28] In both cases, the parameters imply the notion of "effect", "cause" or "change" in the outcome variable for a unit change in the explanatory variable(s). But in exchange for its simplicity, the key disadvantage of the linear regression models is that it must satisfy a number of conditions/assumptions to provide an accurate (unbiased) and reliable (efficient) estimate of the regression function. That is, as linear models impose different forms of constraints on the form of the regression function, it must satisfy their own constraints (and the other subsequent conditions) to justify its usage. Thorough explanation of these assumptions will be provided in the section below and the subsequent chapter. But before we do so, as we have now stepped in the realm of statistical inference, the different estimation processes underlying linear regression models are first described.[29] Specifically, there are two prominent estimators for the linear regression function: the maximum

[25] Linear regression is often mistakenly defined or accredited as the definition of regression. But it is in fact only a simplified type of infinitely possible regression functions.

[26] Parameter is any numerical description of the population. As described before, it is language of statistical inference. It represents the true (parsimonious) characteristics of the population under study. In the context of regression, parameter is the summary of the relation, structure or effects of the explanatory variables on the outcome variable underlying the unknown population the data was generated. Further explanation of statistical inference pertaining to regression is provided in later in the text.

[27] When the explanatory variable is raised to the second power or multiplied (interacted) with another variable then the regression estimate or the parameter estimate is now a function of the values of these variables. Taking the first derivative with respect to the main variable of interest will illustrate this.

[28] The process of taking into account process the effects of other variables will be thoroughly described in later chapters.

[29] In practice, our main interest is to utilize the sampled data to estimate the parameters characterizing the population regression function. To do so, we first need to construct the probable structure of the population regression function that has generated the data. This task can be completed based on our theoretical understanding and prior empirical findings related to the variables of interest (this process will be further explained in

likelihood estimator and the least squares estimator. The explanation of these two estimators and the assumptions under which it provide us with the accurate and reliable (BLUE) estimates are as follow.

Maximum Likelihood Estimation

The maximum likelihood estimation (MLE) is based on the idea that we choose an estimate for the unknown parameter such that the probability of observing the sample values is at the maximum. It is founded on probability theory. Specifically, MLE solves for the regression parameters such that it maximizes the joint distribution of the outcome variable (defined by the parameters and data). This joint distribution is known as the likelihood function. And to obtain the likelihood function we first need to understand the assumptions under which is derived. MLE was founded by R. A. Fisher (1922) who showed that under the following four assumptions (of the population regression function), the method of maximum likelihood can be used to provide accurate and reliable (BLUE) estimates for the regression parameters: (1) Regression is a linear regression model as defined earlier. (2) Conditional distribution of the outcome variable given the explanatory variables is normally distributed. (3) The variance associated with the conditional distribution is constant for all values of the explanatory variables. (4) The conditional distributions of the outcome variable are independent random variables.[30] These assumptions can be collectively represented in the following equation.

$Y_i \sim N(\alpha + \beta x_i, \sigma^2)$

Alternatively,

$Y_i = \alpha + \beta x_i + e_i$ where $e_i \sim N(0, \sigma^2)$ and $E(Y/x) = \alpha + \beta x_i$

Now, by the definition of independence, the probability of obtaining the observed outcome variable values given the explanatory variables (its joint conditional distribution) is the product of observing each value as shown below.

$$(Y_1, Y_2 \dots Y_n | \alpha + \beta x_i, \sigma^2) = f(Y_1 | \alpha + \beta x_i, \sigma^2) \cdot f(Y_2 | \alpha + \beta x_i, \sigma^2) \cdots f(Y_n | \alpha + \beta x_i, \sigma^2)$$

And with the other three conditions stated above, this joint distribution or the likelihood function (*LF*) becomes

$$LF(\alpha, \beta, \sigma^2) = f(Y_1, Y_2 \dots Y_n | \alpha + \beta x_i, \sigma^2) = \frac{1}{\sigma^n (\sqrt{2\pi})^n} e^{\left\{-\frac{1}{2}\sum \frac{[Y_i - (\alpha + \beta x_i)]^2}{\sigma^2}\right\}}$$

The problem of MLE estimation then becomes finding the values (estimates) of the unknown parameters $(\alpha, \beta, \sigma^2)$ which maximizes the likelihood function. This can be achieved through simple calculus by partially differentiating the likelihood function respect to the three parameters and then solving these optimizing

Chapter 6). Once the population regression function is defined, we can then use the observed data to estimate its parameters. But in doing so, we would need to estimate it both accurately and reliably and the underlying assumptions become the key guide.

[30] Independence implies that the correlation is zero. That is the covariance between the random variables is zero. Moreover independence implies that the joint distribution is a product of its marginal distributions.

equations for the parameters simultaneously. This process for (α, β) reduces down to minimizing $[Y_i - (\alpha + \beta x_i)]^2$. Before proceeding to the final form of the MLE estimator, the second estimator known as the least squares estimator is described. As illustrated below the two estimators share very similar derivation process.

Least Squares Estimation

The method of least squares (méthode des moindres carrés) is a mathematical algorithm and curve fitting method introduced during the early nineteenth century by a French mathematician Adrien Legrendre and a German mathematician and physical scientist Johann C.F. Gauss. The authors both applied this method to the problem of determining the astronomical orbits of planets around the Sun. Specifically, they used it as a method to minimize measurement error or the vertical deviation of the estimated curve (defined by $\hat{\alpha}$ and $\widehat{\beta}$) from the observed data points (y_i) as shown below. This process entailed an estimated model with the maximal goodness of fit or "line of best fit" to the observed data points.

$$minimize \sum_{i=1}^{n} (error)_i^2 = \sum_{i=1}^{n} [y_i - (\hat{\alpha} + \hat{\beta} x_i)]^2$$

Yet at this point in history, there was no application of the least squares methods (here on as LS) to the field of statistics i.e. statistical inference. It was only used as a mathematical and descriptive tool to summarize the data. But following Fisher's work on the method of maximum likelihood, researchers started to realize the mathematical similarities with the least squares method. In fact, as you can see from the above equations, the minimization process of the sum of squared deviation and maximization process of the likelihood function for the regression parameters $\hat{\alpha}$ and $\widehat{\beta}$ are identical. The two estimators are the same and will provide the same estimates for these parameters. But this equivalence only holds under the premise that Fisher's MLE assumptions also hold for the LS. Moreover and importantly, as the MLE are BLUE under these conditions, if these assumptions hold, the LS estimates will also be BLUE. This finding (i.e. proof) marked an important milestone in the history of statistical research and was officially acknowledged as the Gauss-Markov Theorem. The name was attributed to Andrey Markov in the early 1900s who rediscovered the potential of LS method by Gauss and Legrendre in statistics for the analysis of regression models. Having now understood the similarities of these two estimators, the final mathematical form for these estimators (solved from the first order equations) are illustrated as follow.

$$\hat{\beta} = (X'X)^{-1}X'Y \quad where\ X'X = \begin{pmatrix} n & \Sigma x_1 & \cdots & \Sigma x_k \\ \Sigma & \Sigma x_1^2 & \cdots & \Sigma x_1 x_k \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma x_k & \Sigma x_k x_1 & \cdots & \Sigma x_k^2 \end{pmatrix} \ and\ X'Y = \begin{pmatrix} \Sigma y \\ \Sigma x_1 y \\ \vdots \\ \Sigma x_k y \end{pmatrix}$$

$\hat{\sigma}^2_{MLE} = (Y'Y - \beta'X'Y)/n \quad and \quad \hat{\sigma}^2_{LS} = (Y'Y - \beta'X'Y)/n - k$ [31]

$std(\hat{\beta}) = \sqrt{\hat{\sigma}^2/X'X}$

Now, as $\hat{\beta}$ is function of *Y* which is distributed normally with mean of zero and constant variance, through the multiplicative property of the normal distribution, $\hat{\beta}$ is also distributed normally as follow.

$$\hat{\beta} \sim N(\hat{\beta}, \frac{\sigma^2}{\sum(x_i - \bar{x})^2})$$

And through substituting the sample $\hat{\sigma}^2$ value, the following term is distributed as a *t* distribution with n – k degrees of freedom.

$$\frac{\hat{\beta} - \beta}{\sigma/\sqrt{X'X}} \sim t_{n-k}$$

Based on this *t* distribution, we can then complete the statistical inference for the true population regression parameter (β) by means of hypothesis test or constructing a confidence interval around the estimated value ($\hat{\beta}$).[32] That is, we can make probabilistic (percentage) statements of the plausible values of the unknown population parameter using our estimates derived from the sampled data.

Before we become comfortable with our conclusions and generalizations, we must assess the accuracy and reliability of our findings. As described earlier, the attainment of accuracy (unbiasedness) and reliability (efficiency) are incumbent upon whether the estimated models meet the underlying regression assumptions. The Gauss-Markov Theorem only holds under the four MLE assumptions addressed by R. A. Fisher (1925). But the assumptions necessary for the LS estimators to provide BLUE estimates continues to be debated. As the LS estimation became an active research area around the 1930s (particularly with the rise of econometric field), the underlying assumptions to ensure BLUE estimates were further investigated. Through the application of rigorous mathematical theories and with the development of electromechanical desk calculators which fastened the matrix algebra calculations, Fisher's four assumptions were contested. The seminal work by Casella and Berger (2002) summarized these conditions as follow.

[31] The MLE estimate of the variance in fact biased in small finite sample (small N) but unbiased asymptotically (N goes to infinity). The reason is because as you can see from the equations, it does not take into account the lost in the degrees of freedom (*k*) through estimating the regression parameters. This necessitates the adjustment to the MLE estimates when conducting inference especially when N is small. This estimate is known as the restricted MLE (RMLE). Least squares estimates on the other hand are unbiased both in finite and infinite sample.

[32] Hypothesis testing and confidence interval estimation are similar but different topics. The difference essentially pertains to the set up and procedure. The hypothesis test stipulates the value of the population parameter we would like to test at a given significance level. The data and estimates are then used to test and confirm this hypothesis. Confidence interval estimation on the other hand does not have this kind of stipulated rules and procedure. It is constructed just like we obtain the point estimate. But both procedures utilizes the same set of information but in different form and procedure with different types of conclusions.

> Under the following fairly general statistical model, the least square estimate of a simple linear regression is optimal in the class of linear unbiased estimates (BLUE): (1) the values x1, x2, …, xn are known, fixed values. (2) the values y1, y2,…,yn are observed values of uncorrelated random variables Y1, Y2,…,Yn. (3) the relationship between the xs and the ys are linear as follows EYi= a + Bxi for i =1,…,n. (4) the variance of Yi is assumed to be constant homoscedastic. (5) no probability distribution of the Yi is assumed (only the first two moments of Yi are assumed). (p.544)

The key distinction addressed by the authors (together with the support from the econometricians) pertains to the normality assumption. Through the application of asymptotic theory[33] and the use of central limit theorem, the authors have shown that the LS estimator converges to a normal distribution. The conditional normal distribution of the outcome variable for the given values of the explanatory variables was perceived as an unnecessary assumption. As the LS estimators require less assumption to achieve the BLUE regression estimates, the econometricians claimed that it is a better (more parsimonious) estimator than the MLE.[34] But today, this point is not free from challenges. There is an increasing acknowledgement of other viable reasons as to why normality is an important condition behind the BLUE properties. These reasons are introduced in the footnote and will be thoroughly explained in the next chapter.[35]

Leaving the normal distribution debate aside for the moment, the assumptions outlined by Casella and Berger (2002) above have been further clarified in many modern textbooks in order to better assist data analysis. For example in Gujarati (2008), the assumptions for MLE and LS estimators to provide BLUE regression estimates are listed as follows:

- Linear regression model where the regression equation is linear in the parameters but not necessarily linear in the variables e.g. Yi = a + bXi + ei
- Fixed X values or X vales independent of the error term. This implies that the covariance between ui and each X variables is zero. cov (ei,X2i)=cov(ei,X3i)=0. This is also known as the exogeneity condition.
- Zero mean value of the disturbance ei, where E(ei/X) = 0 for each i
- Homoscedasticity or constant variance of ei
- No autocorrelation or serial correlation between the disturbances, cov(ei,ej)=0
- Normal distribution of the ei
- The number of observations n must be greater than the number of parameters to be estimated
- There must be variation in the values of X and Y variables.

[33] Analysis of the estimator (i.e. its convergence properties) when the sample size N goes to infinity

[34] Econometric field has addressed that if conditional normality were to be assumed it should be empirically tested e.g. using the Hausman and Heckman tests

[35] As summarized in Fox (2008), there are several reasons why we employ the normality assumption: 1) The error residuals represents the combined influence of a large number of randomly distributed independent variables that are not explicitly introduced in the model e.g. omitted variables, measurement error, etc. The bell shaped and symmetric distribution such as the normal distribution is a good representation of such randomly distributed variable i.e. there is an systemic trend, pattern, skew in the distribution (on average the residuals are zero, there are equal frequency of making positive and negative errors, and the frequency of making large magnitude errors consistently falls). 2) Normality assumption of the outcome variable is basis for other distributions used for the statistical inference including t, chi-square and F distribution. These distributions are derived under the assumption of normally distributed outcome/residuals. 3) Normal distribution (in the family of bell shaped and symmetric distribution) is a comparatively simple distribution involving only two parameters (mean and variance) and its properties are well known and extensively studied in mathematical statistics. It effectively meets the purpose of parsimony underlying statistics. 4) Many phenomena with continuous outcomes often follow the normal distribution. That is, normality is a good approximation of many phenomena.

- No exact collinearity (exact linear relationship) between the X variables
- There is no specification bias. That is the model is correctly specified

In modern statistical practice, these assumptions have become the key points or the "commandments" in guiding our empirical analysis. Gujarati (2008) refers to them as the "classical linear regression model assumptions" (CLRM). Again only when these assumptions jointly hold in our estimated models, the accuracy (unbiasedness) and reliability (efficiency) of our regression estimates are ensured. This will then provide us with the viable statistical inference of the population parameters.

Summary and Conclusion

This chapter re-examined the purpose of statistics which has become a pivotal part of our modern academic research. The notion of parsimony which underlies the field of statistics together with its two key dimensions – descriptive statistics and inferential statistics – were (re)defined. The important role of statistics in our daily lives was also addressed. Special focus was provided on linear regression models, a popular statistical tool used today to describe the relationships inherent between variables. These models also lie at the heart of the VAMs. In doing so, the importance of the underlying assumptions which ensure the accuracy and reliability of the estimates was highlighted. These assumptions became the key guiding principles of our empirical analysis including the VA analyses. Therefore in the following chapter, further explanation of these assumptions will be provided. An intuitive and user-friendly explanation of the overarching ideas underlying these assumptions which is often overlooked in modern statistical text books is illustrated. The different methods to diagnose and correct these assumptions in our estimated models will also be introduced. The understanding of these methods and processes will play a vital part in the implementation of VAMs using data from the Washington State.

CHAPTER 3: STATISTICS AS A SCIENCE OF ASSUMPTIONS – CONDITIONS TO ENSURE ACCURATE AND RELIABLE REGRESSION ESTIMATES

Statistics is a science of parsimony but it is also a science of assumptions. As introduced in the previous chapter, for linear regression estimators to provide accurate (unbiased) and reliable (efficient) linear estimates (BLUE), the estimated model must satisfy a number of assumptions. As these assumptions became the backbone of the models, it was referred to as the key "commandments." In this chapter we therefore take a closer look at these assumptions. In doing so, we first identify the two major groups or classification underlying these assumptions: the first group pertains to the behavior of the residuals, and the second pertains to the behavior of the explanatory variables. This classification is shown in the table below. In the following sections, the explanation of the overarching ideas underlying these two groups of assumptions is provided. The explanation of how the individual assumption plays an important part of these overarching ideas and the consequences on the BLUE property when it is violated the follow.

Table 3.1: Linear Regression Assumptions

I. Properties of the Residuals – Randomly Distributed
Levels of the Residuals:
Mean value of residuals given Xs is zero. E(e/X) =
Expectation and covariance between residuals and Xs is zero. Residuals are independent of Xs. E(e,X)=0, Cov (e,X)=0
Normally distributed residuals
No extreme/outlying influential residuals
Variance (Spread) and Covariance (Dependency) of the Residuals:
Homoscedasticity constant variance of residuals
Independence of residuals i.e. no autocorrelation, serial correlation, clustering or any form of dependencies/covariance between the residuals. Cov(ei,ej)=0
II. Properties of the Explanatory Variables
Clear variation in the values of the Xs
No perfect association and collinearity between the Xs
No extreme high leverage and influential X values

Behavior of the Residuals

If the model were to operate well (i.e. to provide BLUE estimates), we expect it to err in predicting the outcome variable in a random meaningless fashion.[36] The term error is interchangeably used with residuals. Residuals (or error) are the portion of the essence or variation of the observed outcome variable the estimated model fails to capture. It is the representation of the imperfection of the estimated model due to the lack of theoretical understanding, unavailability of the data or difficulty in measuring other factors that could have explained the outcome variable.[37] In ensuring the accuracy and reliability estimates, these errors must be randomly distributed with no consistent, systemic, and regular pattern, fluctuations, drift or structure. It must

[36] Engineering Statistics Handbook (2012). The notion of random error or randomly distributed pattern is an extremely vital concept in the field of statistics. It lies at the heart of experimental design and regression models (as explained in the text). Randomization and randomly distribution was in fact one of the first topics extensively researched by R. A. Fisher who pioneered in experimental design and regression models.

[37] Specifically, the residuals can comprise a host of factors such as omitted variable, different functional form of modeled variables, sampling error (from the data collection process), missing data, measurement error (e.g. miscoding of the variables), etc.

occur primarily as a result of natural imperfection of the model and consequently entail no intrinsic meaning or significance in its pattern.[38] This randomly distributed error is defined in contrast to systemically patterned error which is reproducible discrepancy (bias) that makes our measurement systemically different from the true value. It has consistent, meaningful and significant effects on the outcome variable.[39] When systemic error is present, the regression estimates are not accurate or reliable as we can not only potentially improve the model performance (further capture the variation of the outcome variable) but the estimates can also be potentially distorted and corrupted with the effects of the un-accounted systemic factors. It is therefore vital to achieve residual errors that are randomly distributed.

In formal statistical language, randomly distributed patterned variables are described as independently and identically distributed as normal distribution with mean of zero and constant variance. Applying this to the case of residuals (e_i), this is symbolically represented as[40]

$$e_i \sim iid\ N(0, \sigma^2)$$

As you may have noticed, this equation surmises all the assumptions pertaining to the behavior of the residuals shown in the above table. Moreover, (as shown in the finer classification in the above table), it is based on properties of both the level and spread of the residuals. In other words, for the residuals to portray a randomly distributed pattern, its entire dimension (both the level and spread) must not illustrate any form of systemic and meaningful pattern. To better illustrate how each of these assumption plays a role in constructing the randomly distributed patterned residuals, the following explanation is provided.[41]

[38] For example, in the case of student testing, random errors may arise from irregular mistakes in the marking of tests, miscoding student background information, students happen to be sick on the test date (performing below their potential), etc. These factors all affect the final outcome variable but it has no systemic meaning, significance, or pattern in explaining the variation of the outcome variable.

[39] Examples include skewed scaling of the tests, rounding of measurements in the process of recording, unfair marking of the scorer who is excessively lenient or hard for particular group of students, missing test scores of student with certain backgrounds which has an effect on test scores (e.g. ethnicity), etc. These factors all have a meaningful and consistent association with the outcome variable and if they are left in the residuals it will induce a "non-random" systemic pattern.

[40] i represents the cases, observation, and the lowest unit in the data set.

[41] The origin and derivation of the notion of randomly distributed residuals and the linear regression assumptions can be traced back to the study over the concept of randomization and randomized experiments mainly attributed to the work of R.A. Fisher (1925, 1935). Randomization or random assignment of treatments (the main variable of interest) enables all the factors un-accounted in the model to be (probabilistically) equalized between the treatment and control group. This (statistically) ensures that all the un-accounted factors are missing randomly with no systemic association or pattern with the treatment. This leaves the treatment not to be confounded (mixed or corrupted) with effects of other variables and the reliable and accurate (causal) effect can be estimated. And in estimating the treatment effects, the linear models (linear regression and ANOVA combined) is precisely used for this purpose. And for this reason, the notion of "randomly distributed pattern" of all the factors un-accounted in the model (the residuals) underlies as the overarching idea of the linear regression assumptions. In the ideal setting where the experimenter/analyst can randomly assign treatments/variables to the subjects, the linear regression assumptions are (statistically) met and justified. But in cases where the experimenter cannot randomly assignment the treatments but need to analyze naturally observed non-random data, we must work to reverse engineer and re-create the conditions of randomized experiment (which ensures the reliable and accurate causal estimates) by satisfying the linear regression assumptions. As OECD in NRC (2010) addressed, "a related way of thinking about value-added models [or any empirical analysis using observed data] is that they are an attempt to capture the virtues of a randomized experiment when one has not been conducted." (p.8). Similarly, as Singer in NRC (2010) addressed, "with value-added methods, one is trying to develop analytic fixes or measurement fixes, for what is basically a design problem: students are not randomly assigned to teachers [or schools]." (p.8).

The Mean Value of (the Level of) Residuals Equal to Zero: E(e/X)=0

This condition provides us with the first indication of whether the model is effectively fitting the outcome variable and any factors left un-accounted in the model (residuals) are distributed randomly with no meaningful pattern. If the model is performing well, then the majority of the error should be gravitated around zero leading to an average value of zero. This implies that the model (on average) is capturing the actual observed outcome variable. It does not commit any obvious error that induces a large and systemic deviation of the fitted values from the actual observations. The imperfections of the model are solely attributed to meaningless random error which averages out to zero. But if this condition does not hold and large and consistent error is evident, then this signifies that an important source of variation which can explain the outcome variable is failed to be accounted in the model. A systemic pattern remains in the residuals and the randomly distributed pattern is not achieved.

*The Exogeneity Condition: E(e*X)=0*

The exogeneity condition implies that for the residuals to be randomly distributed, the levels of the residuals should also not illustrate any systemic pattern with any variables including the model fitted values, explanatory variables in the model, explanatory variables not in the model, any functions of these explanatory variables (e.g. nonlinear), etc. If a systemic association is detected with any these variables, then this provides evidence that an important factor is left un-accounted in the model. If the variable is a factor excluded from the model it can now be considered to be incorporated while if it is a factor already included in the model, then a new functional form can be considered. In both cases, the violation of the exogeneity condition implies than an important and meaningful systemic pattern is left in the residual and a randomly distributed pattern is not achieved. Mathematically, the exogeneity condition implies the covariance (e*X) = 0 which under the assumptions of residuals simplifies to E(e*X) = 0.[42] The exogeneity condition can be effectively assessed using different graphical tools such as the scatterplot. If the residuals are randomly distributed with no association with other variables, then the scatterplot will illustrate a flat curve. But if a clear positive or negative trend is detected, this will suggest that an important factor is left un-accounted in the model. As an important systemic pattern is left in the residual, a randomly distributed pattern is not achieved.

Normally Distributed Residuals

The normal distribution assumption implies that if the residuals are randomly distributed, then the estimated model should predict values higher than actual and lower than actual with equal probability. That is, there should be no lean, skew, or pattern in the magnitude and pattern of the errors. Unsymmetrically distributed error (e.g. bimodal) can signify an important binary variable un-accounted in the model.[43] Furthermore, frequency of the error should consistently fall with the size of the error. That is, if the model were to

[42] The key assumptions are the mean of residuals being zero and the independence assumption described later. The lower case x implies the mean centered X values.

[43] This can include gender, binary coded ethnicity, participation in a certain program, attainment of a certain qualification, etc.

effectively capture the pattern of the outcome variable, it should naturally make less errors or "misfits" with big positive or negative sizes. There should not be high frequencies of large model imperfections and errors. These two features together epitomize a bell shaped symmetric distribution such as the normal distribution which is centered around zero.

Absence of Outlying and Influential Observations/Residuals

This assumption implies that if the residuals are randomly distributed, there must be no single (or few) residuals which heavily deviate from the rest of the residuals to exert systemic pressure to disrupt the overall pattern (of residuals gravitated around zero). These extreme values can potentially signify important information of the outcome variable which we fail take into account e.g. extra-ordinary or beat the odds cases which can provide vital and exceptional information in explaining the outcome variable. In such case, an important systemic pattern is induced in the residuals and a randomly distributed pattern is not achieved. On the contrary, if it is due to irrelevant occurrences e.g. miscoded observations, measurement error, randomly missing data, etc. it can potentially be discarded from the data. The detection and handling of outliers therefore demand careful examination.

Homoskedasticity – Constant Spread of Residuals

The homoskedasticity assumption implies if the residuals are randomly distributed the spread of the residuals should also portray no systemic pattern of any form. The spread ("skedasticity" measured by the variance) of the residuals should be constant and equal ("homo"). It should not illustrate any pattern with other variables including the predicted values, variables included in the model, and variable not included in the model.[44] But if the variance varies and illustrate an increasing or decreasing pattern (known as heteroskedasticity), this represents a systemic and structural pattern of the residuals that is yet to be taken into account in the model. It conveys meaningful information and consequently the randomly distributed pattern for the residuals is not achieved.

Independence of Residuals

The independent residuals assumption implies that the residuals themselves should also not illustrate any form of association or "dependency". That is, the covariation of the residuals must be zero.[45] There must be no intrinsic pattern between the residuals. But depending on the phenomena (outcome) being studied, this assumption can be violated. For example, for time series data, the repeated observations (residuals) of the same individual are often always correlated. That is, the same individual is likely to perform similarly across time on the same outcome variable e.g. due to their innate ability and idiosyncrasies. The repeated observations

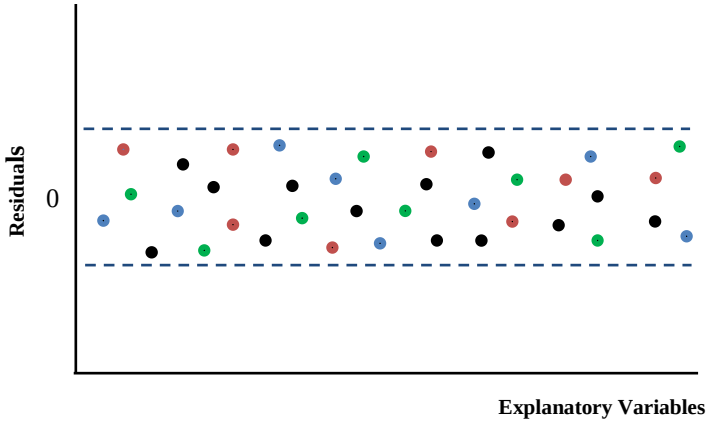
[44] Mathematically, the variance of the residuals is represented in the diagonal elements of the residual variance-covariance matrix. Homoskedasticity implies that the diagonal elements are the same constant.

[45] The covariance of the residuals is represented in the off diagonal elements of the residual variance-covariance matrix. Independence implies that the off diagonal elements are zero.

across time are in essence grouped within the same individual. In the context of VAMs, similar student observations (e.g. test scores) are also clustered within the same teacher, school and district as students with similar demographic background are often grouped together and at the same time share the same learning environment. In both cases the clustering and grouping of the observations induces a form of dependency and systemic pattern between the residuals. Unless this grouping factor is somehow taken into account in our model, it will forbid us from achieving randomly distributed patterned residuals.

To summarize randomly distributed residual patterns are built on a number of key components. A randomly distributed pattern can only be achieved when both the level and spread of residuals convey no systemic meaningful pattern. In practice, these components can in fact be jointly and conveniently represented in a horizontal band shaped scatterplot centered around zero as shown below.

Figure 3.1: Horizontal band shaped randomly distributed residual pattern



As you can see, the residuals are centered around the average value of zero implying that *E(e/X)=0*. The flat line of the residuals signifies no association with any variable which can be represented on the x-axis. The dotted lines represent the 95% spread of the residuals along the estimated line. And the parallel feature or band shape of these lines signify that the variance remains constant (homoscedastic) across the values of the x-axis variable(s). The coloring of the dots which indicate the groups to which the residuals belong also indicate no clustering or dependency across the plot. This horizontal band which illustrates a randomly distributed pattern is clearly distinct from residuals with systemic patterns as shown below. The first figure illustrates heteroskedastic residuals with increasing variance and non-independent clustered pattern while the second second figure illustrates residuals with an inherent increasing pattern most likely due to an important variable (shown on the x-axis) omitted from the model.

Figure 3.2: Heteroskedastic and non-independent residua pattern

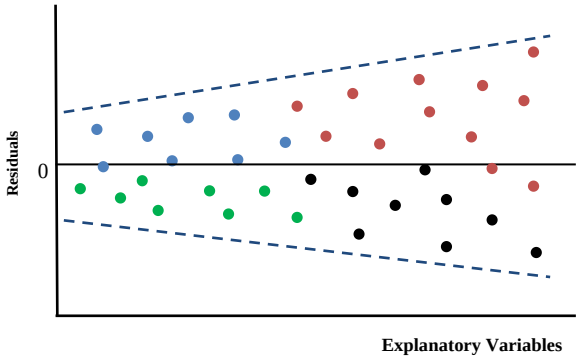
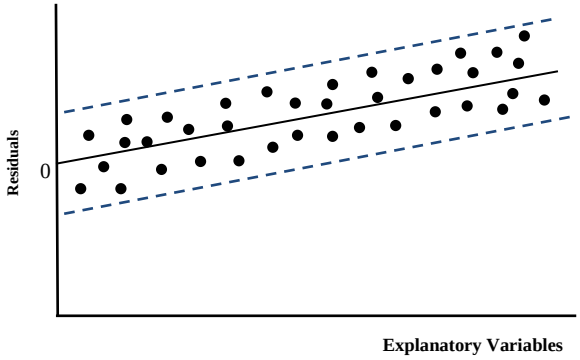


Figure 3.3: Increasing residual pattern



To sum, the horizontal band scatterplot together with the normal distribution provide the two key sources of evidence that signifies the residuals (or all the factors un-accounted in the model) follow a randomly distributed pattern.[46] Moreover, these graphs consequently become the key evidence which will guide and monitor our empirical analysis. When these two figures are consistently achieved in our estimated model, it gives us the best indication that the model estimates are accurate and reliable i.e. BLUE.

Behavior of the Explanatory Variables

The second major group of assumptions for regression models concerns the behavior of the explanatory variables. The explanatory variables determine, explain, and capture the essence of the outcome variable. The explanatory variables must be uniquely defined with their own distinct variation that is not comprised of extreme values that can distort the overall model estimates. The explanation of these features is as follows. First, the explanatory variables must be given, fixed, known, and non-stochastic. They must not be unknown random variables characterized by a probabilistic distribution (like the outcome variable). Second, as described earlier, the variables must be independent and have no systemic relation with the residuals. This condition was referred to as the exogeneity assumption $E(eX)=0$. Third, the explanatory variables must have variation. That is, the spread (variance) must not be zero. If the explanatory variables have no variation, then a regression model defined as the function connecting the conditional means of the outcome variable for different levels (variation) of the explanatory variables will not be defined. If the explanatory variable takes only a single value, then there will only be a single regression "point" modeled (the overall mean of the outcome variable). The function which estimates the approximate change in the average outcome value for a unit increase in the explanatory variable will not be identified.[47] Fourth, analogous to the case of outlying residuals, there must not be explanatory variables with highly deviant and influential values. Such values are referred to as high leverage points. High leverage points can potentially disrupt the model by exerting misleading influence on the parameter estimates. Fifth, the explanatory variables must not illustrate perfect or

[46] Although not clearly illustrated figure, it is assumed that when a vertical slice of residuals is taken for each value of the x-axis variable, the residuals should illustrate a bell shaped symmetric normal distribution.
[47] Mathematically, there will be no converging solution to the first order optimizing equations under the least squares and maximum likelihood framework. Please refer to previous chapter for further details.

close to perfect correlation. When there is a perfect correlation between certain variables, then the regression model cannot distinguish between the variables. As the variables are not uniquely identified, the model cannot provide converging estimates for the parameters. When there is not a perfect correlation but a very strong correlation (above 0.9), the model is subject to the multicollinearity problem where estimates become highly unstable and imprecise with huge standard error values. Multicollinearity will mislead statistical inference with aggravated confidence intervals and underestimation of significant results. This point will be further explained in the section below. To sum, if these properties do not hold in the explanatory variables, the regression estimates are either not defined at all or are not ensured of the BLUE property.

Consequences of Violating the Assumptions on the Statistical Properties of the Regression Estimates

In the previous section, the overarching idea and intuition underlying the two major groups of regression model assumptions were described. But it did not clearly illustrate how it ensures the BLUE property in the estimates. Therefore, in this section, the explanation as to how and why the different assumptions within each group contribute to the attainment of BLUE estimates is provided. We illustrate this by examining how the violation of the different assumptions leads to bias and/or inefficiency (and possibly lack of robustness) of the estimates. As it becomes evident, this process is quite complex and intricate as there are multiple factors that can violate multiple assumptions simultaneously and consequently cause multiple damages to the statistical properties of the estimates. Some of these key sources of violation are as follow.

- Other explanatory variables that could have been included in the model to explain the level or the spread of the outcome variable
- Different functional form of the included explanatory variables (e.g. nonlinear function)
- Measurement error of the outcome and explanatory variables (e.g. due to improper calibration or scaling)
- Missing data that is missing in a non-random fashion (e.g. individuals with particular background having higher tendency of being missing)
- Extreme and influential cases for the outcome variable and explanatory variables (e.g. as consequence of measurement error)
- Clustered and grouped observations due to the nature of the underlying phenomena (e.g. student observation in the same teachers and schools, patient information in the same hospital, repeated observation of the same individuals)
- Similarly defined explanatory variables which induce extremely high correlation (e.g. age and years of working experience)

The explanation of how these factors violate the different assumptions and the BLUE statistical properties of the regression estimates is provided next.

The Violation of E(e/X)=0 Condition, Exogeneity Condition, and More – Implications of Omitting Important Variables

When important factors (such as another explanatory variable, different functional forms, non-randomly missing observations) which systemically affect the outcome variable are omitted from the model, then almost all of the assumptions necessary to construct the randomly distributed residuals are in jeopardy. This

challenges both the unbiasedness and efficiency of the regression estimates. First, as an important source of variation is left into the residuals, this implies that the estimated model is not effectively capturing the outcome variable. The expected value of the residuals is likely not to be zero (or can be improved to be closer to zero). The violation of the *E(e/X) = 0* assumption implies that the estimated model is in fact making unnecessarily more misfits and errors and this will lead to poor model performance. With higher error values, this will lead to larger residual variance estimates. And as the standard error of the parameter estimates are directly proportional to residual variance, this finding will induce inefficiency in the estimates. Second, when the omitted variable illustrates systemic association with the explanatory variables in the model, then this violates the exogeneity condition. As the omitted variable is not under the control of the model as a fixed explanatory variable, it is left free (in the residuals) to distort the parameter estimates. Through its correlation with the modeled variables (*E(e*X) ≠ 0*), the estimated coefficient of the modeled variables is "mixed" or "contaminated" as it is picking up another effect which is not unique of its own. That is, the violation of the exogeneity condition implies that the parameter estimate is inaccurate and biased as it represents the effects of another variable on the outcome variable.[48] This bias is illustrated in the following equations.

Suppose the true or correct population model is the following (all variables are in deviation form from its respective means e.g. $y_i = Y_i - \bar{Y}$)

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$$

But for some reason we mistakenly omit variable x_{2i} and fit the following model

$$y_i = a_1 x_{1i} + v_i$$

Conducting the usual OLS, the estimate for a_1 is

$$\hat{a}_1 = \frac{\sum x_{1i} y_i}{\sum x_{1i}^2}$$

Now substituting the true correct model, we get

$$\hat{a}_1 = \frac{\sum x_{1i}(\beta_1 x_{1i} + \beta_2 x_{2i} + u_i)}{\sum x_{1i}^2} = \beta_1 + \beta_2 \frac{\sum x_{1i} x_{2i}}{\sum x_{1i}^2} + \frac{\sum x_{1i} u_i}{\sum x_{1i}^2}$$

And as $E(x_{1i} u_i) = 0$ (since x_{1i} is fixed and known variable and $E(u_i) = 0$ by definition of residual) and assuming that the two variables are correlated $\sum x_{1i} x_{2i} \neq 0$ (violation of the exogeneity condition),

$$E(\hat{a}_1) = \beta_1 + b_{12}\beta_2$$

[48] This condition is also referred to as the endogeneity condition (in contrary to exogeneity) where the explanatory variable is no longer fixed, deterministic but rather random, stochastic and as function of omitted variable(s).

where $b_{12} = \sum x_{1i}x_{2i}/\sum x_{1i}^2$ is the regression coefficient from a regression of x_2 on x_1. In other words, the $\hat{a}_1$ is a biased estimate of β_1 by the extent of $b_{12}\beta_2$ which is the product of the regression coefficient of the excluded variable on the outcome variable and the regression coefficient of the excluded variable on the included variable. $b_{12}\beta_2$ represents the effect of x_2 on y that is misattributed to the effect of x_1 on y. That is, x_1 mistakenly gets credit for the influence that is rightly attributable to x_2.[49]

Bias can arise from not just from omitted variables but also from omitted functional forms, measurement error, and missing data. Any unaccounted factor which systemically explains the outcome variable and at the same time is associated with the modeled variables will induce bias. Finally, the exclusion of these important variables can also contribute to the violation of other assumptions such as homoskedasticity where the spread of the residuals is determined as a function of the omitted variable; independence of residuals where the grouping variable (e.g. school, firms, hospital) which induces the dependency is a candidate for an omitted variable; and normal distribution where bimodal distributions can be induced from an omitted binary/categorical variable such as gender or ethnicity. The violation of these assumptions can further worsen the bias and can also induce inefficiency (as explained further in the next sections). To sum, exclusion of important explanatory variables is a complex matter as it can induce violations of multiple assumptions simultaneously and cause multiple damages to the statistical properties of the regression estimates. It can induce both bias and/or inefficiency.

Non-Normally Distributed Errors

The violation of the normally distributed residuals has the following implications for linear regression estimates. First, if non-normally distributed residuals such as an asymmetric skewed distribution is present, then the conditional mean values of the outcome variable which defines the regression function may no longer serve as the adequate statistic to describe the data structure. Other statistics/estimators such as the median (also known as the least absolute distance estimator) or the mode instead of the conditional means (regression) will be better statistics to parsimoniously describe the outcome variable. Second, non-normally distributed residuals such as the heavy tailed distribution characterized due to high frequency of large and extreme residuals can induce bias, inefficiency, and sensitivity (no robustness) in the estimates. For example, as described earlier, the extreme outlying residuals can also induce inefficiency as the large residual values lead to excessively large error variance that defines the standard errors of the parameter estimates. Moreover, if there is any slight negative skew on top of the heavy tails, then the conditional mean values will be heightened

[49] The omitted variable bias problem generalizes to the multiple omitted variable case. Suppose we have k explanatory variables, of which the first k_1 are included and the remaining $(k - k_1)$ are omitted. The formula which indicate the bias of the k_1 variables are indicated as follows

$$E(\widehat{\beta_1}) = \beta_i + \sum_{j=k_1+1}^{k} b_{ij}\beta_j \quad for\ i = 1,2,\dots,k$$

where b_{ij} is the regression coefficient of the *ith* included variable in a regression of the *jth* omitted variable which is correlated with this *ith* variable. As evident from this equation, as more variables which are mixed and correlated with the variable of interest, the bias accumulates.

unrealistically due to the excessive frequency and influence of the outlying values. Regression estimates under this condition will be misrepresented and biased (overestimated). Third and finally, (although the assumption of normally distributed residuals is contested by the econometrics field as not necessary), non-normally distributed residuals can invalidate the use of t and F distributions in conducting statistical inference. *t* and *F* distributions are derived under the assumption that the population (or the residual) being studied follows a normal distribution. Thus when the normal distribution assumption does not hold in the data, we cannot empirically justify the use of t and F distributions. To sum, the persistence of non-normally distributed residuals can have severe consequences on the estimates as it can invalidate the very use of regression or induce bias, inefficiency, and sensitivity in the findings.

Non-Constant Residual Variance and Covariance – Heteroskedasticity and Dependence of Residuals

The violation of the homoskedasticity and independence assumption of the residuals will lead to inefficiency of the regression estimates. This is shown as below. First, homoskedastic and independent residuals imply that the variance covariance matrix of the residuals takes the following form.

$$\begin{pmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & \sigma^2 \end{pmatrix} = \sigma^2 \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

The diagonal elements which represent the variance of the residuals are constant and the off diagonal elements which represent the covariance (or dependence) of the residuals must be zero.[50] But when non-constant variance and covariance of residuals are present, the diagonal entries will no longer be constant and the off-diagonal elements will no longer be zero. We designate this matrix as V. When any form of non-constant residual variance covariance (V) is present, the standard error of the conventional regression estimates (based on the least squares or the maximum likelihood) no longer simplifies to the efficient $\sigma^2(X'X)^{-1}$value defined under the homoskedasticity and independence assumptions. Instead, the variance of regression estimates which do not take into account of V becomes

$$var(\beta_{ols}) = E(\beta_{ols} - \beta)^2 = E(\beta_{ols} - \beta)'(\beta_{ols} - \beta) = E\{(X'X)^{-1}X'\varepsilon\}\{\varepsilon'X(X'X)^{-1}\}$$
$$= E\{(X'X)^{-1}X'\varepsilon\,\varepsilon'X(X'X)^{-1}\} = (X'X)^{-1}X'E(\varepsilon\,\varepsilon')X(X'X)^{-1} = (X'X)^{-1}X'var(\varepsilon)X(X'X)^{-1}$$
$$= (X'X)^{-1}X'VX(X'X)^{-1}$$

And this figure is clearly larger than the $\sigma^2(X'X)^{-1}$ which is the most efficient estimate under the assumption that homoskedasticity and independence holds.[51] That is, in the presence of non-constant error variance and covariance, the conventional least squares estimates do not have the smallest variance. It is not the "best" or most efficient, precise, and reliable estimate and this will lead to invalid statistical inference. In practice, the

[50] The above matrix can be conveniently summarized as $\sigma^2 I$ where *I* is an identify matrix with ones on the diagonal and zeros on the off diagonal.
[51] This can be simply shown by inserting σ^2I into the V of the definition of variance. That is, $var(\beta_{ols}) = (X'X)^{-1}X'\sigma^2 IX(X'X)^{-1} = \sigma^2(X'X)^{-1}X'X(X'X)^{-1} = \sigma^2(X'X)^{-1}$

variance covariance matrix of the residuals can take a variety of non-constant structures. That is, they induce different systemic patterns depending on the nature of the underlying phenomena being studied. For example, non-constant spread can arise from its systemic association with another variable. Studies have found that the variance of income distribution across different demographic groups such as gender, ethnicity and age differ consistently. The variation of company profits also shows differences across company sizes.[52] Similarly, the variation of student test scores also often shows differences across gender, ethnicity and geographic locations. For the covariance patterns, as described earlier, in the time series data, observations (and residuals) belonging to the same individual are correlated across time. This is known as autocorrelation. Similarly, observations clustered in the same group (e.g. students clustered in schools) also illustrate dependencies as they share many similar environments. This dependency often entails a block diagonal matrix known as the compound symmetry.[53] And in all cases, if we continue with the conventional regression estimation (without taking the non-constant variance covariance structure into account in the model), inefficiency will be induced in the regression estimates.

Multicollinearity of the Explanatory Variables

As introduced above, the multicollinearity problem or a strong correlation of the explanatory variables will cause the regression estimates to become unstable with very high standard error. That is, it will induce inefficiency in the estimates. This can be shown by re-writing the regression parameter estimates in terms of correlation coefficients. For example in the case of bivariate regression,[54]

$$var(\widehat{\beta_1}) = \frac{\sigma^2}{\sum x_{1i}^2(1 - r_{12}^2)}$$

$$var(\widehat{\beta_2}) = \frac{\sigma^2}{\sum x_{2i}^2(1 - r_{12}^2)}$$

As you can see, as the correlation between the two variables (r_{12}) approaches perfect association (equal to 1) the variance (standard error) of the parameter estimates approaches infinity. In other words, as the correlation increases to very high levels, the efficiency (precision and reliability) of the estimates drastically falls. And this will lead to misleading statistical inference with excessively large confidence intervals and no significant results. Studying the variance estimates for different values of the correlation, it is evident that the variance increases at an exponential rate. For example when r_{12} = 0.50 the variance is 1.33 times the variance when r_{12} is zero but as r_{12} reaches 0.95 it is about 10 times as high and by the time 0.995 it is 100 times as high. The

[52] Gujarati (2008)

[53] Depending on the phenomena being studied, there can be numerous types of non-constant variance covariance matrix V. The study over the structure/types of V has led to an extensive statistical literature known as variance components models. Renown examples include the unstructured matrix, autocorrelation (with different lags), compound symmetry, autoregressive conditional heteroskedasticity, toeplitz, factor analystic, autoregressive moving average, and more.

[54] *xj* are the deviation of Xj from its mean values

cut off correlation value in which we need to start worrying about the multicollinearity problem is said to be 0.9.[55]

To sum, the ultimate goal of any empirical analysis including the VAMs, is to provide the most accurate and reliable (BLUE) estimates of the model. But the above explanation illustrated that this task is complex and intricate as there are multiple and intertwined sources of damages on the statistical properties of the estimates. This is summarized in the table below. In light of this challenge, it becomes crucial to diagnose the estimated model from multiple angles using a variety of devices to detect different forms of violation. Such violations shall then be revised and corrected using different methods until all the assumptions necessary to provide the BLUE estimates are jointly met. These important tasks are known as the regression diagnosis and revision and are described next.

Table 3.2: Sources and consequences of violating the linear regression assumptions

Assumptions	Sources of Violation	Consequence of Violation
I. Properties of Residuals – Randomly Distributed		
Levels of the Residuals:		
Mean value of residuals given Xs is zero. E(ei/X) =	Omitted variable, wrong functional form, non-randomly missing observations, measurement error	Bias
Expectation and covariance between residuals and Xs is zero. Residuals are independent of Xs. E(e,X)=0, Cov (e,X)=0	Same as above	Bias
Normally distributed residuals	Features of underlying phenomena, omitted variable, outliers	Bias + Inefficiency
No extreme/outlying influential residuals	Measurement error, extra-ordinary cases	Bias + Inefficiency + Non-robustness
Variance(Spread)/Covariance (Dependency) of the Residuals:		
Homoscedasticity constant variance of residuals	Features of underlying phenomena, omitted variable	Inefficiency
Independence of residuals i.e. no autocorrelation, serial correlation, clustering or any form of dependencies/covariance between the residuals. Cov(ei,ej)=0	Features of underlying phenomena	Inefficiency
II. Properties of Explanatory Variables		
Clear variation in the values of the Xs	Not relevant but could depend on data collection method	No model estimates
No perfect association and collinearity between the Xs	Measurement error, nature of data, data collection methods	Inefficiency or No estimation
No extreme high leverage and influential X values	Measurement error or extra-ordinary cases	Bias + Inefficiency + Non-robustness

[55] The bivariate regression case can be generalized to the k variable case. In such case, the variance of the kth coefficient is expressed as

$$var(\hat{\beta}_j) = \frac{\sigma^2}{\sum x_j^2 (1 - R_j^2)}$$

where R_j^2 is the R^2 in the regression on the remaining (k – 2) regressions

Regression Diagnosis and Revision

Regression diagnosis is the art of checking the above assumptions necessary for the regression models to provide the BLUE estimates. The reason why it is "art" per se is due to the intricate and intertwined nature of the process in checking the model assumptions. As described above, there are multiple potential sources in violating the assumption and each violation can have multiple forms of damage to the statistical properties of the estimates. We must therefore diagnose the assumptions using multiple devices from multiple angles to look for strong consistency in the findings. These diagnostic tools consist of a variety of graphical devices, numerical summaries, and statistical tests designed at the univariate, bivariate and multivariate dimensions. For example, basic histograms, bar graphs, density/distributional plots, and quantile plots can be used to analyze the univariate features of the residuals and explanatory variables. Scatterplots, cross tabulated tables, and three dimensional or conditional scatterplots can be used to analyze the bivariate and multivariate features. Numerical summaries include the conventional mean, standard deviations, median and mode (univariate statistics) to correlation coefficient, partial and semi-partial correlations, and intra-class correlation coefficient (multivariate statistics). Finally, extensive literature has been devoted to the development of statistical tests designed to test the different assumptions such as the Bush Pagen test of homoskedasticity, Hausman test of exogeneity, Willis test of normality, and more. Once a full diagnosis is conducted we must then consider different remedies to correct for the detected violations. But this process further complicates the matter as there are multiple remedies with different degree of complexity. Moreover, each remedy can potentially correct for one (or more) violation(s) but it can also render new violations. For this reason, it is becomes necessary to fully re-diagnose the model for each and every remedy to make sure new forms of violations are not introduced. And this process must be repeated until all the assumptions are jointly met. The corrective actions consist of a variety of methods. Simple solutions include the transformation of the outcome and/or explanatory variables, re-specification of the model through including new variable(s) or new functional form of existing variables, re-weighting the effects of extreme cases or possibly excluding it from the analysis. These simple solutions can potentially cure the bias and/or inefficiency caused by the presence of non-normal distributed error, omitted variables, and outlying cases. It can also certainly increase the model fit performance to better satisfy the $E(e/X) = 0$ condition. Slightly more advanced methods include weighted least squares (WLS, also known as the generalized least squares or linear mixed model) which essentially transforms the entire model to take the non-constant error variance and covariance problem into account. The WLS estimates provide the efficient estimate in light of the non-constant variance and covariance of the residuals. And some of the highly advanced methods include the Huber-White M-estimation used to correct for non-normal distributed errors and outliers; instrumental variable (IV) estimation which attempts to correct the endogeneity bias by delegating the effects of the endogenous explanatory variable to an new variable (instrument) which is not correlated with the residual; and simulation based methods such as bootstrap and Monte-Carlo to test the stability and robustness of the estimates particularly when the sample size is small. The development of these corrective methods is an active and ongoing research area in today's statistical literature. The sole purpose

again is to resolve the violation of the assumptions necessary to achieve the BLUE estimates we long for. The choice over which corrective method(s) to implement is usually suggested by the diagnosis findings. Simpler methods are always preferred over the complex methods as it is more parsimonious (with less assumptions and mathematical procedures) and easier to interpret. In all cases, diagnosis and revision must work in tandem to complement each other until all the assumptions are jointly met. As Fox (2008) addresses, "taken together, regression diagnostics and corrective methods [revision] greatly extend the practical application of regression models. It is often the difference between a crude, mechanical data analysis and a careful, nuanced analysis that accurately describes the data and therefore supports meaningful interpretation of them." (p.241)

Summary and Conclusion

This chapter highlighted the importance of assumptions in statistics. Specifically, it identified the two major groups of assumptions underlying the linear regression models in providing the BLUE estimates to ensure valid statistical inference. The first group pertained to the behavior of the residuals and the second group pertained to the behavior of the explanatory variables. The notion of a randomly distributed pattern and the idea of unique and identifiable variable underlined as the overarching ideas for the two groups, respectively. And the different assumptions in each group played a collective role in constructing these overarching ideas. But as we opened this black box of assumptions, an intricate and intertwined nexus of association was also revealed. Namely, there were multiple sources which could violate multiple assumptions simultaneously and the violation of each assumption had multiple damages on the statistical properties of the regression estimates –namely both bias and inefficiency. Thus a successful implementation linear regression necessitated the need for a comprehensive diagnosis and revision of the estimated models. And this process was assisted by means of variety of graphical and statistical tools. The diagnosis and revision process for any estimated model must continue until all the assumptions are jointly met which promises the BLUE of the estimates. These key lessons will be used to guide the VA analyses conducted in the second half of this study. Having now understood the definition, mechanics, and role of assumptions underlying the linear regression models as a statistical tool to investigate relationship between variables, we now apply this understanding to the research on VAMs. We will illustrate how the linear regression models will help us achieve the unique effects of teachers on student performance that are BLUE.

CHAPTER 4: TASK 1 – DEFINING THE VALUE-ADDED PARAMETER: FIXED EFFECTS VS. RANDOM EFFECTS

Underlying the definition of VAMs are the two major tasks necessary for its construction and implementation: first to define the teacher effects and second to take factors outside the control of teachers into account in order to improve the quality and precision of the teacher effects. In linear regression language, these two tasks translate to defining the parameter for the teacher VA effects and to removing the mutual dependencies inherent between the teachers and the variables which are outside of their control, respectively. A thorough understanding of these tasks is vital to the successful implementation of VAMs using the actual data. Therefore, in this chapter, a detailed and user-friendly explanation of the first task using the linear regression framework is provided. A detailed explanation of the second task will be provided in the following chapter.

There are two major competing ways for modeling the teacher value-added effects under the linear regression framework – fixed effects and random effects. These two modeling options illustrate two distinct ways of mathematically assigning the parameters to the teachers in order to estimate their unique contribution to student performance. They also involve different sets of statistical assumptions in order to provide the BLUE estimates we long for. But as we proceed to implement these two modeling options in practice, we confront an interdisciplinary divide in justifying its selection. This divide is characterized between the experimental design literature and the observational studies (secondary data analysis) literature. The former is used extensively among researchers in the hard sciences (such as in medicine) while the latter is used extensively among researchers in the social and behavioral sciences (such as economics, education, sociology, psychology, etc.). The mathematical presentation of the two models is equivalent across both disciplines but the nature and control over the data collection used in analyzing the models differ greatly. The former has the capability to randomly select and assign treatments to subjects while the latter does not. This leads to the key differences in the interpretation, application, and most importantly the justification between the two options. In the following sections, the explanation of these differences between the two models is provided. The mathematical presentation and the statistical estimation procedures of the two models then follow.

Experimental Design Interpretation:

Experimental design refers to a specified plan in randomly assigning experimental conditions (treatments or variable of interest) to subjects in order to estimate its effects on the outcome variable of interest.[56] The treatment effects are estimated through investigating whether the average value of the outcome variable between the treatment group and no-treatment (control or placebo) group are statistically different. And this difference in the conditional means given the treatments is precisely estimated using linear regression models.

[56] There are a variety of such plans from simple randomized experiment with one treatments, factorial experiments with more than one treatments, blocking experiments, hierarchical experiments, Latin Square experiments, analysis of covariance experiments, etc. Please refer to Kirk (1975), Fisher (1935), and Lomax (1998) for extensive review of these plans.

For experiments in which the experimenter includes all the treatments that he/she is interested in are referred to as fixed effects model. The experimenter is interested in the magnitude of the effects of each and every treatment on the outcome variable. Thus, if the same experiment were to be replicated, the same treatments would be included in the study. Unlike random effects described below, the conclusions drawn from this model apply only to the selected treatments. The findings do not extend or generalize to other treatments that are not included in the model. The experimenter's attention is "fixed" and specified just to the treatments included in the model. These treatments have been decided (restricted) upon as the main variable of interest. And in the context of VAMs, the treatments or the main variable of interest is the teachers. We are interested in seeing whether significant differences exist in the average student performance across the teachers. However, experiments in which the experimenter selects a "random sample" of treatments from a larger population of all treatments are referred to as random effects model. The experimenter is interested in making generalizations and inferences about the entire population of treatments and not merely those that have been sampled. The experimenter wants to draw conclusions beyond the treatments included in the study. The idea is equivalent to the notion of random sample underlying statistical inference described in Chapter 2.[57] If the same experiment were to be replicated, the same treatments need not be included in the study as the selected treatment is just one treatment which happens to be (randomly selected) in the experiment.[58] Unlike fixed effects model, there is no intrinsic value or interest in the effects of the selected treatments. Instead, the main interest lies in inferring whether any significant variation (differences) exists in the population of all treatment effects.

To sum, and before we proceed to the regression model interpretation, the key point which needs to be highlighted is that the definition and justification of the two models is embedded in the design and data collection method of the study. As a consequence, the decision over the two models is determined at the outset of the study prior to collecting any data and more importantly prior to analyzing the collected data. Moreover, underlying the design of both models is the notion of random assignment of treatments. Each and every subject has the same and independent chance obtaining the treatment. This process ensures that all the factors un-accounted for in the model are missing in a randomly distributed pattern. The residuals are randomly distributed and illustrate no systemic pattern with the treatments. And as described in previous chapters, this ensures the BLUE property of the treatment effects. This feature (the control over the design of the study and random assignment of treatments) is the key anchoring feature that distinguishes the two disciplines and their approach towards the two models. As described next, the control over the data and study design is forgone in the observational studies literature which analyzes "naturally observed" and "un-manipulated" data.

[57] The concept of random selection of treatments from the population of treatments is the same as the random selection of the subjects from the population. Here the researcher is making an inference from the sampled treatments to the population of all treatments just like we make inference from the sample of individual outcomes to the population of all individual outcomes.

[58] In statistical language, the selected treatments are "exchangeable", meaning that from the researcher's point of view any unit in the population could have taken the place of each unit in the observed sample.

Observational Studies Interpretation:

Observational studies which analyze non-experimental naturally occurring data have a completely different approach towards defining fixed and random effects. Unlike the experimental design approach, the justification of either of the two modeling options is determined at the "end" of data collection and more precisely, after estimating the two models. As there is no random assignment of treatments to ensure the BLUE estimates, the two models are chosen based on whether its performance (i.e. the fitted models) satisfies the underlying assumptions necessary to provide the BLUE estimates. This process can be thought of as an attempt to capture the virtues of the random experiment when one has not been conducted.[59] In other words, the models are first tentatively estimated and then reverse engineered (justified) based on its post fitted model performance. For this reason, the underlying statistical assumptions that determine the BLUE become the key guide. In addition to these assumptions other factors which can also affect the justification of the two models are summarized in the following table which is an extension of the work by Rabe-Hesketh and Skrondal (2008), p.124.

Table 4.1: Comparison of the determining factors underlying fixed and random effects

	Fixed effects	**Random effects**
What additional assumptions are required (to the ones introduced in last chapter)?	None for fixed effects.	Random effects must be randomly distributed with mean of zero and constant variance, exogenous to explanatory variables, etc.
Is the model parsimonious?	No, J parameters estimated for each of the J clusters/groups	Yes, one variance parameter for all J clusters/groups
Minimum number of clusters/groups required?	Any number	For estimating within group variance at least 10 or 20
Minimum cluster/group size required?	Must be large; at least greater than equal to 2	None
Can estimate effects of cluster level covariates?	No	Yes
Inference for population of clusters?	No	Yes

As shown above, in addition to the statistical assumptions and data features that affect the BLUE property, research questions and analytical capacity can also influence the decision about choosing between the two models. These points are further described below.

Statistical Assumptions:

The linear regression assumptions described in the previous chapter applies to both fixed and random effects models. But in addition to these assumptions, random effects model requires several other conditions that do not apply to fixed effects models. Among these assumptions, researchers (particularly economists) have emphasized the exogeneity condition (E(treatments*X)=0) which is required under random effects and not for fixed effects. Economists have addressed this condition as the key determinant in justifying whether random effects is a viable modeling option. As Gujarati (2008) describes,

[59] NRC (2010)

> The answer to the question [fixed or random effects] hinges around the assumption we make about the likely correlation between the group specific error [random effects] and the regressor. If group specific errors and the regressors are uncorrelated, the random effects model will be appropriate where as if they are uncorrelated, the fixed effects model is appropriate. (p.606)

Fixed effects model is estimated using the usual linear regression estimators (least square and MLE). It models the treatment variable (the teachers in the case of VAMs) just like any other explanatory variables (Xs) through assigning "fixed" and known 0 or 1 values to represent the belonging/classification of subjects to the respective treatments/groups. No additional assumptions are introduced to the previously described linear regression assumptions in achieving the BLUE estimates. That is, just like any Xs the treatments must be uncorrelated with the residuals. Moreover, as long as the multicollinearity problem is induced, treatments and Xs are allowed to be correlated just like any explanatory variables. On the other hand, random effects models the treatments are just like the randomly distributed residual errors. It must not show any systemic pattern with the explanatory variables. That is, unlike fixed effects model, the treatments and Xs must not be correlated. It must be "exogenous" with $E(treatments*X)=0$. In other words, random effects model makes stronger assumptions for the treatments. When this additional exogeneity condition is violated (and random effects are not randomly distributed), bias is introduced in the model. In such a case, fixed effects model is preferred. But when the exogeneity condition holds, random effects model is preferred as it is more efficient.[60]

In light of these differences, an economist named Hausman (1978) designed a statistical test to assess the exogeneity condition. This test known as the Hausman test builds on the previous explanation that under the null hypothesis of no correlation between the treatment effects and the explanatory variables (exogeneity condition), both fixed effects estimated using the usual least squares and random effects estimated using MLE based weighted least squares are consistent (unbiased),[61] but the least squares is inefficient as it estimates an unnecessarily large set of dummy coefficients in comparison to random effects that only estimates one variance component.[62] Whereas in the alternative hypothesis, least squares is consistent (unbiased), but random effects suffers from the bias induced through correlation of random effects and the explanatory variables. Therefore, under the null hypothesis, the two estimates should not differ systematically and a statistical test can

[60] There are two main reasons why random effects are more efficient. First, fixed effects model based estimation exclusively on the within group variation and completely ignore the between group variation. In essence it is estimating the model for each and every group separately. In contrast, random effects model captures both sources of variation. Thus if there are significant proportion of variation existing between the groups, random effects model captures more variation of the outcome variable and consequently have smaller standard error. That is, it is more efficient. Second, random effects model has larger degrees of freedom. It only has the variance of the treatment (aj) to estimate while fixed effects have dummy variables for each and every teacher to estimate. Random effects therefore have much less parameters to estimate and consequently more degrees of freedom. The degrees of freedom enter the denominator of the standard error calculation. Therefore, the standard error of random effects is much smaller than fixed effects which imply higher efficiency.

[61] Consistency essentially implies the same meaning of unbiasedness. It is referred to the property that the estimator will asymptotically converge (as the sample size goes to infinity) to the true value.

[62] Please refer to previous footnote for further explanation.

be designed based on this difference. Hausman (1978) showed that the following statistic is distributed as chi-squared with *k* degrees of freedom (number of explanatory variables).

$$\left(\beta^{fe} - \beta^{re}\right)' \Sigma^{-1} \left(\beta^{fe} - \beta^{re}\right) \sim \chi_k^2$$

where $\Sigma = Var\left(\beta^{fe} - \beta^{re}\right) = Var\left(\beta^{fe}\right) - Var(\beta^{re})$.[63] If the calculated test statistics are larger than the critical value, the null hypothesis that the effects are uncorrelated with explanatory variables in the model is rejected. Bias is induced in random effects models and thus the use of fixed effects is a better modeling option. And vice versa for the alternative scenario. The Hausman test is extensively used today among economists as the classical test in assessing the exogeneity condition and provides the critical evidence to evaluate whether random effects are a statistically viable modeling option. But it is not without its shortcomings. As Johnston and DiNardo (1997) address, "there is no simple rule to help the researcher navigate past the Scylla of fixed effects and the Charybdis of measurement error and dynamic selection." What the authors imply is that the Hausman test operates under the premise that fixed effects are perfectly specified with no potential bias. But there are always potential sources of bias (such as measurement error and dynamic selection which will introduce systemic pattern in the residuals). The Hausman test does not provide any evidence that fixed effects satisfy the linear regression assumptions to ensure the BLUE estimates. The Hausamn test is a test only for the exogeneity condition required under random effects model. The test alone cannot provide us with complete evidence to decide on the selection between fixed or random effects. As shown in the first row of the table above, the exogeneity condition is one of the several assumptions that distinguish random effects from fixed effects model. As random effects are modeled just like the residual errors, it must illustrate the complete features of a randomly distributed pattern i.e. independently, identically, and normally distributed with mean zero and constant variance. That is, only when all these features for both random effects and the residuals are fully illustrated, the BLUE estimates of random effects are ensured.[64] If there are clear signs of the violation of these assumptions (and cannot be revised), then this provides another evidence in support of fixed effects model given that it meets its own set of assumptions described in the previous chapter.

Data Features and Requirements

Researchers today debate the appropriate data conditions to support the two modeling options. There are no established norms or standards but a number of researchers who specialize in VAMs and multilevel analysis state that random effects should only be used if there are a sufficient number of treatments/groups in the data, typically more than 10 or 20. The reason is because with a small number of groups, the variance between the groups will be poorly estimated. In such a case, fixed effects which assign a binary variable for each group are

[63] This result is proven in Hausman (1978) where the paper proved that the covariance of an efficient estimator with its difference from an inefficient estimator is zero. That is, $Cov(\beta^{fe} - \beta^{re}, \beta^{re}) = Cov(\beta^{fe}, \beta^{re}) - Var(\beta^{re}) = 0$ and thus $Cov(\beta^{fe}, \beta^{re}) = Var(\beta^{re})$ and $Var(\beta^{fe} - \beta^{re}) = Var(\beta^{fe}) + Var(\beta^{re})\ 2Cov(\beta^{fe}, \beta^{fe}) = Var(\beta^{fe}) - Var(\beta^{re})$ as shown above

[64] The residual diagnosis and revision must therefore be conducted for both random effects and the (level 1) residuals.

recommended because it will not exhaust the degrees of freedom (efficiency) of the model. However, if random effects is used merely to make inference regarding the parameters for the explanatory variables and not the between group variance, then a smaller number of groups can suffice.[65] For group sizes (number of subjects/data within each treatment/group), it must be a sufficiently large if fixed effects are to be used. As further explained later, group size is an important determinant of fixed effects estimates and its standard errors. The standard errors are in fact inversely proportional to the square root of the group size. That is, as the group size decreases (which implies less data and information available) for a particular group, the standard error of fixed effects estimates increases (which implies the decrease in the reliability and efficiency of the estimates). Researchers today continue to debate the "sufficient" group size as some suggest a minimum of 5 subjects per group while others suggest a minimum of 20 to 30 subjects per group. However, in random effects models, such a requirement is not necessary in estimating the parameters of the explanatory variables and the variance of the treatment effects. It is only required that there are a good number of groups of size 2 or more.[66] It does not matter if there are small sized groups, even those with only a single observation. Such singleton groups do not provide information on the within groups correlation or how the total variation is partitioned into between and within group variances but they still contribute to the estimation of the parameters of the explanatory variables (assumed to be the same across all groups) and the two variances components as these estimates are based on the entire data sets irrespective of the groups.[67] But when it comes to the predictions of the individual group effects for random effects model, the group size has an important role to play. These predictions are known as the empirical Bayes and it will be fully explained in the section below. As the group size (or the amount of information and reliability within each group) increases random effects prediction gravitates more toward the group mean or fixed effects estimates. That is, as the group size increase for all groups, fixed effects model is preferred. On the other hand, when the group size decreases, the random effects predictions moves away from the un-reliably estimated fixed effects and moves towards the overall mean of all observations (which uses all the information in the data ignoring the grouping structure). Thus in cases when there are a lot of groups with very small group size, random effects model is preferred and justified.

Research Questions and Analytical Capacity

In addition to the statistical assumptions and data features that directly affect the BLUE properties, other less technical factors such the types of research question(s) the two models can analyze also play an important factor in thinking about the two models. With respect to the analytical capacity, random effects model is more attractive as it can explicitly distinguish and estimate the group level explanatory variables. The reason is

[65] Rabe-Hesketh and Skrondal (2008)
[66] Rabe-Hesketh and Skrondal (2008)
[67] Mathematically, the reason for this is as follows. Fixed effects model essentially estimates the model by and for each group separately ignoring any variation that exists between groups. For this reason the amount of data within each group is essential. Random effects model on the other hand capitalizes on both the between and within group variation. As described before this is conducted through the maximum likelihood estimation which simultaneously estimates both variance parameters. The estimation of the between group variation is not directly affected by the group sizes. As long as some information exist for each group, this information can be contributed to the estimation of the between group variance. The within groups variation on the other hand is affected by the sample size just like fixed effects.

because random effects do not suffer from the multicollinearity problem that exists for fixed effects model. In fixed effects model, simple transformation of the set of binary dummy variables can re-create any group level variables.[68] The group level variables cannot be uniquely identified and parameter estimates cannot be provided. One point of clarification is that fixed effects model in essence does take into account all the effects of group level covariates (time invariant individual factors in case of longitudinal panel data analysis where the individuals are the group) through the binary variables but the problem is it cannot separate or distinguish the effects of different group level variables when such information is available.[69] But as there is often scientific interest in the group level variables (e.g. teacher and school characteristics and in the case of longitudinal data, individual background variables e.g. gender, ethnicity, etc.), fixed effects model heavily restrains the types of research questions that can be investigated. Thus when the group level variables are available and there is intrinsic interest in analyzing its effects, random effects provide a better modeling option.

To sum, unlike the experimental design where fixed and random effects are clearly defined and selected at the outset of the study prior to collecting the data, the definition and justification of the two options for the regression model are determined at the end and post fitting the model. The justification of the two models is steered primarily by the statistical assumptions and the data conditions to ensure BLUE estimates and other factors such as the types of research questions the models can analyze. In practice, we must provide a thorough consideration of all these points to determine the selection of either of the two modeling options. Only when these points are thoughtfully and simultaneously balanced, the justification of the two modeling options suffices. If each model successfully satisfies all its respective assumptions and requirements, this provides the justification of both models. Unlike in the case of experimental design, the two models can possibly co-exist.

Mathematical Presentation and Estimation Procedures for Fixed Effects and Random Effects

The definition and justification of fixed and random effects across the two disciplines differ. But the underlying mathematical presentation and the estimation procedures are the same. Both models (for both disciplines) operate under the linear regression framework described in previous chapters. A summary of the mathematical presentation of the two models is provided below, followed by an explanation of the estimation procedures.

[68] Note, the values of the group level variables are repeated across all the subjects within the same group.
[69] Conversely, when group level variables are not available fixed effects can be a useful modeling option. Moreover, it can also represent factors which are impossible or hard to measure.

Table 4.2: Mathematical presentation of fixed and random effects

	Fixed Effects	**Random Effects**
Model Equation	$y_{ij} = x_{ij}\beta + a_j D_j + e_{ij}$	$y_{ij} = x_{ij}\beta + a_j + e_{ij}$
a_j (teacher VA effects)	Fixed and deterministic constant of dummy variables D_j (shown below)	$a_j \sim i.i.d.N(0, \sigma_a^2)$
e_{ij}	$e_{ij} \sim i.i.d.N(0, \sigma_e^2)$	$e_{ij} \sim i.i.d.N(0, \sigma_e^2)$
$var(y_{ij})$	σ_e^2	$\sigma_a^2 + \sigma_e^2$
$var(y_{ij}, y_{i'j'})$	$\begin{cases} \sigma_e^2 & for\ i = i'\ and\ j = j' \\ 0 & otherwise \end{cases}$	$\begin{cases} \sigma_a^2 + \sigma_e^2\ for\ i = i'\ and\ j = j' \\ \sigma_a^2 \quad for\ i \neq i'\ and\ j = j' \\ 0 \quad otherwise \end{cases}$
Dummy Variables	Dummy Coding: $D_J = \begin{cases} 1, \text{if J} = \text{J} + 1 \\ 0 \text{ otherwise} \end{cases}$ for j = 1, …..J a_j estimates interpreted as deviation from the omitted based line teacher/group J.	
	Effect Coding: $D_{J-1} = \begin{cases} 1 \text{ unless j} = \text{J (the last group)} \\ 0 \text{ otherwise} \\ -1 \text{ if j} = \text{J} \end{cases}$ a_j estimates interpreted as deviation from the grand mean (average of all teacher effects). No omitted based line teacher.	

Fixed effects model essentially fits a separate regression line $y_{ij} = x_{ij}\beta + e_{ij}$ for each and every teacher j. This is directed by the dummy variables which in essence separate the data set using its binary values. For example, when the teacher effect of a certain teacher is estimated, all the students taught by this teacher (dummy variable equal to one) enter the estimation but all the other students not taught by the teacher are excluded from the analysis (as they received value of 0 for this teacher dummy variable). For this reason, teacher class size is a vital component of teacher fixed effects estimates.[70] Now, assuming the explanatory variables have the same effects for each teacher, this is presented with a strand of parallel lines where the vertical distance (the difference in the intercepts) from the grand mean ($\hat{\beta}_0$) (or the omitted baseline group under dummy coding) represents the teacher effects ($\hat{a}_j$). This is represented in the left hand figure below for a sample of three teachers. Random effects also illustrate a similar figure but unlike fixed effects the vertical distance (from the grand mean ($\hat{\beta}_0$) or the teacher effects are modeled as a random variable with a normal iid distribution with mean of zero and constant variance. This is shown in the large blue curve in the right figure below. And the residual errors which are calculated as the deviation of the actual outcome variable from its respective (covariate adjusted) teacher means/effects are also distributed similarly with its own constant variance (σ_e^2). This is shown in the series of small blue curves. EB stands for the empirical Bayes estimate of the unobservable individual teacher effects under random effects model. These estimates will be fully

[70] The robustness/sensitivity analysis of fixed effects estimates with respect to class/sample sizes will be thoroughly investigated in later chapters.

described in later section. Finally, as fixed and random effects both show up in the intercept of the model, these models are also referred to as the varying intercepts models.

Figure 4.1: Graphical illustration of fixed effects

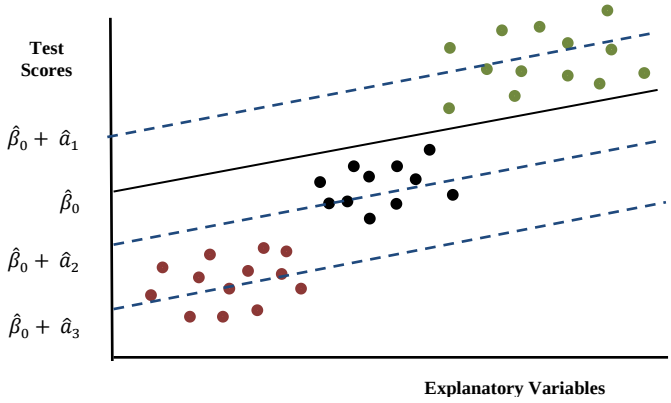
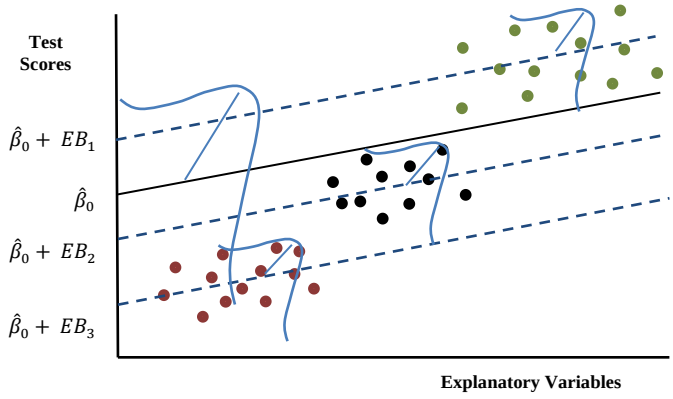


Figure 4.2: Graphical illustration of random effects



The teacher VA parameters for fixed effects model are estimated using the conventional least squares/MLE method. The final form of the least squares estimate is once again provided as $\hat{\beta} = (X'X)^{-1}X'Y$ and its standard error as $std(\hat{\beta}) = \sqrt{\hat{\sigma}^2/X'X}$ where $\hat{\sigma}^2 = (Y'Y - \hat{\beta}'X'Y)/n$ and some of the key components are further clarified as the following.

$$X'X = \begin{pmatrix} n & \Sigma x_1 & \cdots & \Sigma x_k \\ \Sigma & \Sigma x_1^2 & \cdots & \Sigma x_1 x_k \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma x_k & \Sigma x_k x_1 & \cdots & \Sigma x_k^2 \end{pmatrix} \qquad X'Y = \begin{pmatrix} \Sigma y \\ \Sigma x_1 y \\ \vdots \\ \Sigma x_k y \end{pmatrix}$$

Given these matrices, the standard error of the parameter estimate for any particular explanatory variable *i* $std(\hat{\beta}_i)$ becomes $c_{ii}\hat{\sigma}$ where c_{ii} is the element of the *i*th row and the *i*th column of the $(X'X)^{-1}$ matrix which is Σx_i^2. Now, if the variable of interest *i* is the teachers which is coded using dummy variables (with series of repeated (0 or 1) across the students they teach), $\Sigma x_i^2 = n_j$ the teacher's class size. This leads to

$$std(teacher\ fixed\ effects) = \hat{\sigma}/\sqrt{n_j}$$

That is, the efficiency (reliability and precision) of teacher fixed effects is inversely proportional to the teacher size. As described earlier, teacher size implies the amount of data and information available to estimate the teacher effects. Thus as there is more information available in estimating the teacher effects, the efficiency (and also the stability/robustness) of their estimates enhances and vice versa.

Proceeding now to random effects, the usual least squares/MLE method cannot be applied to estimate its parameters. The reason is because the variance covariance of the residuals is no longer constant. More specifically the covariance is no longer zero as dependency in the residuals is introduced (off diagonal elements of the variance covariance matrix is non-zero). And as described in the previous chapter, the violation

of the independence assumption induces inefficiency in the usual estimates. This dependency can be illustrated by first re-writing the model as $y_{ij} = x_{ij}\beta + w_{ij}$ where $w_{ij} = a_j + e_{ij}$. As a_j is now assumed to be a randomly distributed variable with the same properties as the e_{ij}, it can be subsumed under one composite error term. That is, the model can be envisioned as having two error components. And through applying the additive rule of variance, it can be shown that the variance of w_{ij} is sigma $\sigma_a^2 + \sigma_e^2$. The homoskedasticity constant variance still holds but it is now comprised of two components. However, it can be shown that the $w_{ij}s$ belonging to the same group j are now are correlated. For example, if we correlate the outcome of two individuals belonging to the same group j, it is evident that their outcomes (residuals) share the same common error component a_j. Calculation of the covariance between $(y_{ij}, y_{i'j})$ (or alternatively$(w_{ij}, w_{i'j})$) gives the common component of σ_a^2. In other words, for any individuals belonging to the same group, their residuals are associated (dependent) by the magnitude of σ_a^2.[71] The full picture of the variance-covariance matrix is shown below. Each row and column represents a student (or level 1 unit) and the block or brackets signify the groups (teacher or level 2 units) they belong to.[72]

$$\begin{pmatrix} \begin{bmatrix} \sigma_a^2+\sigma_e^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 \\ \sigma_a^2 & \sigma_a^2+\sigma_e^2 & \sigma_a^2 & \sigma_a^2 \\ \sigma_a^2 & \sigma_a^2 & \sigma_a^2+\sigma_e^2 & \sigma_a^2 \\ \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2+\sigma_e^2 \end{bmatrix} & \begin{matrix} 0&0&0&0 \\ 0&0&0&0 \\ 0&0&0&0 \\ 0&0&0&0 \end{matrix} & & \begin{matrix} 0&0&0&0 \\ 0&0&0&0 \\ 0&0&0&0 \\ 0&0&0&0 \end{matrix} \\ \begin{matrix} 0&0&0&0 \\ 0&0&0&0 \\ 0&0&0&0 \\ 0&0&0&0 \end{matrix} & \begin{bmatrix} \cdot&\cdot&\cdot&\cdot \\ \cdot&\cdot&\cdot&\cdot \\ \cdot&\cdot&\cdot&\cdot \\ \cdot&\cdot&\cdot&\cdot \end{bmatrix} & & \begin{matrix} 0&0&0&0 \\ 0&0&0&0 \\ 0&0&0&0 \\ 0&0&0&0 \end{matrix} \\ \begin{matrix} 0&0&0&0 \\ 0&0&0&0 \\ 0&0&0&0 \\ 0&0&0&0 \end{matrix} & \begin{matrix} 0&0&0&0 \\ 0&0&0&0 \\ 0&0&0&0 \\ 0&0&0&0 \end{matrix} & \begin{bmatrix} \sigma_a^2+\sigma_e^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 \\ \sigma_a^2 & \sigma_a^2+\sigma_e^2 & \sigma_a^2 & \sigma_a^2 \\ \sigma_a^2 & \sigma_a^2 & \sigma_a^2+\sigma_e^2 & \sigma_a^2 \\ \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2+\sigma_e^2 \end{bmatrix} \end{pmatrix}$$

This form of dependency is known as compound symmetry. Observations (residuals) in the same group (block) are associated by σ_a^2 while for those across different groups are not associated. Now, in light of these non-independent residuals, we must take this systemic dependency into account to correct for the violation of the least squares assumption and thereby (re)capture the BLUE estimates. This can be done by means of transforming the model by multiplying the entire model with the inverse (reciprocal) of the non-constant error variance covariance matrix. This process in essence re-creates a constant variance covariance matrix condition necessary for the least squares/MLE to provide the BLUE estimates. The usual least squares estimation process can then be applied to the transformed model. This estimation process is known as the generalized

[71] In the context of VAMs, this dependency epitomizes the clustering of similar student performances in the same groups (teachers and schools) due to the self-selection of teachers and schools based on similar family backgrounds (which are associated with student outcomes) and through the sharing of the same learning environments.

[72] By model assumption, the residuals of the individuals belonging to different groups are uncorrelated. This leads to zero values off the blocks.

least squares (GLS) or the weighted least squares (WLS) for weighting the model with the non-constant error variance and covariance matrix.[73] To clarify this process, the following simple proof is provided.

For a given non-constant residual variance covariance matrix $V \neq \sigma^2 \mathrm{I}$ (e.g. the compound symmetry matrix above) which is assumed to be a positive definite matrix,[74] then there exist a matrix G such that $V^{-1} = G^{-1}G$ that is $G = V^{-1/2}$. Transforming the model $Y = X\beta + \varepsilon$ by G yields $GY = GX\beta + G\varepsilon$. We know that its mean is $E(G\varepsilon) = GE(\varepsilon) = 0$ and $var(G\varepsilon) = GE(\varepsilon'\varepsilon)G = GVG' = G(G'G)^{-1}G' = I$ which is precisely a form of homoscedastic and independent error.[75] In other words, the transformed model $GY = GX\beta + G\varepsilon$ resolves the dependent errors problem (covariance not equal to 0) and re-creates the condition in which the usual least squares/MLE can provide the BLUE estimates. Specifically, least squares/MLE or more precisely as the GLS and WLS minimizes the following criteria which weight the usual sum of square residuals by V^{-1}

$$\sum (Y - X\beta)'V^{-1}(Y - X\beta)$$

Alternatively, under the normality assumption of the outcome variable given the explanatory variables, it maximizes the following likelihood function[76]

$$-\frac{N}{2}\log(2\pi) - \frac{1}{2}\log\{\det(V)\} - \frac{1}{2}\sum (Y - X\beta)'V^{-1}(Y - X\beta)$$

And the final form of the GLS estimate and its variance are as follow

$$\beta_{gls} = (X'V^{-1}X)^{-1}X'V^{-1}$$

$$\begin{aligned} var(\beta_{gls}) &= (X'V^{-1}X)^{-1}X'V^{-1}var(\varepsilon)V^{-1}X(X'V^{-1}X)^{-1} \\ &= (X'V^{-1}X)^{-1}X'^{V^{-1}}VV^{-1}X(X'V^{-1}X)^{-1} \\ &= (X'VX)^{-1} \end{aligned}$$

This is the correct and more importantly, the efficient form of standard error in the presence of non-constant variance and covariance of the residuals. It is clearly different from the inefficient least squares estimates which ignores the presence of V which is given by $var(\beta_{ols}) = (X'X)^{-1}X'VX(X'X)^{-1}$. Depending on the structure of V and the values of X, $var(\beta_{ols})$ can overestimate or underestimate the true variance of $var(\beta_{gls})$

[73] This procedure is also referred and subsumed under the linear mixed effects model framework which is an extension of the linear regression framework by incorporating the modeling option of random effects to take into account the clustering/dependency of data at different levels of hierarchy.

[74] This property enables us to calculate an inverse of V which is also positive definite.

[75] To recall *I* is the identity matrix with the value 1 on the diagonal elements and 0 on the off diagonal elements.

[76] det = determinant of the matrix. The key advantage of this latter MLE based estimation is that the dynamic optimization process of MLE provides us with the flexibility and extra capacity to estimate very complicated structures of V with many parameters. While the least squares based estimation can only handle simple transformation without many unknown parameters. As more parameters are provided, it will be close to impossible to simultaneously solve the sets of optimal first order equations with a special algorithm e.g. EM algorithm in which the MLE is based on with its distributional assumption.

and therefore lead to incorrect inferences.[77] This concludes the illustration of the estimation process underlying random effects model. But now, as you may have noticed, although we have obtained the estimate of σ_a^2 which signifies the presence (difference, variance) of any significant teacher effects, this finding alone does not fulfill the purpose of VAMs. The objective of VAMs is to provide a specific numerical value for the performance of each and every teacher. In fixed effects model, this objective was accomplished by assigning a dummy variable for each and every teacher which gave them a unique parameter estimate. And for random effects model, this objective is completed by the empirical Bayes predictions as described next.

Prediction of the Individual Random Effects – Empirical Bayes Estimates as the Teacher Effects Predictions

As described earlier, the teacher effects under random effects model is modeled as a random variable with an independent, identical, and normal distribution with a mean of zero and constant variance. Just like the (level 1) residual errors, no fixed values are assigned to the individual teachers (treatments) as there is no intrinsic interest in its individual estimates.[78] Instead, the interest is focused primarily on the variance of the sampled teacher effects which is used to infer whether statistically significant variance (differences) exist in the population of all teacher effects. However, researchers lead by Mood and Graybill (1963) and soon followed by 30 years of lifetime work by C. R. Henderson (1948, 1950, 1963, 1972, 1975), pondered whether or not a numerical value can be assigned on the individual random effects and if so how.[79] The researchers thought it might seem sensible to take the mean of the random variable a_j. But as $E(a_j)$ is zero by the definition of random variable/effects, no intrinsic or useful information is obtained. Yet, the researchers then thought, once the n_j sampled observations have been observed and we obtain the realized value of $\bar{y}_{.j}$, this mean of the n_j observations in the treatment (or group) j can potentially update our knowledge of the true unknown value of each a_j. For example, suppose the average test score of a particular group or teacher $\bar{y}_{.j}$ (e.g. Mr. Smith) is considerably above the overall teacher average. Then based on this information, we would expect Mr. Smith's true performance α_j (which is defined as the deviation from the grand mean μ) to have a positive value and $\mu + a_j$ to be greater than μ. Similarly, in the opposite scenario, we would expect the α_j value to have a

[77] Existing empirical and simulation studies on heteroskedasticity have shown that OLS without the correction of heteroskedasticity, consistently overestimates the true standard error obtained by the correct GLS procedure. That is, if we do not use GLS and rely on the OLS we would likely to have a wider confidence interval and accept the null hypothesis more often. And studies on autocorrelation has shown that if autocorrelation is positive and the Xs are positively correlated, then it is clear that var(ols) is less than var(gls). That is, the usual OLS variance of parameter estimates underestimates than the GLS which take into account the autocorrelation. Therefore, if we use variance (ols), we shall inflate the precision or accuracy (i.e. underestimate the standard error) of the estimator beta. As a result, in computing the t ratio as t = parameter/standard error, we will be overestimating the t value and thus the statistical significance of the estimated parameter.

[78] The teachers are only a random sample from a population of teachers and if the study were to be replicated, the same teachers need not be included. As the individual teacher effects are not directly observable or estimable it is also referred to as the unobservable, latent or nuisance variable.

[79] Mood and Graybill (1968) were interested in finding a way to assign numerical values to the unobservable and latent concept of intelligence quotients (IQ) for students in particular age group based on their test performance. C. R Henderson and his colleagues focused on the biological study of animal breeding and natural selection. They were interested in assigning values to the latent concept of genetic merit/strength of dairy bull given the milk yields data of his daughters. They were also interested in the identification, ranking, and selection of the strongest genes to explore the natural selection and survival of the fittest in the evolution of species. This ranking and identification is directly on the point with the VAMs research today.

negative value and $\mu + a_j$ to be less than the grand mean of μ. That is, in light of the observed data (information), we can update our belief, expectation, knowledge, and confidence (alternatively, reduce our uncertainty) of the unknown latent/true random effects of each treatment/group. And this process of updating our expectation or beliefs in light of new information is the essence of Bayesian inference.[80] It is a rational approximation of the unobservable true teacher effects a_j in light of receiving pertinent information regarding their performance. Mathematically, this translates to finding $E(a_j|\bar{y}_{.j})$, the expected value of a_j values given the observed mean test score of each teacher j ($\bar{y}_{.j}$). The exact form of the $E(a_j|\bar{y}_{.j})$ or alternatively $E(\mu_j|\bar{y}_{.j})$ which adds the constant grand mean μ is provided below.[81] It can be derived from different starting points: regression models, linear mixed models, and Bayesian inference which are provided in the Appendix. The $E(a_j|\bar{y}_{.j})$ values are referred to as predictions rather than estimates to delineate the fact that we are approximating/modeling the values of a latent unknown variable and not observed known values of an outcome variable (as conducted in the linear regression model). Finally, in practice, the estimate of $E(\hat{\mu}_j|\bar{y}_{.j})$ which substitutes the parameters with the respective estimates from random effects model (the GLS/WLS estimates) is known as the empirical Bayes estimates.

$$E(\mu_j|\bar{y}_{.j}) = \frac{\frac{1}{\tau^2}\mu + \frac{n_j}{\sigma^2}\bar{y}_{.j}}{\frac{1}{\tau^2} + \frac{n_j}{\sigma^2}} = \frac{\frac{1}{\tau^2}\mu}{\frac{1}{\tau^2} + \frac{n_j}{\sigma^2}} + \frac{\frac{n_j}{\sigma^2}\bar{y}_{.j}}{\frac{1}{\tau^2} + \frac{n_j}{\sigma^2}}$$

As you can see, the prediction of the individual random effects illustrates a very interesting pattern. It is in fact a weighted average of the grand mean μ (which is estimated as $\bar{y}_{..}$) and the individual group means $\bar{y}_{.j}$ (which is estimated using the dummy variables under fixed effects model).[82] The weight on the individual group mean is n_j/σ^2, the product of n_j and the precision of the data for each group. Precision is the inverse of variance and is interpreted as the level of reliability, efficiency, and confidence (same as variance but in opposite direction) of the information provided by the data. And the weight on the grand mean is $1/\tau^2$, which is the precision of random effects (the inverse of the amount of variation and uncertainty inherent between the random effects). Now, when the variance of the data within each group rises (its precision fall), the prediction of random effects responds to this by putting less weight on the $\bar{y}_{.j}$ and moving more towards the grand mean μ which is more reliably estimated as it uses the entire data set across all the groups. Similarly, as the size of the group falls (less information provided by the groups), the prediction of the random effects again adjusts by pulling away from this less reliant estimate of $\bar{y}_{.j}$ towards the reliably estimated grand mean. And in the

[80] The Bayesian inference is described further in the Appendix.

[81] Either $E(a_j|\bar{y}_{.j})$ or $E(\mu_j|\bar{y}_{.j})$ can be used interchangeably as a measure of teacher VA effects. Measure of teacher performance (or its relative ranking in lieu to other teachers) will be the same. Yet, the latter term is slightly more useful as it provides a comparable measure to fixed effects estimates $\bar{y}_{.j}$ and as described further below in the text, it provides a clearer linkage between random effects predictions and fixed effects estimates.

[82] To be slightly more precise, the respective means are essentially the "adjusted" grand mean and "adjusted" teacher means after the effects of the explanatory variables have been taken into account.

opposite case of more precise data by the group with larger group size, the prediction pulls more towards the $\bar{y}_{.j}$. Finally, when the variance of the average scores between the groups rise (less precision), the prediction puts less weigh on the grand mean which is estimated also as the mean of all group means (average of the between group variation). In other words, the random effects predictions steer towards the estimates which are more precise, reliant, and based on abundant information. This weighting process is also referred to as "shrinkage", "Stein effect" or "borrowing effect" as for groups with less precise and less informative data, the respective random effects predictions borrows more information from the rest of the groups by moving towards the grand mean which again is estimated using the entire of data. Gelman and Hill (2007) summarize this phenomenon as follows.

> The weighted average [the predictions of random effects] reflects the relative amount of information available about the individual group (the un-pooled or by group fixed effects estimates), on one hand, and the average of all the groups (the pooled or grand mean estimate), on the other…the relative weights determined by the sample size of the groups and the variation within and between groups. (p.254)

The prediction formula for random effects is therefore a powerful mechanism that optimally combines the different sources of information underlying the data. Moreover, in essence, it corrects and compensates for the lack of efficiency and precision of fixed effects estimates (teacher means) due to small group sizes by making multiple adjustments.[83] In practice, this has vital implications as some teachers and schools suffer from a very small number of students. In such a case, the efficiency and reliability of fixed effects teacher effects estimates is very low (high standard errors). These estimates are also very unstable (not robust) as one missing observation or irregular and miscoded observation can drastically alter its estimates. The random effects predictions alleviate these potential problems by steering towards more reliable sources of evidence. It resolves the weakness inherent in fixed effects estimates. The prediction formula therefore provides an effective common ground to compare and contrast fixed effects and random effects predictions. It clarifies the different factors (i.e. the data conditions) distinguishing the two models. And this provides further and important feedback in helping the selection between the two modeling options.

Summary and Conclusion:

This chapter provided the explanation of the first major task underlying the construction of VAMs – to define the VA parameter. In doing so, it identified the two competing modeling options for this purpose – fixed effects and random effects. Yet, opening the black box of this topic illustrated an interdisciplinary divide in providing the *raison de entre* between these models - the experimental design literature which defines and decides on the two options at the outset of the study prior to collecting or analyzing the data; and the

[83] Please refer to footnote 12 for further explanation as to why fixed effects are more inefficient.

observational studies literature which justifies on the modeling scheme after analyzing the data. In the former case BLUE estimates were ensured through the random assignment procedure embedded in the study design but for the latter case, BLUE estimates had to be reverse engineered through checking and revising the estimated model to meet the different assumptions necessary for BLUE. These assumptions in essence allowed the estimated models to capture the virtues of a randomized experiment when one has not been conducted. [84]

As the VA analysis in this dissertation uses non-experimental observational data, the latter approach becomes pertinent for this study. To successfully implement the VAMs, we must therefore conduct a careful, thoughtful, and thorough examination of the underlying the assumptions (and other factors such as sample sizes and types of research questions to analyze) to justify the two models. The explanation of the linear regression assumptions and consequences of violating the assumptions provided in the previous chapter will assist heavily in this process. The final VAM specification is only justified after all the assumptions and other determinants of BLUE are simultaneously taken into account. Further explanation of the diagnosis and revision process together with its actual implementation will be provided in the second part of this dissertation. Before we do so, an explanation of second major task of VAMs: to take factors outside the control of teachers into account is provided next.

[84] NRC (2010)

CHAPTER 5: TASK 2 – TAKING FACTORS OUTSIDE THE CONTROL OF TEACHERS INTO ACCOUNT

The second major task involving the VAMs is to take factors outside the control of teachers into account in order to provide the reliable and accurate (BLUE) estimate of their unique contribution to student performance. This task can be perceived as the way of "leveling the playing field" in order to a make fairer comparisons of teacher effectiveness by separating out the effects of factors outside of the control and responsibility of teachers (e.g. students' prior achievement level, socio-economic background, ethnicity, gender, etc.). Statistically, this task pertains to removing important omitted variables left in the residuals in order to revise the misattribution of the effects of other factors on the teacher effects. It pertains to correcting the violation of the exogeneity assumption described in Chapter 3. And this task can be completed under the multiple linear regression framework which by construction can statistically partition and separate the mutual dependencies inherent between the explanatory variables in explaining the outcome variable. As addressed in Chapter 2, this is a very powerful mechanism which can give rise to the "cause and effect" interpretation. Yet, despite its popularity, this seemingly simple task is in fact quite involved and controversial. A thorough understanding of this process requires a solid understanding of both the theoretical and statistical/mathematical dimensions underlying this process. These dimensions must work collaboratively in justifying which variables are taken into account in the model. But oftentimes, the collaborative roles of these dimensions remain unrecognized as one is prioritized over the other. As the understanding of these dimensions is critical for the successful implementation of VAMs in practice, an explanation of the two dimensions and their collaboration is provided next.

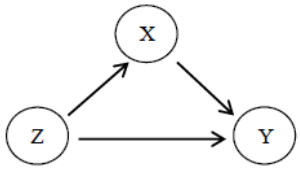
Theoretical Dimension of the Process of Taking Factors into Account – Confounding vs. Intervening Variables

To effectively take factors outside the control of teachers into account, we first need to start with some theoretical and logical concept which describes the causal mechanisms underlying the social phenomena – student achievement and its determinants. Depending on the direction of the influence we perceive (or "theorize") between the variables, we classify two different kinds of control variables: confounding variables and intervening variables. And this classification has important consequences for the interpretation of the parameter estimates. As Fox (2008) describes, "the proper interpretation of bias [and parameter estimates] depends on the nature of causal relationship [arrows of influence we theorize] between the variables." (p.112)

The statistical conditions for the two types of control variables are equivalent. First, the variable must be associated with the explanatory variable of interest (teachers). That is, the correlation must not be zero. Second, the control variable(s) must be associated with the outcome variable but it must also logically explain and determine its values. An arrow of influence must be directed from the variable(s) to the outcome variable. Given these conditions, it is then the duty of our theories to envision a possible direction of influence (causal mechanisms) underlying the associations in the first condition. And depending on the direction of the arrows,

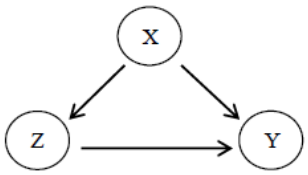
we classify the two types of control variables. A confounding variable is defined when the control variable influences both the variable of interest and the outcome variable. That is, arrows are directed from the control variable (Z) to the variable interest (X) and to the outcome variable (Y) as shown below.

Figure 5.1: Graphical illustration of the confounding variables



This figure illustrates that the variable interest (X) and the outcome variable (Y) are mutually dependent on the confounding variable (Z). In this scenario, if we fail to control the confounding variable Z in the model, then the effect between X and Y represents the genuine or causal effect of X on Y but also the effect of Z on Y through X. That is, the effect of X on Y is mixed and spurious as we misattribute the effect of Z on X. We commit an "under-adjustment" bias of the model estimates. That is, as an important variable Z which systemically explains the outcome variable is omitted and left in the residuals, the residuals are no longer randomly distributed. Moreover, as Z is correlated with modeled explanatory variables (X), the residuals (which now accommodate Z) are now correlated with X. And as described in Chapter 3, this violates the exogeneity condition and induces bias in the model. Now, an intervening variable is defined when the control variable is influenced by the variable of interest and influences the outcome variable as before. Unlike the confounding variable, the arrow is now directed to the control variable (Z) from the variable of interest (X) as shown below.

Figure 5.2: Graphical illustration of the intervening variables



In this scenario, the main variable (X) now has an "indirect effect' on the outcome variable (Y) through the intervening variable (Z). The causal relationship between X and Y is mediated by Z. Thus when the intervening variable is controlled, the total effect of X on Y will be partitioned into this "indirect effect" and "direct effect" or other influences of X on Y that remain after the indirect effect is taken into account. If we focus solely on the direct effect (and fail to report the indirect effect), we will be making an "over-adjustment" for the effect of X. We will be providing a conservative interpretation for the effect of X. But there is no bias committed in the traditional sense (where extraneous effect of another variable is misattributed to the main variable) but we must be cautious in interpreting the results as the variable of interest can have multiple avenues for influencing the outcome variable. To sum, depending on the direction of influence we theorize between the variables, we will be making quite a different interpretation of the parameter estimates.

In the context of VAMs, researchers across all disciplines acknowledge the prevalence of severe confounding variables when it comes to estimating teacher effects. As Dale Ballou in the NRC (2010) "Getting Value Out of Value-Added" report describes,

> Non-random assignment is pervasive in education, results from decisions by parents and school administrators: residential location decisions (often influenced by the perceived quality of local schools); parental requests for particular teachers or other efforts to influence teacher assignment; administrative decisions to place particular students with particular teachers—sometimes to improve the quality of the teacher-student match, sometimes as a form of favoritism shown to teachers or parents. (p. 8)

In other words, student and families (and also by schools and administrators) may systematically (non-randomly) self-select their schools and teachers. And this selection mechanism is (often) determined by a variety of student background characteristics. And as a consequence of this self-selection, schools and teachers serve student populations with very different backgrounds. But moreover, the factors that influence the selection of teachers and schools are also found to influence students' academic achievement. It has long been recognized that the same student background characteristics are strong predictors of student performance: family background characteristics (such as family income, parental education, language spoken at home; economic and cultural resources at home, etc.); demographic characteristics (such as gender, ethnicity, immigration status, etc.); residential and community factors outside the school environment, etc. Formal school and classroom learning environment, policies, resources and leadership are also found to have substantial influence on teachers and student achievement. And this nexus of associations creates the classic case of confounding in estimating teachers' effects as described above: student background variables determine the self-selection of teachers; student background variables affect student performance; and teachers affect student performance. If the effects of these background variables that are outside the control of teachers are not taken into account (separated or filtered) from the teachers, the effects on student performance will be mixed and misattributed to the teacher. That is, teacher effects will be biased.

But when it comes to detailed specification of the causal mechanisms, researchers continue to remain divided. Educators, researchers, and policymakers all agree that there are numerous factors affecting student achievement and learning. But given the interdisciplinary nature of education and also the varying availability of variables in the data set, researchers have envisioned different types of models to capture teacher effects.[85] Moreover, when it comes to detailed specification of the causal mechanisms (not just the main variable but

[85] Please refer to Plecki et al. (2009) for a literature review of different variables used in modeling student performance and teacher quality/effects.

also among the control variables), researchers continue to remain divided.[86] But despite this ongoing debate, clear statistical understanding and proper implementation of confounding and intervening variables is clearly lacking in practice. The notion of confounding variables often dictates the process without acknowledging the possibility of intervening variables. And this lack of awareness is one of the key contributors in aggravating the heated debate over deciding which variables to take into account in the model. In light of this problem, the recent book by Harris (2011) titled "Value-Added Measures in Education – What Every Educator Needs to Know" provides the first bold attempt to find consensus in the VA literature. It provides a statistically sound yet very intuitive approach towards thinking about which variables to take into account in VAMs. This approach is based on the "cardinal rule of accountability" which states "to hold people (teacher and schools) accountable for what they can control." Underlying this rule is the very essence of the confounding and intervening variables described above. The notion of what a teacher "can control" is equivalent to saying what teacher "can affect and influence" or "what they can decide and are responsible for". Statistically, arrows will be directed from the teachers to these factors they can control. And this creates the precise condition for the intervening variable for these "can control" variables. As shown in the table below, examples of these variables include teachers' decision to participate in programs such as professional development and to obtain additional credentials. And studies have found that these variables have a significant effect on student performance, teachers have an indirect effect on student performance via these intervening variables.[87] Conversely, factors that "are outside of control" of teachers or "factors teachers cannot influence" but rather factors teachers are "exposed to and affected by" are exactly the confounding variables. Arrows of influence are directed from these variables to the teachers. As described above and as shown in the table below, there are number of these confounding variables. We must control for these variables in order to avoid misattributing its effects (introduce bias) to the teachers.

[86] For example, in the current VAM research, the following two topics are at the center of heated debate. First, the role of prior test scores. Some researchers such as Sanders et al. (2006) have addressed how other covariates including SES is unnecessary once the (multiple sets of) prior test scores (which is also a function of SES) are taken into account. McCaffrey et al. (2009) on the other hand, disagrees how prior test scores alone cannot take into account of all the SES effects which may have time invariant effects (affect current and prior test scores differently). Coleman and Hoffer (1987) and econometricians also address the danger of the endogeneity problem with the prior test scores under the assumption that the errors are autocorrelated. They suggest the use of instrumental variables or dynamic panel analysis which utilizes 2 year differences in prior scores as instruments. Second, the construction of the socio-economic status index. Studies use different sets of available variables (e.g. FRPL status, parental education, parental occupation types, immigration status, family wealth and resources, etc.) to represent the SES construct and researchers have addressed the danger of over-generalizing with the interpretation SES index/concept as it can entail very different meanings.

[87] Plecki, Elfers, and Nakamura (2012) found that in the State of Washington teacher qualifications, education level, and the institution from which they graduated have a significant effect on student performance. These findings are also found in other states. Please refer to the document for further references.

Table 5.1: Harris (2011) approach in deciding which variables to take into account in VAMs

Uncontrollable		**Controllable**
Measured (account for)	*Unmeasured (account for)*	*Measured (shouldn't account for)*
School Resources:	District leadership	Teacher credentials
Class size	District funding	
Staff positions	District policies	
Teacher experience	Collaboration among schools	
Grade/subject experience	School leadership	
Teacher's aide	School policies	
	Collaboration within schools	
Student Characteristics:		
Prior test scores		Program participation
Race/income		
Mobility		
Variation in achievement		
Student absences		
Courses		

To sum, this framework provides us with a very user friendly approach in synthesizing the roles of our theories and the two types of control variables. That is, it provides a very thoughtful approach in distinguishing between confounding and intervening variables from a wide spectrum of variables which can be taken into account to capture the elusive teacher effects. Having now acknowledged the role of our theories and statistical concepts underlying the taking into account process, we now proceed to the explanation of the statistical/mathematical mechanics underlying this process. The understanding of the mathematical dimension will enable us to be more cautious and aware data analyst in practice.

Statistical/Mathematical Dimension of the Process of Taking Factors into Account – Semi-partial Correlations and Unique Variance Explained

There are two types of statistical/mathematical explanations underlying the taking into account process in order to estimate the accurate and reliable unique teacher effects. Both are administered under the framework of multiple linear regression models. The first explanation is based on the idea of semi-partial correlations and the second is based on the idea of unique variance explained (R squared) by each variable.[88] The former is used to describe the regression estimates as the net effect while the latter is used to describe the estimates as unique variance explained by the variable above and beyond other variables. Although we have become widely accustomed to these interpretations today, clear understanding of these two statistical concepts often does not prevail. As these explanations lie at the heart of the taking into account process, without a clear understanding of these concepts we are not able to successfully implement and monitor this task in practice. The explanation of these concepts therefore demands further clarification as provided below.

[88] The former correlation based interpretation is mainly attributed to the work of K. Pearson while the latter is mainly attributed to R. Fisher who designed the ANOVA or variation decomposition approach towards analyzing linear regression models.

1. Semi-partial Correlations:

In Chapter 2, the linear regression estimates were derived to take the following form.

$$\hat{\beta}_{ls/mle} = (X'X)^{-1}X'Y$$

where X is a matrix with the data values of all the explanatory variables (N x k) and Y is the matrix of the values of the outcome variables (Nx1).[89] But from this multivariate matrix alone, it is not possible to clearly witness the "taking into account" or "controlling process" of the effects of other variables that are confounding or intervening the effects of the variable of interest. The matrix equation masks the vital information that leads to our understanding of this process. In order make this mechanism more transparent and intuitively clear, we need to disassemble the above matrix equation. We start with the bivariate regression (only two explanatory variables) case where the population regression model we could like to estimate is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_3 X_{2i} + u_i$$

And the usual least squares/MLE estimators minimizes the sum of squared residuals with respect to the parameters[90]

$$min \sum u_i^2 = \sum (Y_i - (\beta_0 + \beta_1 X_{1i} + \beta_3 X_{2i}) \; with \; respect \; to \; \beta_0, \beta_1, \beta_2$$

to arrive at the following first order optimizing equations.

$$\bar{Y} = \widehat{\beta_0} + \widehat{\beta_1}\bar{X}_1 + \widehat{\beta_2}\bar{X}_2$$

$$\sum Y_i X_{1i} = \widehat{\beta_0} \sum X_{1i} + \widehat{\beta_1} \sum X_{1i}^2 + \widehat{\beta_2} \sum X_{1i}$$

$$\sum Y_i X_{2i} = \widehat{\beta_0} \sum X_{2i} + \widehat{\beta_1} \sum X_{1i} X_{2i} + \widehat{\beta_2} \sum X_{2i}^2$$

And through simultaneously solving for these equations, we arrive at the following least squares/MLE estimates where the variables are defined as deviation from its respective means e.g. $y_i = Y_i - \bar{Y}$

$$\widehat{\beta_0} = \bar{Y} - \widehat{\beta_1}\bar{X}_1 - \widehat{\beta_2}\bar{X}_2$$

$$\widehat{\beta_1} = \frac{(\sum y_i x_{1i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{1i} x_{2i})}{(\sum x_{1i}^2)(\sum x_{2i}^2) - (\sum x_{1i} x_{2i})^2}$$

$$\widehat{\beta_2} = \frac{(\sum y_i x_{2i})(\sum x_{1i}^2) - (\sum y_i x_{1i})(\sum x_{1i} x_{2i})}{(\sum x_{1i}^2)(\sum x_{2i}^2) - (\sum x_{1i} x_{2i})^2}$$

Inserting the data (respective values of the variables) into the formulas then gives us the numerical value of each estimate. This is the usual output provided by our statistical analysis which we interpret as the "the net

[89] N was the number of individuals/sample in the data and k is the number of explanatory variables included in the model.
[90] This is conducted by partial differentiation of the sum of squared residuals with respect to each parameter.

effect" of each variable on the outcome variable. Although the above equations are used in a majority of modern statistical textbooks to demonstrate the idea of "net effect" and "take into account" it is actually not the correct form for doing so. Even from closely looking at the different components of the equations, there are no clear illustrations and signs that signify the extensively used interpretation. We must further work on the equations in order to define it in terms of semi-partial correlations that provide us with the precise presentation of the taking into account process. To do so, we need to first transform the above equations in terms of pairwise correlation coefficients as shown below.

$$\widehat{\beta_1} = \frac{\sigma_Y(r_{Y1} - r_{Y2}r_{12})}{\sigma_{X_1}(1 - r_{12}^2)}$$

$$\widehat{\beta_2} = \frac{\sigma_Y(r_{Y2} - r_{Y1}r_{12})}{\sigma_{X_2}(1 - r_{12}^2)}$$

These formulas can derived by dividing the numerator and denominator of the earlier equations of $\widehat{\beta_1}$ and $\widehat{\beta_2}$ by $(n-1)^2$ and then simplifying the expression based on the definitions of variance and correlation.[91] And by standardizing the variables, the variance and standard deviation of such standardized variables will be equal to one and this will further simply the formulas (by inserting value 1 into the standard deviation components).[92] In other words, the above expressions illustrate that bivariate linear regression estimates can be defined solely in terms of degree of associations (correlation) inherent between the different explanatory variables and the outcome variable. And more importantly, these formulas have direct resemblance to a special kind of correlation, the semi-partial correlation as shown below.

$$r_{Y(1.2)} = \frac{r_{Y1} - r_{Y2}r_{12}}{\sqrt{1 - r_{12}^2}}$$

Comparing this equation to the linear regression estimates above illustrates the direct resemblance as shown below.[93]

$$\widehat{\beta_1} = \frac{r_{Y1} - r_{Y2}r_{12}}{1 - r_{12}^2} = \frac{r_{Y(1.2)}}{\sqrt{1 - r_{12}^2}} = \frac{1}{\sqrt{1 - r_{12}^2}} \cdot semipartial\ correlation$$

The linear regression estimate $\widehat{\beta_1}$ is directly and positively proportional to the semi-partial correlations only differing by a constant. The same findings apply for $\widehat{\beta_2}$. Now, semi-partial correlation is the correlation between the outcome variable and an explanatory variable of interest from which the effects/association of all

[91] To recollect, the correlation coefficient between two random variables X and Y is defined as the covariance between X and Y divided by the product of the standard deviations of X and Y as shown below. It provides a measure of the degree of association between variables ranging from 0 (no association) to 1 (perfect association).

$$r_{XY} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2 \sum(Y - \bar{Y})^2}} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

[92] Standardization can be conducted through the same procedure of calculating z-scores by taking the deviations from the means and dividing it by the standard deviation

[93] The standardized form of $\widehat{\beta_1}$ is used where σ_{X_1} and σ_Y are equal to one.

the other explanatory variables have been partialled out from the variable of interest. The notation $r_{Y(1.2)}$ is read as the correlation coefficient between Y and X_1 after the variance/influence that X_2 has in common with X_1 has been removed from X_1. The "partialling out" or "removing" the effect is equivalent to subtracting from the explanatory variable of interest (X_1), the estimated X1 ($\widehat{X_1}$) using all the other explanatory variables under the regression framework (regressing X_1 on all the other explanatory variables). This process is called residualizing the X_1 (X_1 – estimated $\widehat{X_1}$) just as we calculate the residuals errors in the conventional linear regression models. As you can see, the outcome variable remains unaltered and only the explanatory variable of interest is residualized giving the name "semi" partial correlation.[94] Now, the residualized explanatory variable is then correlated with the unaltered outcome variable to provide the semi-partial correlation. And this residualizing process which removes the effects of all the other variables from the variable of interest is precisely the idea of "taking into account" or "controlling" for the effects of other variables under the multiple linear regression framework. It is the origin of the "net effect" interpretation we are accustomed to. Through removing the overlapping associations the variable of interest shares with the other variables, we are able to achieve a more pure and finer essence of the variable of interest. And through calculating the degree of association of this purified variable of interest with the outcome variable, we are able to achieve a much better estimate of the unique, accurate, and direct effect of the variable of interest on the outcome variable. This mathematical approach is the first presentation of the taking into account process that underlies the multiple regression estimates.

The key advantage of this presentation based on semi-partial correlation is that the estimates are interpreted as "net effects" on the original scale of the variables. This feature will be forgone under the second illustration which is defined on the variance scale (square of the original scale). On the other hand, the key drawback is that is that we still do not have the full precise picture of the taking into account process as we cannot isolate or pin down the exact magnitude of the confounding effects (or intervening effects) that has been removed from the variable of interest to identify the unique effects. As we cannot see the relative magnitude between the unique effects and the confounding (and intervening) effects, we cannot perceive the severity of the biases or alternatively the extent of improvement we have achieved through controlling the variables. To further delineate the taking into account process, we need to resort to the second presentation based on the proportion of unique variation of the outcome variable explained by each explanatory variable as described next

[94] Semi-partial correlation is different from partial correlation which represents the correlation between the criterion and a predictor after common variance with other predictors has been removed from ***both*** the criterion and the predictor of interest. That is, after removing variance that the criterion and the predictor have in common with other predictors, the partial expresses the correlation between the residualized predictor and the residualized criterion. It answers what is the correlation coefficient between two variables when the influences of a third variable is held constant

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

2. Proportion of Unique Variation Explained

An important part of regression analysis is to consider the goodness of fit of the estimated model in capturing the actual observed values of the outcome variable. The key measure for this purpose is the coefficient of determination or R^2 which is calculated as the ratio of the explained variation by the estimated model and the total variation of the outcome variable as derived as follows. The total variation (sum of squares) is equal to the sum of regression/explained sum of squares and the residual sum of squares,[95]

$$\sum(Y-\bar{Y})^2 = \sum(\hat{Y}-\bar{Y})^2 + \sum(Y-\hat{Y})^2$$

And by dividing both sides by $\sum(Y-\bar{Y})^2$, we obtain the R^2as follow.

$$1 = R^2 + (1-R^2)$$

Higher R^2values indicate a better model fit as more variation (pattern) of the outcome variable is explained and captured by our estimated model. But now, there are further important mechanisms which are masked under this R^2. Focusing first on the case of simple linear regression with only one explanatory variable, if we substitute the usual least square estimate $\hat{\beta}_{ls} = (X'X)^{-1}X'Y$ which is equal to $\sum(xy)/\sum x^2$ (where *x* and *y* are both deviations from its mean) into the $\hat{Y}$ (which becomes $\hat{\beta}_{ls}X$), the regression sum of squares ($\sum(\hat{Y}-\bar{Y})^2$) becomes $\sum(xy)^2/\sum x^2$.[96] And by converting it in terms of the proportion of variation explained (R^2), by dividing it by $\sum(Y-\bar{Y})^2$, the regression sums of squares component becomes $\sum(xy)^2/\sum x^2 y^2$ which is the "square" of the correlation coefficient between the explanatory variable and the outcome variable.[97] In other words, the proportion of total variation explained by the regression model is directly proportional to the degree of association between the outcome variable and the explanatory variable.

Proceeding now to the bivariate regression case (with two explanatory variables), repeating the same mathematical procedures derives the regression sum of squares as follows:[98]

$$\sum(\hat{Y}-\bar{Y})^2 = \hat{\beta}_1^2\sum {x_1}^2 + 2\hat{\beta}_1\hat{\beta}_2\sum x_1 x_2 + \hat{\beta}_2^2\sum {x_2}^2$$

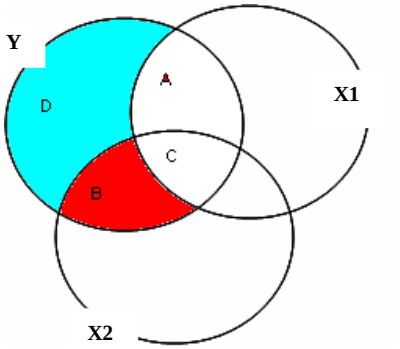
And through substituting the linear regression estimates (described in terms of pairwise correlations) into this equation and dividing by the total sum of squares, the R^2 of this model becomes

$$R^2_{Y.12} = \frac{r_{Y1}^2 + r_{Y2}^2 - 2\,r_{Y1}r_{Y2}r_{12}}{1-r_{12}^2}$$

[95] Variation is defined as the sum of squared deviation of observed scores from the mean of all scores. It is slightly different from variance which divides the variation by the number of observations minus 1 (this is called the degrees of freedom. 1 is subtracted from the number of observations because the mean was already estimated using the observations). Variance is in essence the average variation across the sample.
[96] Full and comprehensive proof is provided in Edwards (1985)
[97] Please refer to footnote 7 for the mathematical equation of the correlation coefficient.
[98] Cohen (1975), Edwards (1985), and Gujarati (2008) provide detailed derivation.

where $R^2_{Y.12}$ is read as the R^2 of the regression model with two explanatory variables X1 and X2. And as you can see, the proportion of explained variance (R^2) for the bivariate regression is also expressed as a function of different combinations of correlation coefficients between the outcome and explanatory variables. This derivation can then be naturally extended to regression models with as many explanatory variables. The reason why this expression is derived is because it can be shown that the R^2 can be partitioned/decomposed into parts that are "uniquely" attributable to each variable and to those that is "commonly" shared between the variables. And this partitioning of the variance has important implications in describing the taking into account process. To illustrate this point, the following Venn or Ballantine diagram is constructed.

Figure 5.3: Venn diagram of bivariate linear regression



Each circle represents the variation of each variable which is standardized to have an area of 1 e.g. A+B+C+D=1 for the outcome variable Y. The area A+C will then represent the R^2 of a simple regression model with only X1 as the explanatory variable ($R^2_{Y.1}$). And as explained above, this area is equal to the square of the correlation coefficient between X1 and the outcome variable Y. The area B+C is defined similarly for the variable X2 ($R^2_{Y.2}$). Following the same idea, the area A+B+C will then be the $R^2$of the multiple regression model when with both X1 and X2 as the explanatory variables ($R^2_{Y.12}$). And the mathematical formula for this was derived above. But having done so, we can then derive the values for the areas A, B, and C alone through simple arithmetic. Areas A and B represents the unique R^2 attributed to the variables X1 and X2, respectively, after removing the area C which represents the common R^2 shared by the two variables. For example,

$$Area\ A = R^2_{Y.12} - R^2_{Y.2} = \frac{(r_{Y1} - r_{Y2}r_{12})^2}{1 - r^2_{12}}$$

$$Area\ B = R^2_{Y.12} - R^2_{Y.1} = \frac{(r_{Y2} - r_{Y1}r_{12})^2}{1 - r^2_{12}}$$

Now, taking a good look at the formulas for areas A and B which represent the R^2 uniquely and purely attributed to the two explanatory variables, respectively, it is in fact the "square" of the semi-partial correlation of each variable with the outcome variable. Earlier, it was shown that the regression estimates can be expressed in terms of semi-partial correlations. In other words, the above finding illustrates that the unique contribution of each explanatory variable in explaining the variation of the outcome variable above and beyond

the other variables or alternatively the increment in the R^2 measure when the variable of interest is added last in the model (after, on top of the other variables) are directly proportional to the regression estimates as illustrated below. And this entitles it to become the second explanation of the regression estimates. Moreover, through the semi-partial correlation, the above explanation is directly proportional to the first "net effects" explanation of the regression estimates. That is, the two approaches provide the same results.

$$\widehat{\beta_1} = \sqrt{unique\ R^2\ contributed\ by\ X_1} \cdot \frac{1}{\sqrt{1 - r_{12}^2}}$$

The key advantage of the R^2 based interpretation of the regression estimates is that we are now able to better delineate the taking into account process. More specifically, we are able to identify the exact degree, magnitude, and extent of the confounding (or intervening) effects that have been removed and "taken into account" through the multiple regression framework in order to provide a more reliable and accurate effect of the variable of interest. Looking at the Venn diagram again, through the process of defining the area A and B, we essentially had to filter, subtract, and separate out the area C which is the amount of common variance that is shared by the explanatory variables in explaining the outcome variable. This area C can be derived by simply subtracting either area A or B from the respective simple linear regression R^2 values described above (area A+C and B+C, respectively). And as described earlier, depending on our theoretical lenses (which defines the causal mechanisms or the directions of the arrows/influence between the variables), this area C is precisely the extent of confounding biases or intervening effects we have longed to identify. For example, if X2 is a confounding variable of X1 (arrow pointing into X1 from X2 and from X2 to Y), then area C represents the magnitude and degree of confounding effects of X2 on the effects of X1 on Y that the multiple regression model (with both variables) corrects by removing it to provide a better and accurate measure of the effects of X1 on Y (as measured by the area A). Conversely, if the confounding X2 is omitted (not taken into account) in the regression model, then area C precisely represents the measure of bias that is disrupting the effects of X1 on Y. It is the measure of the effect that is mistakenly misattributed or mixed to the effect of X1 on the outcome variable.

The above illustration can be naturally extended to the case of multiple explanatory variables (more than 2 variables). For example, in the case of 3 variables, the unique contribution of each variable in explaining the variation of the outcome variable is as follow.

$$r_{Y(1.23)}^2 = R_{Y.123}^2 - R_{Y.23}^2$$
$$r_{Y(2.13)}^2 = R_{Y.123}^2 - R_{Y.13}^2$$
$$r_{Y(3.12)}^2 = R_{Y.123}^2 - R_{Y.12}^2$$

Following the same mathematical procedures, these formulas will then equal the square of semi-partial correlation of the respective variable which will link us to the linear regression estimates and the first

interpretation based on "net effects".[99] And for the general *k* explanatory variables case, the unique contribution of each variable becomes

$$r^2_{Y(k.123\ldots k-1)} = R^2_{Y.123\ldots k} - R^2_{Y.123\ldots k-1}$$

Finally, the statistical significance of the unique variance explained by each variable can be tested using the *F* distribution. Fisher (1922) has shown that under the assumption of a normally distributed population of the outcome variable, the following formula follows a distribution known as the *F*-distribution.[100] This is the mirror image of the *t* statistics used in the first ("net effects") presentation.[101]

$$F = \frac{r^2_{Y(1.23)}}{(1 - R^2_{Y.123}/(n-k-1)} = \frac{R^2_{Y.123} - R^2_{Y.23}}{(1 - R^2_{Y.123})/(n-k-1)}$$

To sum, the statistical presentation of the regression estimates in terms of semi-partial correlation and proportion of variation explained (R^2) deepens our understanding of the taking into account process underlying the multiple linear regression models. As these statistical concepts provided us the exact magnitude of the unique effects/variation explained by each variable and the overlapping confounding biases or intervening effects inherent between the variables, it has put an empirical lens on our theories in considering which variables to take into account. Finally and importantly, by being aware of the precise definition of the semi-partial correlation and the pairwise correlations underlying the regression estimates, this enables us to become even more cautious with data analysis in practice. These basic univariate statistics will give us the ability to obtain the first hand confirmation of our theories prior to estimating any regression models. It will also shield us from any data anomalies that can disrupt the models and advise us with the model revision process when the regression assumptions are violated.

Summary and Conclusion: Towards Successful Implementation of the Taking into Account Process – A Theoretical and Data Driven Approach

In the previous sections, the important roles of our theories and mathematical derivations underlying the taking into account process were highlighted. The interaction and collaboration between these roles were also illustrated. But in the course of statistical history, researchers have failed to acknowledge the importance of both theory and data in the taking into account process also known as the model building processes. One school of thought referred to as the confirmatory data analysis (CDA) has prioritized the role of theory over data while another school of thought called the exploratory data analysis (EDA) has prioritized data over

[99] The mathematical derivation process is the same as the bivariate case but it is more involved. Cohen (1975) and Edwards (1985) provide some illustration of this process.

[100] The F-distribution was initially derived and used under the ANOVA framework which is based on the mean square error. Through simple arithmetic, this F test formula can be re-expressed in terms of R^2as shown in the text. This derivation is provided in Edwards (1985) and Gujarati (2008).

[101] For example, in a simple linear regression case, the square of t statistic will be equivalent to the F statistic. Further illustration of the link between t and F tests is provided in Rutherford (2001)

theory.[102] But today, an increasing number of researchers acknowledge the importance of both theory and data. It is their contention that the two schools of thought can collaborate to solidify the adequacy of the findings i.e. BLUE estimates that have real life meaningful interpretation.[103] As Bosker and Snidjers (1999) address,

> There are two steering wheels [in decision which variables to take into account]: substantive (subject-matter related) and statistical considerations. These steering wheels must be handled jointly. (p. 91)

Specifically, the researchers highlight the importance of an iterative procedure between our theories, exploratory analysis of the data "prior" to fitting the model, and diagnosis and revision of the "post" fitted model (explained in Chapter 3). For example, the work by Neter et al. (2004) suggests a comprehensive four-step strategy for building and estimating a linear regression model: data collection and preparation (with tentative model construction), reduction in the number of variables (as part of exploratory data analysis), model refinement and selection (model diagnosis and revision), and model validation.[104] After a tentative model is built in support of our theories, the authors address the importance of first conducting an exploration of the data e.g. the basic descriptive summaries introduced in Chapter 2 such as means, standard deviation, simple correlations, and simple regression. This descriptive analysis will provide us with the initial confirmation of our theories and whether the linear regression assumptions are likely to be satisfied. It allows us to get a sense of what the data is portraying and more importantly it will shield us from any data anomalies (e.g. outliers) which could disrupt the estimates. This analysis can also suggest the possibility of reducing the list of variables especially when some similarly defined variables are subject to high collinearity. Once the models are then estimated and thoroughly diagnosed and revised (as explained in Chapter 2 and 3), the authors then address the importance of conducting validity of the estimated models by assessing its sensitivity and generalizability. Sensitivity analysis can be conducted through comparing the estimates to the findings from other prior analysis or conducting simulation analysis that can intentionally alter some of the underlying determinants (e.g. sample size) of the parameter estimates. Generalizability or plausibility of the estimates can be evaluated by applying the estimated model to a new data set or a hold out sample to see if the model estimates prevail in other data sets.[105] To sum, these procedures highlight the importance of the collaborative analysis between our theories, data, and estimated models. Only when the model is effectively cross-validated with respect to these factors, the adequacy (BLUE properties) of the estimates which have real life (theoretical) meaning is ensured. The second half of this study will reflect on these procedures as we implement VAMs using Washington State longitudinal data.

[102] The CDA is dictated by our theories where data is used to deduce our specified theories prior to fitting the model while the EDA is navigated by the data to induce possible theories and hypothesis underlying the data. Extensive reviews of these two approaches date back to early 1900s e.g. Tukey (1977) who found the notion of exploratory data analysis.
[103] Gelman and Hill (2011), Muijs (2004), Fox (2008), Gujarati (2008), Raftery (2011), Tukey (1977), Neter et al. (2004) and more.
[104] Detailed explanation and presentation is provided in Chapter 11 Model Building Process of their book.
[105] Hold out sample can be obtained by splitting the data prior to the data or collecting a random sample from existing data.

PART II: VALUE ADDED ANALYSIS USING THE WASHINGTON STATE LONGITUDINAL DATA

In Part I of this study, the explanation of the two major tasks underlying VAMs were provided– defining the VA parameter and take factors outside the control of teachers into account. And to successfully complete these tasks under the linear regression framework, the importance of collaboration between our theories in envisioning the causal mechanisms, evidence provided by the exploration of the data (prior to estimating any models), and evidence provided by the diagnosis of the estimated model performance (post estimating the models) was highlighted. And for the post-estimation diagnosis, the linear regression model assumptions became the guiding principals to ensure the BLUE estimates. In Part II of this dissertation, this study applies these explanations to conduct a thorough teacher VA analysis using the Washington State longitudinal data. This proceeds in the following sequence of chapters.

Chapter 6	Descriptive and Exploratory Analysis of the Data
Chapter 7	Preliminary VA Analysis – based on our theories, existing work, and resu from the descriptive analysis
Chapter 8	Diagnosis and Revision of the Preliminary VA Analysis
Chapter 9	Validation Analysis (i.e. Sensitivity Analysis) of the Revised VAMs
Chapter 10	Policy Analysis Using the Revised and Validated VAMs

Unlike the existing VA research which often hastily jumps into the policy analysis which receives the media attention, this study conducts a more rigorous and comprehensive analysis and evaluation of the models. Only for the models in which the BLUE property is justified, will the findings be applied for the policy relevant circumstances. It is the contention of this study that only through the rigorous diagnosis and revision of the models we can achieve the utmost confidence in utilizing the VA results for the betterment of our education system.

CHAPTER 6: DESCRIPTIVE ANALYSIS OF THE WASHINGTON DATA

This chapter presents the descriptive analysis of the Washington State longitudinal data used in conducting the VA analyses. As described in Chapter 2, descriptive analysis examines, describes and summarizes the data in manageable and simple form. In this chapter, the findings of the descriptive analysis will used to investigate whether the data set confirms and justify the conduct of VAMs. Attention will therefore be made on the conditional mean values of the outcome variable for the different values of the explanatory variables especially the teachers which is the main variable of interest. This will provide us with the initial justification of using the regression models (defined as the conditional mean function as described in Chapter 2) which lies at the heart of VAMs. Attention will also be made on the interconnections between the explanatory variables and the conditional mean values of the outcome variable for different combination of explanatory variables. These findings will provide initial empirical evidence of the presence of confounding or intervening effects between the variables (i.e. with the teachers) which will signify the importance of conducting the second major task of VAMs – to take into factors outside the control of teachers into account. In the following section, the description and the basic (univariate) summaries of the variables comprising the Washington data set is provided. The detailed descriptive analyses of the variables then follow.

The Washington State Longitudinal Data Set

This study utilizes the Washington State longitudinal data set which combines the Comprehensive Education Data and Research System (CEDARS), the S-275, and the Teacher Data made available by the Office of Superintendent of Public Instruction (OSPI) and under the courtesy of the TNE Project at the University of Washington.[106] Of the years made available, 4th and 5th graders between 2006/2007 to 2007/2008 which had the highest percentage of matched cases with 71% is used in this study to conduct VA analyses for 5th grade math teachers.[107] The explanation of the variables and its basic (univariate) descriptive statistics are as follow.

[106] CEDARS contains the students' Washington Assessment of Student Learning (WASL) performance from grade 3 to 10 and student background characteristics such as gender, ethnicity, free-or-reduced price lunch status, English as second language status. The S-275 and the Teacher Data contain information of the teachers including their date of birth (age), gender, ethnicity, highest degree/education attained, teacher preparatory institutions teacher graduated/certified, subjects teachers are certified, and the schools and districts they belong.
[107] Other periods had less than a 50% match rate and was therefore not sought in this analysis.

Table 6.1: Variable description and categorization

	Coding and Categories
Dependent Variables	
5th Grade Math Scores	Contiuous variable
Student Level Variables	
Gender	Student is female = 1, male = 0
Ethnicity	Student is native american = 1, asian = 2, black = 3, hispanic = 4, white = 5
Free or Reduced Price Lunch Eligbility	Student is on FRPL = 1, not on FRPL = 0
4th Grade Math Score	Contiuous variable
Teacher Level Variables	
Teacher Experience	Continuous variable
Highest Education Degree	Teacher has a masters or doctorate = 1, bachelors and others = 0
Gender	Teacher is female = 1, male = 0
Ethnicity	Teacher is white american = 1, otherwise = 0

Table 6.2: Basic descriptive statistics of the variables

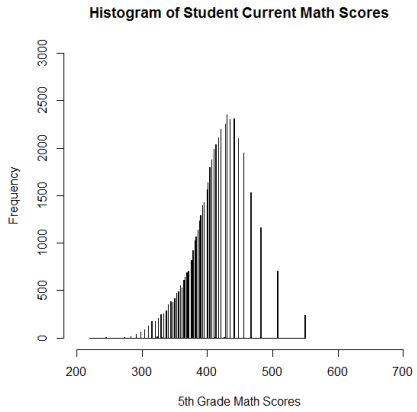
	Mean	***Std Dev***	***Mode***	***Mini***	***Max***
Dependent Variables					
5th Grade Math Scores	408.5	37.70	-	219.0	550.0
Student Level Variables					
Gender	0.488	0.500	0 (51%)	0	1
Ethnicity	4.249	1.319	5 (66%)	0	5
FRPL Status	0.167	0.373	0 (83%)	0	1
4th Grade Math Score	406.1	41.45	-	213.0	550.0
4th Grade Math Score (grand mean centered)	0.000	41.45	-	-193.1	143.9
Teacher Level Variables					
Years of Experience	13.70	9.847	-	0	45
Years of Experience (grand mean centered)	0.000	9.847	-	-13.7	31.3
Highest Education Degree	0.672	0.470	1 (67%)	0	1
Gender	0.745	0.436	1 (74%)	0	1
Ethnicity	0.937	0.242	1 (94%)	0	1
Number of 5th Grade Students	51,161				
Number of 5th Grade Math Teachers	2,864				
Number of Schools (5th grade students and teachers belong)	936				

In the following sections, the descriptive analysis of the variables listed above is conducted starting with the outcome variable of interest followed by the main explanatory variables of interest (the teachers) and other explanatory variables.

The Outcome Variable of Interest

The main outcome variable of this study is the 5th grade mathematics test scores provided by students who were in 5th grade in 2007-08. The distribution of this variable illustrates the following pattern.

Figure 6.1: Histogram of student 5th grade math scores



Two noteworthy features are evident in this distribution: first, a bell shaped symmetric curve resembling the normal distribution and second, a broken distribution particularly towards the upper end of the scale. To explore the first feature, the numerical summaries of the above histogram are calculated to find a mean of 408.5, standard deviation of 37.7, mode of 403 and the median of 407. The closeness of the three central tendency indices provides the empirical evidence of symmetric distribution. The actual symmetric index or skewness value is calculated to be 0.173 which is very close to the perfect symmetry of zero. Symmetry is also evident in the similar spacing of the first quartile is at value 386 and the third quartile is at 430 from the median value. This is indicated in the equal spacing of the top and bottom lines of the boxplot from the dark middle line (median) shown below.

Figure 6.2: Quantile plot of 5th grade math scores

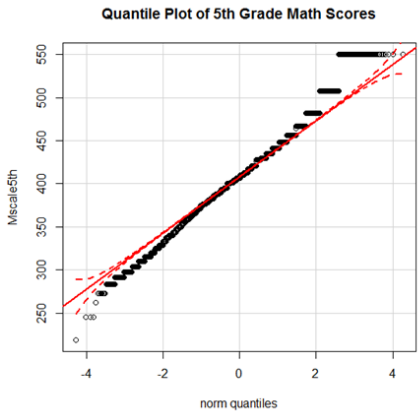
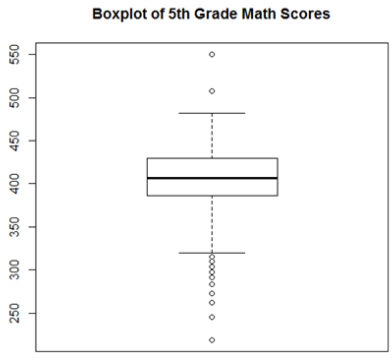


Figure 6.3: Boxplot of 5th grade math scores



To test whether the symmetric distribution resembles the normal distribution, the qqplot shown above is constructed.[108] Except for the few observations toward the upper and lower end of the distribution, majority of the observations lie on the straight red line which indicates the normal distribution. The slightly more deviation towards the upper end confirms the slight positive skewness evident above. Finally, calculating the kurtosis value, the distribution has value 3.838 which is slightly above yet very close to the normal distribution

[108] The quantile comparison plots (qqplots) compares observed data with a theoretical probability distribution. If the observed data closely resembles the distribution, then the data will lie on the 45 degree line.

value of 3. Thus from these findings, it can be said that the 5th grade mathematics test have a good approximation of the normal distribution.

Regarding the second feature – the broken and clustered distribution towards the upper end signifies the design and type of psychometric test characterizing the 5th grade WASL test.[109] As addressed in the WASL technical report (2005), WASL is a "criterion based test" which measures whether the students achieve the specified standards required by the state curriculum. It measures the students' achievement primarily at the passing standard and below and not toward the upper or the overall range of achievement.[110] For this reason, the WASL is comprised of test items (questions) concentrated around the standard and below. That is, there are finer measurements toward the lower end of the ability scale and not towards the upper scale. As there are fewer items to measure students ability toward the upper end of the item difficulty (less items to distinguish the students' ability), students' performance in this area are clustered together. In the context of VAMs, it is important to understand the reason and source for his systematic pattern. The understanding of this pattern will help us in the model building process and residual diagnostics. Keeping the features of the outcome variable in mind, we now proceed to the descriptive analysis of the explanatory variables which will be used to estimate and model the variation of the outcome variable.

Explanatory Variables:

The Main Explanatory Variable of Interest - Teachers

The main variable of interest in this study is the 5th grade math teachers. We would like to investigate whether the teachers have a significant effect on 5th grade students' math performance. That is, whether significant differences in the teacher mean values of their student test score (the conditional means which define the regression function) exist. In Washington State, there are total of 2,864 5th grade math teachers. These teachers teach a variety of difference class size as shown below.

[109] Thorough investigation of the graphics command was conducted to show that it is not due to any technical specification of the graphs e.g. the number of categories/bins used to construct the histogram.

[110] The tests which measure the overall ability of students are known as the general ability test which include the SAT and the GRE.

Figure 6.4: Histogram of teacher class size

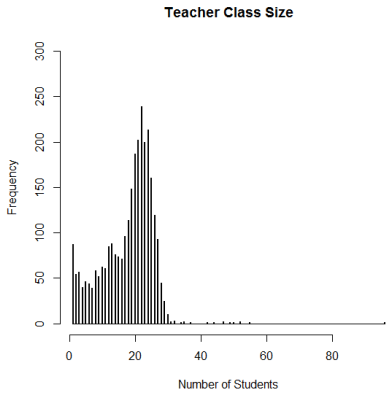
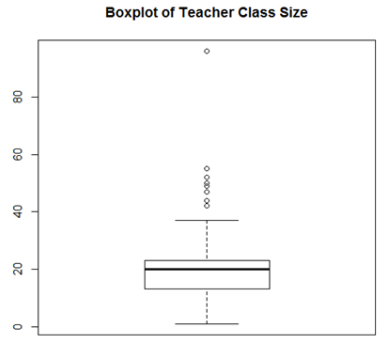


Figure 6.5: Boxplot of teacher class size



The mean value of teacher size is 18 students with a median of 20 and mode of 22. But as you can see from the distribution, there is quite a lot of variation around these central tendency indexes. The standard deviation of 7.73 and the range spans from teachers with only 1 student to 96 students. The latter teacher is indicated as a clear outlying value as shown in the boxplot. Moreover, the distribution indicates a clear non-symmetric pattern with high concentration of teachers toward the lower end of the scale. There are in fact quite a lot of teachers with only 1 student in their class (87 teachers). But as it was described in Part I of this study that the class size is the critical factor in determining the reliability and stability of the teacher effects, the teachers with very small class sizes would impose challenges their VA estimates. This point is further investigated in later chapter. Looking now at the teacher average test scores (the class average test scores) show the following trend.

Figure 6.6: Histogram of teacher average 5th grade math scores

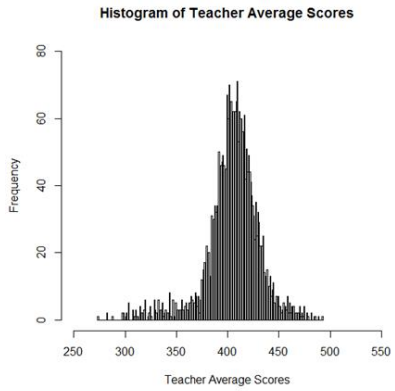
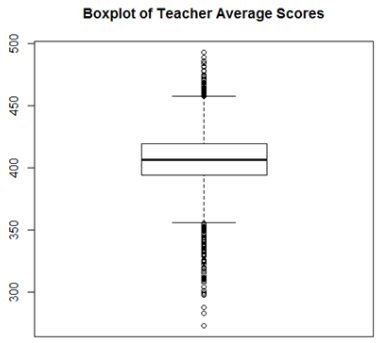


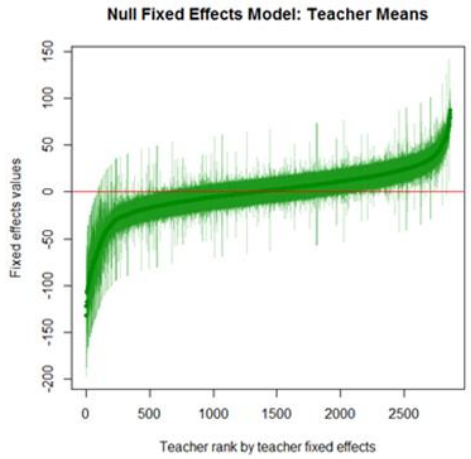
Figure 6.7: Boxplot of teacher average 5th grade math scores



The average of the teacher mean test scores is 405.2 with median of 406.7 and mode of 413 points. The closeness of these three values indicates a symmetric distribution. This is also confirmed with a skewness value of -0.925 which is very close to the perfect symmetry of zero. Yet, there is a clear variation (differences) in teacher means with a standard deviation of 16.5 and a range of 273 and 492.9 points. Furthermore, the symmetry in the distribution implies that there is similar frequency of teachers performing below and above the overall mean (mean of teacher means). That is, similar number of teachers has relatively high means and

relatively low means. To test whether the differences in the teacher means are statistically significant (whether the teachers differences are not due to some random fluctuation, chance, or error), the following figure known as the caterpillar plot is constructed.

Figure 6.8: Null fixed effects model: teacher means



The caterpillar plot presents the results of the simple one-way ANOVA test with the teachers as the fixed treatment factor. The dark green line/dots in the middle represent the deviations of the teacher means from the overall mean of 405.2.[111] These mean values are then supplemented with the standard error or the 95% confidence interval ($\pm$1.96 times the SE estimates) represented with the band (vertical lines) crossing through the teacher effects. Thus, if the 95% confidence interval band does not cross the 0 value, the teacher mean is statistically and significantly different from the overall mean value. And this provides the first evidence of the presence of significant teacher effects. Looking at the plot again, as you can see from the teachers toward the upper end of the spectrum, these teachers have mean values which are significantly greater than the overall mean. On the other hand, the teachers toward the lower end of the spectrum have mean values which are significantly less than the overall mean. In other words, depending on the teacher the student is taught, the students' performance will significantly vary. The differences in mean scores among the total 2,864 teachers are due to systemic differences between the teachers and not due to random error.[112] Teachers are an important variable in explaining student performance.

But the important question then is to ask whether these differences are solely due to teachers or to other factors (outside of their control) which also affect student performance. That is, we would like to know whether the teacher means are significantly high (or low) because the teacher themselves effectively rose (lessoned) their students' performance or because the teachers just happen to be surrounded with high achieving students with

[111] As described in Chapter 4, this is referred to as the effect coding where estimates shown in the plot are deviations of teacher means from the grand mean. The 0 value signifies the grand mean. The original teacher means can be calculated by simply adding these deviations to the overall mean value.

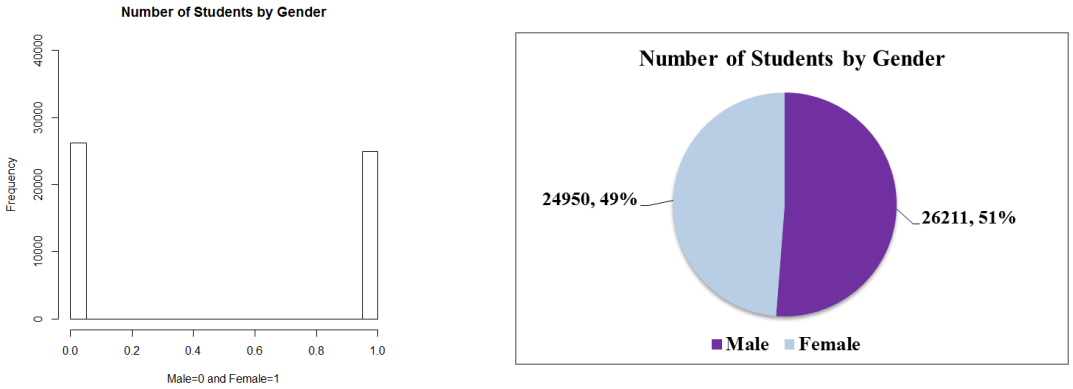
[112] This finding implies that the ANOVA F test which evaluates whether all the teacher means are jointly equal to the grand mean (all deviations equal to zero) is rejected.

favorable background (and vice versa). This challenge is the *raison de entre* and the second major task of the VAMs as described in Chapter 5. In order to investigate whether this challenge is addressed by the data, the descriptive analysis of the student background variables is provided next. The exploration of the nexus of association of these variables with the teachers and the possible confounding effects then follows.

Explanatory Variables Outside the Control of Teachers – Student Background Variables

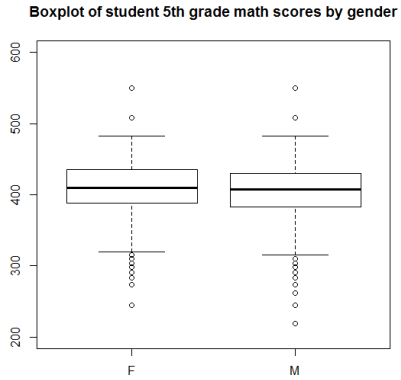
The Washington data consists of the following background variables that students bring to the teachers and classrooms: gender, ethnicity, free-or-reduced priced lunch status, and prior 4th grade test score. The descriptive analyses of these variables are provided below starting with students' gender.

Figure 6.9: Histogram and pie chart of student gender



There is a fairly equal number of male and female 5th grade students in the State of Washington with 51.2% and 48.8% of the total students, respectively. Looking at their average test score performance shows the following result.

Figure 6.10: Boxplot of student 5th grade math scores by gender



The female students have a slightly higher average score with 411.2 than the male students with 405.8. The standard deviation is virtually the same with 37.7 and 37.5, respectively. Other numerical summaries for the

two groups also show close resemblance as described in the footnote.[113] To test whether the difference in the mean scores is statistically significant, simple one-way ANOVA and simple linear regression are estimated to show the following results.[114]

Table 6.3: Estimates of one-way ANOVA and simple linear regression with student gender

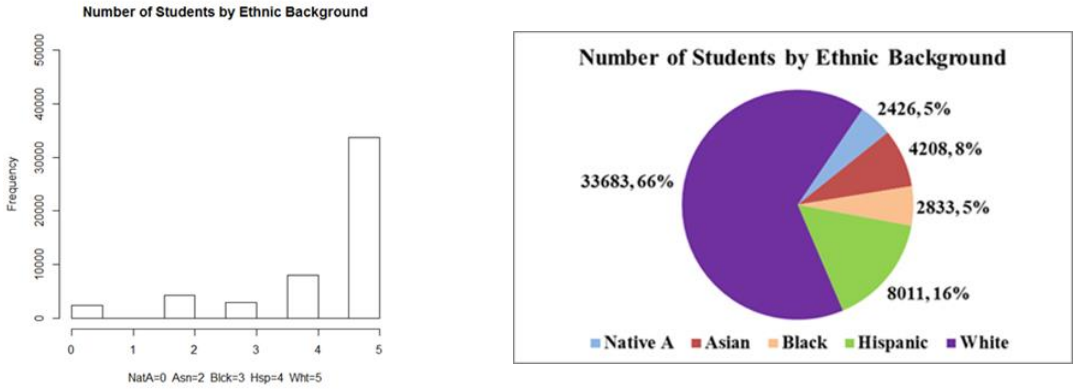
	Df	*SS*	*MS*	*F value*	*p value*	
factor(Gender)	1	69517	369517	261.3	0.00	***
Residuals	51159	72342157	1414			
	Estimate	*SE*	*t value*		*p value*	
(Intercept)	405.87	0.232	1747.4		0.00	***
Student is Female	5.38	0.333	16.16		0.00	***
R square	0.005					
Adj R square	0.005					
Note: *** significant at 0.1%; ** at 1%; * at 5%						

The finding above indicates that the difference between the two groups is statistically significant at the 0.1% level. That is, the higher average performance of female students of 5.38 points is not due to random error (measured by the standard error/variance) but rather due to systematic difference within gender. Student gender is an important factor in explaining students' math performance. This finding in fact goes in opposite direction of the existing studies which identifies male students performing higher than female students in mathematics. One key reason for this is likely due to the different grades and ages assessed (e.g. NAEP assesses 8th grade and PISA assesses 15 year olds). This point can be further explored in the future. Looking now at the students' ethnic background illustrates the following result.

[113] Median of 407 and 410; Mode of 430 and 435; range of (219, 550) and (245, 550) for male and female, respectively.

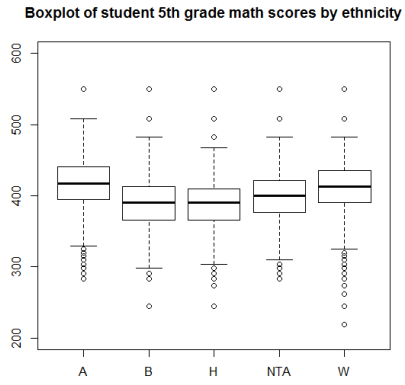
[114] ANOVA and dummy variable regression provide the same results. For this reason, the two methods are subsumed as linear models (LM). Yet, regression is slightly more informative as it gives the parameter estimate which is the differences in the conditional mean values of the different treatments.

Figure 6.11: Histogram and pie chart of student ethnicity



The most frequent ethic background among the 5th grade students is the White students who comprise of 65.8% of the total students. This is followed by Hispanic, Asian, Black and Native American students. Looking at their average math performance shows the following result.

Figure 6.12: Boxplot of student 5th grade math scores by ethnicity



As indicated by the solid black line in the middle, the average performance of White ethnic background students with 413.9 exceeds all the other ethnic groups (Native American students with 399.5; Black American students with 389.5; and Hispanic American students with 389.4) except for Asian students with 419.5. Despite the large differences in the number of students, the spread of each distribution shows similar patterns with standard deviation of 36.4, 36.3, 38.8, 34.8, and 34.1 for White, Native American, Asian, Black and Hispanic students, respectively. This similarity is also indicated with the similar boxplots (which indicate the middle 50% of the distribution (25th to 75th percentile) as shown above. To test whether the differences in the mean scores are statistically significant, we again conduct a simple one-way ANOVA and simple linear regression as shown below.

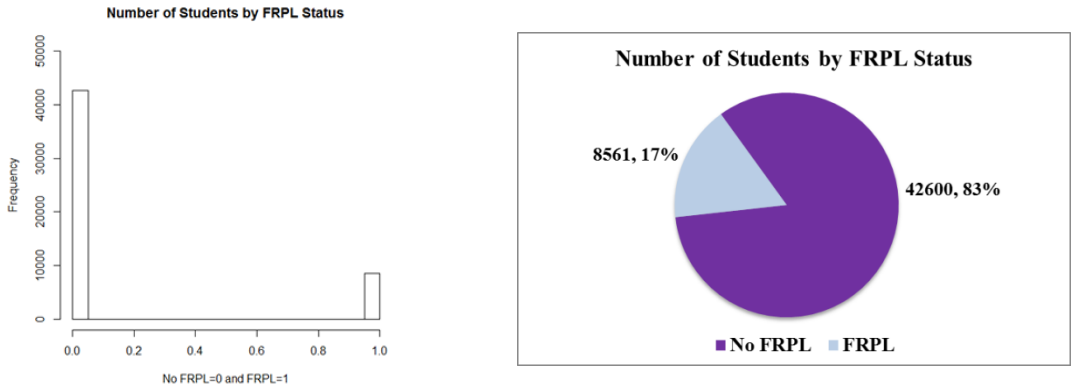
Table 6.4: Estimates of one-way ANOVA and simple linear regression with student ethnicity

	Df	*SS*	*MS*	*F value*	*p value*	
factor(Ethnicity)	4	5633302	1408325	1074	0.00	***
Residuals	51156	67078372	1311			
	Estimate	*SE*	*t value*		*p value*	
(Intercept)	399.58	0.735	543.5		0.00	***
Student is Asian	19.90	0.923	21.55		0.00	***
Student is Black	-10.09	1.002	-10.08		0.00	***
Student is Hispanic	-10.19	0.839	-12.14		0.00	***
Student is White	14.32	0.761	18.82		0.00	***
R square	0.077					
Adj R square	0.077					

Note: *** significant at 0.1%; ** at 1%; * at 5%
Note: Base group is Native American students

The finding above indicates that the differences in the means values are statistically significant at the 0.1% level. That is, the high performances of White and Asian students are not due to random error but due to systematic differences inherent between the ethnic groups. Together with student gender, student ethnicity is an important factor in explaining students' math performance. This finding is consistent with the existing studies which also identified significant ethnicity gap in student performance within the U.S. and abroad.[115] Proceeding now to the students' free-or-reduced priced lunch (FRPL) status shows the following result.

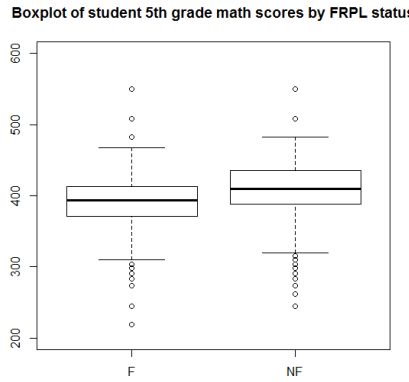
Figure 6.13: Histogram and pie chart of student FRPL status



The total number of 5th grade students who do not receive FRPL exceeds the total students who receive FRPL by a large margin with 83.3% in comparison to 16.7% of the total students. Looking at their average math performance shows the following result.

[115] These findings are evident in the NAEP study for the US finding and the PISA study for the international findings.

Figure 6.14: Boxplot of student 5th grade math scores by FRPL status



The students who do not receive FRPL have a higher average performance of 411.7 than the students who do receive FRPL with 392.5 points. Despite the large differences in the sample sizes, both groups have a relatively similar variation with standard deviation of 37.48 and 34.54 for without FRPL and with FRPL students. To test whether the difference in the mean scores are statistically significant, we again conduct a simple one-way ANOVA and simple linear regression as shown below.

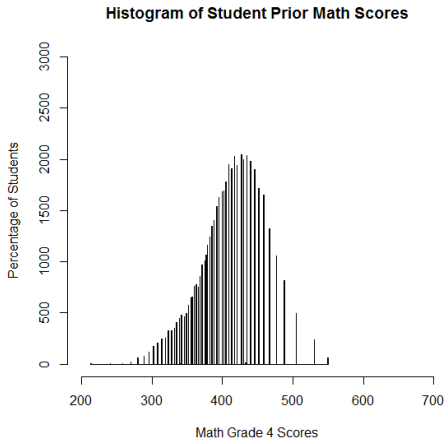
Table 6.5: Estimates of one-way ANOVA and simple linear regression with student FRPL status

	Df	*SS*	*MS*	*F value*	*p value*	
factor(FRPL)	1	2644994	2644994	1931	0.00	***
Residuals	51159	70066680	1370			
	Estimate	*SE*	*t value*		*p value*	
(Intercept)	411.71	0.179	2296.2		0.00	***
Student with FRPL	-19.26	0.438	-43.95		0.00	***
R square	0.036					
Adj R square	0.036					
Note: *** significant at 0.1%; ** at 1%; * at 5%						

The finding above indicates that the difference between the two groups is statistically significant. That is, their higher performance of students without FRPL is not due to some random error but rather due to systematic difference between the two groups. FRPL status is also an important factor in explaining students' math performance. Comparing the coefficient estimates with the previous findings also shows that the difference in mean scores of -19.26 is almost 5 times the magnitude of gender with 5.38. FRPL status is often used as a proxy of students' socio-economic background (SES) as students who officially approved as in need of economic assistance are eligible. This finding supports the extensive literature on the effects of SES on student achievement. Both NAEP and PISA have found SES as the variable with the largest significant effect. These studies also illustrate how other explanatory variables (defined at the student, school and country level variable) lose its effect and significance once SES was taken into account in the model. Finally, looking at the

students' prior test scores (4th grade) or the initial academic knowledge and challenge students bring to the teachers show the following results.

Figure 6.15: Histogram of 4th grade (prior) math scores



The prior year's test scores (4th grade WASL) illustrate the same features to the current test score (5th grade WASL) described earlier. First, there is a clear indication of the symmetric bell shaped normal distribution pattern and second, the broken distribution towards the upper end which signifies the criterion based test property of the WASL. The mean is calculated as 406.1 points with a median of 405, mode of 427 and standard deviation of 41.5. The symmetric distribution is confirmed with skewness value of -0.008 which is virtually equal to the perfect symmetry of zero.[116] The qqplot plot shown below provides evidence in support of normal distribution as majority of the observations lie on the 45 degree line. The kurtosis value is 3.201 which again is very close to the normal distribution with 3.

Figure 6.16: Quantile plot of 4th grade math scores

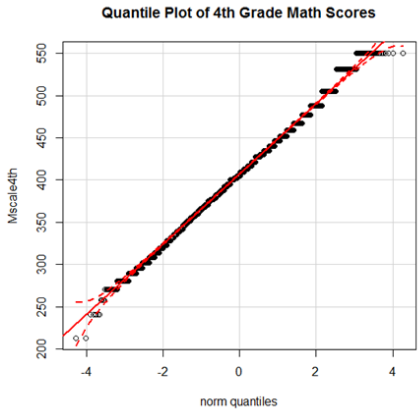
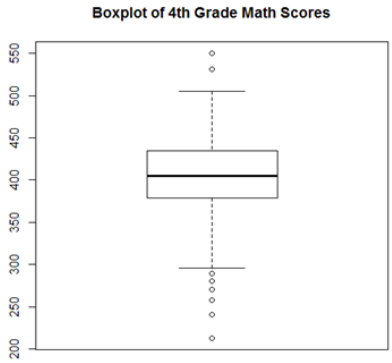


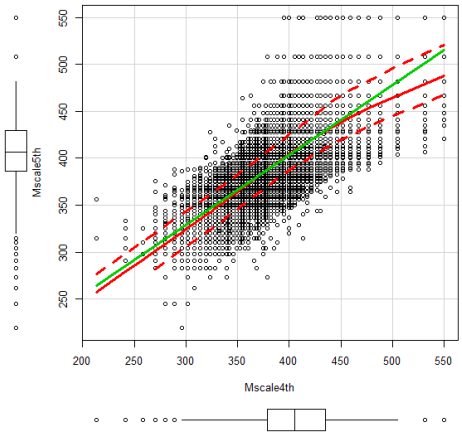
Figure 6.17: Boxplot of 4th grade math scores



To examine the association between the initial and current test scores, the following scatterplot is constructed.[117]

[116] Symmetry is also indicated with the closeness of the three central tendency indices and the equal spacing of the quartile values of 379 and 435.
[117] As the prior test score is a continuous variable (and not a binary variable as in the previous cases), we cannot calculate the conditional mean for each value of the prior test scores. The corresponding indicator for this purpose is the correlation coefficient as shown in the text below.

Figure 6.18: Scatterplot of 4th grade and 5th grade math scores



As you can see, a clear positive association between the two test scores is illustrated. The green line indicates the straight line approximated by the simple linear regression (estimated below). The red curve represents the non-parametric lowess curve (with the dotted lines indicating the 95% variation around the line) which is a more advanced estimation method of the regression function as described in Chapter 2. As you can see both curves illustrate a strong (almost perfect) positive association between the two variables. The non-parametric curve also hints a possible nonlinear relation as shown with the curve towards the upper end of the figure. To provide numerical summary of this association, the correlation coefficient is calculated to confirm the strong positive association with a value of 0.819. And finally to test whether this association is statistically significant, the simple linear regression model is estimated to show the following result.

Table 6.6: Estimates of simple linear regression with student prior test scores

	Estimate	*SE*	*t value*	*p value*	
(Intercept)	105.80	0.941	112.5	0.00	***
Mscale4th	0.745	0.002	323.4	0.00	***
R square	0.672				
Adj R square	0.672				
Note: *** significant at 0.1%; ** at 1%; * at 5%					

The regression coefficient of 0.745 is statistically and significantly different from zero as indicated with a very high t-value of 112.5 and with very small p-value of 0.00. That is, the highly positive relation with the two test scores is not due to some random error but rather due to systematic and structural effect of the prior test score on the current test score. Prior test score is an important factor in explaining students' current math performance. Students who initially perform at a high level are likely to continue to perform at a high level. On the other hand, students who initially perform at a low level (bringing initial academic challenges and difficulty to the 5th grade teachers) are likely to continue to perform at a low level.

To sum, the descriptive findings in the above sections provided us with the first confirmation of our initial expected theories underlying the variables and the first empirical justification of the VAMs. The outcome variable of interest – the 5th grade test scores – illustrated clear and significant variation among the students. And the majority of the explanatory variables used to explain this variation illustrated statistically significant associations in line to our theories and previous findings. Specifically, the main variable of interest – the teachers – illustrated significant effects/differences in the mean outcome of the student performance. The student background variables including their gender, ethnicity, FRPL status, and prior test scores also illustrated clear and significant effects on the student test scores. Among them, the FRPL status and prior test scores illustrated the most profound effects. But the next important question is to know whether these significant effects especially the teacher effects are purely due to the teachers themselves or to some other extraneous factors outside of their control. That is, we must investigate whether the teacher effects are accurate and unbiased. To explore this challenge, we must advance the previous descriptive analysis by examining whether the student background variables are distributed unevenly across the teachers and whether the interconnections of these variables with the teachers induce differences in student performance. These two conditions will provide empirical evidence of the presence of confounding effects defined in Chapter 6.[118] And this analysis will provide further evidence in support of the VAMs which will take into account these factors outside the control of teachers to provide the reliable unique effects of teachers. To explore these conditions, the following multivariate analysis is conducted.

Multivariate Descriptive Analysis of Teachers, Student Background Characteristics, and the Outcome Variable

The descriptive summaries of the distribution of the student background characteristics across the 5th grade math teachers (which will also be referred to as the student peer effects) show the following results.

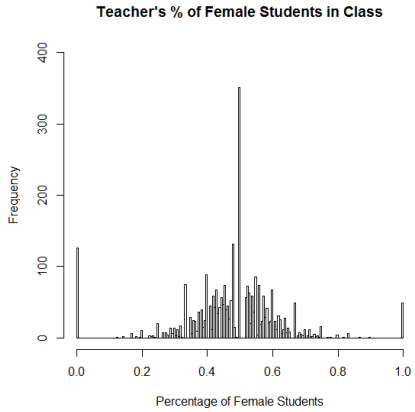
[118] But as addressed in Chapter 6, depending on the theoretical understanding of the variable, the interconnections between the variables can also be perceived as "intervening effects". This point will be further explained in next chapter.

Table 6.7: Descriptive summary of types of students in teachers' classes

	Mean	*SD*	*Median*	*Mode*	*Range (mini, max)*
% of Students on FRPL	0.174	0.176	0.133	0	(0,1)
% of White Ethnicity	0.650	0.271	0.714	1	(0,1)
% of Female Students	0.476	0.157	0.500	0.5	(0,1)
Average Initial Performance	402.4	28.2	403.6	402	(270 , 519)
Total Number of Teachers	2864				

The above table indicates very similar trend to the overall descriptive statistics defined at the student level shown earlier. Yet, looking at the entire distribution of these variables across the teachers reveals interesting pattern. Starting with the gender balance (percent of female students) across the teachers shows the following results.

Figure 6.19: Histogram of percentage of female students in teachers' class



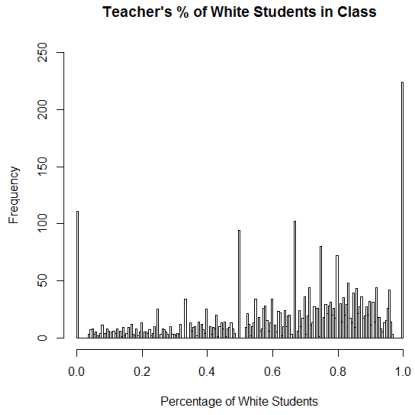
The above figure illustrates that the teacher mean of 47.6% female students in the class is driven by the high frequency of teacher with gender balanced class (50% female and male students). These teachers are the mode and also the median. But as you can see from the figure, there is clear variation with a standard deviation of 15.7 percentage points and relatively large spikes at the minimum and maximum values of 0% and the 100%. The former teachers only have male students and the latter teachers only have female students in their class. But as it was shown earlier that female students perform significantly higher than male students, the above finding entails that if we do not take into account of this variation in gender balance, we would misattribute the effects of gender to the teachers. We would overestimate the effects of teachers with high concentration of female students (and vice versa). This confounding effect can be roughly approximated through a three variable multivariate descriptive analysis between teachers, students' gender, and 5th grade math performance. But as there are close to 3,000 teachers, we must reduce the number by categorizing them into smaller groups based on the gender balance in the class. The mean test scores for these groups can then calculated as shown below.

Table 6.8: Teacher mean scores by percentage of female students in class

	Mean	*SD*	*N*
Less than 25%	370.94	40.79	177
25 to 50%	406.86	22.06	1590
50 to 75%	409.85	20.03	1033
More than 75%	382.27	46.96	64
Less than 50%	403.26	26.85	1767
More than 50%	408.24	23.41	1097

As you can see, teachers with high mean performance also have high concentration of female students.[119] But given the positive female effect on student performance, the high mean performance of the teachers is mixed between the teachers themselves and the high concentration of female students. Unless we separate the effects of students' gender from the teachers, we cannot capture the teachers' unique contribution on their student performance. Moreover, from the descriptive analysis alone, we cannot determine whether the mean differences are statistically significant. To achieve these tasks, we must resort to the VAMs which are based on the multiple linear regression framework. Before we do so, we first proceed with the descriptive analysis of other student background variables. Looking at the distribution of student ethnicity across the teachers shows the following results.[120]

Figure 6.20: Histogram of percentage of White students in teachers' class



As you can see, the distribution of student ethnicity across teachers indicates a highly deviant and un-symmetric pattern. The mean is 64.9% White students but the standard deviation is 27.1 percentage points (which is much bigger than the gender distribution). Moreover, the highest frequencies are at the maximum value followed by the minimum values. That is, a lot of teachers teach classes comprise of only White

[119] The mean score of the more than 75% is slightly lower but this is most likely due to the very small sample size of 64 students. The mean value is not very reliable. Aggregating the percentages to less or more than 50% shows provides more reliable estimates that confirm our expectation.

[120] For better illustration, the variable is transformed into a binary (white and non-white) variable. In the VA analysis, the original coding with the four categories will be used.

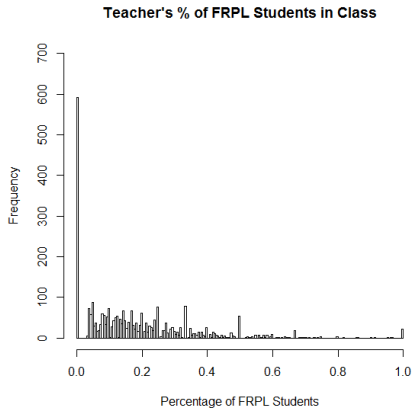
students and only non-White students. This indicates a sign of polarization of the of student ethnic background balance across the teachers. More importantly, as it was shown earlier that White ethnic background students perform significantly higher than other students (except for Asian students), if we do not take into account this variation in student ethnicity, we would misattribute its effect to the teachers. We would overestimate the effects of teachers with high concentration of White students (and vice versa). This can be roughly approximated as follow.

Table 6.9: Teacher mean scores by percentage of White students in class

	Mean	*SD*	*N*
Less than 25%	384.12	27.05	339
25 to 50%	397.98	25.76	432
50 to 75%	409.14	22.70	849
More than 75%	410.69	23.65	1244

As you can see, teachers with high mean performance also have high concentration of White ethnic background students. But given its positive effect on student performance, the high mean performance of the teachers is potentially mixed between the teachers themselves and the high concentration of White students. Unless we separate and remove the effects of students' ethnic background from the teachers, we cannot capture the reliable unique estimate of their contribution on student performance. Together with the previous findings of student gender, this provides further empirical evidence to conduct VA analyses. Looking now at the distribution of the percentage of students receiving FRPL across the teachers shows the following result.

Figure 6.21: Histogram of percentage of FRPL students in teachers' class



The mean percentage of students receiving FRPL in teachers' classes is 17.4%. But looking at the entire distribution illustrates a widely variant distribution around the mean. The standard deviation is 17.6 percentage points which is in fact larger than the mean value itself. There is a general downward sloping trend in the frequencies where the majority of the teachers have very low percentage of students on FRPL. But there is also a huge spike of frequency at the 0% value which surpasses the rest by a large margin. Approximately 20% of

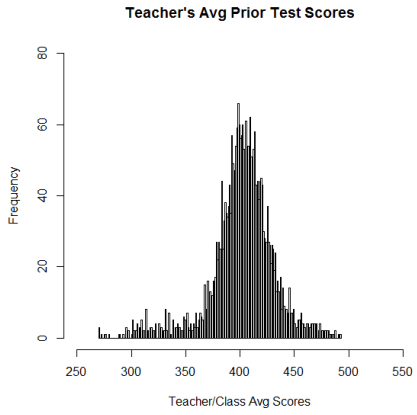
the teachers teach a class where none of the students receive FRPL. Now, as it was shown earlier that FRPL status has a profound and significant effect on student performance, if we do not take into account this variation in FRPL status, we would misattribute its effect to the teachers. We would underestimate the effects of teachers with high concentration of FRPL students (and vice versa). This confounding effect can be roughly approximated as follow.

Table 6.10: Teacher mean scores by percentage of FRPL students in class

	Mean	*SD*	*N*
0% on FRPL	410.40	37.34	592
Less than 25%	409.68	24.80	2182
25 to 50%	392.92	20.83	555
50 to 75%	383.24	27.06	92
More than 75%	375.71	33.46	35

As you can see, teachers with low mean performance also have high concentration of FRPL students. But given its negative effect on student performance, the high mean performance of these teachers is mixed between the teachers themselves and the high concentration of FRPL. Again, unless we separate the effects of students' FRPL status, we cannot capture the reliable estimate of teachers' unique contribution on student performance. Finally, looking at the distribution of the students' prior test shows the following result.

Figure 6.22: Histogram of teacher average 4th grade math scores



The mean prior test scores across the teachers portray a symmetric distribution around the mean of 402.4 points. Symmetry of the distribution is confirmed with the skewness value of -0.84253.[121] Looking at the spread of the distribution, a clear variation is indicated with standard deviation of 28.2 points and a range of 270.0 and 492.8 points. That is, some teachers are exposed to high concentration of students with low initial achievement level (as shown on towards the lower end of the distribution) while some are exposed to high concentration of students with higher initial achievement level (as shown towards the upper end of the

[121] The mode is at 402.0 points and the median is at 403.6 points and the closeness of these summaries with the mean supports the symmetry of the distribution

distribution). There is clear imbalance in the initial achievement and academic challenges students bring to the 5th grade teachers. Now as it was shown earlier that prior test scores have a profound and highly significant effect on current test scores, if we do not take into account this variation in prior test scores, we would misattribute its effect on the teachers. We would overestimate the effects of teachers with high concentration of high achieving students (and vice versa). These confounding effects can again be approximated as follow.

Table 6.11: Teacher mean scores by (class/teacher) average prior test scores

	Mean	*SD*	*N*
Less than 350	334.88	24.09	137
350 to 402.4	393.50	13.18	1220
401.2 to 450	418.51	12.70	1404
More than 450	454.95	15.74	103
Note: 402.4 is the mean prior test score			

As you can see, teachers with high mean performance are surrounded with students with high initial performance. But given the positive effect of initial achievement on current student performance, the high mean performance of these teachers are again mixed between the teachers themselves and the effects of the initial achievement. That is, unless we control for initial achievement, we cannot capture the reliable and precise estimate of teachers' unique contribution on student performance.

To sum, the findings in this section provided further empirical evidence and justification of the VAMs. The analysis illustrated that the student background variables which have significant effects on the student performance are distributed unevenly across the teachers. Calculating the mean performances of the students for the different groups of teachers (based on the student background variables) illustrated that high performing teachers were also surrounded with high achieving students – female, White ethnic background, without FRPL, and with high initial achievement (and vice versa). That is, data illustrated that the teacher effects is mixed and confounded with the effects of these student background variables. And without taking into account of these variables, we are unable to achieve the reliable and precise estimate of the teacher effects. This is precisely the raison de entre of the VAMs defined in the first part of this study. Now, as shown in the Appendix, extending the above analysis for groups of teachers defined by different combinations (not just one) of student background variables illustrates further complexity in the web of confounding effects. This finding further necessitates the VAMs which can simultaneously take into account multiple factors in a single model in order to unravel the complex nexus of associations. The models will not only provide a more reliable estimate of the unique contribution of each variable on the outcome but also the statistical significance of each estimate. Before we proceed to the VA analyses in the next chapter, the descriptive analysis of other potential control variables – the teacher characteristics and schools fixed effects, are provided next.

Other Explanatory Variables – Teacher Characteristics and School Variable

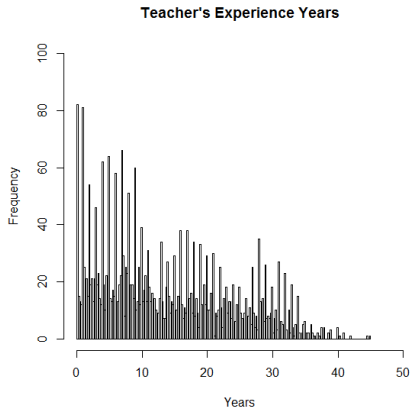
The above section explored the student background variables at the factors outside the control of teachers. But as it was illustrated in Chapter 5 using the table by Harris (2011), teacher background variables such as gender, ethnicity, and experience level are also outside of teachers' control and should be taken into account as confounding variables. The teacher education variable, on the other hand, is under the control and discretion of the teachers and should be considered as an intervening variable. The descriptive statistics of these variables for the 5th grade math teachers in the Washington data are therefore provided as follow. The summary of other teacher level variables i.e. student background variables aggregated at the teacher level (peer effects) was illustrated in the previous section.

Table 6.12: Descriptive summary of teacher characteristics variables

	Mean	*SD*	*Median*	*Mode*	*Count*	*Range (mini, max)*
Teacher is Female	0.750	0.433	1	1	male =716, female=2148	(0,1)
Teacher is White Ethnicity	0.935	0.246	1	1	non-W=185, White=2679	(0,1)
Teacher with MA Degree	0.663	0.473	1	1	no MA= 964, MA=1900	(0,1)
Teacher Experience	13.6	9.89	11.2	1	-	(0, 45)
Total Number of Teachers	2864					
Total Number of Schools	936					

As you can see, of the total of 2,864 5th grade math teachers, the majority of them are female (75%) with White ethnic background (93.5%) and slightly over half of the teachers have a MA degree (66.3%). The mean years of experience is 13.6 years but there is fair amount of variation with standard deviation of 9.89 years and range from less than 1 year to 45 years. As you can see from the histogram below, there is in fact a clear positive skew in the distribution with higher concentration of teachers with less years of experience. The mode is in fact 1 year of teaching experience.

Figure 6.23: Histogram of teacher experience



Now, to examine the association of these variables with student performance, the conditional mean values of student test score and the correlation coefficient (for the continuous experience variable) is calculated.

Table 6.13: Teacher mean scores by teacher characteristics

Teacher Mean Scores by Teacher Characteristics						
	Mean	*SD*	*Median*	*Mode*	*Range*	*N*
Teacher is Male	405.3	22.83	406.6	388	(283, 475)	716
Teacher is Female	405.1	26.60	406.7	413	(273, 493)	2148
Teacher is Non-White	397.2	25.73	400.2	330	(304, 461)	185
Teacher is White	405.7	25.62	407.3	413	(273, 493)	2679
Teacher Without MA	403.7	26.68	405.2	390	(298, 489)	964
Teacher With MA	405.9	25.17	407.4	413	(273, 493)	1900
Teacher Experience	0.111*					

Note: * this value is the correlation coefficient between teacher experience and teacher means scores

There seem to be no difference in the average performance between male and female teachers. Yet, White ethnic background teachers and possession of MA degree show higher mean performance. Teacher experience is also found to have positive correlation of 0.111 with student performance. To check whether these mean score differences (and correlation) are statistically significant, we conduct simple ANOVA and linear regression as shown below.

Table 6.14: Estimates of one-way ANOVA and simple linear regression with teacher characteristics variables

Teacher Gender:						
	Df	*SS*	*MS*	*F value*	*p value*	
factor(teacherGender)	1	8	8.4	0.013	0.91	
Residuals	2862	1891853	661			
	Estimate	*SE*	*t value*		*p value*	
(Intercept)	405.26	0.961	421.78		0.00	***
Teacher is Female	-0.125	1.110	-0.113		0.91	
R squared	0.0000					
Adj R Squared	-0.0003					
Teacher Ethnicity:						
	Df	*SS*	*MS*	*F value*	*p value*	
factor(teacherEthnicity)	1	12488	12488	19.02	0.00	***
Residuals	2862	1879373	657			
	Estimate	*SE*	*t value*		*p value*	
(Intercept)	397.22	1.884	210.838		0.00	***
Teacher is White Ethn	8.495	1.948	4.361		0.00	***
R squared	0.0066					
Adj R Squared	0.0063					
Note: *** significant at 0.1%; ** at 1%; * at 5%						

Teacher Higher Degree:						
	Df	*SS*	*MS*	*F value*	*p value*	
factor(teacherMA)	1	2998	2998	4.542	0.033	*
Residuals	2862	1888863	660			
	Estimate	*SE*	*t value*		*p value*	
(Intercept)	403.73	0.827	487.94		0.00	***
Teacher with MA Degree	2.165	1.016	2.131		0.033	*
R squared	0.0016					
Adj R Squared	0.0012					
Teacher Experience:						
	Estimate	*SE*	*t value*		*p value*	
(Intercept)	401.27	0.811	495.03		0.00	***
Teacher Experience	0.288	0.048	5.959		0.00	***
R squared	0.0123					
Adj R Squared	0.0119					
Note: *** significant at 0.1%; ** at 1%; * at 5%						

The above analyses indicates that the difference in the mean scores between the White and non-White teachers is statistically significant at the 0.1% level. The association of teacher scores and teacher experience is also strongly significant. The difference in the mean scores between teacher with and without MA degree is also significant but only at the 5% level. And finally, the difference in mean scores between male and female teachers is not significant. In the context of VAMs, as many of these features are outside the control of

teachers, if we do not take these factors into account, we can again misattribute its effects on the teachers. The above findings therefore provide empirical justification for the VAMs to take into account these variables simultaneously and to separate its effects from the teacher variable. The nexus of association inherent between these teacher characteristics variables and the rough approximation of the confounding effects are provided in the Appendix.

School Fixed Effects

In addition to the teacher characteristics, the schools and its characteristics (e.g. the school working environment, facility, resources, etc.) can have significant effect on student performance. And as these schools and its features are often outside the control of teachers, it becomes a viable confounding variable which we would need to take into account to capture the unique teacher effects.[122] The WA data unfortunately do not comprise the variables characterizing the schools. It only provides the index which identifies the school in which the 5th grade students and 5th grade math teacher belong. Analogous to teacher fixed effects model, the school index variable is a binary dummy variable with values 0 or 1. Based on this index variable, there are a total of 936 schools in which the 5th grade students and 5th grade math teachers belong. The distribution of the number of 5th grade students and 5th grade math teachers assigned to these schools show the following results.

Figure 6.24: Histogram of school student size (number of 5th grade students)

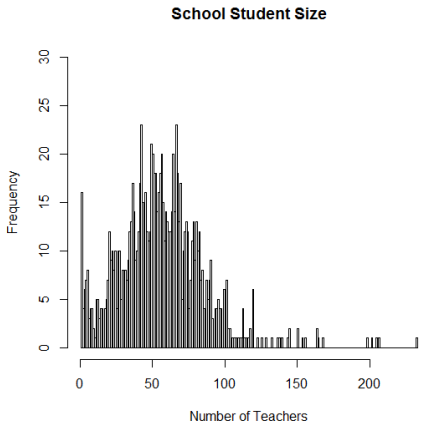
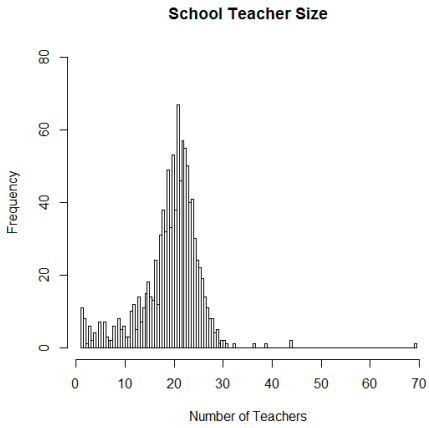


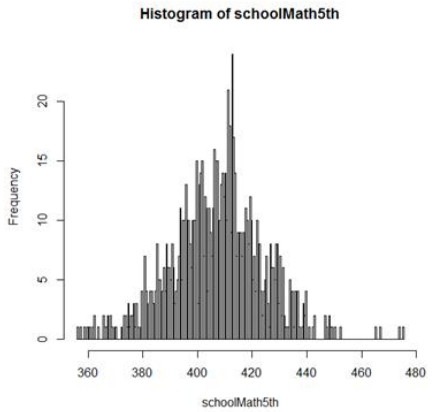
Figure 6.25: Histogram of school teacher size (number of 5th grade math teachers)



The mean number of 5th grade students in these schools is 54.6 with a median of 53 and mode of 43 students. There is a relatively large variation around the mean with a standard deviation of 29.7 and range of 1 to 233. And for the number of 5th grade math teachers, the mean is 19.4 teachers and standard deviation of 5.88 and range of 1 to 69. The correlation of the two indices indicates strong association of 0.60 which naturally reflects how schools with more students have more teachers. Looking now at the student average performance across these schools illustrate the following results.

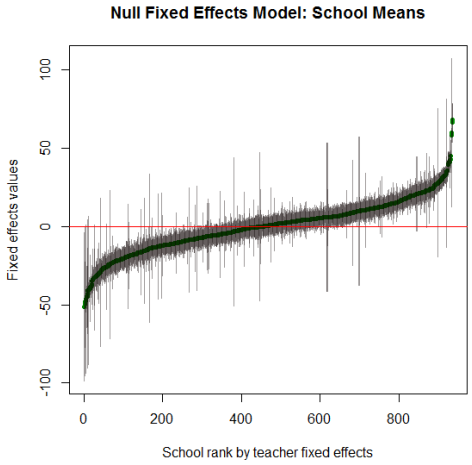
[122] For example in most circumstances, teachers are selected and hired by the school or stationed under the discretion of the government. Teacher often do not have the control over this hiring process. The school environment e.g. types of students attending, school financing, type of teachers working, etc. are also determined by factors beyond their control.

Figure 6.26: Histogram of school average 5th grade math scores



The schools have a mean student performance of 407.1 with a median of 407.6 and a mode of 413 points. The closeness of these values confirms the symmetric distribution illustrated above. But clear deviation around the mean is indicated with standard deviation 16.4 and a range of 356.0 and 475.3 points. That is, some teachers (situated toward the upper end of the scale) work in high performing schools while some teachers (lower end of the scale) work in low performing schools. To test whether the school mean differences are statistically significant, we again construct the caterpillar plot which illustrates the findings of the simple one-way ANOVA tests as shown below.[123]

Figure 6.27: Null fixed effects model: school means



As you can see from the schools toward the upper end of the spectrum, these schools have mean scores which are significantly greater than the overall mean 407.1 as the 95% confidence interval bands do not cross the 0 value (overall mean).[124] On the other hand, as you can see from the schools toward the lower end of the spectrum, these schools have mean scores which are significantly less than the overall mean. Thus, the differences in mean scores among the total 936 schools are due to systemic differences between the schools

[123] Linear regression model with school dummy fixed effects could have also been used for this purpose.
[124] The average deviation/mean score is 2.423103e-14 with standard deviation of 16.44763 and the average 95% confidence interval is 4.264529 with standard deviation of 3.505544.

and not due to random error. That is, school is an important factor in explaining student performance. Students' performance significantly matter depending on the school the student attends.

Now, as teachers do not have direct control over the school and its characteristics, unless the school variable is taken into account, its effects can be misattributed to the teachers. We must therefore remove and separate its effects from the teachers in order to capture their unique contribution on student performance. But unlike the student peer effects variables, we cannot categorize the teachers into few groups (e.g. teachers with 0 to 25% of their students on FRPL) based on the school variable which is a binary variable. This makes it difficult to obtain the rough approximation of the confounding effects shown with the other variables. If we still like to investigate the confounding effects we must calculate the teacher means scores across the entire 934 schools. But this is uneconomical and very tedious. Moreover, as we would also like to investigate further avenues of confounding effects through triangulating the school variables with other variables, this will double and triple the calculation of 934 mean scores. This is the fundamental limitation of descriptive analysis. This limitation further necessitates the VAMs which are based on the multiple regression models that can incorporate all these variables simultaneously in one model to estimate the conditional mean values for all variables including each of the 2,864 teachers. Finally, there is an important caveat in utilizing school fixed effects. School fixed effects are widely favored in the econometric literature as it said to control for all the time invariant unobservable and un-measurable factors defined at the school level that affect student performance. It is said to move the estimates one step closer to causal estimates. But in plain language, we must keep in mind that the 0 or 1 binary indicator aggregates all possible school features that are time invariant. This can comprise of number of different factors: whether the schools' success is due to their working environment, atmosphere, climate factors, type of school (charter, private, public), school budget and finance amount, etc. The relative contribution of these features is not going to be known. School fixed effects provide a crude, abrupt, rough aggregated measure of the effects schools have on student and teacher performance. It is very vague when it comes to interpreting the results. A possible data collection of the school specific features would improve the situation.

Summary and Conclusion:

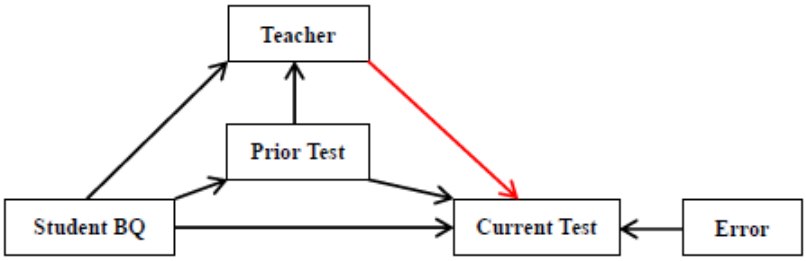
This chapter conducted the descriptive analysis of the Washington State data which will be used for the VA analyses. The findings provided us with the first hand confirmation and empirical justification of the VAMs. Statistically significant variation and differences between the teachers mean scores (without any adjustment with the explanatory variables) were found. This finding provided us with the first indication of the presence of significant teacher effects. The explanatory variables outside the control of teachers were then analyzed to depict significant associations with the teachers. The conditional mean values for difference combination of teachers and these variables also illustrated clear differences. These findings provided us with the empirical evidence of the presence of confounding (or intervening) effects underlying the teacher means scores.

Although the descriptive analysis provided us with the first hand empirical evidence of VAMs, it fell short of completing the main objective of providing reliable and accurate estimates for each and every teacher. The fundamental limitation of descriptive analysis is that it is only capable of conducting the analysis for few variables at a time. As the number of variables increased, it became harder to track and interpret the different interconnections underlying the variables. It was impossible to graphically visualize the relation when more than 3 or 4 variables were involved. Furthermore, only a rough approximation of the confounding effects was illustrated. In order to indicate the confounding effects, we had to reduce the number of teachers by classifying them into four groups based on the values of student characteristics in their class. This grouping process significantly reduced the amount of important variation inherent between the teachers (from 2864 teachers to 4 groups). But this process contradicts the fundamental purpose of this study which is to provide the reliable and accurate estimate of each of the 2864 teachers' contribution on student performance. And finally, the descriptive analysis failed to provide the statistical significance of these teacher effects after taking into account some of the confounding effects. These challenges all necessitate the need for more complex and advanced statistical tools namely the VAMs. As described repeatedly, VAMs will simultaneously take into account as many control variables to provide the statistically reliable and accurate estimate of the unique contribution of each teacher on their student performance. Before we proceed to the VA analyses, final note regarding the descriptive analysis is that as the multiple regression models captures the conditional means of the outcome variable after taking into account multiple variables, it is in essence a natural extension of the descriptive analysis of the conditional means presented above. Descriptive findings are therefore expected to be reflected in the regression estimates to provide further validation of the models. Moreover, when data anomalies or irregularities are found, it will provide important feedbacks in order to revise the estimated models. The descriptive findings will continue to have important roles as we proceed with the VA analyses.

CHAPTER 7: PRELIMINARY VALUE ADDED ANALYSIS

This chapter conducts the VA analysis of 5th grade math teachers in the State of Washington. The VA analysis extends the previous descriptive analysis by simultaneously taking into account all the multiple factors outside the control of teachers in order to better estimate the accurate and reliable unique contribution of teachers on student performance. But in addition to the empirical support illustrated in the previous chapter, this chapter also adopts the theoretical framework provided by Harris (2011) in Chapter 5 and the prior existing work on VAMs in order to construct the VAM specification. This process reflects on the importance of both data and theory in the model building (variable selection) process as described in Chapter 5. Specifically, two main groups of VAMs will be constructed and analyzed in this chapter – the basic VAM and the extended VAMs. Underlying these models is the Harris (2011) "cardinal rule of accountability", which states to take into account factors (as confounding variables) if they are outside the control of teachers. The basic VAM adopts this rule to take into account (only) the student background factors outside of the control of teachers. This relatively simple model was analyzed as one of the first and prominent VAMs during the early stage of VAMs research.[125] The graphical presentation of this model is illustrated below. The student background (BQ) variables comprise of students' gender, ethnicity, and FRPL status. And the teacher VA effects are represented with the red line.

Figure 7.1: Theoretical model diagram of basic VAM

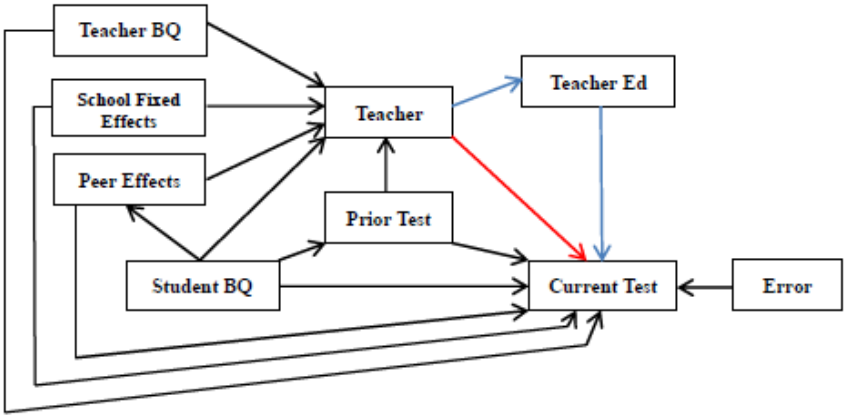


The extended VAMs extend the basic VAM by taking into account other variables outside the control of teachers defined primarily at the higher levels of hierarchy e.g. teacher and school levels. In the Washington data, these variables include the teacher variables (namely teachers' gender, ethnicity and experience), the school fixed effects variable, and the student peer effects (which are the teacher level average of the student background variables). The teachers' education level (whether or not they hold a master's degree) will be considered an intervening variable as the decision to obtain the higher degree is under the discretion (control) of the teachers. For the models in which this variable is controlled, attention will be provided to recognize the direct and indirect (via the teacher education variable) teacher effects. The direct teacher VA estimates in this model will be perceived as a slightly conservative measure of its effectiveness. The graphical presentation of

[125] These models are also referred to as the contextual VAMs. It lies at the foundation of other complex models such as the cross classified layered models. Please refer to McCaffrey et al. (2003) for more explanation. These models have been used by almost all the leading researchers in VAMs. Please refer to the following footnote 4 for further references.

the extended VAMs is illustrated below. The direct effect is represented with the red line while the indirect effect is represented with the blue lines.

Figure 7.2: Theoretical model diagram of basic VAM



This study acknowledges some of the caveats and debatable simplifications underlying these extensively used models. As notified in Chapter 5, the models treat the prior test score as a fixed, deterministic, and known variable and not as an endogenous variable with an unobservable random error term. An endogenous variable implies that it needs to be modeled and explained (just like the current test scores) and the random error term represents all the unknown factors (aside from factors included in the model) which can potentially affect the variable. If the prior test score is an endogenous variable (e.g. with the presence of an autoregressive error term), then this will induce bias in the model estimates.[126] This study acknowledges this possibility and the variety of technically advanced tools to resolve this problem (as described in the footnote) but it will be left as the future next step and area of further consideration.[127] Second, this study will focus primarily on the analysis and interpretation of the teacher VA parameter. The main concern is to identify all the confounding (and possibly intervening) effects of the teachers in order to achieve the most accurate and reliable VA estimates. The study might fall short to acknowledge other theoretically relevant mechanisms (direct and indirect effects) inherent in the other explanatory variables. But in light of these simplifications and unlike the majority of the existing VA research, this study will conduct rigorous assessment of the above models to make sure that these simplifications (and potentially others) will not consistently harm the model estimates. The analysis of the above models conducted in this chapter marks only as the initial and "preliminary" findings as thorough diagnosis and revision of these estimates in accordance to the linear regression framework will be provided in the following chapter. Only when the complete empirical evidence to ensure the accuracy and reliability of the

[126] This can be illustrated through simple substitution of the endogenous prior test score into the model. Given the autoregressive error terms, the prior test score will be correlated with the error term of the current test score (through the lagged error) and consequently violate the exogeneity condition.

[127] The key solutions for a endogenous variable such as the prior test score is to find an instrumental variable (exogenous shocks/variables) which is associated with the prior test score but not to the current test score. A policy change which affected the prior test score but not the current test score might be a viable option. If such variables are not available but if more than 3 years of data are available, then the dynamic panel analysis which utilized two year lagged test scores as the instrument can also be used (please refer to the work by Arello and Bond (2002) for further explanation). And finally, if the source of the variation in the prior test score can be known and be measured, it can always be controlled and conditioned this source as an explanatory variable. This option is often the simplest but requires the availability of such variables or further data collection.

estimates are provided, the model estimates will be considered for real life application. Keeping these points in mind, this chapter proceeds as follows.

In the first section below, the basic VAM will be analyzed using both fixed and random effects models. The extended VAMs with the teacher characteristics variables will then be analyzed. Other extended VAMs with school fixed effects, student peer effects (averages of the student level variables at the teacher level), and random/varying slope models for the prior test score and FRPL (given its profound significant effects)[128] then follow. The extended VAMs will be analyzed using random effects model as fixed effects model is subject to the multicollinearity problem described in Chapter 4. Finally, following the work of Tekwe et al. (2004), Hanushek and Rivkin (2008), McCaffrey et al. (2009), and many others, the VA estimates across all the models will be compared and contrasted to investigate its consistency and similarity under different model specifications.

Basic Value-Added Models

The estimates of the basic VAMs using fixed effects and random effects are provided below. Brief literature review of these two modeling options is provided in the footnote.[129]

[128] In these models, the effect (slope) of the FRPL and prior test score variables will be estimated for/within each and every teacher.

[129] The VA analyses conducted in this chapter reflect on number of previous leading researchers on VAMs. The use of fixed effects at different levels of hierarchy to take into account the unobservable time invariant factors builds on the extensive econometric work lead by the seminal work of Hanushek and Rivkin (2005) who pioneered in the usage of panel data analysis in education data. Their work is followed by others such as Cloetfleter et al. (2006, 2007), Rockoff (2004), Harris & Sass (2005, 2009), Sass (2006, 2008), and Goldhaber (2000, 2010). The use of random effects VAMs builds on the grand breaking work of Raundenbush and Bryk (2002) who pioneered the hierarchical linear modeling (HLM) in analyzing education data sets. HLMs are founded on random effects model and is utilized extensively among statisticians such as McCaffrey et al. (2003, 2004, 2005, 2007), Sanders et al. (1996, 2000, 2006), and Webster (2005). It is also the basis of number of advanced modern statistical models such as cross-classified models and Bayesian models which draws a lot of attention today.

Table 7.1: Estimates of basic VAM using fixed and random effects

Basic Fixed Effects VAM			
	Estimate	*SE*	*t value*
Intercept	387.86	6.793	57.10
Student Characteristics:			
Student is Female	2.358	0.184	12.83
Student is Native American (base)	-	-	-
Student is Asian	4.196	0.560	7.49
Student is Black	-1.436	0.610	-2.36
Student is Hispanic	0.419	0.523	0.80
Student is White	1.600	0.459	3.49
Student is on FRPL	-1.667	0.272	-6.13
Students' Prior Test Score*	0.708	0.003	271.81
Teacher Fixed Effects:			
Teacher ID1	15.48	7.827	1.98
Teacher ID2	23.70	7.826	3.03
Teacher ID3	27.03	8.575	3.15
Teacher ID4	25.04	8.161	3.07
Teacher ID5	16.85	8.382	2.01
Teacher ID6	9.30	8.047	1.16
… (continue to TID 2864)			
School Fixed Effects	No		
R Square	0.725		
Adjusted R Square	0.709		
AIC	456186		
BIC	481582		
N of Students	51161		
N of Teachers	2864		
Degrees of Freedom	48290		
Note: * Prior test scores are grand mean centered			

Basic Random Effects VAM			
	Estimate	*SE*	*t value*
Intercept	405.93	0.464	874.80
Student Characteristics:			
Student is Female	2.441	0.183	13.40
Student is Native American (base)	-	-	-
Student is Asian	4.472	0.546	8.20
Student is Black	-1.167	0.593	-2.00
Student is Hispanic	0.155	0.506	0.30
Student is White	1.836	0.449	4.10
Student is on FRPL	-1.795	0.266	-6.70
Students' Prior Test Score*	0.719	0.002	290.40
School Fixed Effects	No		
Variance Components:			
Between Teacher Variance	50.17		
Within Teacher Variance	414.93		
Intraclass Correlation	0.108		
AIC	456751		
BIC	456839		
Number of Students	51161		
Number of Teachers	2864		
Degrees of Freedom	51152		
Note: * Prior test scores are grand mean centered			

Looking first at the effects of student background variables for fixed effects results illustrate that female students on average outperform male students by 2.36 points. Similarly, students with FRPL on average perform 1.67 points less than students without FRPL; White ethnic background students on average outperform Native American students by 1.60, Black students by 3.04 and Hispanic students by 1.18 points.[130] Furthermore, one point increase in students' prior test score is associated with 0.708 points increase of the average current test scores. These estimates are all found to be statistically significant. The significance of the prior test scores is again evident. Looking now at random effects estimates shown on the right, virtually the same findings are illustrated for the student variables. The findings for these two models both conform to the results shown in the previous descriptive analysis. But unlike the descriptive findings, VA estimates simultaneously take into account all the mutual dependencies inherent between the variables and the 2864 teachers. That is, it takes into account more forms of confounding and intervening effects as illustrated in the theoretical model diagram of the basic VAM presented above. This process can be witnessed by comparing the VA estimates to the simple linear regression findings (which do not control for other variables) from the previous chapter. The results of the simple linear regression are re-summarized below.

[130] The actual adjusted mean values for each category of variable can be calculated by simply adding these estimate values to the overall adjusted mean (intercept).

Table 7.2: Summary of simple linear regression estimates with student background variables

Summary of Simple Linear Regression Estimate for Student Variables					
	Estimate	*SE*	*t value*	*p value*	
Studet is Female	5.377	0.333	16.16	0.00	***
Student is Native A	-	-	-	-	
Student is Asian	19.90	0.923	21.55	0.00	***
Student is Black	-10.09	1.002	-10.08	0.00	***
Student is Hispanic	-10.19	0.839	-12.14	0.00	***
Student is White	14.32	0.761	18.82	0.00	***
Student with FRPL	-19.26	0.438	-43.95	0.00	***
Prior Test Score	0.745	0.002	323.4	0.00	***
Note: *** significant at 0.1%; ** at 1%; * at 5%					

The magnitude of the parameter estimates and significance (*t* values) of all the student background variables have fallen tremendously. The gender effects halved and the ethnicity and FRPL effects have fallen to almost the one tenth of the simple regression estimate values. Looking at the theoretical model shown in Figure 7.1 again, these reductions are attributed primarily to the different indirect effects of these variables. The total effect (illustrated in the simple linear regression) is now diverged into the indirect effects of these variables through the prior test scores and the teachers and the remaining direct effects of these variables.[131] But for the prior test score, the reduction in its estimates is attributed to both the removal of the confounding effects (by the student background variables) and the intervening indirect effect through the teacher variable. Finally, the direction and magnitude of these changes all closely reflect the findings illustrated in previous chapter. No odd or inconsistent findings were identified.

Looking now at the estimates of the main variable of interest – the teachers VA estimates, fixed effects and random effects (empirical Bayes predictions) illustrate the following results.

[131] The indirect can be obtained by running a series of sequential regressions and comparing its estimates to the simple linear regression. This analysis (although not reported) has shown that once the prior test score is controlled, the three variables have lost most of its direct effect. That is, the variables have a tremendous indirect effect on the current test score through the prior test score.

Figure 7.3: VA estimates of basic fixed effects VAM

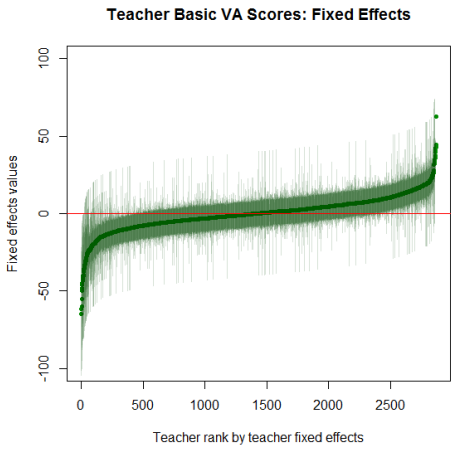
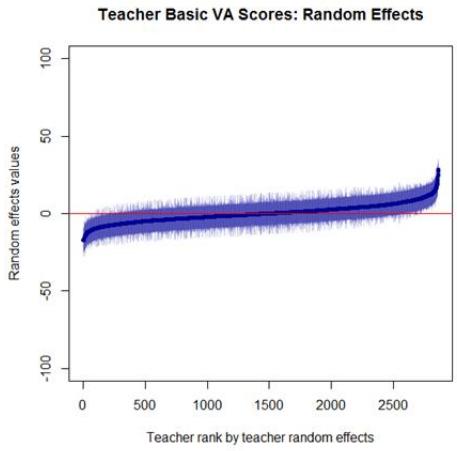


Figure 7.4: VA estimates of basic random effects VAM



These caterpillar plots portray the teacher VA effects as the dark circular dots in the middle of the object. These values are deviations of the adjusted teacher means from the adjusted overall mean.[132] The adjusted overall means for the fixed and random effects are 405.56 and 405.93, respectively.[133] Looking at fixed effects estimates in further detail, the mean value of fixed effects is 0.0051 with a spread of (standard deviation) of 10.61 as shown left hand histograms below. And the mean value of the standard error is 5.74 with standard deviation of 3.27 as shown on the right figure below.

Figure 7.5: Histogram of basic fixed effects VA estimates

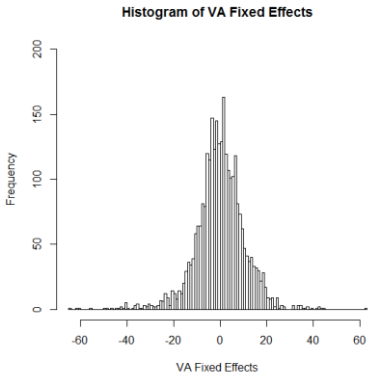
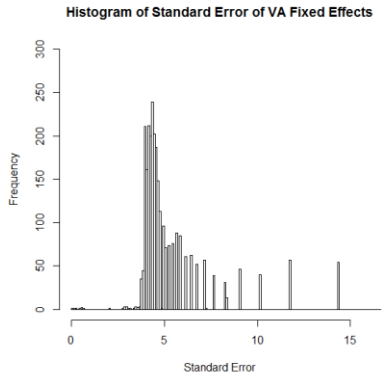


Figure 7.6: Histogram of standard error of basic fixed effects VA estimates



As we move beyond the means and standard deviations, it is also evident from caterpillar plot and histogram that there are many teachers whose VA estimates clearly deviate from the adjusted grand mean (of 0) with the 95% confidence interval bands not crossing the value zero. This implies that their VA estimates are statistically different from the adjusted grand mean. That is, the teachers (situated in the upper end of the scale) add

[132] The word "adjusted" implies that we have now taken into account the factors outside the control of teachers. It is the teacher means after the effects of other factors have been separated from the teachers. It is to distinguish from the teacher means illustrated in the descriptive analysis chapter.

[133] VA effects can be interpreted either in the deviation form or in terms of the adjusted teacher mean scores which can be obtained by simply adding the deviations to the overall adjusted mean (intercept). Both figures portray the same information and ranking of teacher effectiveness. The VA estimates values are also supplemented with the standard error or the 95% confidence interval ($\pm$1.96 times the SE estimates) as shown with the band (vertical lines). Again if the band crosses the value 0, then this gives evidence that the VA estimates are not significantly different from the overall adjusted mean.

significant value to their students' performance which is beyond the overall average performance. In fact there are quite a few teachers add almost 50 more points on the math test than the overall average. This can be interpreted as follow: if a random student is taught by these teachers, the student's performance on average would increase by 50 points irrespective of the student's background. On the other hand, as shown toward the lower end of the plot, there are quite a lot of teachers who significantly reduce the students' average performance by close to 50 points from the overall average. This can be interpreted as follows: if a random student is taught by these teachers, the student's performance on average would decrease by 50 points irrespective of the student's background. Both cases illustrate that teachers matter in determining students' academic performance. Looking now at random effects estimates (empirical Bayes predictions), it is clearly evident that the VA values are more moderate (smaller magnitude) and more stable (with smaller standard error bands) than fixed effects estimates. The entire distribution looks as if it has gravitated towards the value 0. This finding epitomizes the "shrinkages property" underlying the empirical Bayes estimates described in Chapter 4. This property is thoroughly examined in the Appendix. Looking now at the histograms below, the mean of random effects is 0.00 with a much smaller spread (standard deviation) of 5.67 in comparison to the 10.61 of fixed effects.

Figure 7.7: Histogram of basic random effects VA estimates

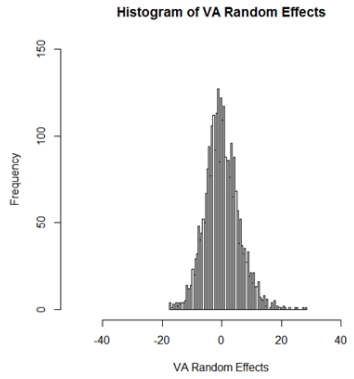
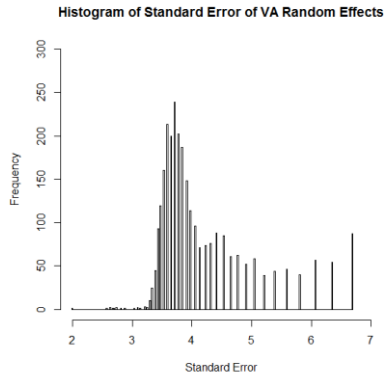


Figure 7.8: Histogram of standard error of basic random effects VA estimates



Looking at the standard error estimates on the right, it is evident that the magnitude of the error is also smaller than fixed effects value with a mean value of 4.165. The standard deviation of the errors is even much smaller (closer to four times) with 0.827 in comparison to the 3.274 of fixed effects.[134] In other words, although random effects estimates are gravitated towards 0, given the shorter confidence interval bands, there are still quite a lot of teachers whose VA estimate are significantly different from 0. As you can see from the teachers situated toward the upper end of the plot, these teachers add more than 20 more significant points on students' average performance than the overall average. On other hand, as shown on the other end of the plot, there are number of teachers who add 20 significantly less average points than the overall average. Again, if a random student is taught by the former group of teachers, the student's performance on average would increase by 20

[134] This difference is clearly illustrated the difference in the confidence interval bands of the caterpillar plots. Random effects model looks as if it trimmed the large confidence intervals of the fixed effects.

points irrespective of the student's background (and vice versa). Teachers are important determinant of students' academic performance.

The Confounding Effects Taken into Account

One of the primary objectives of VAMs (the second major task) is to take into account the factors outside the control of teachers which are confounding the teacher effects. We now investigate whether this objective was fulfilled by comparing the VA estimates to the unadjusted teacher mean values which do not taken into account the confounding effects. These models are referred to as the null models and the deviation of the (unadjusted) teacher mean values from the (unadjusted) grand mean for the fixed and random effects are presented below.

Figure 7.9: Null fixed effects model: teacher means

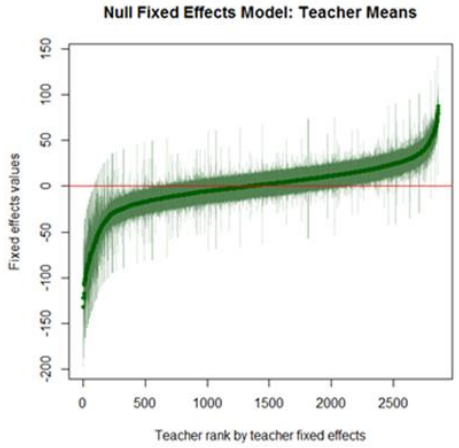
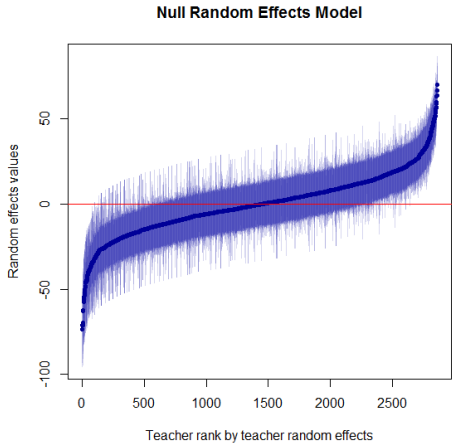


Figure 7.10: Null random effects model



Comparing these figures to the previous VA caterpillar plots (Figure 7.3 and 7.4), it is clearly evident that the overall magnitude of the VA estimates for both models are much smaller than the corresponding null model estimates. The mean values of the deviations from the unadjusted grand means are also very close to zero, but its standard deviations of 25.706 and 17.657 for the null fixed and null random effects models are close to 2.5 times and over 3 times larger than the respective basic VA estimates. The confidence interval bands are also longer than the ones for the VA estimates.[135] These findings all imply that through taking into account of the factors outside the control of teachers, the VA estimates have become more moderate in comparison to the null model estimates. The positive VA estimates are less positive in comparison to the corresponding (unadjusted) teacher means and the negative VA estimates are less negative than the corresponding (unadjusted) teacher means. The entire distribution of the VA estimates have shrunk and gravitated towards the value 0 (the adjusted overall mean). This reduction in the estimate is in fact perfectly consistent and aligned to the descriptive findings illustrated in the previous chapter. The descriptive analysis found that more teachers with high mean scores were surrounded with high performing background students than teachers with low mean

[135] The SE or the 95% CI interval band is also much larger for the null model with mean of 9.364706 and 5.317763 for the fixed effects null model and 7.964329 and 2.503585 for random effects null model.

scores. Thus, the VA results shown above illustrate that through taking into account the confounding effects of the student background variables, the overestimated teacher means have now been corrected to become more moderately positive while the underestimated teacher effects have now less negative. The confounding effects have been removed from the teacher means and more reliable estimates of teachers' unique effects have been attained. The models have successfully fulfilled its task. But to fully ensure that the estimates are in fact BLUE, we still must comprehensively assess the estimated model to prove that all the regression assumptions have been met.

To sum the basic VAM, fixed and random effects estimates are compared and contrasted below. This analysis follows suit with the work of number of leading VA researchers such as Tekwe et al. (2004), Raundenbush (2005), Hanushek and Rivkin (2008), Jakubowski (2008), McCaffrey et al. (2009), who have used the correlation analysis as the mean to provide numerical estimate of the similarity and consistency of the VA estimates (ranking) across different models.[136] The correlation coefficient for the two VA estimates and its standard error bands are calculated as follow.

Table 7.3: Correlation of basic fixed effects and basic random effects VA estimates and standard errors

Correlation of Basic FE and RE Estimates and SE		
	Basic FE	Basic RE
Basic FE	1.0000	-
Basic RE	0.8884	1.0000
	Basic FE SE	Basic RE SE
Basic FE SE	1.0000	-
Basic RE SE	0.8871	1.0000

As you can see, fixed and random effects illustrate a highly positive correlation value of 0.888 for the VA estimates and 0.887 for the standard error estimates. This finding provides a strong indication of consistency in the estimates across the two models. Many teachers identified as highly effective based on a high fixed effects VA estimates are also identified as highly effective with a high random effects VA estimates (and vice versa). These findings support the previous findings in Tekwe et al. (2004) who found high correlation of close to 0.90 between the two models (with the basic VAM specification) using 5th grade math scores in Florida. And Jakubowski (2008) also found high correlation of above 0.90 between the two models using a

[136] From a policy perspective, the magnitude of the VA estimates is in fact not the entire interest. If the main policy objective is to identify the effective teachers we can all learn from and the non- effective teachers we can all assist, the consistency of the VA estimates across the different models is more important. That is, the similarity of the relative ranking or positioning of the teachers across the VA estimates is more important. If a teacher is identified as highly effective with relatively high VA estimate using one VA specification, we hope that this teacher is also estimated as highly effective under another VA specification. In light of the ongoing debate facing fixed vs. random effects and the fact that there is never a one perfect theory/model to depict the VAMs, the comparison of the estimates across the different model is considered as the viable and effective approach especially to gain confidence and assurance of the findings.

high school exam data (for reading and math-science) in Poland. These findings all gives support and confidence to the VA findings illustrated above.

Extended Value-Added Models

This section extends the basic VAMs to take into account other potential factors outside the control of teachers. These factors include the teacher characteristics variables, the school fixed effects, varying effects prior test score and FRPL across the teachers, and student peer effects (teacher level averages of the student background variables). The estimates of the extended model with the teacher characteristic variables are first provided as follow.

Extended VAM with Teacher Characteristics

Table 7.4: Estimates of extended VAM with teacher characteristics

Random Effects VAM with Teacher Characteristics			
	Estimate	*SE*	*t value*
Intercept	403.09	0.880	457.90
Student Characteristics:			
Student is Female	2.441	0.183	13.40
Student is Native American (base)	-	-	-
Student is Asian	4.487	0.546	8.20
Student is Black	-1.135	0.593	-1.90
Student is Hispanic	0.197	0.506	0.40
Student is White	1.821	0.449	4.10
Student is on FRPL	-1.791	0.266	-6.70
Students' Prior Test Score*	0.719	0.002	290.20
Teacher Characteristics:			
Teacher is Female	0.730	0.380	1.90
Teacher is White	1.828	0.679	2.70
Teacher has MA Degree	0.227	0.353	0.60
Teacher Experience	0.032	0.017	1.90
School Fixed Effects	No		
Note: * Prior test scores are grand mean centered			

Variance Components:	
Between Teacher Variance	49.92
Within Teacher Variance	414.87
Intraclass Correlation	0.107
AIC	456743
BIC	456867
Number of Students	51161
Number of Teachers	2864
Degrees of Freedom	51148

The estimate of the student background variables illustrate very similar magnitude and significance to the values shown in the basic VAM. Female students, White and Asian students, and prior test scores are shown to have positive significant effects and FRPL status showed negative significant estimates. Looking now at the teacher characteristics variables, teachers' White ethnicity is found to have a statistically significant positive effect of 1.828 points on students' average performance in comparison to the non-White teachers. Teachers' female gender and experience are also found to have positive and slight significant effect on students' average performance. And teachers' education level (with or without MA degree) is also shown to positive association but the estimates are not found to be significant. These findings are consistent and aligned to the descriptive findings provided in the previous chapter. The simple linear regression estimates of each variable are re-summarized as follow.

Table 7.5: Summary of simple linear regression estimates with teacher characteristics

Summary of Simple Linear Regression Estimates for Teacher Characteristics					
	Estimate	*SE*	*t value*	*p value*	
Teacher is Female	-0.125	1.110	-0.113	0.91	
Teacher is White Ethnicity	8.495	1.948	4.361	0	***
Teacher with MA Degree	2.165	1.0159	2.131	0.0332	*
Teacher Experience	0.288	0.048	5.959	0	***

Comparing the VAM estimates to the simple linear regression estimates illustrates that the magnitude of the VAM estimates and its significance level is much smaller. In accordance to the model diagram shown in Figure 7.2, this reduction can be attributed to the indirect effect of these variables through the teachers and the confounding effects due to the student background variables. Regarding the indirect effects, as described earlier, the total effect of the variables has now been diversified into the indirect effect through the teachers and the remaining direct effect (shown as the estimates in the table). But for the teachers' education level variable, the reduction in its estimate is attributable to the confounding effects by the teachers. The above estimate for this variable therefore illustrates a more accurate and reliable effects of teacher education level on the student performance.[137]

Looking now at the main variable of interest, variance components estimates illustrate that more variation of teacher performance is explained by the model. This is indicated with the reduction of the between teacher variance to 49.92 from the 50.17 of the basic VAM. As the between teacher variance is a key determinant of the empirical Bayes VA estimates, this reduction will translate to smaller magnitude of the VA estimates as shown below.

Figure 7.11: VA estimates of extended VAM with teacher characteristics

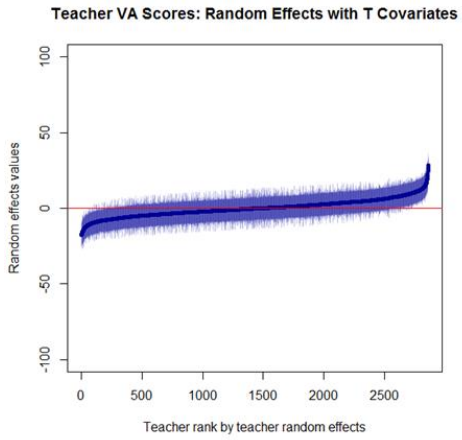
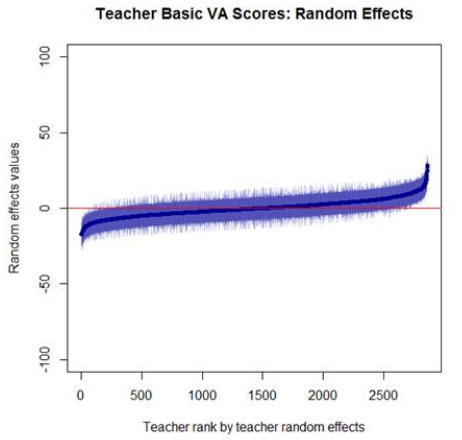


Figure 7.4: VA estimates of basic random effects VAM



[137] Again this study acknowledges that depending on our theories there could have been other mechanisms affecting the decision to obtain higher education e.g. it could have been related to the other teachers' background characteristics.

The caterpillar plot for the extended VAM shown in Figure 7.11 illustrate similar pattern of the basic VAM shown again (from Figure 7.4) on the right. But as expected, the mean value of extended VA estimates is -1.353436e-13 which is much smaller than the mean of -1.207604e-11 for the basic model.[138] The standard deviation of the VA estimates also decreased to 4.161 from 5.671. And the standard error (band length) also decreased with a mean of 4.1612 from mean of 4.1655.[139] Now, in accordance to the model diagram this reduction again shed light on the confounding effects of the teacher ethnicity, gender, and experience which was removed from the teacher effects. That is, by taking these variables into account, the extended VAM has corrected the misattribution of these effects on the teacher effects. It thereby improves the accuracy and reliability of the teacher VA effects. But with respect to the teacher education variable, through taking into account this variable, some of the total effects of the teachers (i.e. its indirect effect) have been diversified and absorbed by the teacher education variable. The teacher VA effect shown above is therefore a slightly strict and conservative measure of teacher effectiveness by the extent of the indirect effect.[140]

Extended VAMs with School Fixed Effects

The extended VAM which controls for school fixed effects shows the following results.

Table 7.6: Estimates of extended VAM with school fixed effects

Random Effects VAM with School Fixed Effects			
	Estimate	*SE*	*t value*
Intercept	107.78	4.335	24.86
Student Characteristics:			
Student is Female	2.400	0.183	13.14
Student is Native American (base)	-	-	-
Student is Asian	4.403	0.552	7.98
Student is Black	-1.416	0.601	-2.36
Student is Hispanic	0.429	0.515	0.83
Student is White	1.590	0.452	3.52
Student is on FRPL	-1.824	0.268	-6.80
Students' Prior Test Score*	0.719	0.002	288.39
School Fixed Effects:			
School ID1 (base group)	-	-	-
School ID2	12.628	10.576	1.19
School ID3	1.587	5.717	0.28
School ID4	15.599	7.966	1.96
School ID5	8.268	5.591	1.48
School ID6	12.188	7.648	1.48
School ID7	12.188	10.576	0.02
…(continue to SchoolID 936)			
Note: * Prior test scores are grand mean centered			

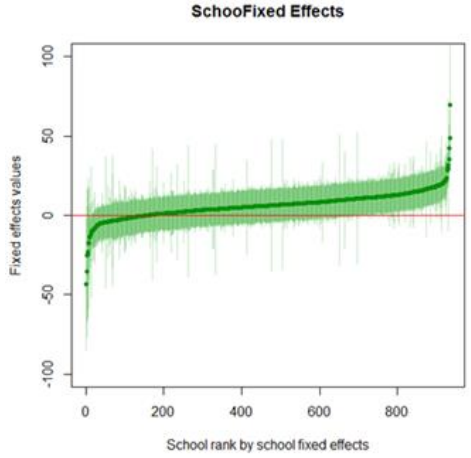
Variance Components:	
Between Teacher Variance	11.18
Within Teacher Variance	414.85
Intraclass Correlation	0.026
AIC	456568
BIC	464925
Number of Students	51161
Number of Teachers	2864
Number of Schools	936
Degrees of Freedom	50217

The school fixed effects model also presents very similar estimates (both magnitude and significance) for the student level variables in comparison to the basic VAM. Female students, White and Asian students, and prior test scores are shown to have positive significant effects and FRPL status showed negative significant

[138] Comparing the values for each teacher also showed that the extended model estimates is smaller for all teachers.
[139] The standard deviation also decreased to 0.8249 from 0.8279.
[140] To further explore this point, the above model was estimated without the teacher education level. The teacher VA estimates was then compared to illustrate no significant differences. This provides evidence that the indirect effect is very small and insignificant.

estimates. Looking now at the caterpillar plots of the entire school fixed effects illustrates the following results.

Figure 7.12: School fixed effects estimates



The mean of school effects is 6.190 with standard deviation of 7.714. And the mean of standard error (band) is 6.138 with standard deviation of 2.543. It is evident from the figure that schools (situated toward the upper end of the scale) add significantly more points on the student performance than the schools situated on the other end of the plot. That is, if a random student attends these schools, the students' performance is likely to increase significantly irrespective of their background (vice versa for the schools on the opposite end). And looking now at the teacher VA estimates, as shown in the variance components table, there is significantly large reduction in the between teacher variance. The between teacher variance is 11.18 which is close to 80% reduction from the 50.17 of the basic model. And this reduction translates to even smaller empirical Bayes VA estimates as shown below. But before we proceed to interpret the estimates, few caveats with this finding (i.e. the reduction in the VA estimates) are summarized in the footnote.[141]

[141] The reason for this reduction in the between teacher variance and VA estimates is not only due to the important role of school in explaining the student performance but also due to the huge number of school fixed effects parameter being fitted in the model i.e. the 936 binary (0,1) school variables. That is, a lot of structure and parameters are used to explain the data (as shown in the reduction of the degrees of freedom by 936). Mathematically, with more parameters being estimated the least squares minimization process will automatically entail more variance explained (better model fit) regardless of whether each variable actually contributed in explaining the outcome variable. This is known as the overfitting problem where the model fit has been inflated due to excessive use of parameters in explaining the data. And this is often criticized as the problem of R square index and AIC index which both do not penalize for this excessive variables used in the model. As you can see, looking at the AIC index it illustrates a huge improvement in the model fit with a reduction of close to 200 points (10 points being significant reduction). But the BIC index, on the other hand, penalizes for any added variables which do not significantly contribute in explaining of the variation of the data (variables which do not increase the R squared values as much). And as you can see, the BIC index of school fixed effects mode has increased by close to 8000 points to 464925 from 456839 of the basic random effects model. Thus school fixed effects model is in some way overfitting the data excessively large number of school parameters. And this also explains why the between teacher variation has fallen dramatically.

Figure 7.13: VA estimates of extended VAM with school fixed effects

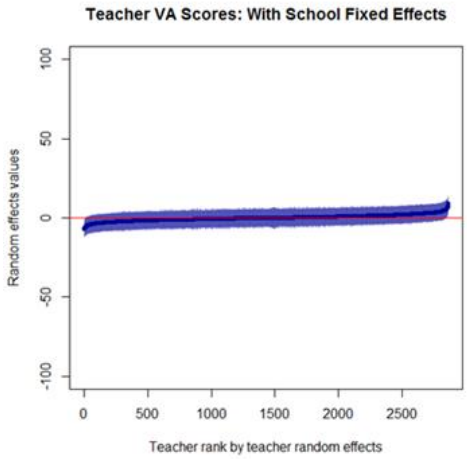
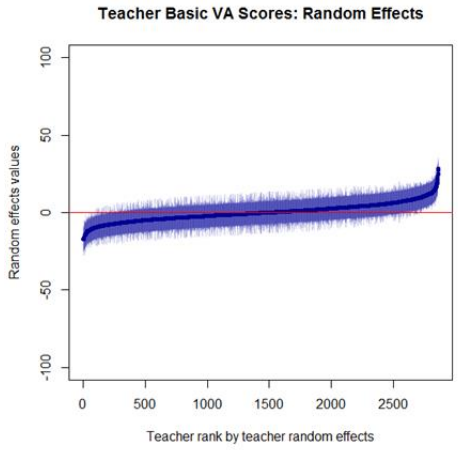


Figure 7.4: VA estimates of basic random effects VAM



As you can see, fixed effects model illustrates a clear reduction in the magnitude and standard error (band) of the VA estimates compared to the basic model. The mean of VA estimates is 1.092822e-13 (with standard deviation of 1.862) and the mean of standard error (bands) is 2.769 (with standard deviation of 0.209). And these estimates are again much smaller than the basic model estimates provided earlier. In accordance to the model diagram shown in Figure 7.2, this reduction is epitomizes the confounding effects by school fixed effects variable. Intuitively, as described in the last chapter, school fixed effects represent all the time invariant unobservable factors which affect teachers' performance and outside the control of teachers such as the type of school (charter, private, public), school budget, school working environment, school resources and facilities, etc. The above estimates illustrate that all these factors as a whole have significant role in explaining (confounding) the teacher VA effects. And by taking these confounding effects into account, we were able to achieve a more accurate and reliable teacher VA estimates. But as addressed earlier, from fixed effects alone we cannot identify which of the specific school features have affected the teacher VA estimates the most or the least. It only provides a very crude and aggregated measure of the school effects. The understanding of the individual and relative effects of school characteristics on student and teacher performance requires further data collection of these variables.

Extended VAM with Random Slope for Prior Test Scores and FRPL Status

Given the profound significant effects of prior test score and FRPL status, the possibility of varying (different) effects of these variables across the teachers is investigated. These models are referred to as the random slope models.[142] The estimates of these two models are as follow.

[142] These models can be envisioned as estimating the effects (slope) of these variables for and within each of the 2864 teachers. This give rise to the variation/differences in the effects of these variables across the teachers.

Table 7.7: Estimates of extended VAM with random slope for prior test scores

Random Effects VAM with Random Slope for Prior Test Score			
	Estimate	*SE*	*t value*
Intercept	406.044	0.461	880.80
Student Characteristics:			
Student is Female	2.410	0.182	13.30
Student is Native American (base)	-	-	-
Student is Asian	4.622	0.543	8.50
Student is Black	-1.074	0.589	-1.80
Student is Hispanic	0.239	0.502	0.50
Student is White	1.875	0.446	4.20
Student is on FRPL	-1.820	0.264	-6.90
Students' Prior Test Score*	0.724	0.003	237.10
School Fixed Effects	No		
Note: * Prior test scores are grandmean centered			

Random Slope for Prior Test Scores	
Variance Components:	
Between Teacher Variance	46.71
Within Teacher Variance	404.59
Random Slope Prior Test Variance	0.008
Intraclass Correlation	0.103
AIC	456389
BIC	456495
Number of Students	51161
Number of Teachers	2864
Degrees of Freedom	51151

Table 7.8: Estimates of extended VAM with random slope for FRPL status

Random Effects VAM with Random Slope for FRPL			
	Estimate	*SE*	*t value*
Intercept	405.93	0.465	873.40
Student Characteristics:			
Student is Female	2.444	0.183	13.40
Student is Native American (base)	-	-	-
Student is Asian	4.465	0.546	8.20
Student is Black	-1.156	0.593	-1.90
Student is Hispanic	0.157	0.505	0.30
Student is White	1.822	0.449	4.10
Student is on FRPL	-1.847	0.265	-7.00
Students' Prior Test Score*	0.719	0.002	290.20
School Fixed Effects	No		
Note: * Prior test scores are grand mean centered			

Random Slope for FRPL	
Variance Components:	
Between Teacher Variance	52.73
Within Teacher Variance	414.78
Random Slope FRPL Variance	1.052
Intraclass Correlation	0.113
AIC	456743
BIC	456849
Number of Students	51161
Number of Teachers	2864
Degrees of Freedom	51151

The student level variables again illustrate similar findings to the previous findings. Focusing our attention on the variance components estimates, it is evident that the between and within teacher variance for the random slope model with prior test score has decreased compared to the 50.17 and 414.93 of the basic models, respectively. That is, the random slope parameter has effectively captured the variation of the between teacher performance and within teacher student performance. To investigate whether this model improvement due to the random slope parameter is statistically significant, the AIC and BIC values are compared to the basic model values to illustrate a decrease in magnitude of more than 300. This is signifies that the random slope parameter (the variation in the effects of prior test scores across teachers) is statistically significant.[143] That is, there is significant difference in the effects of prior test scores across teachers. Looking at the estimates of the FRPL model, the model does not illustrated any improvement in explaining the between and within teacher variation. The AIC and BIC model fit also illustrate that the random slope of FRPL is not statistically

[143] Based on Raftery (2005) reduction in AIC and BIC value of over 10 is considered statistically significant effects.

significant. That is, there is no significant difference in the effects of FRPL across the teachers.[144] Looking now at the teacher VA estimates for these two models illustrate the following results.

Figure 7.14: VA estimates of extended VAM with random slope for prior test scores

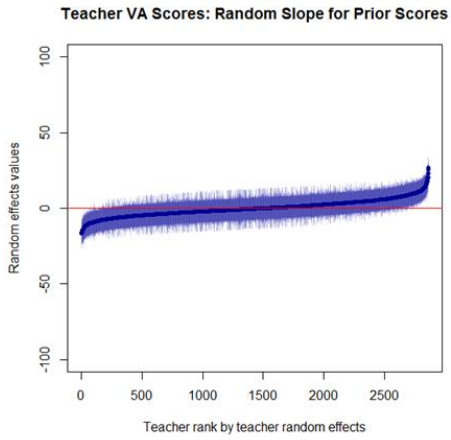
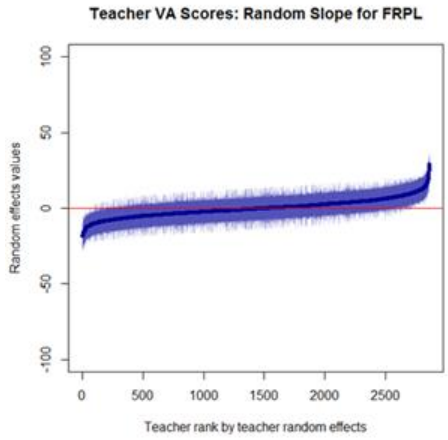


Figure 7.15: VA estimates of extended VAM with random slope for FRPL status



Focusing the prior test score model, the mean of the VA estimates has decreased to 5.77489e-12 in comparison to the basic model with -1.20760e-11. The model also illustrate a much smaller standard deviation (spread) of the VA estimates with 5.341 in comparison to the 5.671 of the basic model.[145] And the standard error of the estimates illustrates not much different a mean of 4.17704 in comparison to the basic model with 4.1655. In accordance to the theoretical model in Figure 7.2, this reduction is again attributable to the confounding effects of the prior test scores. But unlike the previous cases, the precision and accuracy of the confounding effects are enhanced. That is, the (confounding) effects of the prior test scores have been estimated for and within each of the teachers' classes. These unique and teacher specific confounding effects has then been removed from the corresponding teacher effects on the student performance. That is, by taking these more precisely estimated confounding effects into account, this model further improved the precision and accuracy of the teacher VA estimates.

Extended VAM with Student Peer Effects

Given the student background variables, the teacher level averages of these variables are calculated to represent the student "peer effects" characterizing the teachers' classroom. The estimates of the VAM which take into account these variables are illustrated as follow.

[144] To obtain the specific estimate of the effects of these variables for each teacher, we must apply the empirical Bayes prediction formula (applied previously to the random intercept) to the random slope parameter described earlier.

[145] For the FRPL model, the mean of the VA estimate is -8.009559e-13 with standard deviation of 5.814. And the mean standard error is 4.266638.

Table 7.9: Estimates of extended VAM with student peer effects variables

Random Effects VAM with Peer Effects			
	Estimate	*SE*	*t value*
Intercept	264.53	2.927	90.38
Student Characteristics:			
Student is Female	2.453	0.184	13.33
Student is Native American (base)	-	-	-
Student is Asian	3.609	0.536	6.73
Student is Black	-1.120	0.582	-1.93
Student is Hispanic	0.569	0.496	1.15
Student is White	1.841	0.444	4.15
Student is on FRPL	-1.998	0.272	-7.36
Students' Prior Test Score*	0.681	0.003	266.22
Teacher Level Averages:			
Teacher/Class Avg Test Score	0.350	0.007	49.06
Percentage of Female Students	-0.973	1.124	-0.87
Percentage of FRPL Students	9.208	0.994	9.27
Percentage of White Students	-3.249	0.617	-5.27
Teacher Class Size	-0.022	0.019	-1.16
School Fixed Effects	No		
Note: * Prior test scores are grand mean centered			

Variance Components:	
Between Teacher Variance	14.95
Within Teacher Variance	414.34
Intraclass Correlation	0.035
AIC	454930
BIC	455063
Number of Students	51161
Number of Teachers	2864
Degrees of Freedom	51147

As you can see, there is an overall reduction in the magnitude and significance of the student background variables (defined at the student level) estimates. As these variables were used to calculate and determine the peer effect variables, the reduction in the estimates represents the strong indirect effect of these variables via the peer effects variables. This is represented with the arrow pointing into the peer effects variable in the theoretical model diagram shown in Figure 7.2. Looking now at the peer effect variables, it is evident that the teacher (class) average current test scores have a profound significant effect on student performance. That is, teachers and classrooms who achieve high performance as a whole unit have significant positive peer effect on each of the classmates' individual performance. The percentage of White students in the class illustrate negative (but with less significance) peer effect while the percentage of students on FRPL in the class is have a positive peer effect on student performance. These are slightly peculiar findings as it contradicts with the student level estimates for these variables. One possibility maybe due to the very small number of FRPL students found in each class and other non-FRPL student are creating a positive peer environment by helping the FRPL students. Another possibility maybe due to the strong indirect effects of these two variables which are subsumed under the class average current test scores.[146] The percentage of female students (class gender balance), on the other hand, are not shown to have significant effects.

[146] Further diagnosis and examination of these variable estimates will be considered in the next chapter.

Looking now at the between teacher variance estimates, the amount of between teacher variance explained has fallen tremendously to 14.95 which is over 70% reduction from the 50.17 of the basic VAM.[147] And this reduction is reflected in the smaller magnitude of the empirical Bayes VA estimates shown below.

Figure 7.16: VA estimates of extended VAM with student peer effects

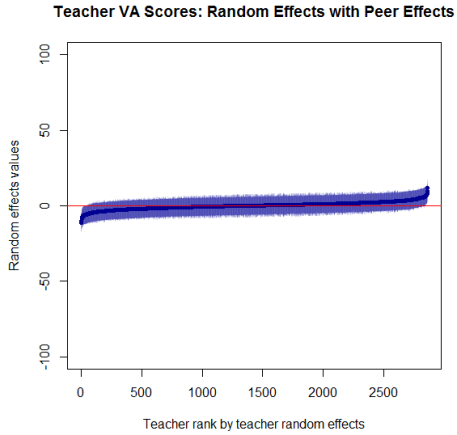
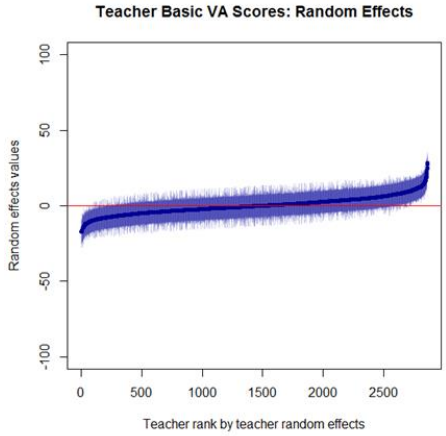


Figure 7.4: VA estimates of basic random effects VAM



As you can see, compared to the basic VAM estimates shown again on the right, the VA estimates have shrunk tremendously together with its standard error (bands length). The mean value of VA estimates of the extended model is now 2.16539e-12 which is smaller than the -1.20760e-11 of the basic model. Moreover, the standard deviation (spread) of these estimates is much smaller with 2.3541 compared to the basic model's 5.671.[148] In accordance to the theoretical model diagram in Figure 7.2 again, this reduction in the VA estimates represents the profound confounding effects of the peer effects which were removed from the teacher effects. As these peer effects are outside the control of teachers, by separating its effects from the teachers we improved the accuracy and reliability of the VA estimates. Finally, as shown in the Appendix, adding the teacher characteristics variables to the peer effects model did not alternate the results shown above. This illustrates the relative strong and resilient effect of the peer effects variables on the teacher VA and student performance.

Summary and Conclusion – Comparison of All VA Estimates

To sum, a variety of VAMs were estimated in this chapter. The basic and extended VAMs all illustrated significant VA estimates of teachers on student performance. The extended model improved the basic VAM by taking into account further confounding effects delineated in our theoretical models. Attention was also made to the possible intervening effects which could over-adjust lead to conservative measure of the VA and variable estimates. In light of these findings, this concluding section follows suit with the previous correlation analysis and the number of previous studies by the leading VA researchers to evaluate consistency and

[147] The model fit indices also illustrate this vast improvement in explained variance as the AIC and BIC indices have decreased by a magnitude of over 1700 for both the AIC and BIC illustrating the statistical significance of the newly added variables (based on Raftery (2005), reduction of 10 BIC points is considered a statistically significant effect/improvement).

[148] The standard errors have decreased to a mean value of 3.049918 (with standard deviation of 0.2864397) from the 4.1655 (and 0.8279) of the basic model.

similarity of the VA estimates across the different models. These correlation estimates are summarized as follow.

Table 7.10: Correlation of VA estimates of the basic and extended VAMs

Correlation of Value-Added Estimates of Basic and Extended VAMs								
	Basic FE	Basic RE	RE T Cov	RE Peer Eff	RE TCov & Peer	RE Rslope FRPL	RE Rslope Prior	RE School FE
Basic FE	1.000	0.888	0.885	0.622	0.622	0.888	0.853	0.609
Basic RE	0.888	1.000	0.997	0.774	0.773	1.000	0.986	0.689
RE with T Cov	0.885	0.997	1.000	0.774	0.775	0.997	0.983	0.686
RE with Peer Effects	0.622	0.774	0.774	1.000	0.999	0.773	0.779	0.540
RE with T Cov and Peer	0.622	0.773	0.775	0.999	1.000	0.772	0.779	0.538
RE with Rslope FRPL	0.888	1.000	0.997	0.773	0.772	1.000	0.986	0.689
RE with Rslope Prior	0.853	0.986	0.983	0.779	0.779	0.986	1.000	0.678
RE with School FE	0.609	0.689	0.686	0.540	0.538	0.689	0.678	1.000

Note: FE=fixed effects; RE = random effects; T Cov=teacher covariates/characteristics; Peer Effect= teacher level avg of student characteristics; Rslope= random slope

The VA estimates across the different models illustrate very high correlation. The average correlation across the models is 0.798.[149] This finding supports the earlier work by Tekwe et al. (2004), Hanushek and Rivkin (2008), Goldhabor (2010), Raundenbush (2005) and more. It also effectively responds to the concerns raised by Rothstein (2011) who addressed a minimal 0.7 to 0.8 correlation is necessary to establish confidence in the VA estimates.[150] Looking at the table above again, it is evident that the highest correlations are found between random effects models namely the basic random effects, with teacher covariates and the two random slope models. On the other hand, the lowest correlation are evident with the peer effects and school fixed effects models, which were the two models illustrating the most reduction in the between teacher variance and the smallest VA estimates. From the statistical and mathematical perspective, these findings are not surprising. The models which were similarly specified (with similar rigor, similar number of restrictions (variables), and similar modeling approaches) illustrated the highest correlations. The school fixed effects and peer effects imposed the most restrictions and structure to the data (most variables/parameters) and together with the (possible) over-fitting problem has led to the largest increase in the variation explained (the best model fit). And this lead to the most conservative VA estimates which deviated most from the rest of the models.[151] This pattern in the VA estimates was also witnessed in Tekwe et al. (2004) who also found very high correlation (above 0.9) for models with similar sets of variables controlled (regardless of whether it is fixed or random effects). But once the peer effects variables (average demographics and intake variables) were controlled the correlation fell to the 0.7 range and below (lowest of 0.65).[152]

In the nutshell, the above findings provide a strong indication of the similarity and consistency of the VA estimates across the models. It illustrates that the order (and ranking) of the estimates are not heavily altered by

[149] The average correlation of the standard error estimates are 0.948.
[150] Rothstein (2011) re-analyzed VA analyses of the MET project (administered by the Gates Foundation) to find correlation of only 0.5 to 0.6 when different sets of control variables were considered.
[151] The average test score variable in the peer effect variable might also be subject to the reverse causality problem. This point will be further taken up in the next chapter.
[152] The results are provided on Table 5 of Tekwe et al. (2004).

the variables incorporated in the model or the modeling options (fixed or random effects) use to estimate the VA effects. Teachers identified as highly effective (with high VA estimates) in one model are also identified as highly effective (with high VA estimates) in other models (and vice versa). And this consistency gives us a sense of confidence and assurance in the VA estimates. Now, the majority of the existing VA research terminate their analysis at this stage and apply the findings for policy relevant scenarios involving high stakes decisions (e.g. to provide bonuses to the effective teachers or fire the non-effective teachers). But this study finds this approach extremely dangerous and misleading. As addressed in the title of this chapter, the analysis conducted in this chapter is still in the "preliminary" and "premature" stage of the VA analyses. And more so, the findings are far from being ready to be applied for the real life policy setting. No empirical evidence or proofs are provided in this chapter to ensure accuracy (unbiasedness) and reliability (efficiency) of the estimates. As described in Chapter 2, these statistical properties lie at the heart of providing us with the trustworthy and viable statistical inference (the generalizations and conclusions) of the parameter estimates. It is a fact that possible confounding effects have been taken into account in accordance to our theoretical models. But as addressed in Chapter 5, we must now follow up and integrate the findings with empirical evidence (post fit analysis) that the extra variables have actually improved the model and more importantly, has not induced new forms of problems (violations) in the model. We must now fully diagnose whether the estimated models have effectively satisfied all the linear regression assumptions in ensuring the "best (efficient) linear unbiased estimates" (BLUE). And if the assumptions are violated, we must seek different solutions to revise these violations. It is only when the rigorous diagnosis and revision are completed and the regression assumptions have been successfully met, the true confidence, assurance, and solidarity of the VA estimates are achieved. And such estimates can be then be applied for the real life setting to inform and improve our education policies.

CHAPTER 8: DIAGNOSIS AND REVISION OF THE PRELIMINARY VAMS

This chapter conducts the diagnosis and revision of the preliminary VA analyses conducted in the previous chapter. It evaluates whether the linear regression model assumptions necessary to ensure the BLUE estimates are satisfied. And in cases when the assumptions are violated, the appropriate revisions will be implemented on the estimated model. For each revision conducted, the model will then be fully re-diagnosed to make sure new forms of violations are not introduced. This iterative procedure between (re) diagnosis and revision will continue until all the assumptions are successfully achieved to ensure us the BLUE estimates. And to reiterate, the only model in which statistical inference can be conducted is the one which provides us with the BLUE estimates. This is the only model we can confidently interpret and possibly apply in the real life setting. The two major groups of linear regression model assumptions one pertaining to the residuals and the other pertaining to the explanatory variables are summarized again below. As thoroughly explained in Chapter 3, the overarching idea underlying the residuals is that it must be randomly distributed with no systemic and meaningful pattern. This ensures that the model effectively captures the essence and pattern of the outcome variable and the effects of un-accounted omitted factors are not misattributed to modeled variables. And for the explanatory variables, the variables must illustrate unique and distinguishable variation without the problem of multicollinearity. This enables the linear regression model to provide reliable and converging estimates. And as described in Chapter 4, the random effects model (RE) must satisfy these assumptions defined at each level of the hierarchical data structure and model specification i.e. student and teacher levels.[153] The consequences of failing to detect and revise the violations on the statistical properties of the estimates are summarized again below.

Table 8.1: Summary of the linear regression model assumptions and consequences of violation

Assumptions	FE	RE	Consequences of Violation
I. Properties of Residuals – Randomly Distributed			
Levels of the Residuals:			
Mean value of residuals given Xs is zero. E(ei/X) =	x	x	Bias
Expectation and covariance between residuals and Xs is zero. Residuals are independent of Xs. E(e,X)=0, Cov (e,X)=0	x	x	Bias
Normally distributed residuals	x	x	Bias + Inefficiency
No extreme/outlying influential residuals	x	x	Bias + Inefficiency
Variance(Spread) and Covariance (Dependency) of the Residuals:	x	x	
Homoscedasticity constant variance of residuals	x	x	Inefficiency
Independence of residuals i.e. no autocorrelation, serial correlation, clustering or any form of dependencies/ covariance between the residuals. Cov(ei,ej)=0	x	x	Inefficiency
		x	
II. Properties of Explanatory Variables – Unique Variation		x	
Clear variation in the values of the Xs	x	x	No model estimates
No perfect association and collinearity between the Xs	x	x	Inefficiency or No est
No extreme high leverage and influential X values	x	x	Bias + Inefficiency

[153] This point will be further explained in later sections.

III. Higher Level Residuals and Explanatory Variables			
Randomly distributed higher level residuals (same assumptions as I above)		x	Same as above
Higher level explanatory variables with unique variation with no multicollinearity problem (same assumptions as in II above)		x	Same as above

In line to the previous chapters, this chapter extends the descriptive findings and theoretically constructed models by providing further empirical evidence to justify and validate the different VAMs. As Fox (2008) addresses,

> Taken together, regression diagnosis and corrective methods [revision] greatly extends the practical application of linear models. It is often the difference between a crude, mechanical data analysis and a careful, nuanced analysis that accurately describes the data and therefore supports meaningful interpretation of them. (p.241)

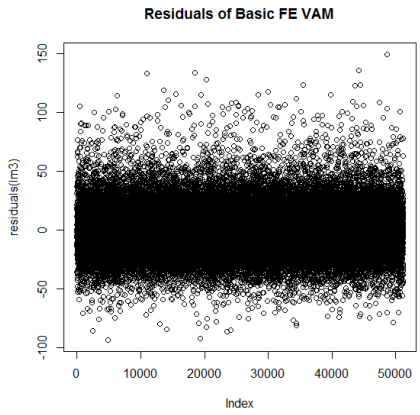
In the following sections, the diagnosis and revision of the basic fixed effects VAM are first conducted. The diagnosis and revision of the basic and extended random effects VAM then follow.

Diagnosis of the Basic Fixed Effect VAM

Diagnosis of the E(e/X) = 0 Condition and the Normality Assumption

The *E(e/X) = 0* condition provides the initial indication of whether the model is effectively capturing and fitting the actual data observations. If the model is performing well, it should error in a random meaningless fashion which implies that the errors on average have no system value (*E(e/X) = 0*). The condition signifies that all the important factors which systemically explain the outcome variable is taken into account in the model and all factors left un-accounted in the residual is left missing at a random fashion. To diagnose this condition, the following figure which plots the residuals by its unique student identification number is provided

Figure 8.1: Plot of basic fixed effects residuals



The plot indicates a high concentration of the residuals around zero. It provides a strong initial indication that the expected/average value of the residual is zero (*E(e/X)=0*). The basic model provides a good first indication of capturing all the systemic factors in the model. To further examine this condition and the normality assumption, the following histogram and qqplot are constructed.

Figure 8.2: Histogram of basic fixed effects residuals

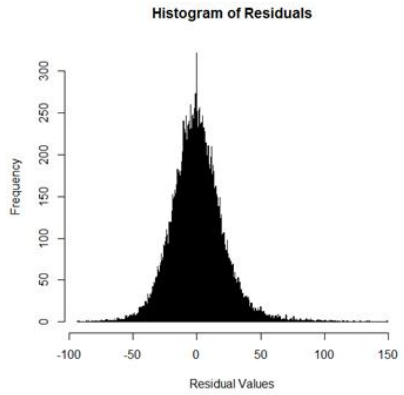
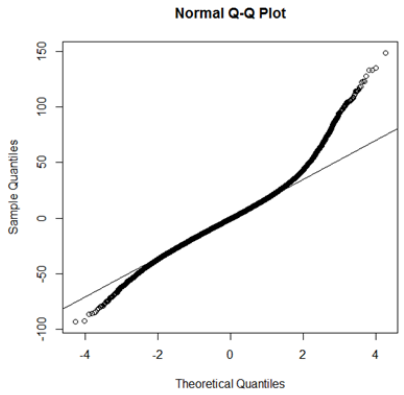


Figure 8.3: Normal quantile-quantile plot of basic fixed effects residuals



The normal distribution assumption of the residuals implies that if the residuals are randomly distributed, then it should make less frequent error of larger sizes and equal and balanced frequencies of both positive and negative errors. There should be no systemic trends and pattern in the error distribution. Looking at the histogram on the left, the symmetric and bell shaped curve of the normal distribution is clearly illustrated. The distribution is also concentrated around the zero value which gives another empirical evidence of the *E(e/X)=0* condition. But conducting a closer examination reveals that the normal distribution assumption is in fact violated with the presence of a peaked and heavy tails in the distribution. This is indicated with the kurtosis value of 5.1869 which is significantly larger than 3 of the normal distribution.[154] These features are also confirmed in the quantile probability plot (qqplot) on the right where number of residuals deviate from the 45 degree line (signifying the normal distribution) toward the lower and upper end.[155] Finally, the Pearson chi-square normality test also significantly rejects the null hypothesis of normality with chi-square value of 1808.344 and p-value of 0.[156] The presence of the heavy tails signifies that the model is making excessively large amount of large misfits. As described in Chapter 3, this can be attributable to several factors such as the extreme and outlying observations and omitted variables. And these factors can both introduce bias and

[154] The skewness is calculated to be 0.5331 which is closely in line to the 0 of the normal distribution.

[155] The qqplot compares the quantile values (or the shape) of the residuals to the quantile values of a hypothetical distribution (in this case the normal distribution with the same mean and standard deviation of the residuals). If the distribution of the residuals resembles the normal distribution, then the data points should lie on the straight 45 degrees line. The deviation of the plot at the lower and upper end of the plot signifies that the lower percentile (e.g. the 10^{th} percentile) is reached a much smaller value for the residuals than a hypothetical normal distribution while the higher percentile (e.g. 90^{th}) is reached at much higher value for the residuals than the normal distribution. This in fact, signifies the peaked and heavy tail distribution described earlier. There is more observations/frequency left at the tails of the distribution than a standard normal distribution. Taken together, the normal distribution of the residuals is violated.

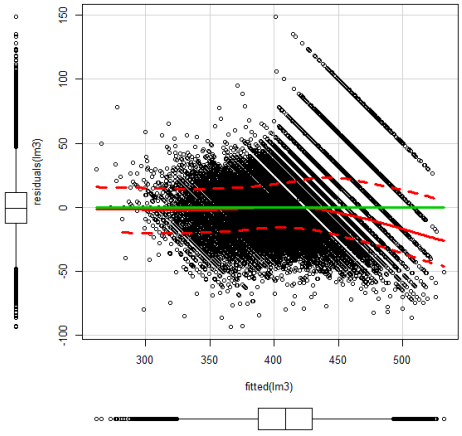
[156] Other normality tests such as the Anderson-Darling normality test and Cramer-von Mises normality test were also conducted. Both tests violated the normality assumption at a high significance level.

inefficiency in the model. Thus, with the presence of this severe peaked and heavy tailed distribution, the statistical properties of the basic fixed effects model estimates are questioned.

Diagnosis of the Exogeneity Condition

The exogeneity condition implies that if the residuals are randomly distributed, then the levels of the residuals should not entail systemic patterns with any variables: the estimated/fitted values, variables within the models, functions of these variables, and variables not within the models. If a systemic pattern in detected, this suggests important variable(s) (which can systemically explain the outcome variable) is failed to be taken into account in the model. To diagnose the exogeneity condition, the scatterplot of the residuals with the fitted values is first provided as follow.

Figure 8.4: Scatterplot of basic fixed effects residuals

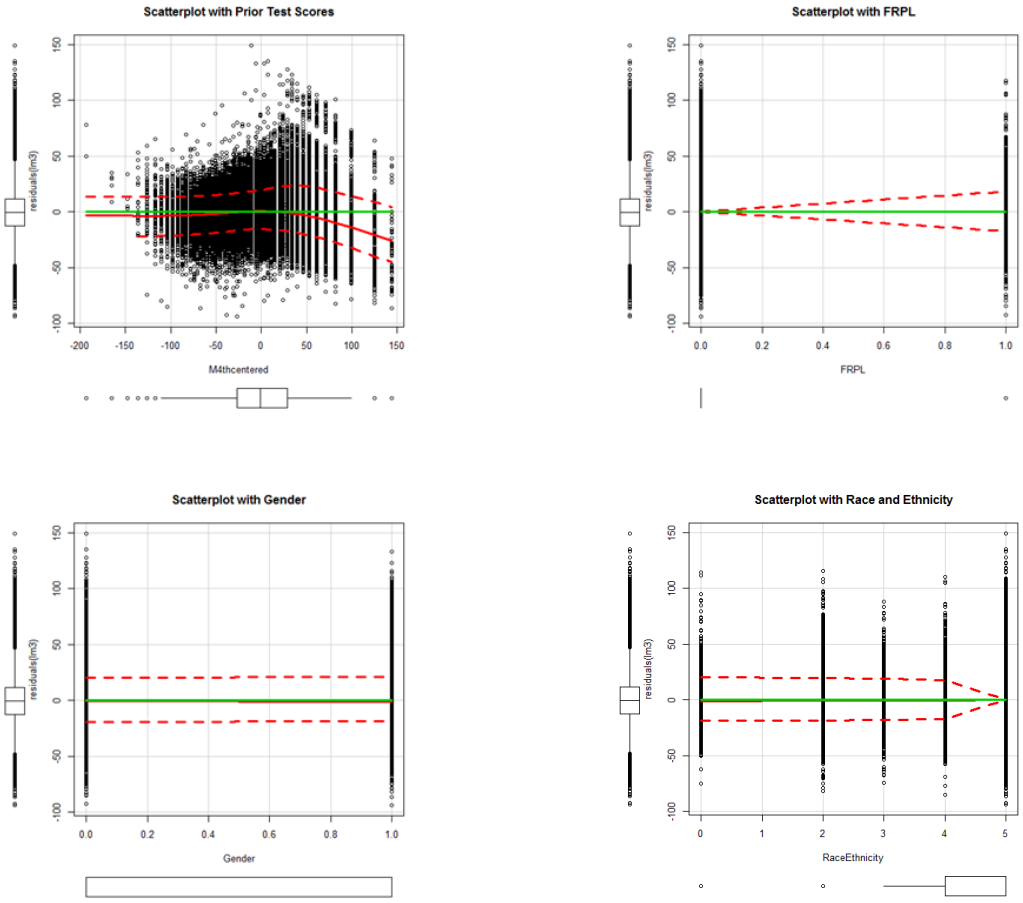


As you can see, the green simple linear regression line illustrates that there is no association between the residuals and fitted values. This is indicated with a perfectly horizontal line (with slope = 0) and provides an indication in support of the randomly distributed residuals. The plot is also concentrated around zero value giving further evidence of the *E(e/X) = 0* condition. But looking now at the red non-parametric regression function which better captures the underlying pattern of the data, a downward sloping non-linear trend toward the upper end of the figure is detected. The non-parametric curve does not provide evidence is support of the randomly distributed pattern. But as the basic fixed effects model does not incorporate any form of non-linearity, it fails to take into account this systemic trend. The omitted non-linear effect is misattributed to the linear effect of the variable or any variables correlated with this effect. The exogeneity condition is violated and the basic model is subject to an omitted and miss-specified functional form bias.[157] To identify the source

[157] As you may have noticed, the broken distribution shown toward the upper end of the scale which epitomizes the property of the criterion based test characterizing the WASL described in Chapter 6. The residuals which are the difference between the actual test score and the model fitted values carries on the basic properties of the actual test score (outcome variable) i.e. the type of variable, scaling, etc. But as shown in the descriptive analysis, this systemic trend does not violate the normality distribution i.e. the symmetry as equal frequency of observations above and below the mean is retained. It also did not contributed to the peaked and heavy tails of the distribution. Moreover, as this test property equally applies to all the students regardless of their background (all students take the same test), this systemic pattern does not confound the student variables included in the model. No bias is introduced in the model. As it becomes evident later, this broken distribution is entrenched in the data set and consistently shows up in different diagnostic graphs.

of this nonlinear pattern, the scatterplot of the residuals with respect to each of the explanatory variables underlying the fitted model is constructed as follow.

Figure 8.5: Scatterplot of basic fixed effects residuals by student background variables

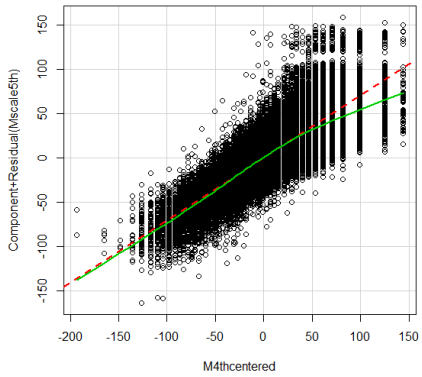


It is evident that the scatterplot of the prior test scores closely resembles the plot of the fitted values shown above. The nonlinearity is most likely inherent with the prior test score. For the categorical variables, both the linear regression and non-parametric curves merge to illustrate a perfectly horizontal flat line. This illustrate that there is no obvious non-linear pattern associated with these variables.[158] Now, to further examine and validate the nonlinearity with the prior test scores, the component residual plot (CRplot) which plots the partial residuals instead of the residuals with respect to the explanatory variable is constructed below.[159] Similar to the scatterplots, if the residuals are randomly distributed, then the CRplots will illustrate a horizontal band pattern in the direction of the matching linear regression and non-parametric regression curves.

[158] The dotted red lines which illustrate the variation of the residuals from the mean values of each category illustrate clear differences. These differences are (most likely) attributable to the differences in the sample size of each category.

[159] The partial residuals are calculated through adding the estimated partial relationship between the outcome variable and the explanatory variable (product of parameter estimate and the explanatory variable) to the residuals of the model. This provides an even better essence of the association between the explanatory variable and the outcome variable (sum of modeled association and un-modeled association). Plotting this partial residual with respect to the explanatory variable can then illustrate even better association underlying the two variables i.e. of any un-modeled association.

Figure 8.6: Component residual plot of basic fixed effects residuals and prior test scores

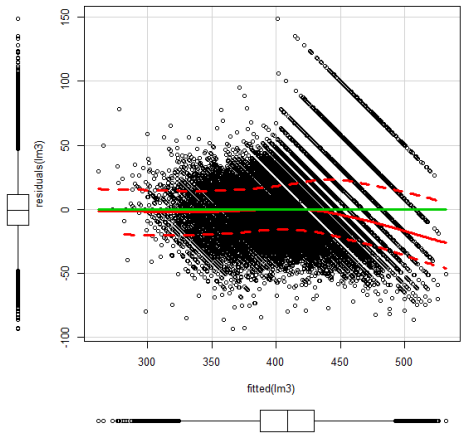


As you can see, there is again a non-linear non-parametric curve (this time in green) which deviates from the straight simple regression line. That is, CRplot confirms nonlinearity prior test score effect which yet to be modeled. To sum, the findings above provide empirical evidence that the basic fixed effects model fail to take into account all the important systemic effects explaining the outcome variable. The exogeneity condition is violated and bias is introduced in the estimates.

Diagnosis of the Homoskedasticity Assumption

The homoskedasticity assumption implies that if the residuals are randomly distributed, then the spread of the residuals must also portray no systemic pattern. The spread must be constant and homoskedastic with any variables. To diagnose this condition, the scatterplot of the residuals is examined again below.

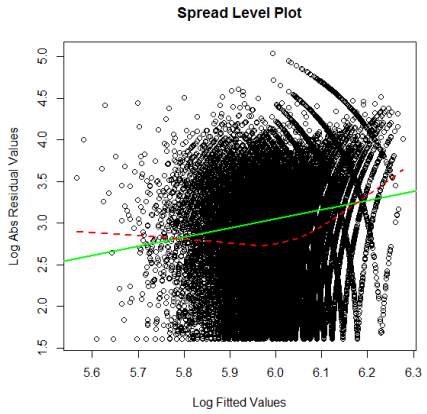
Figure 8.4: Scatterplot of basic fixed effects residuals



The red dotted lines represent the non-parametric curve which captures the variation of the residuals across the fitted values. Specifically, it captures the 95% confidence interval of the residuals along the solid red non-parametric regression curve. As you can see from these dotted lines, there is a slight increase in the variation

of the residuals toward the upper end of the fitted values. The dotted lines are not equally spaced and parallel to the non-parametric regression curve. That is, as the fitted values increases, the variation of the residual also increases. The homoskedastic assumption is violated and the un-modeled systemic trend is evident in the spread of the residuals. To provide further evidence of this violation, the spread-level plot (SLplot) designed by Tukey (1977) which is used to investigate the presence of non-constant variance also constructed to show the following results.

Figure 8.7: Spread level plot of basic fixed effects residuals



The SLplot plots the log of absolute value of the residuals and the log of the fitted value. When constant variance is detected, no systemic pattern between the two variables is detected. But as you can see from the figure above, there is a clear upward trend with slope of 1.092 for the basic fixed effects model. The SL plot confirms increase in the variance of the residuals for increase in the fitted values and the violation of the homoskedasticity assumption. Finally, a statistical test for the homoscedasticity condition known as the Pagen Busuch test (which regresses the variance of the residual with all the explanatory variables to find any non-constant systemic pattern) was conducted to significantly reject the null hypothesis of homoscedasticity (parameter estimates are statistically different from zero) with high chi-square value of 3228.74 and a p-value of 0.00. To sum, all these findings provide empirical evidence that the homoskedasticity assumption is violated in the basic fixed effects model. And as described in Chapter 3, the model estimates are inefficient and the statistical inference is misled.

Diagnosis of the Independence Assumption

The independence assumption implies that if the residuals are randomly distributed, then the residuals themselves should not illustrate any systemic pattern, resemblance, similarity, and relation. But as described in Chapter 3, in VA and education research, the independence assumption is questioned as similar students are clustered (self-select) under the same teacher and schools.[160] To diagnose the independence condition, a descriptive measure known as the intra-class correlation coefficient (ICC) can be used. The ICC is equal to the

[160] Other typical setting when the independence assumption is the time series data where repeated observations of the same individuals are always correlated. That is, the repeated observations are grouped and clustered within the same individual.

average correlation between values of two randomly drawn observations (e.g. students) in the same randomly drawn group (e.g. teacher). Mathematically, the ICC is represented as follow.

$$ICC = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_y^2}$$

where σ_a^2 is the between group variance (variance of group means) and σ_y^2 is the within group variance (average variance of observations within groups). As evident from this equation, ICC is interpreted as the fraction of the total variance that is due to the total variance between groups.[161] Calculating the ICC for the basic fixed effects model, value of 0.178 is found. This implies that the observations/residuals within teachers are on average positively correlated. This value is relatively high in comparison to other educational research where Bosker and Snidjers (1999) found ICC value low as 0.05 in many education data sets. Now, to test whether this ICC or dependency pattern is statistically significant, a model specification test between the constrained/restricted model without the teacher VA parameter and the unconstrained/unrestricted model with the teacher VA parameter (the basic fixed effects model) can be conducted. This test enable us to evaluate whether the teacher VA (or any grouping variable) as a whole have a combined significant effect.[162] That is, it evaluates whether the systemic dependency modeled through the VA parameter is statistically significant.[163] The results of this test are as follow.

[161] This ICC equation was first derived by Fisher (1935). It is important to note that while ICC is viewed as a type of correlation, it is different from the conventional Pearson correlation coefficient. The Pearson correlation coefficient is defined for a pair of two different variables/measures without any grouping structure (e.g. measure of association between height and weight of all first grade students). ICC, on the other hand, is defined only for one variable/measure across a data set structured into different groups (e.g. measure of association of height of first grade students within their first grade teacher/class). The mathematical procedure in calculating the two types of correlation is very similar but one key difference is that in the ICC, the variable (data) are centered and scaled using a pooled (weighted) mean and standard deviation across the different groups whereas in the Pearson correlation, each of the two variables (data) is centered and scaled by its own mean and standard deviation. For example, in the simplest case with a data consisting of N groups with a pair of observations within each group, the ICC becomes

$$\frac{1}{N\sigma^2}\sum_{n=1}^{N}(x_{n1}-\bar{x})(x_{n2}-\bar{x}) \text{ where } \bar{x}=\frac{1}{2N}\sum_{n=1}^{N}(x_{n1}+x_{n2}) \text{ and } \sigma^2=\frac{1}{2N}\{\sum_{n=1}^{N}(x_{n1}-\bar{x})+\sum_{n=1}^{N}(x_{n2}-\bar{x})\}$$

As you can see, unlike the Pearson correlation, the mean $\bar{x}$ and variance σ^2 are defined through pooling the estimates (weighted average) across the different groups. The ICC shown in the text is the direct extension of this procedure for the case when any number of groups and observations within groups are considered. Mathematical proof is provided by Fisher (1935).

[162] That is, the test examines whether all the teacher VA estimates are jointly equal to zero.

[163] As described by Bosker and Snidjers (1999), given the ICC formula, the ICC (dependency) is positive and significant if and only if the between teacher variance (on the numerator) is positive and significant. Therefore, testing the significance of the ICC amounts to the same as testing the significance of the between teacher variance. Now, in addition to the model comparison test provided in the text, other statistical tests such as the ANCOVA with the teacher as the fixed treatment can also be used. The analysis will give the same finding. The results for the teacher effects of ANCOVA was in fact represented in the caterpillar plot which illustrated signs of significant teacher effects as teachers toward the upper and lower end of the plot clearly deviated from the grand mean (with the 95% confidence interval not crossing the value 0). But from the caterpillar plot alone, we cannot conduct a joint/combined test to evaluate whether all the teacher effects (as a whole) are zero and we must resort to the model comparison general F test.

Table 8.2: Statistical test of the teacher VA parameter – model comparison test (of basic fixed effects VAM)

Model Comparison Test - Statistical Test of the Teacher VA Parameter							
	Df	*RSS*	*Diff Df*	*SS*	*F value*	*p value*	
Restricted Model	51153	23663949	2863	3703639	3.1297	0.000	***
Un-restricted Model	48290	19960310					
Note: *** significant at 0.1%, ** at 1%, * at 5%							

As you can see, the unrestricted model (basic fixed effects VAM) illustrate better and statistically significant fit then the restricted model without the teacher VA parameter. This entails that the teacher VA (the grouping variable) which engenders the systemic dependency pattern (non-zero ICC) is statistically significant. And in order to achieve the randomly distributed residual pattern and the reliable estimates, we must take this dependency pattern into account. As described in Chapter 3, failing to do so (as in the constrained/restricted model) the estimates will be inefficient and unreliable and the statistical inference will be misled. Thus, the above findings illustrate the empirical justification and importance of conducting VAMs i.e. the first task of defining the teacher VA parameter in order to take into account the clustering and dependency of the observations.[164]

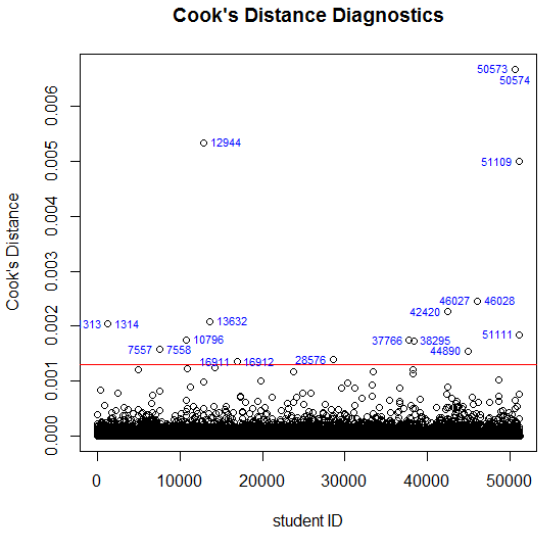
Diagnosis of the Outlying and Influential Residuals

The absence of the outlying and influential residuals implies that if the residuals are randomly distributed, then there should not be any residuals (cases or observations) which can single handedly exert systemic pressure to heavily alter and disrupt the model estimates. Such cases can be due to extra-ordinary and exceptional cases which entail important information or due to measurement error and human miscoding which entail misleading information. We must therefore handle the influential observations with care and close examination. To investigate the presence of these influential cases, influential diagnosis index called the Cook's distance is calculated to show the following result.[165]

[164] A point of further clarification in interpreting the ICC results is as follow. The ICC is defined only through taking into account of the grouping variable which engenders and is responsible of the dependency and clustering of the observations. But through acknowledging and incorporating this grouping variable in the model, we essentially take into account (correct and revise) the un-accounted dependency pattern. In other words, the grouping variable helps us to both identify the presence of systemic dependency (ICC) and at the same time take into account of this pattern in the model. In the linear regression framework, the grouping variable both diagnoses and revises the violation of the independency assumption simultaneously. In the context of VAMs, the grouping variable is precisely the teacher VA parameter. The VAMs (through the teacher VA parameter) detects the presence of systemic dependency in the residuals and explicitly takes into account this pattern to provide the efficient and reliable estimates. Given the significant ICC values, the model without the teacher VA parameter (the constrained and restricted model in the model comparison test) which ignores the dependency pattern is then the problematic model that is subject to the inefficiency and misleading inference. Thus in essence, the VAMs correct this constrained model by taking into account of the dependency through the teacher VA parameter. The presence of non-zero and significant ICC value provides the empirical evidence and justification to conduct the VAMs or any multilevel analysis. On the other hand, if the ICC is zero and insignificant, the grouping variable becomes irrelevant and the simple linear regression model without the teacher VA parameter prevails. Finally, as the ICC is defined by the two variance components, addition of other variables (defined at both levels) can potentially further reduce this systemic dependency pattern. It can improve the model by further taking into account and explaining away the dependency pattern and thereby further alleviating the violation of the independence assumption. Keeping these points in mind will help us interpret the revision process in the coming sections.

[165] For an observation to be highly influential, it must be comprised of an outlying residual error value (deviation from the estimated model) and also outlying explanatory variable values. The latter is referred to as high leverage values. When these two extreme features are combined, highly

Figure 8.8: Cook's distance diagnostics of basic fixed effects VAM



As you can see from the figure, several cases with relatively large value of the Cook's distance index are identified. Setting 0.012 as the cutoff value (as shown on the horizontal line), 20 observations were identified with index values above this cutoff.[166] These highly influential cases are exerting excessive influence on the model estimates. And depending on the underlying cause of this high influence measure, these cases are misleading the estimates through introducing both bias and inefficiency.[167] The statistical properties of the basic fixed effects are again questioned. Further examination of the underlying features of these influential cases will be sought in later section as we decide on how to handle and revise them.

Diagnosis of the Properties of the Explanatory Variables

Proceeding now to the diagnosis of the second major group of assumptions of the linear regression models - the behavior of the explanatory variables illustrates the following results. First, as shown in the following descriptive summaries, there is a clear variation in each of the variables. This is represented with non-zero values for the standard deviation (SD) and range (mini, max). That is, all the students do not possess the same single value for the explanatory variables and this enable us to proceed with the linear regression analysis.[168]

---

influential case which can drastically (and unrealistically) alter the model parameter estimates is defined. The Cook's distance measure is essentially a product of the residual value and the leverage value for each case/observation.

[166] There are no established norms or thresholds with respect to the Cook's distance measures to define influential cases. The influential cases are defined relative to the rest of the observations. Please refer to Fox (2008) for further explanation.

[167] As described in Chapter 3, depending on the specific values and direction of the explanatory variables for these cases, the heavy influence on the parameter estimates will translate into bias in line to these underlying values. In the case, if the cases are due to miscoding and measurement error we would like to abandon this effect. But for the exceptional cases we will like to cite and note its effects. Now, the extreme residual value will automatically contribute to the increase in the standard error of the model and thereby induce inefficiency

[168] As defined in Chapter 2, linear regression model is the function connecting the conditional means of the outcome variable for different values of the explanatory variable(s). Given a clear variation of the explanatory variables this function can be defined and estimated. But if the variable has no variation and takes only a single value, then the regression function or the conditional mean is simply the mean of the outcome variable. That is, it is only a point. The regression function or the curve/line connecting the different conditional means for different values of the explanatory variable is no longer defined.

Table 8.3: Descriptive summary of student background variables

	Mean	**SD**	**Mode**	**Mini**	**Max**
Gender	0.49	0.50	0 (51%)	0	1
Ethnicity	4.25	1.32	5 (66%)	0	5
FRPL	0.17	0.37	0 (83%)	0	1
Prior Test Score	406.1	41.45	-	213.0	550.0

Second, looking at the following correlation table of the explanatory variables, there is no indication of perfect correlation and the multicollinearity problem (correlation over 0.90). The variance inflation factor (VIF) which is an index used to calculate the multicollinearity problem also shows that all the variables have values much smaller than 2 which is the value that warns of the presence of the multicollinearity problem. The values of each variable are uniquely identified and the regression estimation process can proceed to provide converging and reliable estimates.

Table 8.4: Correlation and variance inflation factor of student background variables

	Gender	Race Ethnicity	FRPL	Prior Test Score
Gender	1			
Race Ethnicity	-0.00306	1		
FRPL	-0.00314	-0.11367	1	
Prior Test Score	0.04677	0.10303	-0.20205	1
	VIF			
Gender	1.0024			
Race Ethnicity	1.1772			
FRPL	1.1192			
Prior Test Score	1.1185			

Finally, as shown above, there were several cases with high leverage values (extreme values of the explanatory variables) which led to the high Cook's distance influential measure. These cases can potentially induce misleading influence on the estimates i.e. introduce bias and inefficiency. These cases are already identified in the above analysis and it will be dealt with in the following sections.

To sum, the diagnosis of the basic fixed effects VAM detected number of violations of the linear regression assumptions required to ensure the BLUE estimates. The explanatory variables illustrated clear and uniquely defined variation but the residuals did not illustrate the randomly distributed pattern as number of important systemic patterns was left accounted in the model and into the residuals. The estimates of the basic model are therefore subject to both bias and inefficiency and it should not be interpreted or used for real life circumstances. In light of these problems, the following section introduces different methods to revise these violations in order to achieve the BLUE estimates.

Revisions of the Basic Fixed Effects VAM

Revisions of the Peaked and Heavy Tailed Non-Normal Distribution

They key and probably the most economical and simplest solution to revise the peaked and heavy tailed non-normal distribution of the residuals is to transform the model such that less emphasis is placed on the extreme values situated in the heavy tails.[169] This can be achieved by descending down the power transformation towards zero e.g. power of 1/2, 1/4, 1/8, etc. For example, for a sample of observations -81, -16, -4, 0, 4, 16, 81, transforming the values with power of 1/2 will lead to -9, -4, -2, 0, 2, 4, 9. As you can see, larger values are reduced more than the smaller. As we further descend down the power towards 0 e.g. power of 1/4, 1/8, 1/16, etc. the large extreme gets reduced even more in comparison to the smaller values. That is, there is much less weight and emphasis placed on the extreme values. But when the power takes on the value zero, all the sample values become 1. This is not a useful outcome as the original data (variation) is eliminated and analysis with the data can no longer be conducted. But a useful and important mathematical property is that as power approaches zero, it approaches the natural log of the original sample values. In other words, the natural log transformation is essentially the smallest power transformation. This is a very useful transformation which is widely used in different academic disciplines such as physics as the transformed values all take on positive values and in the case of linear regression models, the estimated coefficients are interpreted as the change in the percentage of the average outcome variable for a unit increase in the explanatory variable.

Now, given the relatively severe peaked and heavy tails of the residual distribution, the natural log transformation of the outcome variable which puts the minimal weight on the extreme tail values seems like the best and most viable solution. If we are skeptical of such option, we can always consider a variety of power transformation and compare the results. But such procedure is very time consuming. A more precise and slightly more advanced method to find the optimal power transformation would be to use the Box-Cox transformation estimation which utilizes the maximum likelihood estimation to find the power value which maximizes the fit to the observed data.[170] The Box-Cox transformation is estimated for the basic fixed effects model to show the following results.

[169] More advanced methods include the Huber-White M-estimation which puts less weight on the cases situated in the residuals through a iterative estimation process. Please refer to Fox (2008) for demonstration.

[170] Further explanation of Box-Cox transformation is provided in Fox (2008).

Table 8.5: Box-Cox power transformation estimates (of basic fixed effects VAM)

Box-Cox Power Transformation Estimates			
Power Est	*SE*	*Lower Bound*	*Upper Bound*
-0.0030	0.0248	-0.0515	0.0455
	LR Test	*df*	*p-value*
lambda = (0)	0.0146	1	0.904
lambda = (1)	1610.0	1	0.000

The Box-Cox transformation estimation provides the optimal power value of -0.003. The Likelihood Ratio test also shows that the null hypothesis power (lambda) of 0 is not statistically rejected but for power (lambda) of 1 it is statistically rejected at a very high significance level. The Wald lower and upper bound estimates also surround the value 0. In other words, the estimates give strong support of a power transformation that is very close to 0. But as just described, such transformation is precisely the natural log transformation. That is, the Box-Cox transformation provides empirical support that the natural log transformation of the outcome variable as the optimal solution to resolve the peaked and heavy tail problem. Now, before we jump into this solution, we first consider the remedies for the other violations as similar or conflicting methods can often arise.

Revisions of the Non-Constant Variance of the Residuals

The violation of the homoskedasticity assumption can also be revised by means of transformation. One of the widely used methods is suggested by Tukey (1977) who showed that power transformation of 1 – b (where b is the slope of the simple regression in the spread-level plot shown earlier) is an effective "variance stabilizing transformation" of the residuals. The SL plot of the basic model shown above had a linear regression line with a slope of 1.092. Based on this estimate, the power transformation of 1 – 1.092 = -0.092 is suggested as the variance stabilizing transformation. But as just witnessed with the Box-Cox transformation, such small power transformation equates to the natural log transformation. In other words, the natural log transformation is an effective multi-task method in revising the peaked heavy tails and the non-constant variance of the residuals. Now, there is an extensive literature on other methods in correcting the non-constant variance as described in the footnote.[171] But as these advanced methods require additional assumptions, it has the additional risk of introducing other violations to the model. The natural log transformation is therefore the best first option. If the non-constant variance continues to persist, other methods can be considered.

[171] An alternative method is to study the form of the non-constant variance to mathematically model this underlying structure. Once a good model for the variance pattern is achieved, the original model can be transformed by means of dividing the entire model with the variance function/model. This method (also known as generalized least squares (GLS) or weighted least squares (WLS)) essentially recreates the constant variance assumption of the linear model. The transformed model can therefore be estimated with the usual least squares method. The main drawback of this method is that the correct functional form of the variance must be achieved. If the modeling of the variance is mistaken (can never be perfect), then the GLS estimates will be biased. A more advanced yet less risky method is to use the Huber-White or "sandwich" robust standard errors. This method simply substitutes the residual values calculated from the linear regression model into the standard error equation used in the GLS and WLS. It is a short cut method. The key advantage is that this method does not necessitate us to find the structure of the non-constant variance which has the risk of mis-modeling the underlying pattern as described previously.

Revisions of the Violation of the Exogeneity Condition

Based on the descriptive analysis above, the key method to correct for the violation of the exogeneity condition is to incorporate the nonlinear prior test score into the model. This can be achieved by adding a squared (power of 2) prior test score term into the model. The significance of this newly added term can then be tested and examined using the CRplots. If further misspecification of this variable is evident, higher power transformations (e.g. cubic or quadratic) can also be considered.

Revision of the Dependent Residuals

The dependency of the residuals can be revised through taking into account the grouping variable which engenders the dependency and clustering pattern. As described above, in the context of VA research, the grouping variable is precisely the teacher VA parameter and the fixed effects VAM has (already) effectively taken the dependency pattern into account. The problematic model subject to inefficiency is the model without teacher VA parameter. In essence, the fixed effect VAM has revised this model by explicitly modeling the dependency pattern through the teacher VA parameter.

Revisions of the Outlying Influential Cases

The outlying influential cases can be handled using a variety of methods. The simplest method would be to exclude the outlying influential observations from the data set and re-estimate the model. But in doing so, we must be careful to examine the underlying features of these observations as it can potentially reveal vital and important information or be due to measurement error and random miscoding. The exclusion of the observations is recommended particularly for the latter case. A more computational advanced method is the Huber White robust estimation (M-estimation) which utilizes an iterative weighting mechanism that puts less emphasis on the extreme values situated in the tails. This method can also resolve the heavy tail problem described above. This option was sought with the fixed effects model but as confirmed with John Fox who designed the R command for this method, the iterative mechanism could not provide converging estimates due to the over parameterization of the model with close to 3,000 teacher dummies variables and the very large data set of over 50,000 students. This option will be sought again in the future when the computing capacity of the R software improves.

To sum, a series of different revisions were identified to correct the different violations of the linear model assumptions. The revision process can become quite intricate and intertwined as some remedies (e.g. the natural log transformation) served as multiple cures while some violations (e.g. heavy tails) entailed multiple remedies. Moreover, one revision could also possibly introduce new forms of violations. A careful revision process which iteratively re-diagnoses the model for each revision is therefore necessary. Keeping this point in mind, the revisions will be implemented in the following sequence. First, the multi-task natural log

transformation is considered. Once the re-diagnosis of the transformed model is conducted, the inclusion of non-linear prior test score followed by the correction of influential cases will be conducted. Re-diagnosis for each of these revisions will also be conducted. This iterative procedure will continue until all the linear regression assumptions to ensure BLUE are jointly achieved at its best effort.

Implementation of the Revisions of the Basic Fixed Effects VAM

The residuals of the natural log transformed basic fixed effects VAM to illustrate the following results.

Figure 8.9: Histogram of natural log transformed fixed effects residuals

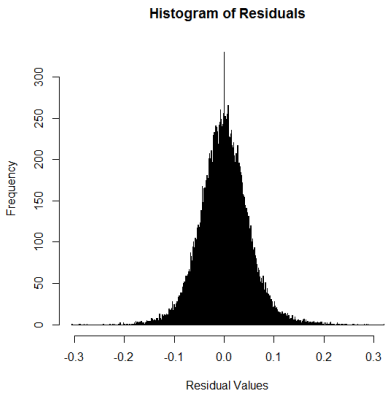
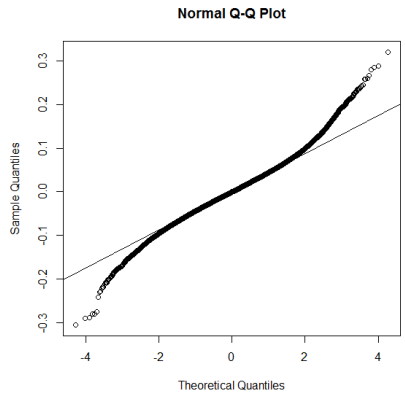


Figure 8.10: Normal quantile-quantile plot of natural log transformed fixed effects residuals



In comparison to the original basic model, there is a clear sign of improvement in the peaked and heavy tails of the distribution. The qqplot shown on the right also confirms this improvement with fewer cases deviating from the 45 degree normal distribution line. To provide numerical indications of this improvement, the skewness and kurtosis indices calculated to show the following results.

Table 8.6: Skewness and kurtosis values of the natural log transformed and basic (original) fixed effects residuals

Revised Transformed Model		**Original Model**	
Skewnes	0.134318	Skewness	0.533078
Kurtosis	4.387769	Kurtosis	5.186971

As you can see, the revised model illustrates a significant improvement in leveling the peaked and heavy tails and also the slight positive skew of the distribution. The residuals are now much closer to the normal distribution with skewness of 0 and kurtosis of 3. Thus, the natural log transformation effectively revised the first violation of non-Normally distributed residuals. Now, to diagnose if the transformation improved the second violation – the non-constant residual variance, the scatterplot of the residuals and the SL plot is constructed again as follow.

Figure 8.11: Scatterplot of natural log transformed fixed effects residuals

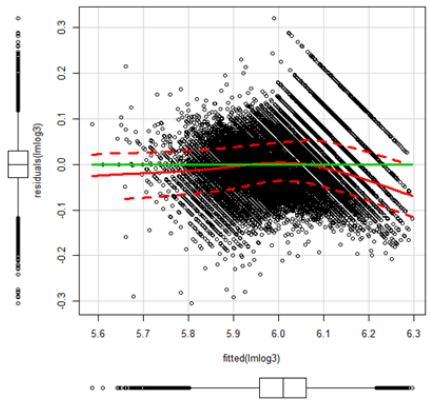
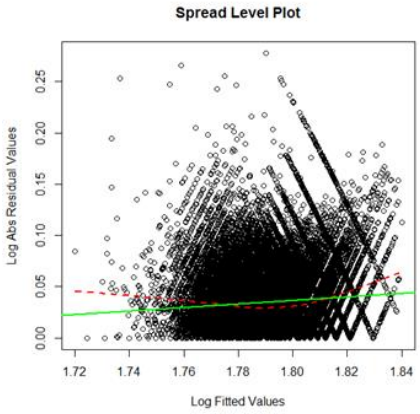


Figure 8.12: Spread level plot of natural log transformed fixed effects residuals



As you can see from the red dotted lines in the scatterplot on the left, the increase in the variance toward the upper end of the scale is no longer evident. The dotted lines are now more parallel along the non-parametric curve indicating an equally spaced constant variance of the residuals. And looking at the SL plot on the right, the slope is now much flatter with a value of 0.1772 compared to the 1.092 of the original model. It is now much closer to the flat horizontal pattern signifying constant homoscedastic variance. And with the slope of 0.1772, this entails a variance stabilizing power transformation of 1 – 0.1772 which is virtually equal to power of 1 with no transformation. These findings all illustrate that the natural log transformation effectively revised the non-constant variance of the residuals.

Now to make sure no other forms of violations are introduced the natural log transformed model is re-diagnosed. Looking at the residual scatterplot shown below on the left, a clear non-linear trend left un-accounted in the model continues to be evident. Constructing the scatterplot and the CRplot with respect to the explanatory variables, the non-linear trend is again identified with the prior test scores as shown below. These findings all provide empirical evidence that the exogeneity condition continues to be violated.

Figure 8.13: Scatterplot of natural log transformed fixed effects residuals and prior test scores

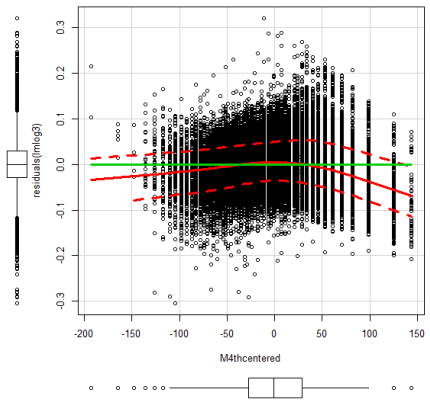
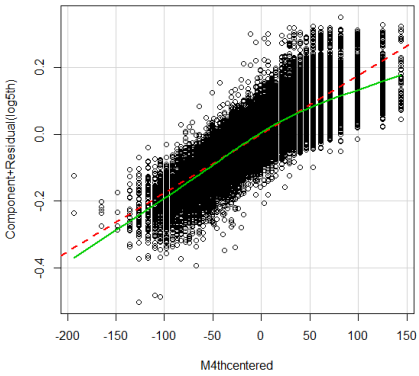


Figure 8.14: Component residual plot of natural log transformed fixed effects residuals and prior test scores



The ICC, on the other hand, continues to illustrate positive value of 0.211 with high statistical significance with F statistics of 3.234 and p-value of 0.000. A clear positive dependency of the residuals within teacher is detected and has been effectively taken into account in the VAM. It is also evident that the ICC has increased slightly from the previous un-revised model of 0.107. This is in line to our expectation as the as the natural log transformation reduced the magnitude of the residuals with extreme values by putting less weight. It has made these residuals more similar to the rest and has thus increased the dependency value. And finally the Cook's distance influential index was calculated (although not provided in text) to identify the similar group of cases potentially disrupting the model with misleading influence on the estimates. Thus, the natural log transformation fulfilled its purposes but other violations identified earlier continue to prevail. Given this diagnosis, we therefore first incorporate the non-linear effect of the prior test scores into the natural log transformed model in an attempt to revise the violation of the exogeneity condition. This shows the results shown below. The influential cases will be dealt subsequently.

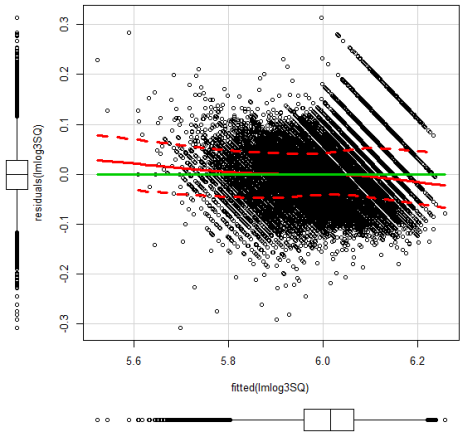
Table 8.7: Estimates of the natural log transformed with non-linear prior test score fixed effects VAM

	Estimate	*SE*	*t-value*	*p-value*
Intercept	5.9626	0.0161	369.64	0.000
Prior Test Score	0.0018	0.0000	284.27	0.000
Prior Test Score Squared	0.0000	0.0000	-43.236	0.000
Student is Female	0.0058	0.0004	13.175	0.000
Student is Native American	-	-	-	-
Student is Asian	0.0103	0.0013	7.762	0.000
Student is Black	-0.0026	0.0014	-1.803	0.071
Student is Hispanic	0.0018	0.0012	1.422	0.155
Student is White	0.0038	0.0011	3.475	0.001
With FRPL	-0.0036	0.0006	-5.553	0.000

The non-linear squared prior test score illustrates a statistically significant effect of -3.958261e-06. This negative coefficient portrays the concave association between the prior test score and current test score. It

implies that the change in the percentage of current test score for a unit increase in prior test score is increasing at a diminishing rate. The model fit indices is also illustrated significantly better performance with the AIC and the BIC decreasing by more than 10 points in comparison to the model without the non-linear term.[172] And looking now at the residuals of this model illustrates the following result.

Figure 8.15: Scatterplot of natural log transformed with non-linear prior test scores fixed effects residuals



As you can see, the un-modeled non-linear trend is no longer evident in the residuals. The revised model has effectively taken the non-linear pattern into account and the residuals now illustrate a much closer resemblance of the randomly distributed pattern indicated with the horizontal band. The linear regression line and the non-parametric curve are also more closely aligned. These findings both illustrate that there are no other obvious systemic pattern omitted in the model and left in the residuals. This finding is also illustrated with the CRplots of the prior test scores which also illustrate the randomly distributed horizontal band and the close conformity of the two regression curves as shown below. The plots indicate that there are no other important functional forms of the prior test scores (e.g. cubic or quadratic) which are omitted from the model. To sum, these findings all provide empirical evidence that the exogeneity condition is now better satisfied in the model.

[172] Raftery (2005) illustrated that the reduction of the BIC index by more than 10 points is considered significant effect/model improvement.

Figure 8.16: Component residual plot of natural log transformed with non-linear prior test scores fixed effects residuals and prior test scores

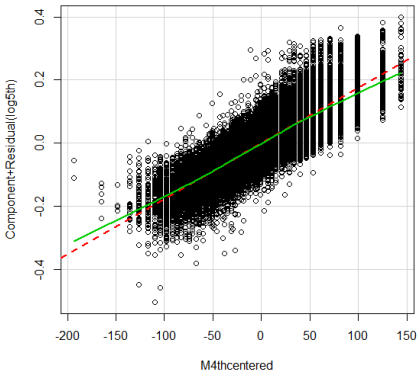
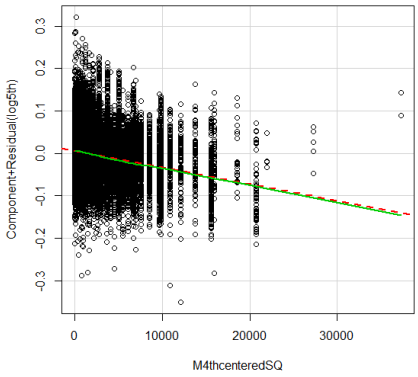
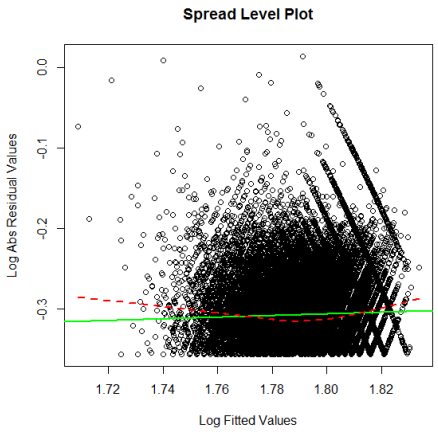


Figure 8.17: Component residual plot of natural log transformed with non-linear prior test scores fixed effects residuals and prior test scores squared



Now, to make sure that other forms of violation are not introduced, the revised model is re-diagnosed to illustrate skewness and kurtosis values of 0.2944 and 4.4405, respectively. These values are virtually equal to the values shown in the previously natural log transformed model. The spread of the residuals also remained constant and homoscedastic as shown in the parallel dotted lines along the red non-parametric curve in the scatterplot above and the flat horizontal spread level (SL) plot shown below. The linear regression line for the SL plot (in green) in fact has a flatter slope of 0.10495 than the previous case. The variance stabilizing transformation is now 1 – 0.10495 =0.89505, which again signifies no transformation (power of 1).

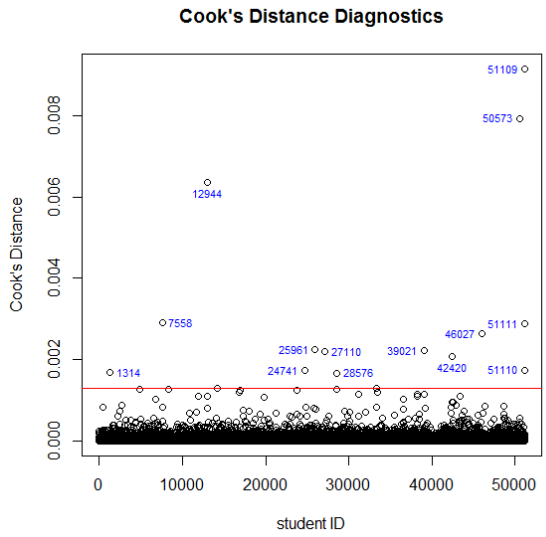
Figure 8.18: Spread level plot of natural log transformed with non-linear prior test scores fixed effects residuals



The ICC continues to illustrate positive value of 0.198 with high statistical significance with F statistics of 3.264 and p-value of 0.000. The positive dependency of the residuals within teachers has again been effectively taken into account in the model. The slight reduction in the ICC value in comparison to the previous natural log transformed model with 0.202 signifies that the dependency pattern of the residuals have been further taken into account (explained away) by the newly added squared prior test scores. That is, there is now less resemblance among the residuals within teachers. Thus the addition of the non-linear prior test score

did not affect the previously corrected violations. But looking now at the Cook's distance influential diagnosis, several cases which are potentially disrupting the estimates are identified as shown below

Figure 8.19: Cook's distance diagnostics natural log transformed with non-linear prior test scores fixed effects VAM



Setting 0.0013 as the cutoff value (as the red horizontal line), 20 observations were identified with index value above this cutoff. As you can see, a number of cases which were identified as influential in the original (un-revised) model are identified again as influential in the revised model. To handle these cases, the underlying features of these cases are studied as follow.

Descriptive analysis of the influential cases illustrated that the corresponding students all underperform the overall average performance of 408.5. There were slightly more male students and students not on FRPL.[173] The ethnicity variable, on the other hand, illustrated balanced frequency of White and non-White students.[174] And comparing the 4th grade and the 5th grade test scores, there were three students who indicated an abrupt and unrealistic change of over 100 points. This could potentially signify some miscoding or mis-measurement with respect to these students. But one of the key patterns underlying these cases was that the size of the class/teacher they belonged was extremely small. All the students belonged to a teacher/class with less than 5 students and many of them were in fact the only student belonging to his/her teacher. And this finding explains some of the reasons why these students exerted a lot of influence on the estimates. First, as described in Chapter 3, teacher fixed effects VA estimates (or the adjusted teacher means) are determined by the sample size within teachers' classes. Thus, the few students exerted a lot of influence in determining their respective teacher VA estimate. They had a high leverage on their teacher variable. Second, as also described in

[173] Given male students on average performed lower than female students, the deletion of these underperforming male students would decrease the gender gap. And as students on FRPL on average performed lower than other students, the deletion of these underperforming FRPL students would decrease the gap between the two groups.
[174] The deletion of the cases should therefore not change in the effects of ethnicity and the ethnicity gap.

Chapter 3, the efficiency and reliability (standard error) of the teacher VA estimates is inversely related to the teacher class size. With small sample sizes, the reliability of the teacher VA estimates is very low. The residuals which are calculated as the deviation of student performance from the teacher VA estimate (teacher adjusted means) are then also unpredictable and unreliable. It has a higher chance of taking extreme residual values. As the influential measure is calculated as the product of leverage value and residual value, these two conditions lead to the high influence measure for the 20 cases. And this association of highly influential cases and small group sizes was also found in the influential diagnosis analysis by Bosker and Snidjer (1997) using the secondary school data in the Netherlands. Having now understood the unreliability and instability associated with these cases, the natural log with prior test score model which excludes these cases is estimated to show the following results.

Table 8.8: Estimates of the natural log transformed with non-linear prior test score fixed effects VAM without influential cases

	Estimate	*SE*	*t-value*	*p-value*
Intercept	5.9628	0.0161	370.27	0.000
Prior Test Score	0.0018	0.0000	284.81	0.000
Prior Test Score Squared	0.0000	0.0000	-43.706	0.000
Student is Female	0.0057	0.0004	13.074	0.000
Student is Native American	-	-	-	-
Student is Asian	0.0103	0.0013	7.721	0.000
Student is Black	-0.0027	0.0014	-1.845	0.065
Student is Hispanic	0.0017	0.0012	1.382	0.167
Student is White	0.0038	0.0011	3.451	0.001
With FRPL	-0.0036	0.0006	-5.573	0.000

The estimates above changed very slightly from the original model in accordance to the direction indicated by the descriptive analysis. This is not surprising given the huge data set of over 50,000 students. Yet, the above findings imply that the estimates are more robust and stable than the original estimates. Re-diagnosing this model illustrates skewness and kurtosis values of 0.299 and 4.391 which as expected illustrate a slight improvement in the peaked and heavy tail problem. But in comparison to the natural log transformation (which affected the entire 51116 observations), this improvement (which only affects 20 observation) is very small. Looking now at the scatterplot and SL plot illustrates the following results.

Figure 8.20: Scatterplot of natural log transformed with non-linear prior test scores fixed effects residuals without influential cases

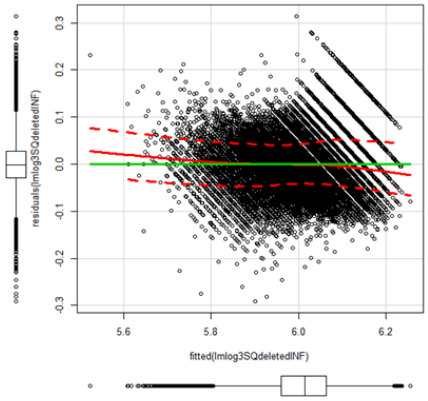
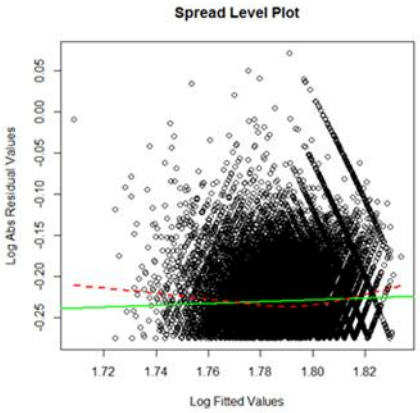


Figure 8.21: Spread level plot of natural log transformed with non-linear prior test scores fixed effects residuals without influential cases



As you can see, virtually the same result shown in the previous case (without excluding the influential cases) is illustrated. The horizontal band shaped pattern with the flat and merged linear regression and non-parametric curves signify that the residuals are randomly distributed. This implies that there are no obvious factors which can systemically explain the outcome variable is omitted from the model and the exogeneity condition is satisfied. The parallel dotted lines indicate the presence of constant homoscedastic variance of residuals. This is also supported with the SL plot which continues to show no pattern with a flat slope of 0.1063. The ICC also remained virtually the same with 0.194 with statistical significance with F statistics of 3.264 and p-value of 0.000. Thus the exclusion of the 20 influential cases did not heavily affect or introduce new forms of violations to the previously revised model. The estimates of the natural log transformed with squared prior test scores are found to be quite robust to the influential cases

Summary of the Diagnosis and Revision of the Basic Fixed Effects VAM

To sum, the above sections implemented the diagnosis and revisions of the preliminary basic fixed effects VA analyses conducted in the previous chapter. The diagnosis detected number of violations of the linear regression assumptions underlying the model. The preliminary findings were therefore subject to both bias and inefficiency and the statistical inference (generalizations and conclusions) based on these findings were misled. A variety of revisions were then implemented to correct the violations. For each revision, a complete re-diagnosis of the model was conducted to make sure new forms of violations were not introduced. This iterative procedure between revision and diagnosis led to a natural log transformed model with a non-linear prior test score term (with or without the influential cases) as the model which best satisfied all the linear regression assumptions simultaneously. All the factors left un-accounted in the model (residuals) were distributed randomly and each of the explanatory variables used in the model illustrated its unique and distinguishable variation. This model provided us with the best approximation of the BLUE estimates we longed for. And these estimates give us the utmost confidence and comfort to interpret and possibly apply the

findings. Given the natural log transformation, the coefficient estimates are now interpreted in terms of percentage change in the outcome variable for a unit increase in the explanatory variable. For example, being taught by a particular teacher will increase the average performance of the students by (magnitude of the teacher VA estimate)% in comparison to the overall adjusted mean. Finally, following suit with the previous chapter, the comparison of the main parameter estimate of interest – the teacher VA estimates across the different models is provided below.

Table 8.9: Correlation of VA estimates of the original and revised fixed effects VAMs

	Basic FE	Log Transf	Log Nonlinear	Log NonL No Influ
Basic FE	1			
Log Tranformed	0.9847	1		
Log Transformed Nonlinear	0.9872	0.9686	1	
Log Transf Nonlinear No Influential	0.9821	0.9593	0.9951	1

As you can see, the VA estimates are highly correlated across the models with values over 0.95. This consistency and similarity of the estimates illustrates that the different models identified similar values (and relative ranking) of the teacher VA estimates. This provides us with some relief and assurance in the model estimates i.e. the revision processes did not drastically or un-realistically alter the findings. But this does not imply that we can proceed with the original model estimates and forego the diagnosis and revision processes. It is only the final natural log transformed with non-linear prior test scores (with or without influential cases) which ensures us the BLUE estimates. It is only this final model which gives us the true confidence and assurance to conduct statistical inference and possibly apply the findings for policy relevant purposes.

Diagnosis and Revision of the Basic Random Effects VAM

As described in Chapter 4, in comparison to fixed effects model, random effects model must satisfy additional assumptions in order to ensure the BLUE estimates. Specifically, the residuals defined at each level of the data structure (student and teacher level) must be randomly distributed and the explanatory variables also defined at each level of the data structure must have its unique and distinguishable variation with no multicollinearity problem. In diagnosing these assumptions, the seminal work by Hilden-Minton (1995) has shown that residuals defined at level-1 can be estimated such that they are un-confounded by the residuals at level-2 but the other way around is not possible.[175] The correct specification at level-1 must take precedence over the specification at level-2.[176] In the following sections, the diagnosis and revision of the level 1 model specification (student level) of the basic random effects VAM is first conducted. The diagnosis and revision of the level 2 model specification (teacher level) then follow. The Hausman test which can evaluate the

[175] This is intuitively makes sense as in the case of the basic 2-level model, the random intercepts or slopes values are essentially the outcome of fitting the level-1 specification for/by all the groups. The level-2 specification then tries to explain the heterogeneity in these values across groups with the second level covariates.

[176] This procedure is supported in the work by Raundenbush and Bryk (2002) and Snidjers and Bosker (1999).

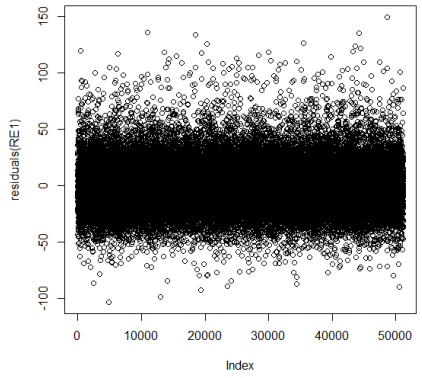
exogeneity condition of the level 2 residuals will also be conducted. Finally, comparison and contrast of the teacher VA estimates across the different models will be conducted.

Diagnosis of the Level-1 Specification of the Basic Random Effects Model

Diagnosis of the E(e/X) = 0 Condition and the Normality Assumption

Following suit with the basic fixed effects VAMs diagnosis, the plot of the level-1 residuals by its unique student identification index is provided as follow.

Figure 8.22: Plot of level-1 basic random effects residuals



As you can see, the plot indicates the high concentration of the residuals around zero providing strong indication that the expected/average value of the residual is zero (*E(e/X)=0*). It illustrates that the model is capturing and fitting the actual data observations well and no obvious signs of important systemic patterns left un-accounted in the residuals. A solid initial sign of randomly distributed residual pattern is evident. To examine entire distribution of the residuals, the following histogram and qqplot are constructed.

Figure 8.23: Histogram of level-1 basic random effects residuals

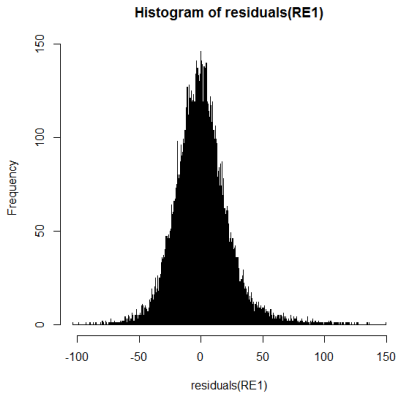
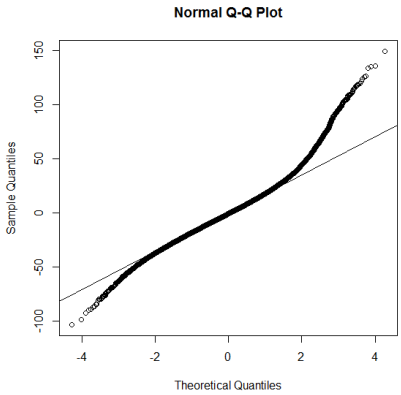


Figure 8.24: Normal quantile-quantile plot of level-1 basic random effects residuals



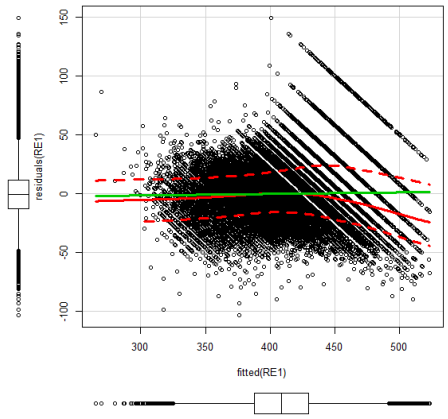
The histogram illustrates a symmetric and bell shaped curve in line to the normal distribution. As described earlier, this signifies that the model is making less frequent larger size errors and equal and balanced frequency of negative and positive misfits. That is, obvious systemic patterns are not illustrated in the errors. The

distribution is also centered around zero providing further evidence of the *E(e/X)=0* condition. Yet, closer examination of the distribution again reveals the violation of the normality assumption with the presence of a peaked and heavy tails. This is illustrated with a kurtosis value of 5.27651 and number of residuals deviating from the 45 degree line (signifying the normal distribution) toward the lower and upper end of the qq plot.[177] Thus the basic random effects model is also subjective to both bias and inefficiency engendered by the peaked and heavy tailed distribution.

Diagnosis of the Exogeneity Condition

To examine whether the levels of the residuals entail no systemic pattern with any variables, the following scatterplot with the fitted values are constructed.

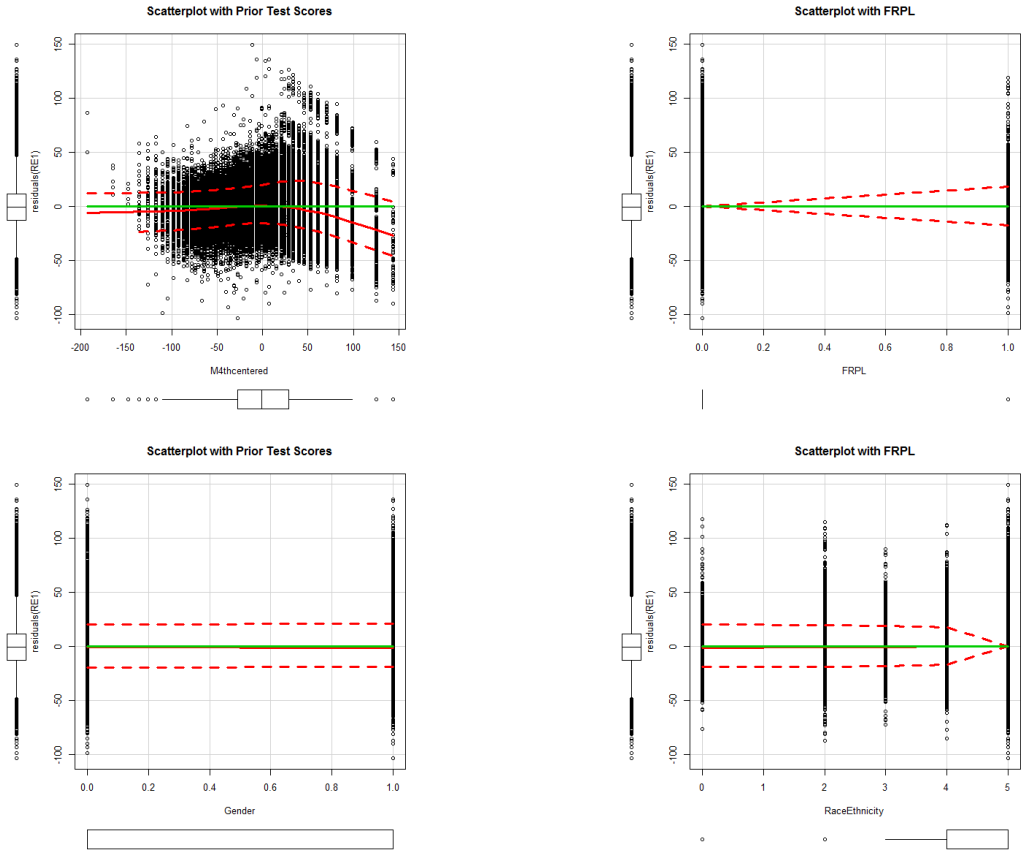
Figure 8.25: Scatterplot of level-1 basic random effects residuals



As you can see, the non-parametric (in red) clearly deviates from the linear regression line (in green) line to detect an un-modeled non-linear pattern in the residuals. The horizontal band shaped pattern is not illustrated and the residuals are not randomly distributed. Moreover, the non-linear trend is misattributed to the linear effect of the respective variable (or any other effects which are correlated to the non-linearity). The exogeneity condition is violated and bias is evident in the basic random effects model. To investigate the potential source of this nonlinear pattern, the scatterplot is reconstructed for each explanatory variable underlying the fitted model to show the following results.

[177] The skewness value is 0.567439 which is not much different from the value of 0 of the normal distribution.

Figure 8.26: Scatterplot of level-1 basic random effects residuals by student background variables

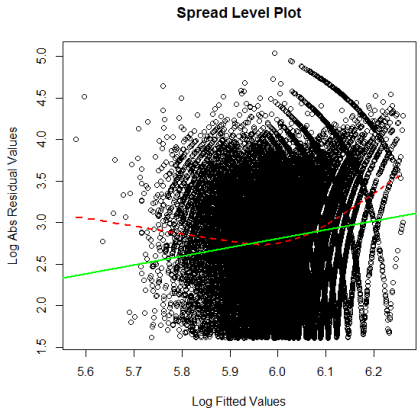


As you can see, very similar findings illustrated with fixed effects model are shown. The un-modeled non-linear pattern is clearly attributed to the prior test scores. And the categorical variables illustrate no omitted systemic pattern with the residuals as the two regression lines are both flat with slope of zero.

Diagnosis of the Homoskedasticity Assumption

To examine whether the spread of the residuals illustrate no systemic pattern (constant homoscedastic variance), the red dotted lines of the scatterplot plot (which illustrate the 95% spread of the residuals around the red non-parametric curve) and the spreads level plot are examined. Looking first at the scatterplot (with the fitted values shown above), the red dotted lines expands outwards toward the upper end of the fitted values. There is an increasing systemic trend in the spread and homoskedasticity assumption is violated. This is also confirmed with the SL plot (shown below) which also illustrates an increasing pattern with a slope of 1.04973. Thus with the violation of the homoskedasticity assumption, inefficiency is evident in the basic random effects model.

Figure 8.27: Spread level plot of level-1 basic random effects residuals



Diagnosis of the Independence Assumption

To examine whether the residuals themselves illustrate any systemic pattern, relation and resemblance, the intra-class correlation coefficient (ICC) of the basic random effects model is calculated to illustrate a value of 0.108. This implies that the observations/residuals within teachers are again on average positively correlated. And to test the statistical significance of this dependency, the likelihood ratio (LR) test which compares the log likelihood of the model with the teacher VA parameter (the basic random effects model) and without the teacher VA parameter is conducted to illustrate high statistical significance with LR statistics of 2255.08 and p-value of 0.000.[178] Alternative statistical test based on the BIC index suggested by Raftery (2005) is also conducted to illustrate a large reduction in the BIC value 2408 points for the model with the teacher VA parameter.[179] Thus, the teacher VA parameter and the dependency/clustering pattern (engendered by the teacher VA parameter) are statistically significant. And as clarified above, this significant dependency pattern has (already) been taken into account the basic random effects VAM through the teacher random VA effects. The constrained model without the teacher VA parameter which ignores the dependency is susceptible to the inefficiency in the estimates and misleading statistical inference. The ICC findings above therefore give the empirical evidence and justification to conduct the random effects VAM.

Diagnosis of the Outlying and Influential Residuals

The Cook's distance influential diagnosis command for random effects model (which is based on the linear mixed model framework) which can operate under a large data set (like the Washington data) is currently not

[178] The likelihood ration (LR) test is analogous to the general F test described earlier. The LR test is based on the comparison of the log likelihoods of the two models while the F test is based on the R squared values. The underlying idea and intuition is the same but the LR test enable us to conduct statistical tests for more complex models and parameters e.g. the different kinds of non-constant error variance and covariance. The F test on the other hand can only be used for simple multiple regression models. The LR test is asymptotically equivalent to other likelihood based tests such as the Wald test and Lagrange Multiplier (LM) test.

[179] As addressed earlier, Raftery (2005) has illustrated that difference/reduction of more than 10 BIC points is considered statistically significant effect/improvement in model fit.

available with the R statistical package.[180] But as it was shown earlier in the influential diagnosis of fixed effects model, the exclusion of the small number of cases (20 or so) did not alternate the estimates at all. As advised and confirmed with Bosker and Snidjers over email, for a large data sets like the WA State data where the ratio of the number of level 1 units (students) and the level 2 units (teachers) is very large, influential diagnosis for level 1 residuals is not an effective strategy as taking into account of the few observations merely changes the results. Instead, the scholars addressed the importance of conducting level 2 influential diagnosis. In doing so, highly influential level 2 residuals are often attributable to groups with very small sample sizes which are often comprised of few outlying level 1 units. This finding was also illustrated in fixed effects diagnosis where the influential cases (students) all belonged to very small teachers/classes (all had less than 5 students in the class). The outlying level 1 unit within these small sample sized influential level 2 units (or the entire group if it only has one student sample) can then be taken into account to revise the model. Following suit with this advice (and as a substitute for the influential diagnosis), the sensitivity analysis the model estimates with respect to the small sample size of level 2 units will be thoroughly conducted in the following chapter.

Diagnosis of the Properties of the Level 1 Explanatory Variables

The variables used for the level 1 specification in the basic random effects model are identical to fixed effects model. The same findings shown in fixed effects model with the clear variation within each variable with no sign of multicollinearity equally apply to the random effects model

To sum, the diagnosis of the level 1 residuals of the basic random effects model illustrated very similar violations shown in fixed effects model. As the assumptions necessary for the random effects model to provide BLUE estimates were not fully met, the estimates of the basic model are both biased and inefficient. The statistical inference based on this model is thus misleading. In order to correct these violations, the following section implements the different revisions introduced earlier. And in doing so, re-diagnosis of the model for each revision will be carefully conducted. This iterative process between (re)diagnosis and revision will continue until all the assumptions are jointly met in the model.

Revision of the Level 1 Specification of the Basic Random Effects VAM

Revision of the Peaked and Heavy Tailed Non-Normal Distribution and the Non-Constant Variance of the Residuals

To correct for the non-normal distribution characterized with a peaked and heavy tails and the non-constant variance of the residuals, the natural log transformation of the outcome variable is implemented. The skewness and kurtosis values of this revised model illustrate the following results.

[180] This point was confirmed with Tom Snidjers over email who contributed to the design of the R command. Further development of the influential diagnosis commands for multilevel data sets is currently underway.

Table 8.10: Skewness and kurtosis values of the natural log transformed and basic (original) random effects level-1 residuals

Revised Transformed Model		Original Model	
Skewness	0.312536	Skewness	0.567439
Kurtosis	4.545761	Kurtosis	5.276509

As you can see, the severe peaked and heavy tails has receded and the slight positively skew has also improved. The residuals have taken a huge step towards the normal distribution. Looking now at the scatterplot of the residuals with respect to fitted values and the spread level (SL) plot illustrate the following results.

Figure 8.28: Scatterplot of natural log transformed level-1 basic random effects residuals

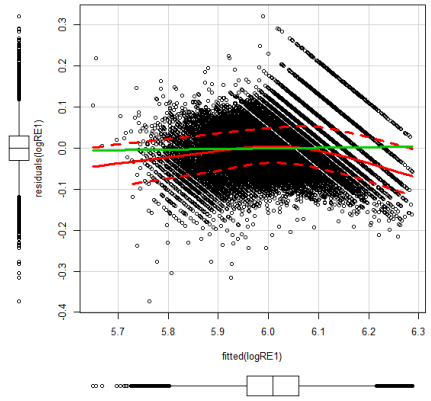
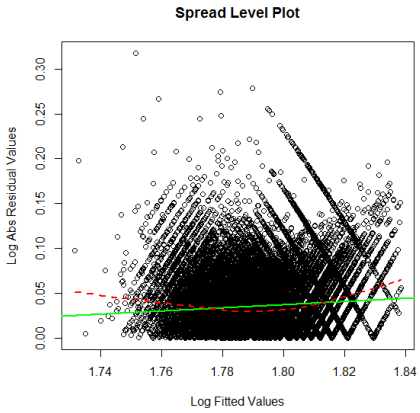


Figure 8.29: Spread level plot of natural log transformed level-1 basic random effects residuals



As you can see from the dotted non-parametric lines, the increase in the spread of the residuals particularly toward the upper end of the fitted values has now disappeared. The dotted lines are now more parallel to the non-parametric regression curve. The SL plot has also flattened with a slope of 0.17028 compared to the 1.04973 of the original model. And based on the Tukey rule, there is no longer a need for variance stabilizing transformation of 1 – 0.17028, which is very close to 1 (no transformation). These findings both illustrate that the homoskedasticity assumption has now been retained in the model. The natural log transformation has again fulfilled its purposes. Now, re-diagnosing the other assumptions, the ICC value of the revised model continues to illustrate a positive value of 0.112 with high statistical significance with LR statistics of 2328.16 and p-value of 0.000 continues to be illustrated.[181] The slight increase in the ICC value is again in line to our expectation as the natural log transformation reduced the magnitude of the residuals with extreme values, making it more similar to the rest of the residuals and thus increasing the dependency. But for the exogeneity condition, it continues to be violated as described next.

[181] The statistical significance is confirmed with the BIC value which is smaller by 2447 in comparison to the corresponding revised model without the teacher VA.

Revision of the Violation of the Exogeneity Condition

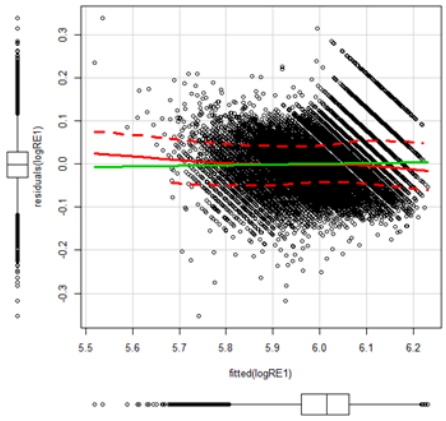
Looking at the above scatterplot again, the un-modeled non-linear function continues to be clearly evident. An important systemic pattern continues to be left un-accounted in the residuals and the exogeneity condition is violated. As the diagnosis above illustrated that this systemic pattern is most likely due to the prior test score, we now incorporate a nonlinear function of prior test score (with power raised to 2) into the log transformed model to show the following results.

Table 8.11: Estimates of the natural log transformed with non-linear prior test score random effects VAM

	Estimate	*SE*	*t-value*
Intercept	6.0100	0.0011	5381.0
Prior Test Score	0.0018	0.0000	303.0
Priror Test Score Squared	-0.00000398	0.0000	-45.0
Student is Female	0.0059	0.0004	14.0
Student is Native American			
Student is Asian	0.0110	0.0013	8.0
Student is Black	-0.0020	0.0014	-1.0
Student is Hispanic	0.0012	0.0012	1.0
Student is White	0.0043	0.0011	4.0
Student on FRPL	-0.0039	0.0006	-6.0
Variance Components:			
Between Teacher Variance	0.0003	0.017362	
Within Teacher Variance	0.0023	0.048397	
AIC	161399		
BIC	161302		

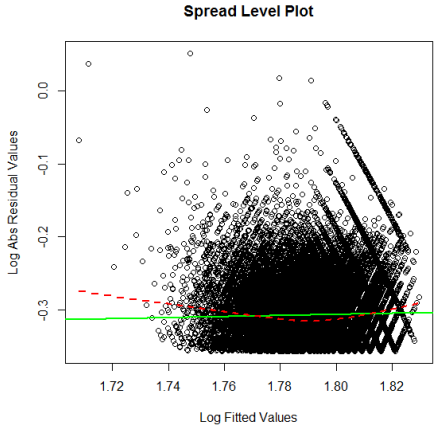
As you can see, the nonlinear prior test scores have a significant negative effect on student performance. The increase in the current test score for a unit increase in prior test score is increasing at a diminishing rate. This is represented with a concave non-linear association which is consistent the un-modeled systemic pattern in the residuals. The model fit with the squared term also shows clear significant improvement with both the AIC and BIC indices decreasing by more than 10 points to 161399 and 161302 from 456751 and 456839 of the original basic model. The decrease in the effects of the linear prior test scores also signifies how the non-linear effect was misattributed to this variable prior to the revision. This improvement is also reflected in the following scatterplot of the residuals.

Figure 8.30: Scatterplot of natural log transformed with non-linear prior test scores level-1 basic random effects residuals



As you can see, there is no longer the un-modeled non-linear trend. There is now a better conformity of the non-parametric curve and the straight linear model line which indicates that there are no obvious signs of other un-modeled systemic patterns in the residuals. The regression curves are both flat and the horizontal band which symbolizes the randomly distributed pattern is illustrated. Now, to make sure other forms of violations are not introduced, the skewness and kurtosis is re-calculated to be 0.312536 and 4.045761 which is virtually the same as the previous natural log transformed model. The normal distribution pattern is retained. The ICC also illustrates very similar value to the previous log transformed model with 0.114 and high statistical significance with LR statistics of 2472.67 and p-value of 0.000.[182] And as evident from the parallel red dotted lines of the scatterplot and the flat SL plot (with slope of 0.07343) shown below, the constant homoskedstic variance is also maintained.

Figure 8.31: Spread level plot of natural log transformed with non-linear prior test scores level-1 basic random effects residuals



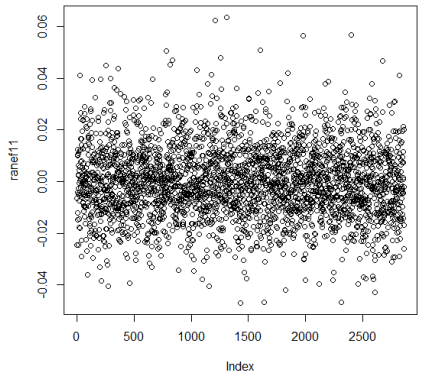
[182] The statistical significance of dependency/ICC is also confirmed with the reduction in the BIC value of 2578 in comparison to the model without the teacher VA grouping variable.

To sum, the diagnosis and revision of the level 1 specification of the basic random effects model lead to a natural log transformation with non-linear prior test score as the final model specification which best satisfied all the assumptions necessary to provide the BLUE estimates.[183] But as addressed above, this completes only the first half of the diagnosis and revision process of random effects model. We must now repeat the above analysis for the residuals and variables defined at the second level of the data structure (teacher level). Given the final level 1 revised model, we now proceed to the diagnosis and revision of the level 2 specification.

Diagnosis of the Level-2 Specification of the Level 1 Revised Basic Random Effects Model

For random effects model to ensure the BLUE estimates the residuals defined at the second level of the multilevel data structure must also be randomly distributed and the explanatory variables defined at this level must also illustrate unique variation with no multicollinearity problem.[184] Following suit with the level 1 diagnosis, the plot of the level-2 residuals by its unique teacher identification index is first provided as follow.

Figure 8.32: Plot of level-2 residuals of level-1 revised basic random effects



The plot indicates a high concentration of the residuals around zero providing strong indication that the expected/average value of the level 2 residual is zero ($E(aj/X) = 0$). The basic model seems to effectively capture the variation of the outcome variable defined the second level of data structure. To examine the overall distribution of the residuals, the following histogram and qqplot are constructed.

[183] Other more advanced methods e.g. the Huber White M-Estimation is also available to even better improve the heavy tails (high kurtosis values). This method was again attempted with the random effects model but it failed to provide converging estimates due to the large sample size. The implausibility of implementing this method was confirmed with John Fox over email who developed the R command. Further work on these methods will be left for the future together with the improvement of the R software.

[184] Now in accordance to the explanation and mathematical presentation of random effects model provided in Chapter 4, the level two between teacher variation was modeled with the single variance component (σ_a^2). Unlike the fixed model and the level 1 residuals shown above, the residuals for the second level cannot be simply calculated by taking the difference of the actual observed values and the model fitted values. Instead, the level 2 residuals must be estimated with the empirical Bayes predictions (of the random intercept term) after taking into account the variables in the level 1 specification. The empirical Bayes predictions can be used to represent the teacher VA estimates but it can also be used (in fact was initially used) for this purpose of diagnosing the assumptions underlying random effects model.

Figure 8.33: Histogram of level-2 residuals of level-1 revised basic random effects

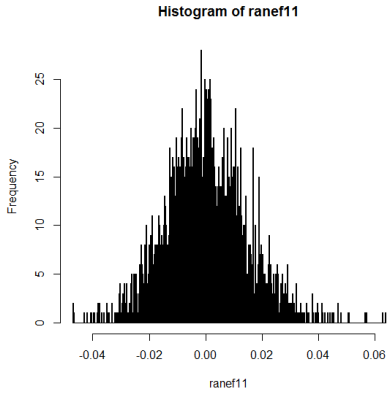
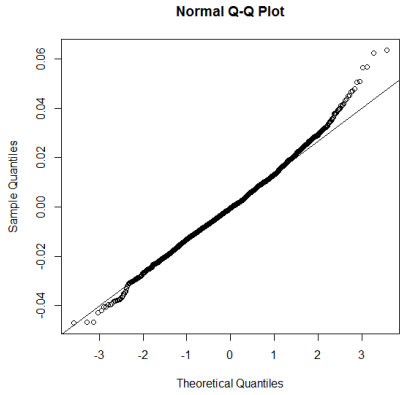


Figure 8.34: Normal quantile-quantile plot of level-2 residuals of level-1 revised basic random effects



The histogram of the level 2 residuals also illustrates a symmetric and bell shaped curve closely in line to the normal distribution. The skewness and kurtosis values are calculated to be 0.2416815 and 3.667213 which are virtually equivalent to the normal distribution. Unlike the level 1 residuals, peaked and heavy tailed distribution is not evident. This finding is confirmed with the qqplot shown on the right where majority of the residuals are closely aligned along the 45 degree line. To examine the homoskedasticity and exogeneity assumptions, the following scatterplot and spread level plots are constructed.

Figure 8.35: Scatterplot of level-2 residuals of level-1 revised basic random effects

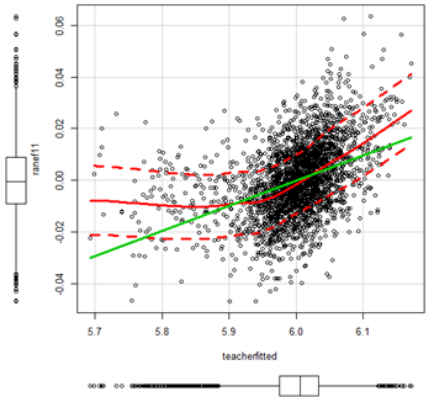
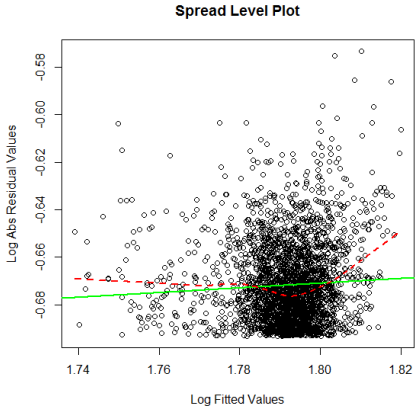


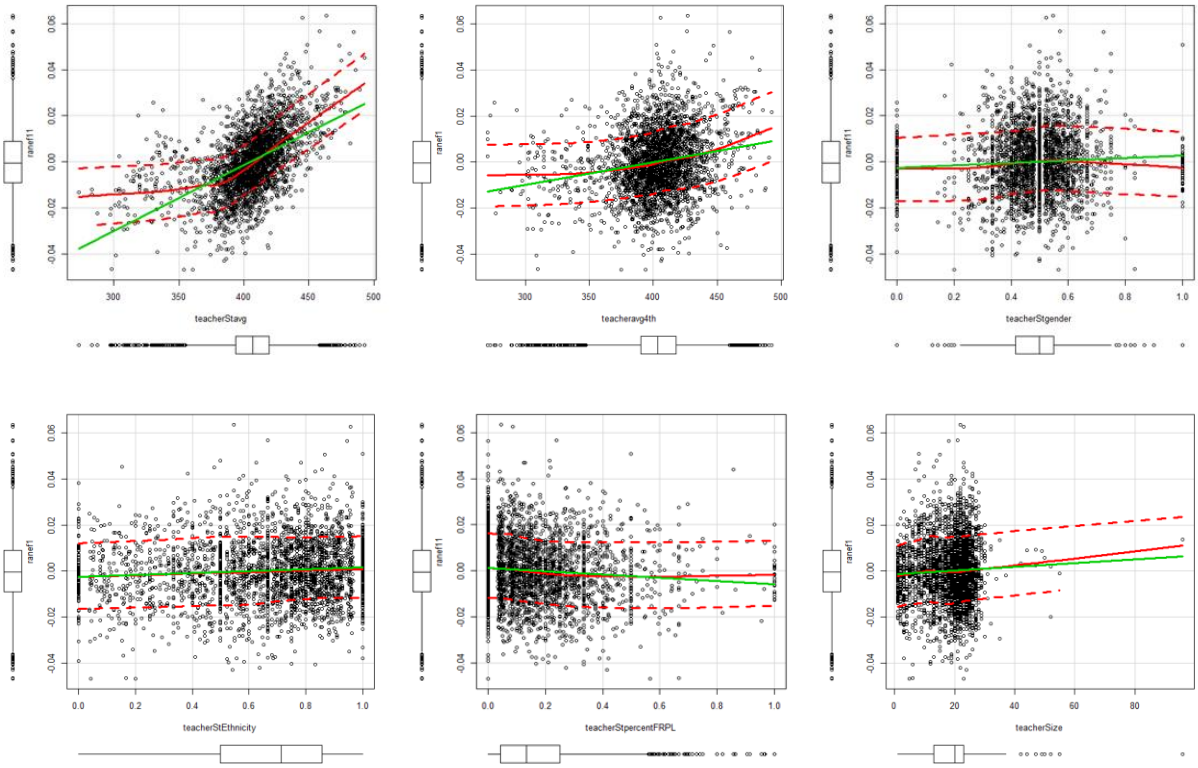
Figure 8.36: Spread level plot of level-2 residuals of level-1 revised basic random effects



Looking first at the dotted non-parametric curves, there is no clear increasing or decreasing trend in the spread of the residuals. The dotted lines are parallel to the non-parametric regression curve and the SL plot illustrates no systemic trend with a flat linear regression line with slope 0.09962. Both of these findings illustrate that the constant homoscedastic variance assumption is satisfied in the model. But for the exogeneity condition, as you may have noticed from the scatterplot on the left, there is a clear increasing and possibly non-linear systemic trend in the residuals. The horizontal band characterizing the randomly distributed pattern is clearly not indicated. As this systemic pattern is not taken into account in the model, the model is not only failing to capture the underlying pattern of the data (poor model performance) but the systemic effect is misattributed to the variables which are correlated to this pattern. The exogeneity condition of $E(ajX)= 0$ is violated and bias is

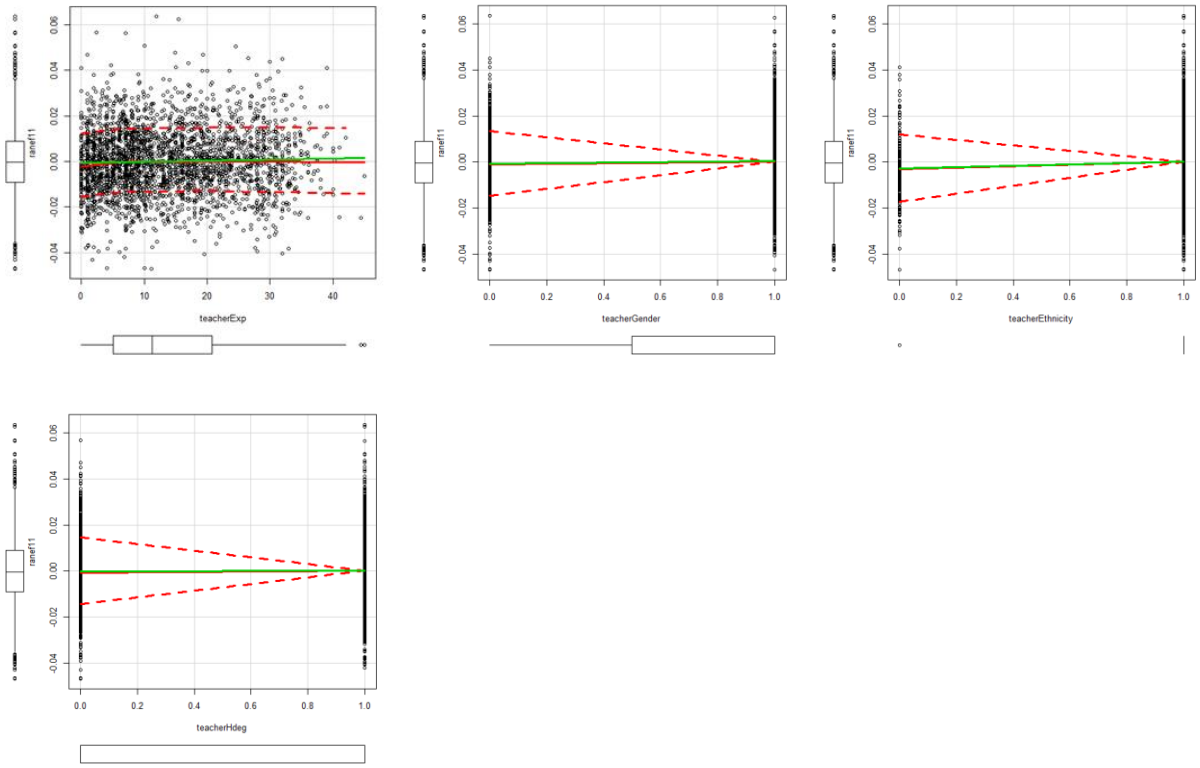
introduced in the model. To investigate the source of this unaccounted systemic pattern, the above scatterplot is re-constructed for the two sets of teacher level variables introduced in the previous chapter in conducting the extended VAMs – the student peer effects variables (teacher level averages of the student background variables) and teacher characteristics variables. The results are as follows, starting with the student peer effects – from left to right: teacher/class average current (5^{th}) test scores, teacher/class average prior (4^{th}) test scores, percentage of female students in class, percentage of White students in class, percentage of FRPL students in class, and teacher class size.

Figure 8.37: Scatterplot of level-2 residuals of level-1 revised basic random effects by teacher level variables



And for the teacher characteristics variables the scatterplots are illustrated as follows, from left to right: teachers' experience, teachers' gender, teachers' ethnicity, and teachers' education level (whether or not with a MA degree).

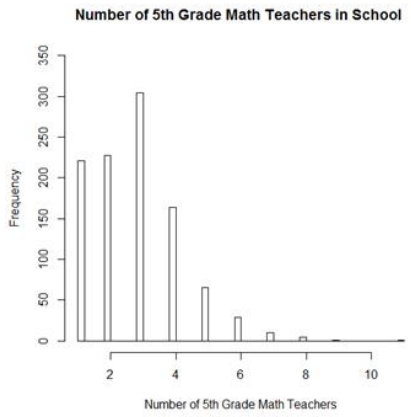
Figure 8.37: (continued)



From these figures, it is evident that the un-modeled increasing pattern is most likely due to the teacher average current (5^{th} grade) test scores and/or the teacher average prior (4^{th} grade) test score. But looking at other plots, although the magnitude is not as profound as the average test scores, systemic trends are also evident. The horizontal band with matching linear regression and non-parametric curves are not clearly identified. For example, percent of female students illustrate a slight non-linear trend, the percentage of FRPL students illustrates a slight weak decreasing trend, and teacher class sizes illustrate an increasing pattern. Furthermore, although hard to envision, the conditional mean residual values for the different categories of the teacher characteristics variables were also found to be different. In a separate analysis, simple linear regression of the residuals with each of the teacher characteristics variables was conducted to illustrate the statistically significant differences. Thus these findings illustrate that both the peer effects variables and teacher characteristics variables are also contributing to the un-modeled systemic trend left in the residuals. And this also provides us with the empirical justification to conduct the extended VAMs (estimated in the previous chapter) as the means to correct the violation of the exogeneity condition of the basic random effects model. Finally, to examine the possibility of dependency among the level 2 residuals, the ICC of the residuals (the teacher VA estimates) within schools was calculated to illustrate a positive value of 0.253 with statistical

significance with LR statistics of 176.77 and p-value of 0.00.[185] This implies that the teacher VA estimates of the 5th grade math teachers working in the same school are similar and correlated. The random distributed pattern of the level 2 residuals is not achieved and we should incorporate the school grouping variable in order to take into account of this dependency pattern. But before we consider this modeling option, further diagnosis of this finding raises several points of concern. First, the reliability of the between school and within school variance components (underlying the ICC) is very low as it is based on very small number of observations/teachers. As shown in the histogram below, there are only few 5th grade math teachers working in the same school. The mean is 2.79 teachers and the mode is 3 teachers.

Figure 8.38: Histogram of school teacher size
(number of 5th grade math teachers)



As the variance components are based on the number of 5th grade math teachers working in each school, this questions the reliability, precision, and stability of the estimates. For example, 1 teacher who happens to achieve unusually high VA (possibility due to small number of outlying student observations) can heavily distort the variance estimates. The low reliability of the variance estimates also explains the relatively low significance value of the ICC in comparison to the student and teacher level ICC calculated earlier.[186] Second, as it will be explained later in this chapter (and in the Appendix), the diagnosis of the school fixed model in the previous chapter introduced new violation in the level 2 residuals assumption. Namely, a complex heteroskedastic pattern was illustrated and inefficiency was introduced. The unstable variance pattern can be potentially due to the poor fit of the school fixed effects (large residuals) originating from the small number of observations to estimate each school effect. Thus this finding together with the small teacher sample size in schools demand further attention and potentially more complex analysis such as the generalized least squares to model the heteroskedasticity pattern and Bayesian statistics which can respond to the small sample size problem by means of simulation. Few other issues regarding incorporating the schools (as the third level) are

[185] The level 1 unit is now the teachers with its teacher VA estimates and level 2 unit (the grouping variable) is the schools. The BIC value in comparison to the model without the school VA grouping variable also illustrated a reduction of 168 points which confirms the statistical significance of the dependency/ICC.

[186] The level 1 revised model illustrated significance of ICC at the student and teacher level with LR statistics of 2472.67 which is close to 15 times larger than the teacher and school level ICC of 177. The reduction in the BIC value of 2578 at the student and teacher level is also approximately 15 times larger than the teacher and school level with 168.

also described in the footnote.[187] Given these challenges and points of concern, school effects were not considered in this study. Responding to the challenges will be left as the future next step.

To sum, the diagnosis of the level 2 residuals illustrated that the normal distribution and the homoskedasticity assumptions were effectively satisfied. Yet, a clear violation of the exogeneity assumptions was illustrated. All the systemic patterns at the level 2 specification were not taken into account in the model and were left in the residual. And this consequently introduced further bias in the unrevised models. That is, the statistical inference based on the original basic random effects model or the level 1 revised (only) model are both misleading. Based on these findings, the following section conducts the revision of the level 2 specification in order to better meet the necessary assumptions to ensure the BLUE estimates.

Revision of the Level 2 Model Specification of the Level 1 Revised Random Effects VAM

In light of the above diagnosis, the two groups of teacher level variables are incorporated into the level 1 revised basic random effects model in attempt to revise the violation of the exogeneity condition. But as you may have noticed, the addition of these variables also defines the extended VAMs analyzed in the previous chapter. Thus, the extended VAMs not only responds to different types of research questions but also serve the important role of correcting the violation of the exogeneity condition shown in the basic random effects model. And as we proceed to these extended VAMs, we must now diagnose the second major half of the regression assumptions – the behavior of these teacher level explanatory variables. The teacher level variables just like the student level variables must illustrate its unique and distinguishable variation with no multicollinearity problem. Looking at the descriptive summaries of the teacher level variables in Chapter 6 again, clear variation within each of the variables is illustrated. This is represented with the non-zero values for the standard deviation and range. And looking now at the following correlation table, the associations between the different variables are illustrated.

[187] Other potential issues and challenges in incorporating the school variable are as follow. First, under the random effects model framework, incorporating school grouping variable as the level 3 random effects introduces a new set of assumptions in ensuring the BLUE estimates. Namely, just like the student and teacher residuals/effect, the school (random) effects must also be randomly distributed with no systemic pattern in its level, variance, and covariance. But the development and understanding of this complex three level hierarchical model is still an ongoing research. Statistical software such as HLM, SASS, and STATA are capable of accommodating these models but the R software is yet to incorporate these models. Moreover, the understanding of how the assumptions could interact/influence across the different levels is lacking and the diagnosis tools to test and revise the different assumptions are also not well developed. Second, in calculating the ICC at the teacher and school levels, only the 5th grade math teachers were considered. Constraining the teachers only to the math teachers is another key reason for the small sample size problems. Dependency in teacher performance can also arise from their association (interaction, peer effects) with other (non-math) teachers in the school. For example, Ponisciak and Byrk (2005) conducted a multivariate analysis using both reading and math scores and teachers to find positive correlation of math teacher VA estimate and reading teacher VA estimates. With the inclusion of the reading teachers, the sample size to reliably calculate the ICC can increase. As reading data is also available in the WA data set, this extension can be considered in the future. Finally, further investigation and literature review on the different mechanisms underlying teachers' selection of schools (or being selected by the schools) will also improve our understanding and analysis of the dependency pattern. This will enable us to identify the key factors which determine/explain the dependency problem. And through incorporating these factors we can further model (explain away) the dependency pattern and alleviate the violation of the independence assumption.

Table 8.12: Correlation of teacher level variables

	T Avg Current	T Avg Prior	% FRPL St	% White St	% Female St
Teacher Avg Current Scores	1				
Teacher Avg Prior Scores	0.91879	1			
% of FRPL Students	-0.35975	-0.39200	1		
% of White Students	0.31396	0.34027	-0.46086	1	
% of Female Students	0.19893	0.18976	0.01112	-0.04418	1
Teacher is Female	-0.00210	-0.02134	0.01432	-0.02010	-0.00416
Teacher is White	0.08125	0.07103	-0.05964	0.15347	-0.00854
Teacher with MA Degree	0.03981	0.03962	-0.03044	0.06737	0.00135
Teacher Experience	0.11071	0.10904	-0.08208	0.11589	0.03006
Teacher Class Size	0.29896	0.30102	-0.08375	0.07370	0.17195
	T is Female	T is White	T with MA	T Experience	T Class Size
Teacher is Female	1				
Teacher is White	0.02214	1			
Teacher with MA Degree	-0.01194	0.02624	1		
Teacher Experience	-0.03279	0.05339	0.12327	1	
Teacher Class Size	-0.02847	0.01834	0.03993	0.02968	1
Note: T = teacher					

Majority of the variables illustrate moderate to low correlation values with less than 0.50. But as highlighted in purple, there is a very high correlation between the two teacher level average test scores. The correlation is 0.91879 and this introduces the problem of multi-collinearity. The regression model will not be able to distinguish the difference between these two variables to provide reliable and unique converging estimates.[188] Thus in light of this problem, two sets of revised extended models with one centered on the average 5th grade test scores and another on the average 4th grade test scores are analyzed. The estimates of these two models are as follow.

Revised Models Using the Teacher Average 5th Grade Test Scores

The extended model using the teacher average current (5th grade) test score together with peer effects and teacher characteristics variables are estimated to show the following results.

[188] To validate this point, an extended model which incorporated both the average 5th grade test score and the 4th grade test score was estimated to show no converging estimates. The statistical software could not provide converging estimates.

Table 8.13: Estimates of the level-2 (and level-1) revised random effects VAM using teacher average 5th grade test scores

	Estimate	*SE*	*t value*
Intercept	5.668	0.0071	801.1
Student Characteristics:			
Student is female	0.0060	0.0004	13.60
Student is Native American	-	-	-
Studet is Asian	0.0089	0.0013	7.00
Student is Black	-0.0020	0.0014	-1.50
Student is Hispanic	0.0020	0.0012	1.70
Student is White	0.0043	0.0011	4.10
Student on FRPL	-0.0044	0.0006	-6.80
Prior Test Score	0.0017	0.0000	278.5
Prior Test Score Squared	0.0000	0.0000	-44.3
Teacher Characteristics:			
Percentage of FRPL Students	0.0214	0.0024	9.00
Percentage of White Students	-0.0088	0.0015	-5.90
Percentage of Female Students	-0.0043	0.0027	-1.60
Teacher Avg Current Test Score	0.0008	0.0000	49.30
Teacher Class Size	-0.0001	0.0000	-1.80
Teacher is Female	0.0007	0.0007	1.10
Teacher is White	0.0013	0.0012	1.10
Teacher with MA Degree	0.0002	0.0006	0.40
Teacher Experience	0.0000	0.0000	-1.10

Note: Prior test score and its squared function are grand mean centered

Variance Components:	
Between Teacher Variance	0.000089
Within Teacher Variance	0.002342
Intraclass Correlation	0.036718
AIC	163203
BIC	163027
Number of Students	51161
Number of Teachers	2864
Degrees of Freedom	51142

The average current test score has a positive effect with a very high significance level of 49.3. It effectively captured the un-modeled increasing systemic pattern underlying the un-revised level 2 residuals. Other peer effects variables also illustrate some significant effects but at a much less degree of significance. The teacher characteristics variables, on the other hand, do not show much significant effect.[189] Now, to further investigate whether the revised model served its purpose and did not introduce new forms of violations, the residuals of the revised model are re-diagnosed to illustrate the following results.

[189] The small magnitude and low significance level of these variables can potentially be attributed to the indirect effect through the average current test scores. When the average current test scores were excluded from the model, the magnitude and significance of the estimates were much bigger.

Figure 8.39: Scatterplot of level-1 residuals of the level-2 revised random effects using teacher average 5th grade test scores

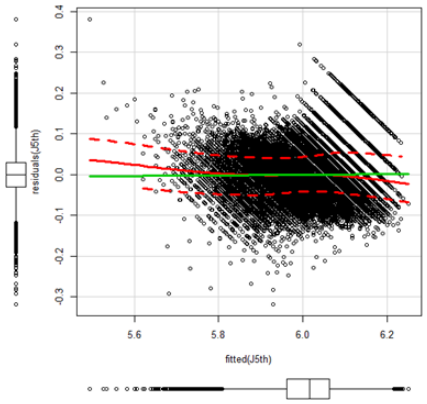
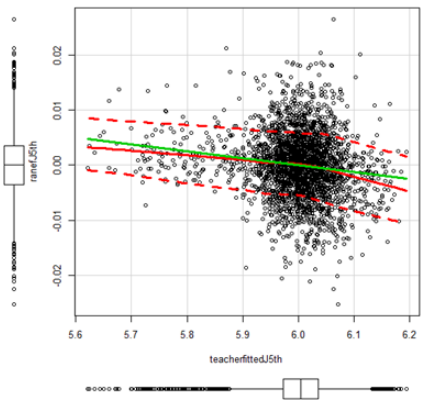


Figure 8.40: Scatterplot of level-2 residuals of the level-2 revised random effects using teacher average 5th grade test scores



As you can see from Figure 8.40 on the right, the upward trend which underlined the un-revised level 2 residuals has now disappeared. This illustrate that the revised model effectively served its purpose to capture the un-modeled systemic trend. The figure illustrates a slight downward trend but the slope of the green regression line is fact estimated to be of -0.01269 which is much closer to zero in comparison to the slope of 0.09962 of the unrevised (level 1 revised only) model.[190] And for other assumptions, the normal distribution pattern is retained with skewness and kurtosis of 0.02299315 and 3.848716 and the constant variance pattern is mostly achieved with a parallel non-parametric dotted lines. And finally, the level 1 residuals as shown in Figure 8.39 continue to illustrate a good approximation of the randomly distributed pattern achieved earlier. The skewness and kurtosis are retained with 0.3063105 and 4.524643, respectively, and constant variance is illustrated with the parallel non-parametric dotted lines. The ICC is also has a positive value of 0.0367 with high statistical significance with LR statistic of 336.65 and p-value of 0.000.[191] But in comparison to the the previous level 1 (only) revised model with ICC value of 0.114, the ICC and its significance is much smaller. This signifies that the dependency pattern of the residuals within the teachers have now been further taken into account (explained away) by the newly added level 2 variables. There is now much less dependency and resemblance among residuals within the teachers.

Revised Models Using the Teacher Average 4th Grade Test Scores

An extended model using a linear function of the teacher average prior (4th grade) test score together with other peer effects and teacher characteristics variables was first estimated and diagnosed. The parameter estimate of the linear average prior test score had a profound positive and significant effect with a t-value of 11.5. Yet, the diagnosis of the level 2 residuals continued to illustrate the positive trend. In light of these findings, a non-linear squared function of the average prior test score was then added to the model. The parameter estimate of the squared term showed a positive and (mildly) significant effect with a t-value of 1.45. The diagnosis of

[190] The slight downward sloping trend shown in the figure can be mainly attributed to the differences in the scaling of the x and y axis variables.
[191] The statistical significance is also confirmed with the BIC value which is smaller by 390 points in comparison to the corresponding model without the teacher VA parameter.

level 2 residuals also continued to show positive trend but to a much lesser degree. A cubic function of the average prior test score was attempted to make sure no further systemic trends are omitted. But the estimate of the cubic term did not have a significant effect with t-value of 0.40. The model fit also did not show significant improvement. The diagnosis of the level 2 residuals also illustrated no clear difference. Given these findings, the squared average prior test score model with other peer effects and teacher characteristic variables was chosen as the best and final model. The estimates of this model illustrate the following results.

Table 8.14: Estimates of the level-2 (and level-1) revised random effects VAM using teacher average 4th grade test scores

	Estimate	*SE*	*t value*
Intercept	5.984	0.0554	108.0
Student Characteristics:			
Student is female	0.0058	0.0004	13.14
Student is Native American	-	-	-
Studet is Asian	0.0107	0.0013	8.21
Student is Black	-0.0017	0.0014	-1.24
Student is Hispanic	0.0017	0.0012	1.43
Student is White	0.0040	0.0011	3.74
Student on FRPL	-0.0035	0.0006	-5.48
Prior Test Score	0.0018	0.0000	283.8
Prior Test Score Squared	0.0000	0.0000	-43.3
Teacher Characteristics:			
Percentage of FRPL Students	0.0003	0.0031	0.11
Percentage of White Students	-0.0004	0.0020	-0.21
Percentage of Female Students	0.0102	0.0034	2.95
Teacher Avg Prior Test Score	-0.0002	0.0003	-0.61
Teacher Avg Prior Test Score Squared	0.0000005	0.0000	1.45
Teacher Class Size	0.0001	0.0001	0.93
Teacher is Female	0.0018	0.0009	1.98
Teacher is White	0.0032	0.0016	1.99
Teacher with MA Degree	0.0003	0.0008	0.37
Teacher Experience	0.0000	0.0000	0.49
Note: Prior test score and its squared function are grand mean centered			

Variance Components:	
Between Teacher Variance	0.00028
Within Teacher Variance	0.00234
Intraclass Correlation	0.105907
AIC	161579
BIC	161393
Number of Students	51161
Number of Teachers	2864
Degrees of Freedom	51141

As you can see, the average prior test score has a positive non-linear effect with a significance level of 1.45. That is, it has a convex pattern (U shaped) resembling the pattern illustrated in the un-revised model. Thus the revised model effectively captured the features of the un-modeled systemic pattern underlying level 2 residuals. And for the other variables, the teacher characteristics namely gender and ethnicity are found to have some positive significant effects and the percentage of female students are also found to have some significant effects.[192] And to make sure other assumptions are not violated, the level 1 and level 2 residuals are re-diagnosed to illustrate the following results.

[192] Just like in the previous case, when the average prior test scores were excluded from the models the other peer effects and teacher characteristics variables showed larger and more significant findings. This potentially illustrates the different indirect effects these variables have through the average prior test scores.

Figure 8.41: Scatterplot of level-1 residuals of the level-2 revised random effects using teacher average 4th grade test scores

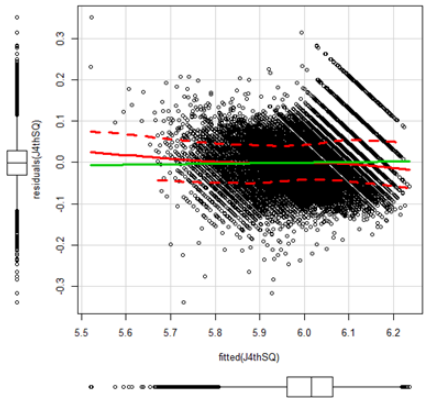
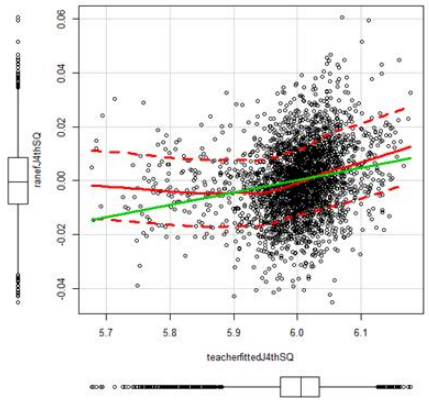


Figure 8.42: Scatterplot of level-2 residuals of the level-2 revised random effects using teacher average 4th grade test scores



As you can see from Figure 8.42, the upward trend which underlined the un-revised level 2 residuals has now clearly faded away. The slope of the green linear regression line is now 0.04563 which is less than half of the original model with 0.09962. The revised model has effectively captured the underlying data pattern and a much better approximation of the randomly distributed pattern is now achieved.[193] Looking now at the other assumptions of the level-2 residuals, the normal distribution pattern is retained with skewness and kurtosis of 0.199585 and 3.650629 together with the constant variance illustrated with the equal spaced and parallel dotted lines in the scatterplot. And finally, as shown Figure 8.41 above, the level 1 residuals virtually remained unchanged from the previous level 1 revised model. It continues to illustrate a good approximation of the randomly distributed pattern. The normal distribution pattern is retained with skewness and kurtosis values of 0.314596 and 4.549646, respectively and constant variance is again illustrated with the parallel non-parametric dotted lines. The ICC also continues to illustrate a positive value with 0.105 and high statistical significance with LR statistic of 2293.72 with p-value of 0.000.[194] The reduction in the ICC value in comparison to the previous level 1 (only) revised model with 0.114 again signifies that the dependency of the residuals within the teachers have now been further explained and taken into account by the newly added level 2 variables.

The Hausman Test – Evaluating and Comparing the Two Models

The two revised models illustrated a clear improvement in capturing the un-modeled systemic pattern evident in the level 2 residuals and alleviating the violation of the exogeneity condition. We now conduct the Hausman test (described in Chapter 5) to further evaluate and solidify whether the exogeneity condition is satisfied in the two models. The test results for the two models together with the original and other extended models are provided as follow.

[193] The slight increasing trend in the figure can again be mainly attributed to the difference in the scaling of the x and y axis variables. The slope of the regression line is again 0.04563 which is virtually equal to zero. Moreover, as addressed in the text, incorporation of the higher power the average prior test scores (cubic term) was not found to have a significant effect. It did not illustrate any further evidence of un-accounted systemic pattern with the average prior test scores.

[194] The BIC value is also smaller by 2371 points in comparison to the corresponding model without the teacher VA parameter confirming the statistical significance.

Table 8.15: Hausman test results (original and revised random effects VAMs)

Hausman Test Results		
	chi-square	*p-value*
Original Un-revised Model	206.1	0.000
Natural Log with Non-linear Prior	163.1	0.000
Avarage 5th with Peer and T Characteristics	13678.9	0.000
Avaerge 4th SQ with Peer and T Characteristics	0.980	0.964
Teacher Characteristics Only	156.4	0.000
Peer and TC without Teacher Average 5th	110.2	0.000
Teacher Average 5th Scores Only	15892.6	0.000

Note: The last five models all uses a log tranformation with squared prior test score for level 1 specification. T = teacher, TC = teacher characteristics, SQ = squared/nonlinear term, Average = teacher average

As you can see, the original un-revised model illustrates a clear and significant violation (rejection) of the exogeneity condition (null hypothesis). Looking at the level 1 revised model (natural log transformation and non-linear prior test scores), the extent of violation of the exogeneity condition has decreased with a much smaller chi-square value of 163.1. The revision effectively took into account some of the systemic pattern left in the residuals but not enough to fully satisfy the exogeneity condition. Looking now at the two level 2 revised models, an interesting pattern is revealed. The model based on the average prior (4^{th} grade) test score and its squared function illustrates no violation of the exogeneity. The exogeneity condition for the level 2 residuals (null hypothesis) is not rejected with a very small chi-square value of 0.980 and very large p-value of 0.964. In fact, it is the only model which satisfies the exogeneity condition. There are no important systemic factors left unaccounted in the residuals which are correlated with the explanatory variables and no bias is introduced in this model. This finding validates the previous revision process based on the average prior test scores. But looking now at the model with the average current (5^{th} grade) test score, a clear and highly significant violation of the exogeneity condition is illustrated. The chi-square value is the largest with the value of 13678.9. But this finding contradicts the previous diagnosis which illustrated a good approximation of the randomly distributed pattern in the level 2 residuals. This contradictory finding is in fact due to another source of endogeneity (violation of the exogeneity) known as the reverse causality problem in which the Hausman test is capturing and now clearly detected in the scatterplots. Reverse causality is evident when an explanatory variable which is used to explain the outcome variable is also explained/modeled by the outcome variable. That is, the explanatory variable is no longer fixed and determined (as required in the regression models) but rather endogenous with a random error term (just like the outcome variable). And in such case (as shown in the footnote) exogeneity condition is violated and bias is introduced in the model.[195] Now in the

[195] Mathematically, if Y = b1X + e and X = b2Y + z where e and z are both error terms then through a simple substitution Y = b1(b2Y + z) + e and this illustrates that Y on the right hand side as an explanatory variable is endogenous (E(eY) is not 0) as it is a function of e. Thus X which is a function of Y is indirectly associated with the e and the E(eX) = 0 is violated.

above model, the average current test score which is modeled as an explanatory variable is clearly determined and a function of the individual students' current test score which is the outcome variable. The model is thus subject to this reverse causality problem and suffers from bias. This problem does not visually show up in the diagnostic plots (as it requires both theoretical and mathematical understanding of the variables) but the Hausman test effectively detects this relation. And this explains why the Hausman test significantly rejects the exogeneity condition in this model. As you can see from the last three rows of the table, for models without the average current test scores, the exogeneity condition is better satisfied. But once the average current scores are included in the model, the exogeneity condition is strongly and significantly rejected with high chi-square value of 15892.6.

To sum, the Hausman tests provided empirical evidence that the revised model based on the average prior test score as the only model which satisfied the exogeneity condition at the level 2. More importantly, this finding makes this model as the only random effects model which effectively satisfied all the regression assumptions required at both levels to ensure the BLUE estimates. That is, this is the only model in which we can confidently interpret, conduct statistical inference, and possibly apply the findings in real life circumstances. Finally, the analysis conducted above also (indirectly) provides us with the diagnosis and revision of the extended VAMs conducted in the previous chapter. The final model illustrates that the estimates provided in the previous chapter (extended model with only the teacher characteristics or only the peer effects) are "preliminary" and do not ensure us of the BLUE estimates. The diagnosis of the random slope models and school fixed effects model were also conducted (as shown in the Appendix) but these models continued to suffer from different model violations. The random slope models continued to illustrate the clear violation of the exogeneity conditions due to the un-modeled increasing systemic patterns left in the residuals. The school fixed effects models, on the other hand, did a better job of modeling the systemic pattern but it introduced new form of violation particularly with a complex form of non-constant variance which expands non-linearly for increasing fitted values. This model is likely to be putting too much structure on the data (with the large number of school dummy variables) without significantly explaining the variation of the outcome variable. This leads to increase in the residual values and its variance. Both models are relatively complex models and demand further work with the possibility of applying more advanced methods to ensure the BLUE estimates.[196] This line of work will be left for the future.

Summary of Random Effects Models – Comparison of the VA Estimates

In the above sections, a full diagnosis of the basic random effects model (at each level of the multilevel model specification) was conducted. Different violations of the regression assumptions were detected and different revisions were implemented to correct these violations. This iterative process between the diagnosis and

[196] For example, the GLS and WLS can be used to capture the non-constant variance for fixed effects model given that the source and structure of the non-constant variance is well studied and modeled. The Huber White M-estimation and Bayesian methods could handle the teacher with particularly small group sizes to improve random slope models.

revision lead to a final model with a natural log transformed outcome variable with a non-linear prior test scores for the level 1 specification and a linear and non-linear (squared) teacher average prior tests score with other peer effects and teacher characteristics variables for the level 2 specification. This model best satisfied all of the regression assumptions to provide us with the BLUE estimates we longed for. Focusing on this final revised model, we now compare the main estimate of interest – the teacher VA estimates across the different models investigate the similarities and consistency in the estimates. This is shown in the following correlation table with the estimates of the final revised model highlighted in purple.

Table 8.16: Correlation of VA estimates of the original and revised random effects VAMs

z						
	Basic	Log	Log SQ	Peer 5th	Peer 4th SQ	
Basic Un-revised	1					
Log Transformed	0.99042	1				
Log with Nonlinear Prior	0.99234	0.98105	1			
Peer with Avg 5th Score	0.77145	0.76801	0.77426	1		
Peer with Avg 4th and Squared	0.96856	0.95960	0.97357	0.88172	1	
Peer 5th with T Characteristics	0.77087	0.76747	0.77362	0.99926	0.88101	
Peer 4thSQ with T Characteristics	0.96698	0.95803	0.97205	0.88064	0.99843	
Teacher Characteristics Only	0.98944	0.97825	0.99715	0.77471	0.97213	
Random Slope FRPL	0.99222	0.98095	0.99982	0.77362	0.97315	
Random Slope Prior	0.98654	0.97200	0.98989	0.77060	0.96408	
School FE	0.68906	0.67868	0.68165	0.53390	0.66987	
	Peer 5th + TC	Peer 4thSQ + TC	T Charact	Rslope FRPL	Rslope Prior	School FE
Peer 5th with TCharcteristics	1					
Peer 4thSQ with T Characteristics	0.88144	1				
Teacher Characteristics Only	0.77538	0.97456	1			
Random Slope FRPL	0.77299	0.97164	0.99695	1		
Random Slope Prior	0.77007	0.96265	0.98700	0.98998	1	
School FE	0.53246	0.66689	0.67797	0.68122	0.67398	1

Note: All the peer effects, teacher characteristics, random slope and school effects model uses a log tranformation with squared prior test scores for the level 1 specification.

As you can see, the VA estimates across the majority of the models illustrate very high correlation. The final revised model shows very high correlation with values over 0.9 for the majority of the models. This illustrates the consistency and similarities of the VA estimates (and its ranking) across the different models. Yet, as addressed before, this does not mean we can forego the diagnosis and revision processes. Again, it is only the final revised model which provides us with the BLUE estimates which we can comfortably interpret. Finally, the problematic models with the teacher average current (5^{th}) test scores and school fixed effects illustrate the lowest correlations. This gives us further positive indication as the models which illustrated the improvement in accuracy and reliability in accordance to the revision illustrated high correlations while the models in which the revisions were found to be ineffective and introduced more problems illustrated the lowest correlations. This provides further validation of the diagnosis and revision conducted in the above sections.

Summary and Conclusion

This chapter conducted a thorough diagnosis and revision of the preliminary VAMs estimated in the previous chapter. The diagnosis detected number of violations of the regression assumptions necessary to ensure the BLUE estimates. Both bias and inefficiency were therefore found in the preliminary estimates and the

interpretation and conclusions provided in the previous chapter were misled. Different sets of revisions were therefore implemented to correct the violations. For each revision implemented, the model was fully re-diagnosed to make sure other violations were not introduced. This iterative procedure lead to a natural log transformed with non-linear prior test score fixed effects model and natural log transformed with non-linear prior test score with linear and nonlinear teacher average prior test scores, peer effects, and teacher characteristics variables for random effects model as the best model which achieved all the assumptions. These models provided us with the BLUE estimates which we longed for. This chapter successfully achieved its objective.

In light of this achievement, the main estimate of interest – the teacher VA estimates of the two final revised models is now compared to illustrate the following results.

Table 8.17: Correlation of VA estimates of the final revised fixed effects and random effects VAMs

Correlation of VA estimates of the final revised fixed and random effects VAMs		
	Log NonL FE	Peer + TC RE
Log with Non-linear prior FE	1	
Log NonL with Peer Effects + T Charact RE	0.84085	1

Note: The peer effects incorporates the average prior test scores and its squared function and not the average current test scores.

As you can see, the two final models have a high correlation of 0.84085. Despite the different modeling scheme, estimation method, and the assumptions required the BLUE estimates, the two VA estimates are highly associated. This gives us great confidence as the best models with the highest accuracy and reliability illustrate very similar and consistent findings. Similar group of teachers are identified as relatively high performing and similar groups of teachers are identified as relatively low performing teachers. And this finding above closely follow suit with the results of Tekwe et al. (2004) which also found correlations between 0.6 to 0.9 for the fixed and random effects model under different specifications (with different sets of variables).[197] It also agrees with the other related studies by Ponisciak and Bryk (2005), Goldhaber (2010), Hanushek and Rivkin (2005) which also used the same methodology but for different grades and subjects to find high correlation up to 0.7 to 0.8. However, unlike these existing studies, this study conducts a more thorough and rigorous evaluation of the VAMs to ensure the accuracy and reliability of the VA estimates. It gives us more confidence and assurance in the findings.

Having now achieved the accuracy (unbiasedness) and reliability (efficiency) of the VA estimates, we are quite ready to apply these findings for policy relevant purposes, the reason why VAMs has gained heated attention

[197] Tekwe et al. (2004) also found the highest correlation between models with similar modeling option (fixed and random effects) and similar specification (i.e. similar sets of explanatory variables).

at the inception. But before we do so, another important analysis which will further solidify findings is needs to be conducted. This is the robustness and sensitivity analysis of the VA estimates with respect to the teacher/class sample sizes. The motivation for this analysis and the findings are provided in the next chapter.

CHAPATER 9: ROBUSTNESS AND SENSIVITY ANALYSIS OF VAMS WITH TEACHER SAMPLE SIZES

In the previous chapter, the unbiasedness and efficiency of the VA estimates were successfully achieved. These two statistical properties which lie at the heart of statistical inference enabled us to draw the most viable conclusions and generalization of the VA estimates. But as described in Chapter 2, in addition to these two properties, other statistical properties such as the robustness is also an important feature and criteria we look for in an estimate. The robustness (or insensitivity) provides us with the measure of stability, poisedness, or firmness of the estimates in lieu to alterations and volatility of the determining factors underlying the estimates. Estimates which are robust give us more confidence and assurance in the measures and the inference we draw from the measures. In the context of VAMs, it was explained in Chapter 4, that the teacher sample sizes (class sizes) have a profound and important role in determining the VA estimate for both the fixed and random effects. It not only determined the VA estimate (the point estimate) but also the standard error of the VA estimates. It is therefore an important determinant of the statistical inference we conduct with the VA estimates. The first sign of the sensitive effects the teacher sample sizes can have on the VA estimates was witnessed in the influential diagnosis of fixed effects model shown in the previous chapter.[198] The diagnosis found that students and teachers belonging to very small classes exerted very high and misleading influence on the overall model estimates. And these influential cases can potentially introduce both bias and/or inefficiency in the estimates. Thus, for the reasons just described, it is critical to conduct a robustness and sensitivity analysis of the final VAMs defined in the previous chapter before we consider interpreting or applying the results. The importance of robustness has in fact already been raised and extensively studied in the VAMs literature today. Yet, researchers continue to struggle and debate over the "sufficient sample size" of the teachers in conducting VA analyses. As addressed by Bell et al. (2008), "currently there are few sample size guidelines referenced in the literature…one rule of thumb for designs in which individuals are nested within groups calls for a minimum sample size of 30 units within each group." This threshold is extensively cited in the VAM research community for example by the Economic Policy Institute (2010) and the National Research Council (2010). But other studies, on an ad hoc manner, specify the minimum sample size to be 5 students (Harris and Sass, 2009), 10 students (Clarke and Wheaton, 2007), and 20 students (Koedel and Betts, 2009). While other more technically advanced studies (e.g. the Monte Carlo simulation studies by Bell et al., 2008) even specify 1 student as a feasible sample size given that there are large number of higher level (e.g. teachers) observations.

In light of this ongoing debate and the importance of teacher sample size in VA analyses, this chapter conducts a comprehensive robustness and sensitivity analysis of the VA estimates derived from the final fixed and random effects model from the previous chapter. It will first conduct a detailed descriptive analysis of the VA

[198] This is also illustrated in the work by Bosker and Snijders (1999).

estimates and its standard errors with respect to the teacher sample size variable. It will then closely follow the analysis and methodology shown in Theall et al. (2008) and Goldhaber (2010) by calculating the percentile point difference and correlation coefficient between the original model and different revised models (which take into account different sets of small sample sized teachers by excluding them from the analysis).[199] The analysis starting with fixed effects model followed by random effects model are as follow.

Fixed Effects Model Estimates and Teacher Sample Sizes

To examine the relation between fixed effects VA estimates and the teacher sample size, the following figures are constructed.

Figure 9.1: Plot of fixed effects VA estimates and teacher size

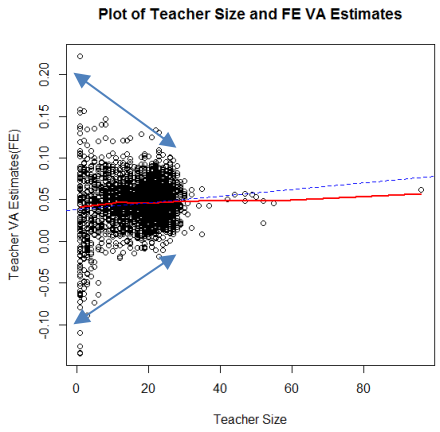
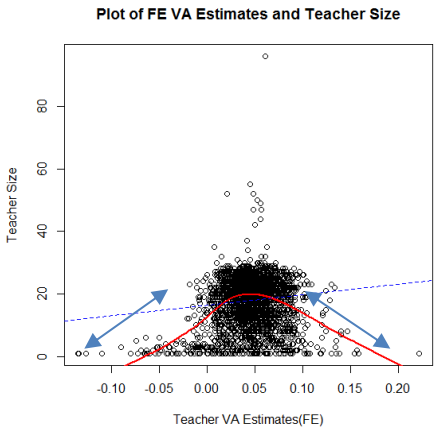


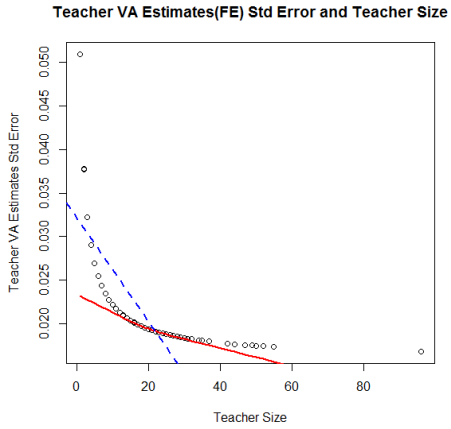
Figure 9.2: Plot of teacher size and fixed effects VA estimates



Looking at Figure 9.1 on the left, as shown with the arrows, there is a larger dispersion of the VA estimates for teachers with smaller sample sizes. This pattern is further clarified by alternating the x and y axis variables as shown in Figure 9.2 on the right. As you can see, the non-parameter curve (in red) illustrate a clear concave pattern where moderate and average sized VA estimates are associated with large sample sizes while extreme VA estimates are associated with very small sample sizes. This finding confirms with some of the existing multilevel analysis literature which cite the volatile and extreme estimates for groups (such as teachers) characterized with small amount of observations.[200] Finally, the same pattern shown above was also evident in the VA analysis by Goldhaber (2010) who also used the Washington data but for different grades and time periods. Looking now at the standard error of the VA estimates, a clear downward trend is illustrated as follow.

[199] A more technically advanced simulation analysis such as the Monte Carlo methods and the Bayesian methods require additional distributional assumption of the VAMs. These analyses require high level of statistical background and will be left for the future.
[200] Mass and Hox (2004), Clarke and Wheaton (2007), and Theall et al. (2008)

Figure 9.3: Plot of standard error of fixed effects
VA estimates and teacher size

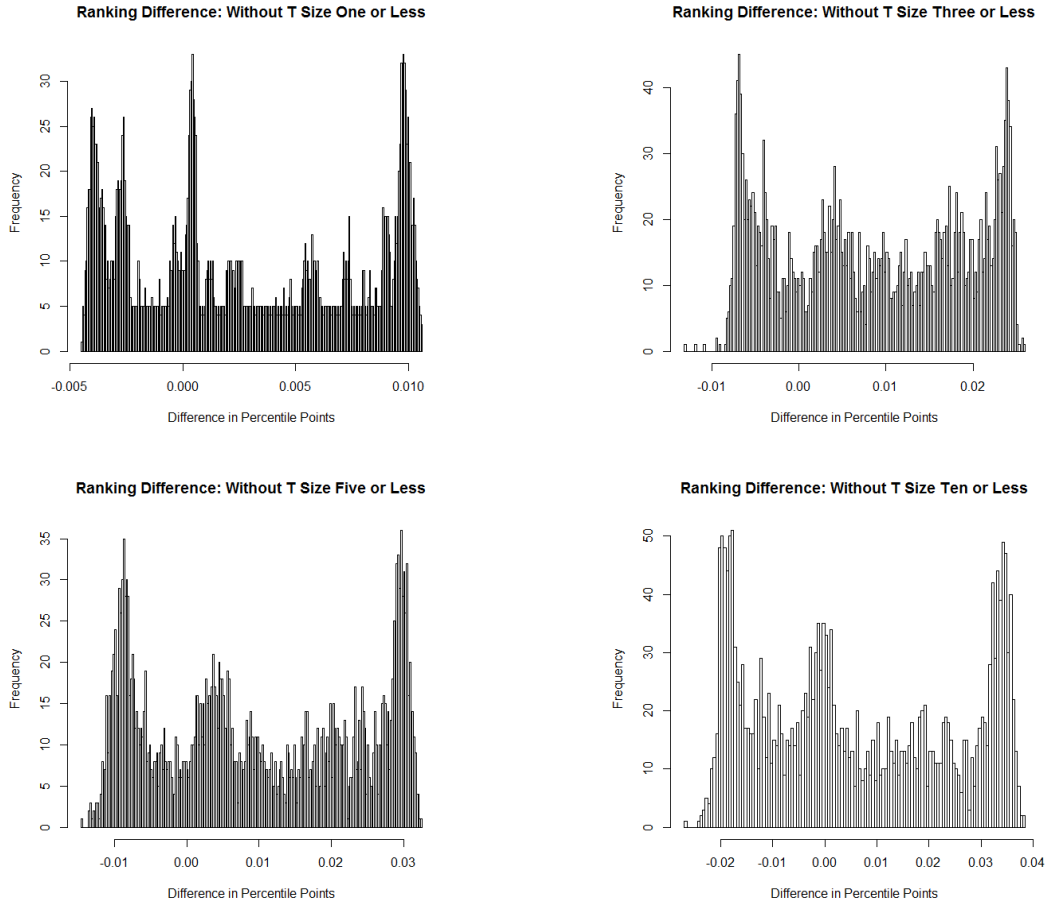


This finding is in line with our expectation as it was shown in Chapter 4 that the standard error of fixed effects VA estimates is inversely related to the square root of the teacher sample size. Therefore, if the teacher sample size decreases (with less information and data) then the standard error of the VA estimate increases (less reliable and less precise).[201] These findings hint that the exclusion of the small sample sized teachers from the VA analyses could potentially alter the VA estimates (i.e. the relative ranking) and the inference drawn in the previous chapter. In order to investigate this robustness or sensitivity, the percentile point change and the correlation coefficient of the VA estimates and its standard errors between the original model (based on the entire data) and the different set of revised models (based on data sets which exclude small sample sized teachers starting with only 1 student, less than or equal to 3 students, less than or equal to 5 students, and less than or equal to 10 students) are calculated.[202] Looking first at the percentile point changes of the VA estimates illustrate the following results.

[201] A close look at the graph shows that there are fewer data points in comparison to the point estimate plot. This is due to the fine value estimates (very small decimal points) which are rounded up and therefore clustered into the same data point.

[202] The results may initially seem obvious from the descriptive findings shown above, but the linear regression models underlying fixed effects model taken into account a variety of unforeseen variations and associations inherent between the teachers (themselves) and the explanatory variables. The effects of these variables and the variance components estimates (which also determine the standard error estimates of the VA estimates) are estimated using the entire data set jointly. Thus the exclusion of the small sample sized teachers (or any teachers) can alter the rest of the teacher estimates and other variables in unpredictable ways.

Figure 9.4: Histogram of percentile ranking difference of fixed effects VA estimates after excluding different sets of small sample sized teachers



As you can see from the figures, the percentile point differences are extremely small across the different comparisons. The mean values are virtually zero with 0.002699, 0.008478, 0.009769, and 0.006023, respectively in ascending order of group sizes. There is a slight increase in the mean percentile changes as we take into account teachers with less than or equal to three or five students but the margins are extremely small at the third decimal point. It is much smaller in comparison to the the inter-temporal studies of VA estimates which illustrated significant changes from 0.20 to even 0.80 percentile points.[203] The above findings therefore illustrate that the relative ranking of the teacher VA estimates did not change at all after taking account the different groups of small sample sized teachers. Looking now at the correlation coefficients of the VA estimates across the different models illustrate the following results.

[203] Inter-temporal studies refer to comparing the VA estimates across time.

Table 9.1: Correlation of fixed effects VA estimates after excluding different sets of small sample sized teachers

	Original	Without 1 or less	Without 3 or less	Without 5 or less	Without 10 or less
Original	1				
Without T Size 1 or less	1.00000000	1			
Without T Size 3 or less	0.99999258	0.99999258	1		
Without T Size 5 or less	0.99998584	0.99998584	0.99999662	1	
Without T Size 10 or less	0.99997935	0.99997935	0.99999053	0.99999496	1

As you can see, very high correlations of over 0.9 (actually close to 1) are found. This confirms the strong consistency and similarity of the estimates shown with the percentile changes. Again, the relative VA estimates and ranking for the other teacher did not alter after taking in account the small sample sized teachers. This strong consistency is also illustrated in the standard errors estimates are illustrated as follow.

Table 9.2: Correlation of standard errors of fixed effects VA estimates after excluding different sets of small sample sized teachers

	Original	Without 1 or less	Without 3 or less	Without 5 or less	Without 10 or less
Original	1				
Without T Size 1 or less	1.00000000	1			
Without T Size 3 or less	1.00000000	1.00000000	1		
Without T Size 5 or less	1.00000000	1.00000000	1.00000000	1	
Without T Size 10 or less	0.99979115	0.99979115	0.99979113	0.99979110	1

The standard errors also illustrate extremely high correlation where some of them are perfectly correlated with value 1. The standard error or the reliability and precision of the VA estimates did not alter after taking into account the small sample sized teachers. As the standard errors determine the confidence interval surrounding the VA estimates, this finding also illustrate that the statistical inference of the VA estimates are robust to small sample sized teachers.

To sum, the above findings provide strong evidence that the VA estimates are robust and insensitive to the small sample sizes of teachers. Both the VA estimates and the standard error were not heavily altered and this enhances our confidence and assurance of the inference and conclusions drawn from the final revised fixed effects VAM. Moreover, it provides another case study to the ongoing debate on the sufficient sample size for multilevel analysis.

Random Effects Model Estimates and Teacher Sample Sizes

Repeating the above analysis for random effects VA estimates illustrate the following results.

Figure 9.5: Plot of random effects VA estimates and teacher size

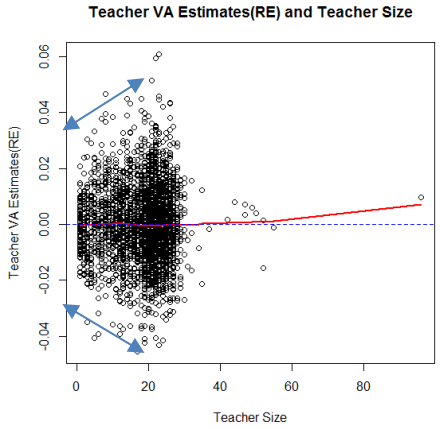
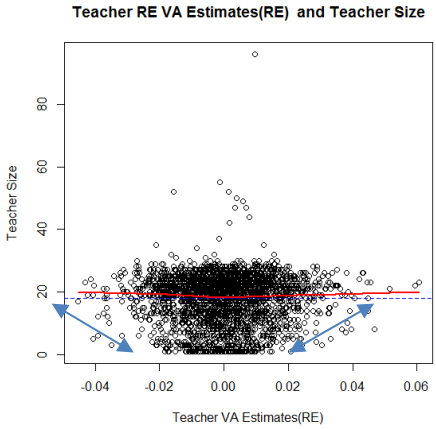
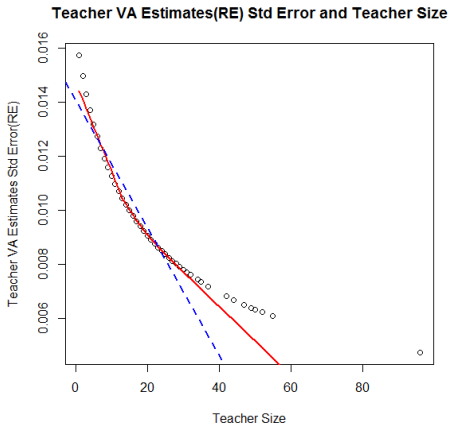


Figure 9.6: Plot of teacher size and random effects VA estimates



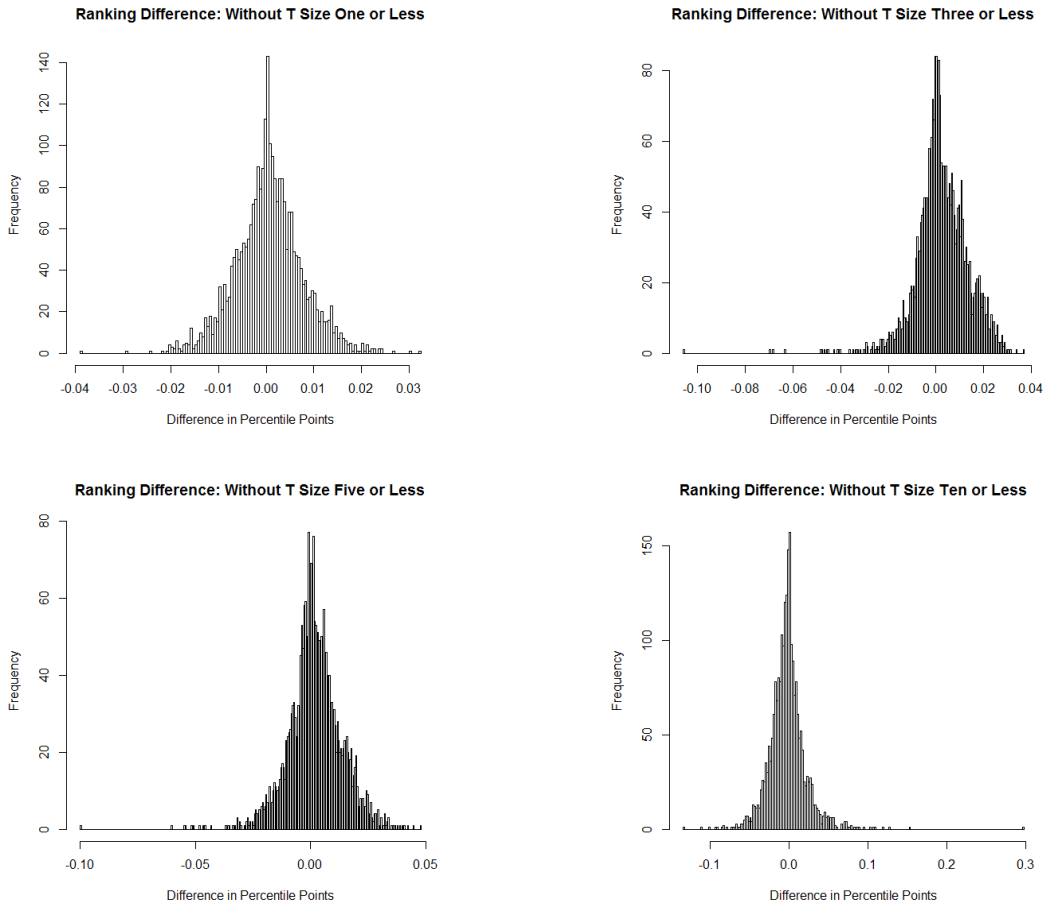
The above figures illustrate the opposite trend shown above with fixed effects models. As you can see from the arrows (which are now directed in the opposite direction), the VA estimates for teachers with very low sample sizes illustrate much less dispersion. The values are in fact now concentrated more towards the value 0 (the overal mean of teacher VA estimates). But as described in Chapter 4, this finding is in line with our expectation as it precisely illustrates "shrinkage" or "borrowing effect" underlying random effects empirical Bayes VA estimates. The empirical Bayes VA estimate for teachers with small sample sizes (low precision with not much information) are adjusted ("shrunk") more towards the overall mean which is estimated reliably with the entire data set and away from the fixed effects adjusted teacher means which is estimated unreliably with the small sample size. That is, teachers with small sample sizes "borrow" information from the reliably estimated overall mean by leaning towards its value. Looking now at the standard errors of the estimates, the negative trend is again evident as shown below. But taking a closer look at the scale and range of the y-axis, it is evident that the standard errors of random effects are 2 to 3 times smaller than fixed effects values. Moreover, the values are much less dispersed.

Figure 9.7: Plot of standard error of random effects VA estimates and teacher size



Now, to investigate the robustness and sensitivity of the VA estimates with respect to the small sample sized teachers, the percentile point changes in the VA estimates after taking into account (excluding) different sets of small sample sized teachers are calculated to show the following results.[204]

Figure 9.8: Histogram of percentile ranking difference of random effects VA estimates after excluding different sets of small sample sized teachers



As it was also evident in fixed effects model, the percentile point differences are extremely small across the different comparisons. The mean percentile changes are virtually zero with 0.000565, 0.002419, 0.001873, and -0.002316, respectively in ascending order of group sizes. This finding again illustrate that the relative ranking of the teacher VA estimates did not change after taking account the small sample sized teachers. These mean values are also much smaller than the values shown in fixed effects model. This implies that random effects VA estimates were even more stable and insensitive to the small sample sized teachers. This again shed light on the shrinkage effects underlying empirical Bayes predictions which adjust for the imprecision of the VA estimates due to small sample sizes. Looking now at the correlation coefficients of the VA estimates

[204] Just like in the fixed effects model, the effects of the exclusion of the small sample sized teachers cannot be predicted from the descriptive summaries alone. The linear (mixed) models underlying the random effects take into account of a variety of associations and variations inherent between the teachers and the student level variables. Moreover, both variance components (between and within teacher variances) are jointly estimated using all the students and teachers under the MLE framework. The exclusion of the small sample sized teachers (or any teachers) can alter the rest of the teacher VA estimates and other variables in unpredictable ways. And as the empirical Bayes estimates are determined as a function of the variance components estimates, it will also alter in unpredictable ways.

illustrate very high correlations of over 0.9 (actually close to 1) which confirm the strong consistency and similarity of the estimates.

Table 9.3: Correlation of random effects VA estimates after excluding different sets of small sample sized teachers

	Original	Without 1 or less	Without 3 or less	Without 5 or less	Without 10 or less
Original	1				
Without T Size 1 or less	0.99979676	1			
Without T Size 3 or less	0.99950684	0.99970341	1		
Without T Size 5 or less	0.99942656	0.99974808	0.99977229	1	
Without T Size 10 or less	0.99730795	0.99730302	0.99669162	0.99771313	1

And for the standard error estimates, the strong consistency in the estimates is also found as shown below.

Table 9.4: Correlation of standard errors of random effects VA estimates after excluding different sets of small sample sized teachers

	Original	Without 1 or less	Without 3 or less	Without 5 or less	Without 10 or less
Original	1				
Without T Size 1 or less	0.99999978	1			
Without T Size 3 or less	0.99999830	0.99999913	1		
Without T Size 5 or less	0.99999584	0.99999698	0.99999905	1	
Without T Size 10 or less	0.99999199	0.99999318	0.99999574	0.99999815	1

To sum, similar to the case of fixed effects model, random effects model also illustrated strong and in fact slightly more robustness and insensitivity of the VA estimates to the small sample sized teachers. The VA estimates (its ranking) and its standard errors did not illustrate any changes after taking into account different groups of teachers with small sample sizes. These findings enhance our confidence in the estimates and the inference drawn from random effects VAM.

Summary and Conclusion

This chapter extended the unbiased and efficient estimates achieved in the previous chapter to evaluate its robustness with respect to the sample sizes of the teachers. The analysis illustrated that the very small sample sized teachers (which can induce both unreliability and bias to the model estimates) did not heavily alter or mislead fixed and random effects VA estimates. Both the VA estimates (its relative ranking) and the standard error for the two models were found to be robust and insensitive to the different groups of small sample sized teachers. These findings together enhanced and solidified our confidence and assurance of the inference and conclusions based on the models. It ensured the statistical property of robustness on top of the unbiasedness and efficiency achieved in the previous chapter.

Together with previous chapters, this chapter completes the diagnosis, revision, and validation of the VAMs. The fullest and most comprehensive evaluation of the VAM estimates has now been conducted. And the utmost accuracy, reliability, and now robustness of the estimates have been achieved. In light of this

achievement, we are finally ready to apply the findings to inform policies which can potentially improve the Washington State's education system.

CHAPTER 10: POLICY ANALYSIS – IDENTIFYING THE TEACHERS WHOM WE CAN LEARN FROM AND TEACHERS WHOM WE CAN HELP

Having now conducted a thorough diagnosis, revision, and robustness analysis of the VAMs, we are now ready to apply these findings to inform potential policies to improve the education system in Washington State. As it is evident in today's heated media, one of the main purposes of VAMs is to identify the so called "effective" teachers with high VA ranking and "non-effective" teachers with low VA rankings in order to better target policies and interventions. For the former teachers, we could potentially learn invaluable lessons underlying their success and for the latter teachers, we could potentially provide assistance. In the following sections, this chapter first applies the VA estimates from the final and revised fixed and random effects models to conduct the conventional VA analysis to identify the effective and non-effective teachers. The underlying features of these groups of teachers will then be studied using the available data. But unlike the majority of the existing studies, this study also engages in an unprecedented and innovative task. It extends the conventional analysis through incorporating student background or equity related indicators to identify the so called "exemplary" and "beat the odds" teachers who achieve high VA estimates and increase the average student performance regardless of the high concentration of underachieving students in their classes. That is, these teachers achieve both quality and equity simultaneously through raising the performance particularly of the students who have historically underachieved. On the other hand, the "non-exemplary" teachers who could not achieve high VA despite of being surrounded with high concentration of high prior achieving students will also be identified. These teachers performed below what was expected. The underlying features of the exemplary and non-exemplary teachers will then be closely examined and studied. This extended analysis greatly improves the application of the VA estimates for policy relevant purposes. Through making the teachers of extraordinary success and in need of significant improvement even more vivid, it assists policy makers in making more effective and efficient targeted policies and interventions. It gives us and policy makers the confidence and assurance of the teachers whom we can possibly draw invaluable lessons and teachers whom we can provide immediate assistance. Yet, there are only few studies to date which extends the conventional VA analysis through incorporating equity related indicators.[205] This study therefore attempts to contribute and highlight the importance of this type of analyses.

Conventional Value-Added Analysis – Identifying and Understanding the Effective and Non- Effective Teachers

In order to identify the effective and non-effective 5th grade math teachers, this study follows suit with previous efforts such as Ballou (2005), Goldhaber (2010), Aaronson et al. (2007), McCaffrey et al. (2009), and Lipscomb et al. (2010) to use the percentile rankings of the VA estimates. Teachers situated in the top ten percent (above the 90th percentile) of the VA ranking will be defined as "effective" teachers while teachers

[205] The three existing studies conducting this analysis are the Achievement Gap Initiative at Harvard University, the Value Added Research Center at the University of Wisconsin Madison, and the Programme for International Student Achievement (PISA) by the OECD. These three studies will be examined again later in this chapter.

situated in the bottom ten percent (below the 10th percentile) will be defined as "non-effective" teachers. This procedure is conducted using the VA estimates from the final and revised fixed and random effects models achieved in Chapter 8. The results are illustrated as follow.

Figure 10.1: Plot of fixed effects VA estimates with 10th and 90th percentile lines

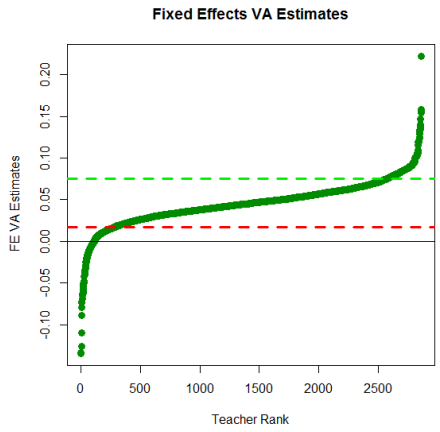
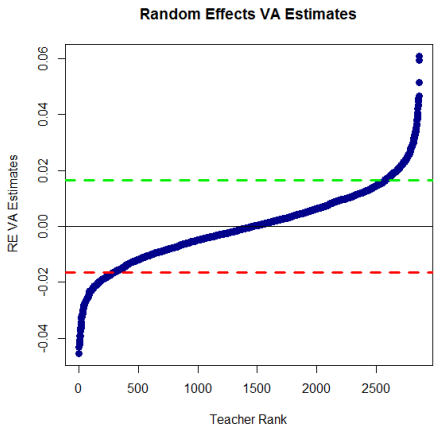


Figure 10.2: Plot of random effects VA estimates with 10th and 90th percentile lines



The effective teachers are situated above the green dotted line (which indicates the 90th percentile value: 0.0755392 for fixed effects and 0.0165925 for random effects) and the non-effective teachers are situated below the red dotted line (which indicates the 10th percentile value: 0.0173361 for fixed effects and -0.0165923 for random effects).[206] There are total of 287 effective and non-effective 5th grade math teachers for both models. Comparing the two groups of teachers between the two models illustrated over 70% match rate for the effective teachers and over 60% match rate for the non-effective teachers. That is, fixed and random effects models consistently identified very similar group of teachers as effective and non-effective. In the following section, we now compare and contrast these two groups of teachers defined across the two models through studying their underlying features. Namely, the teacher characteristics and types of students these teachers teach (student peer effects variables) are explored.

Teacher Characteristics of the Effective and Non-Effective Teachers

The descriptive summary of the teacher characteristics variables underlying the effective and non-effective teachers are as follow.

[206] For fixed effects model, the dummy variable coding is used instead of treatment coding. Fixed effects VA estimates in the plot are deviations of the adjusted teacher means from the base line teacher adjusted mean (teacher ID 1) and not the adjusted grand teacher mean. For this reason, the scale is different from random effects VA estimates which are defined as the deviation from the adjusted grand mean. The magnitudes may be different but the relative ranking of the teachers will not change because of the coding method. The same group of teachers will be identified as effective and non-effective for both coding method. The dummy coding method was used due to technical problem in extracting the teacher ID numbers under the treatment coding of the R statistical software.

Table 10.1: Descriptive summary of teacher characteristics variables for effective and non-effective teachers

Characteristics of Effective and Non-Effective Teachers					
		Mean		Count	
		Effective	*Non-Effective*	*Effective*	*Non-Effective*
5th Grade Math Teacher Character:	**Model:**				
Teacher is Female	FE	0.798	0.732	F = 229, M = 58	F = 210, M = 77
	RE	0.774	0.732	F = 222, M = 65	F = 210, M = 77
	Overall T	0.750		F = 2148, M = 716	
Teacher is White Ethnicity	FE	0.937	0.902	W = 269, non-W = 18	W = 259, non-W = 28
	RE	0.927	0.923	W = 266, non-W = 21	W = 265, non-W = 22
	Overall T	0.935		W = 2697, non-W = 185	
Teacher with MA Degree	FE	0.641	0.610	MA = 184, no MA = 103	MA = 175, no MA = 112
	RE	0.659	0.641	MA = 189, no MA = 98	MA = 184, no MA = 103
	Overall T	0.663		MA = 1900, no MA = 964	
Note: Overall T = overall and entire teachers					

Characteristics of Effective and Non-Effective Teachers											
		Mean		SD		Median		Mode		Range (mini, max)	
		Effective	*Non-Effective*	*Effective*	*Non-Effective*	*Effective*	*Non-Effective*	*Effective*	*Non-Effective*	*Effective*	*Non-Effective*
5th Grade Math Teacher Character:	**Model:**										
Teacher Experience	FE	14.45	12.53	10.08	10.22	13.00	9.50	1	0	(0 , 40)	(0 , 44.5)
	RE	14.10	14.07	9.93	10.77	12.80	11.80	4	0	(0 , 39)	(0 , 44.5)
	Overall T	13.57		9.89		11.20		1			
Teacher Class Size	FE	14.63	11.83	8.24	8.81	15.00	12.00	22	1	(1 , 28)	(1 , 35)
	RE	18.15	18.49	6.27	6.76	20.00	20.00	22	22	(1 , 28)	(1 , 35)
	Overall T	17.86		7.73		20.00		22			
Note: Overall T = overall and entire teachers											

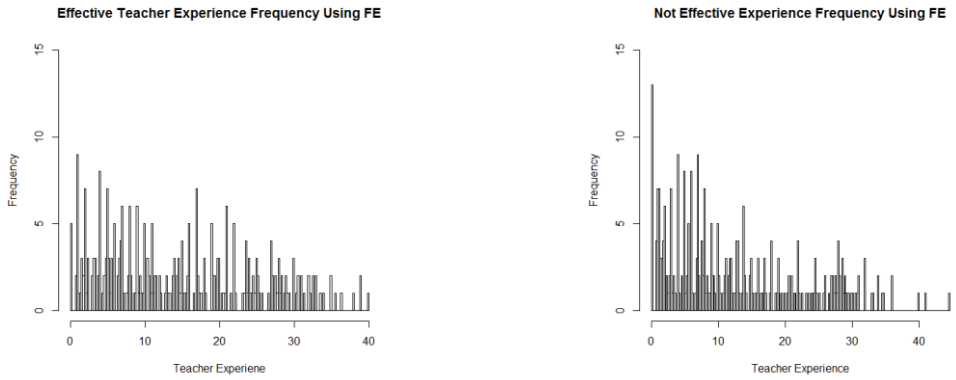
Looking at the two groups of teachers identified under the fixed effect model, it is evident that there are more female, White, with MA degree, and experienced teachers for the effective teachers than the non-effective teachers.[207] As these characteristics were shown to have a positive and significant effect on student performance, this finding implies that the effective teachers' are generally characterized with more advantaged background in raising their students' performance. But looking at the findings under the random effects model, it is clearly evident that these differences in teacher characteristic variables have decreased and in some cases no longer evident. That is, the effective teachers are no longer characterized with more advantaged backgrounds in raising student performance. This finding is in fact perfectly consistently and in line to our expectation. As illustrated in Chapter 8, the final revised random effects model extends the basic VAM specification (of the fixed effects) to take into account the effects of the teacher characteristics variables on student performance. By doing so, it further levels the playing field in estimating the teacher VA estimates by removing and filtering the effects of these variables from the teachers. For this reason, there is no longer significant difference in the prevalence of these variables across the two groups of teachers. As you can see

[207] The teacher class size, on the other hand, illustrates slightly smaller class sizes for the non-effective teachers. This is opposite in the general finding with class size where smaller classes are linked to higher achievement. But studies have found that this finding differs depending on the grade level being assessed e.g. smaller class size at the elementary level is not necessarily linked to higher achievement. Further work and research on this point is needed before we can confidently interpret the results.

from the table, the mean values for the two groups are now very close to the overall teacher means. The random effects estimates implies that the difference in the teacher VA estimates not attributed to the teacher characteristics variables but to other un-measured and unobservable features of the teachers that is beyond the available data. In essence, it provides a more strict and conservative but more accurate measure of teacher effectiveness.[208]

Looking at the descriptive summary again, it is also evident that there is clear variation in the distribution of the teacher characteristics variables within each group of teachers as characterized with the non-zero standard deviation and range values. To better illustrate this feature (i.e. for the continuous teacher experience variable), the entire distribution of the variable based on the fixed effects model is illustrated below. The distribution based on the random effects model (provided in the Appendix) also illustrated similar pattern.

Figure 10.3: Histogram of teacher experience for effective and non-effective teachers



Both groups of teachers illustrate a very similar distribution and variation of experience level. They both illustrate a slightly downward sloping curve where there are relatively less teachers with high levels of experience. But from the underlying variation, it is also evident that some teachers are classified as effective with even higher experience level (than the average of effective teachers) while quite a few teachers are classified as effective despite of very low experience level. Similarly, some teachers are classified as non-effective despite having very high years of experience while some having very low years of experience. Thus looking into the finer details and deviations from the mean values illustrates interesting and unexpected findings. Keeping these findings in mind, we proceed to the examination of the types of students characterizing the two groups of teachers.

Types of Students Underlying the Effective and Non-Effective Teachers

The descriptive summary of the types of students (student peer effects variables) underlying the effective and non-effective teachers are as follow.

[208] This finding does not imply that the fixed effects are statistically inaccurate (unbiased). Instead, fixed and random effects are based on two different theoretical models - the basic VAM for the fixed effects and extended VAM for the random effects. It responds to different theoretical and research questions. And as shown in Chapter 8, both models provide statistically reliable and accurate estimates for the respective models. This illustrates the importance of the dual role of theory and data in the model building process.

Table 10.2: Descriptive summary of student peer effects variables for effective and non-effective teachers

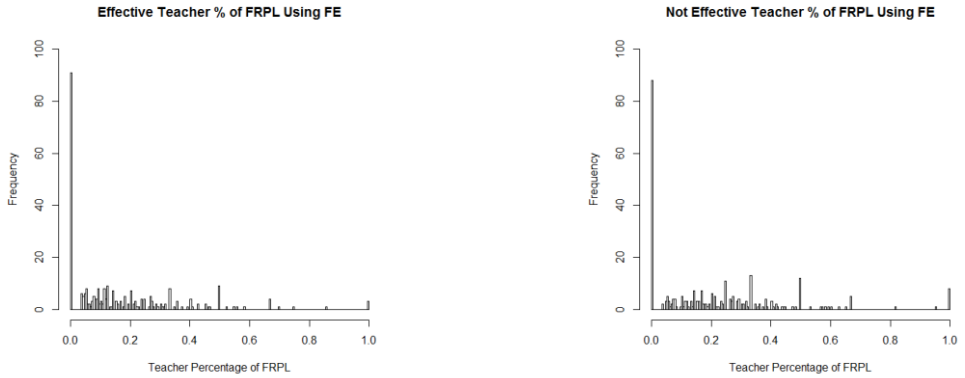
		Characteristics of Effective and Non-Effective Teachers									
		Mean		SD		Median		Mode		Range (mini, max)	
		Effective	*Non-Effective*	*Effective*	*Non-Effective*	*Effective*	*Non-Effective*	*Effective*	*Non-Effective*	*Effective*	*Non-Effective*
5th Grade Student Charactistics:	**Model:**										
% of Students on FRPL	FE	0.157	0.209	0.192	0.230	0.091	0.160	0	0	(0,1)	(0,1)
	RE	0.182	0.173	0.169	0.156	0.143	0.138	0	0	(0,1)	(0,1)
	Overall T	0.174		0.176		0.133		0		(0,1)	
% of White Ethnicity	FE	0.652	0.585	0.284	0.318	0.704	0.600	1	1	(0,1)	(0,1)
	RE	0.614	0.641	0.279	0.254	0.667	0.680	1	0.667	(0,1)	(0,1)
	Overall T	0.650		0.271		0.714		1		(0,1)	
% of Female Students	FE	0.456	0.413	0.200	0.262	0.500	0.440	0.5	0	(0,1)	(0,1)
	RE	0.480	0.465	0.127	0.142	0.500	0.476	0.5	0.5	(0,1)	(0,1)
	Overall T	0.476		0.157		0.500		0.5		(0,1)	
Average Initial Performance	FE	409.2	349.0	38.61	40.29	411.3	389.7	-	-	(270 , 519)	(289 , 508)
	RE	404.1	404.9	30.13	29.71	404.2	406.1	-	-	(272 , 483)	(302 , 487)
	Overall T	402.4		28.2		403.6				(270 , 519)	

Looking at the two groups of teachers identified under the fixed effect model, it is evident that, on average, there are more non-FRPL, White, female, and high prior achieving students for the effective teachers than for the non-effective teachers. As these characteristics were shown to have a positive and significant effect on student performance, this finding implies that the effective teachers are on average teaching more students coming from high achieving background. It can be said that they are on average working in classes with more positive peer effects and positive learning environment. Together with the teacher characteristics variables, the fixed effects illustrate that effective teachers are, on average, characterized with more advantaged conditions in raising student performance. But looking at the findings under the random effects model, it is again clearly evident that these differences in the types of students and peer effects significantly decreased and in many cases no longer existent. The effective teachers are no longer teaching more students who have shown to achieve high academic performance and the positive peer effects from these students. This finding is again perfectly consistently and in line to our expectation as the final revised random effects model also takes into account the effects of the student peer effects. It further levels the playing field in estimating the teacher VA estimates by removing and filtering the effects of these variables from the teachers. And for this reason there is no longer clear difference in these variables across the two groups of teachers.[209] The difference in the teacher VA estimates not attributed to the student peer effect variables (or to the teacher characteristics variables) but to un-measured and unobservable features of the teachers that is beyond the available data. The random effects model again provides a more strict and conservative but more accurate measure of teacher effectiveness. But moving beyond the average summaries, interesting variation is again illustrated in the different variables.

[209] As you can see from the table, the mean values for the two groups are again very close to the overall teacher mean values.

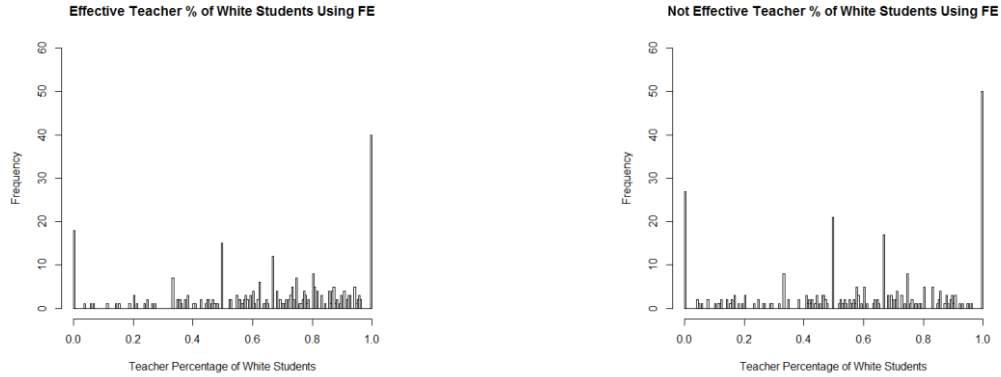
Looking at the descriptive summary table again, there is clear variation in the distribution of the student peer effect variables within each group of teachers. This is again characterized with the non-zero standard deviation and range values and the following histograms which depict the entire distribution pattern.[210]

Figure 10.4: Histogram of percentage of FRPL students for effective and non-effective teachers



Both groups of teachers illustrate similar distribution pattern of the percentage of FRPL students with a downward sloping curve with increasingly fewer teachers teach a class with high percentage of FRPL students. But taking a closer look at the distribution again, it is also evident that there are quite a lot of teachers who are classified as effective with absolutely no (0%) FRPL students in their class while some teachers are classified as effective despite having high percentage of FRPL students whom we know have historically underperformed. Similarly, some teachers are classified as non-effective despite having no FRPL students (who have achieved high prior academic performance) while some teachers are classified as non-effective with high percentage of FRPL students in their class. Thus looking into the variation of the distribution illustrates interesting and unexpected findings. This pattern and interpretation also evident for the other student peer effect variables as shown below.

Figure 10.5: Histogram of percentage of White students for effective and non-effective teachers



[210] The distribution/histograms based on the fixed effects model is provided in the text. The random effects model also illustrated similar distribution and is provided in the Appendix.

Figure 10.6: Histogram of percentage of female students for effective and non-effective teachers

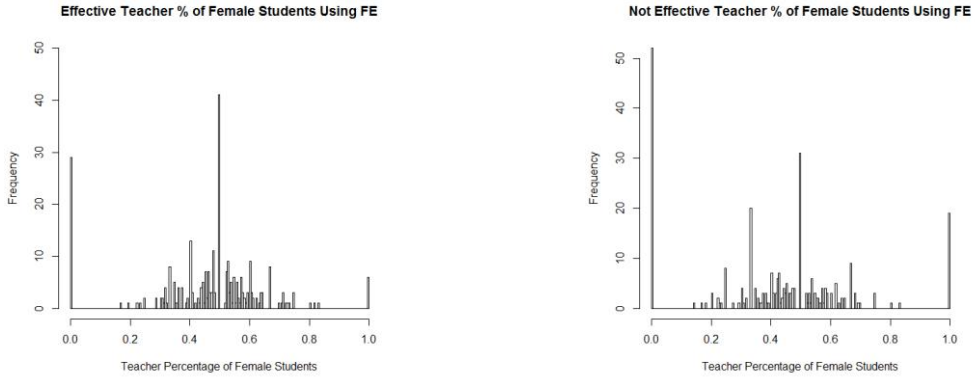
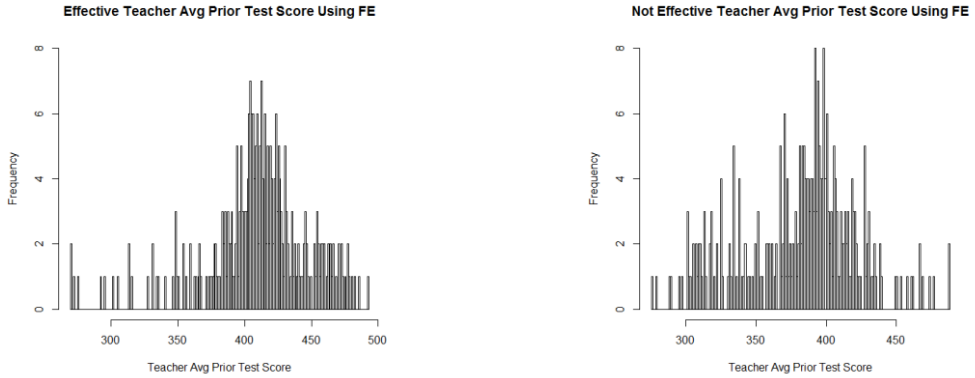


Figure 10.7: Histogram of teacher average prior test scores for effective and non-effective teachers



To sum, studying the underlying characteristics of the effective and non-effective teachers based on the fixed effects illustrated that the effective teachers, on average, were characterized with more advantaged backgrounds and student composition in raising their student's performance than the non-effective teachers. But the random effects model which (statistically) took into account the effects of teacher backgrounds and student peer effects no longer illustrated the difference in the prevalence of these variables between the two groups of teachers. The difference in the teacher VA and effectiveness measures were no longer attributed to these factors but to other un-measured and un-observable factors underlying the teachers. Random effects improved the accuracy and the assurance in the identified effective and non-effective teachers. Yet, clear variation in the teacher characteristics and student peer effects variables underlying the two groups of teacher was also evident. Number of un-expected teachers who achieved (failed to achieve) high VA performance despite surrounded with high concentration of academically challenging (achieving) backgrounds and conditions were identified. In the following section, we further and constructively explore this un-expected group of teachers.

Quadrant Analysis – Identifying the Exemplary Teachers and Non-Exemplary Teachers

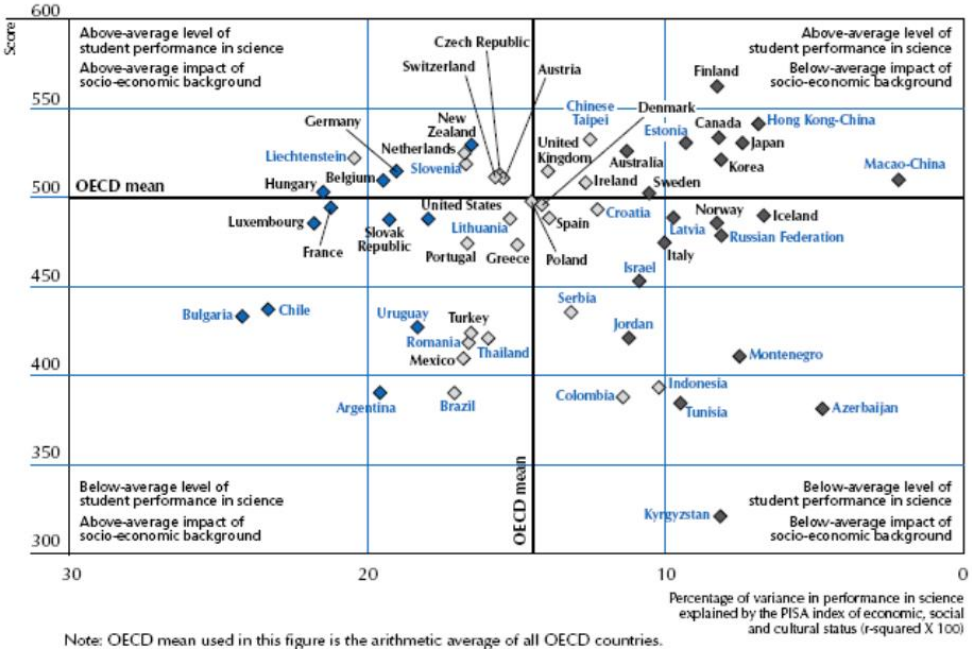
In the previous section, the examination of the underlying characteristics of the effective and non-effective teachers illustrated clear variation within the two groups of teachers. This finding hinted us and gave us with

first empirical evidence of the presence of un-expected and un-ordinary group of teachers. For example, some effective teachers managed to achieve high VA and raise students' average performance despite of teaching a challenging group of students who have historically underperformed in math i.e. students on FRPL, non-White students, male students, and students with low initial achievement. As these teachers were able to raise the performance particularly for these underachieving students, they were able to contribute towards reducing the achievement gap evident in these groups (categories) of students. That is, these teachers were able to achieve both quality and equity, simultaneously. On the other hand, some non-effective teachers could not achieve high VA and raise students' average performance despite of teaching a group of students who have historically shown to achieve high academic performance i.e. students without FRPL, White ethnic background students, female students, and students with high initial achievement. They have performed below what was expected. The former group of teachers can be considered as the "beat the odds" or "exemplary" teachers while the latter group of teachers can be considered as the "non-exemplary" or in some way the "unacceptable" teachers. Thus, by moving beyond the average values and looking into the finer detail of the variation in the results, we are able to identify exceptional and interesting groups of teachers. The identification of these teachers would enable us to obtain even more invaluable lessons (from the former teachers) and provide immediate assistance (for the latter teachers). It will enable us and policy makers to make more effective and efficient targeted policies and interventions.

In order to constructively and systematically identify these two groups of teachers, this section extends the conventional VA analysis with a new equity related dimension. But this type of analysis is not at all common in the current VA literature. There are only three noteworthy projects which include the Achievement Gap Initiative at Harvard University which utilizes achievement gap between different student ethnic groups as the equity indicator; the Value Added Research Center (VARC) at the University of Wisconsin Madison which utilizes the initial achievement level (the prior test score) as the indication of inequity in the achievement level and learning challenges students bring to the classroom; and the Program for International Student Achievement (PISA) by the OECD which constructs a socio-economic index through collecting a comprehensive list of student background indicators such as students' parental education, income, wealth and educational resources at home, etc. In order to identify the exemplary and non-exemplary teachers, these three projects then use a two dimensional quadrant plot with the VA estimates on the y-axis and the equity indicator on the x-axis.[211] These plots enable us to visualize the entire spectrum of teachers and effectively identify the exemplary teachers who fare exceptionally well on both dimensions and the non-exemplary teachers who fare low on the VA estimate but high the equity component (by unable to raise the performance of the prior high achieving students). A sample quadrant plot from the PISA study is provided below.[212]

[211] The PISA which does not conduct VA analyses (as it does not comprise of prior test scores) uses the country (and school) means on the y-axis. Some of the analysis also use adjusted means after taking into account the student background characteristics.
[212] The y-axis is the country mean student test scores and the x-axis is the effects of PISA socio-economic index on student performance. Please refer to PISA (2007) for further explanation.

Figure 10.8: Quadrant plot from PISA (2007)



Note: OECD mean used in this figure is the arithmetic average of all OECD countries.

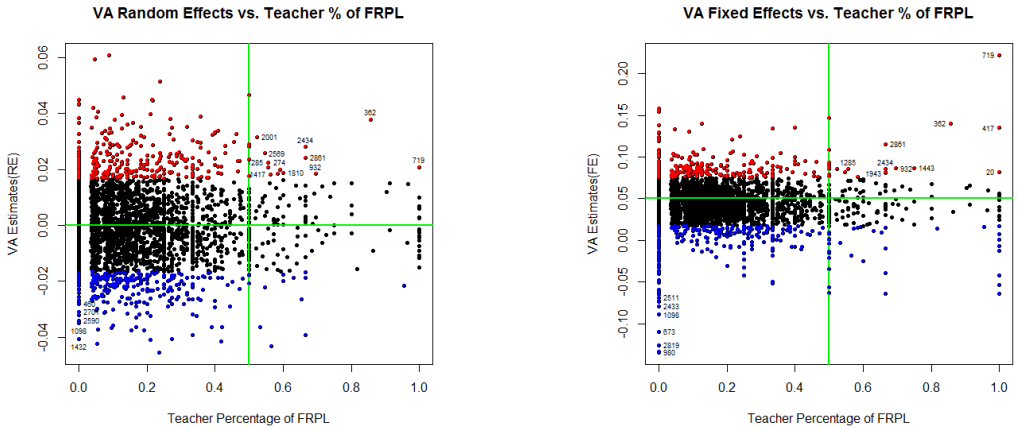
In the following sections, this study follow suit with the three studies to conduct quadrant analyses for Washington's 5th grade math teachers. It will utilize both fixed and random effects VA estimates obtained in Chapter 8 and the different equity related student background indicators made available in the data set.[213] Specifically, these indicators will include the student FRPL status, student ethnicity (as used in the Achievement Gap Initiative), student gender, and students' prior test score performance (as used in the VARC project). The exemplary and non-exemplary teachers identified in the different quadrant analysis will then be compared and contrasted.

Quadrant Analysis 1: VA Estimates vs. Percentage of FRPL Students

The quadrant analysis of the teachers' VA estimates and the percentage FRPL students in teachers' class illustrate the following results. The teachers colored blue are the effective teachers in the top tenth percentile of the VA ranking and the teachers colored in red are the non-effective teachers in the bottom tenth percentile of the VA ranking.

[213] This study acknowledges the extensive literature and different definitions implied by the word "equity". In this study and following suit with the three projects, the equity related indicator will be defined and referred to the effects (differences or gaps) of the different student background variables have on the student achievement.

Figure 10.9: Quadrant analysis 1



The two plots (based on fixed and random effects VA estimates) illustrate a very similar pattern with high concentration of teachers with low percentage of FRPL in their classes.[214] But as you can see from the top right first quadrant, there are several teachers who deviate from the overall trend to achieve high VA and high student average performance regardless of the high concentration of students on FRPL who were shown to significantly underperform the students without FRPL. As these teachers raised the performance particularly of these historically underachieving FRPL students, they have contributed to reducing the achievement gap between the FRPL and non-FRPL students. That is, these "exemplary" teachers achieved both quality and equity simultaneously. On the other hand, as you can see from the teachers in the bottom left third quadrant, these teachers could not achieve high VA and high student performance regardless of the high concentration of non-FRPL students who have shown to illustrate high prior academic performance. That is, these "non-exemplary" teachers could not deliver results despite of teaching high achieving group of students. The best candidates (top 20 or so) for the "exemplary" and "non-exemplary" teachers are listed in the table below. Some of these teachers are also labeled in the Figure 10.9 above.

[214] The difference the y-axis scale may slightly confuses the similarity of the plots, but adjusting the scales (to same range) further clarifies the similarity in the two plots.

Table 10.3: Exemplary and non-exemplary teachers based on VA estimates and percentage of FRPL students

Exemplary and Non-Exemplary Teachers Based on VA and % of FRPL Students							
Random Effects				Fixed Effects			
Exemplary		*Non-Exemplary*		*Exemplary*		*Non-Exemplary*	
274	1875	76	1528	20	1883	180	1528
362	1970	143	1683	215	1943	557	1660
719	2001	158	1700	362	2001	673	1822
932	2085	259	1771	417	2085	692	1876
1285	2430	460	2080	719	2430	863	2180
1380	2434	723	2342	932	2434	980	2366
1417	2437	1037	2590	1285	2569	1070	2433
1608	2569	1098	2608	1380	2583	1098	2511
1715	2650	1217	2670	1443	2730	1304	2652
1810	2861	1432	2701	1608	2861	1432	2819

As shown in the highlighted cells, random effects and fixed effects models identified a number of the same teachers as exemplary and non-exemplary teachers. For the exemplary teachers, over 50% of the top 20 candidates were matched across the two models. And for the non-exemplary teachers, there were only 15% matching rate. But as shown in the Appendix, expanding the list to the top 30 to 40 non-exemplary teachers, a much higher matching rate were found. These findings give us confidence and assurance in the identified teachers. In the following sections, the above quadrant analysis is repeated for the percentage of White students in teachers' classes, percentage of female students in teachers' classes, and the teacher average prior test scores. The exemplary and non-exemplary teachers will be identified in accordance to the VA findings provided in Chapter 8 where White students, female students, and students with high prior test scores had a significantly higher performance. For example, teachers who achieved high VA despite of teaching a class with high concentration of non-White, male, and low prior achievement students will be identified as "exemplary". On the other hand, teachers who could not achieve high VA despite of teaching a class with high concentration of White, female, and with high prior test scores will be identified as "non-exemplary" teachers. The quadrant analyses together with the best candidates (the top 20) for the two groups of teachers are provided as follow.

Quadrant Analysis 2: VA Estimates vs. Percentage of White Students

Figure 10.10: Quadrant analysis 2

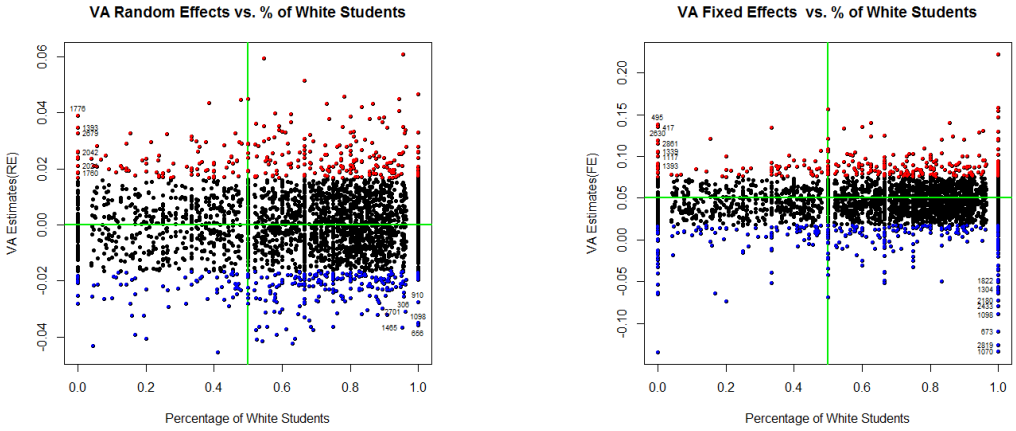


Table 10.4: Exemplary and non-exemplary teachers based on VA estimates and percentage of White students

Exemplary and Non-Exemplary Teachers Based on VA and % of White Students							
Random Effects				Fixed Effects			
Exemplary		*Non-Exemplary*		*Exemplary*		*Non-Exemplary*	
274	1881	123	1465	417	1881	180	1995
1117	1970	306	1481	495	1943	330	2013
1138	2021	476	1580	604	2021	673	2100
1171	2042	656	1746	814	2430	927	2180
1393	2191	784	1830	1117	2502	1070	2366
1743	2430	910	2013	1339	2583	1098	2433
1760	2650	1070	2551	1393	2630	1304	2520
1776	2679	1098	2701	1716	2650	1448	2566
1810	2793	1276	2819	1776	2679	1660	2819
1855	2861	1361	2829	1855	2861	1822	2829

Quadrant Analysis 3: VA Estimates vs. Percentage of Female Students (Class Gender Balance)

Figure 10.11: Quadrant analysis 3

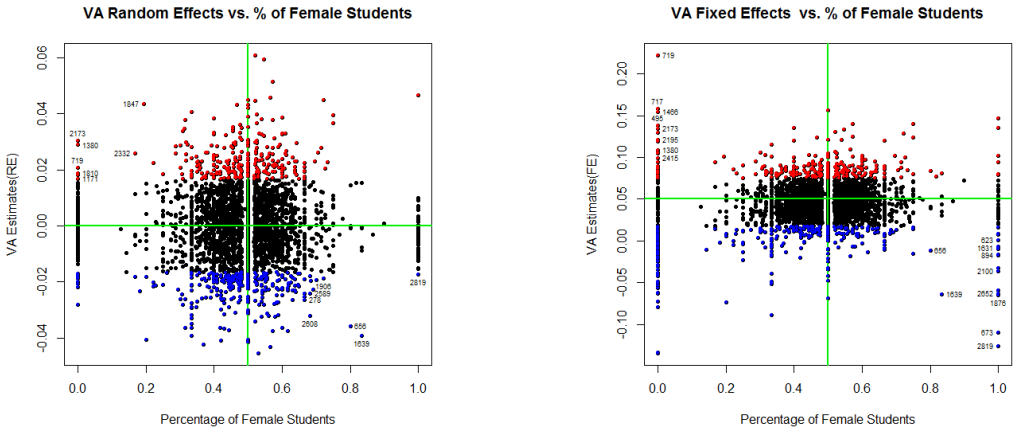


Table 10.5: Exemplary and non-exemplary teachers based on VA estimates and percentage of female students

Exemplary and Non-Exemplary Teachers Based on VA and % of Female Students							
Random Effects				Fixed Effects			
Exemplary		*Non-Exemplary*		*Exemplary*		*Non-Exemplary*	
204	1380	152	1818	20	1772	623	1876
560	1393	259	1906	249	1774	656	2063
620	1810	278	2074	495	2107	673	2100
719	1847	419	2086	717	2173	821	2168
932	1881	656	2342	719	2195	863	2177
1171	1976	793	2439	814	2415	894	2520
1221	2173	1108	2589	1339	2502	927	2615
1270	2195	1572	2608	1380	2606	1304	2652
1285	2332	1639	2760	1466	2630	1631	2819
1350	2445	1813	2819	1716	2730	1639	2827

Quadrant Analysis 4: VA Estimates vs. Average Prior Test Scores

Figure 10.12: Quadrant analysis 4

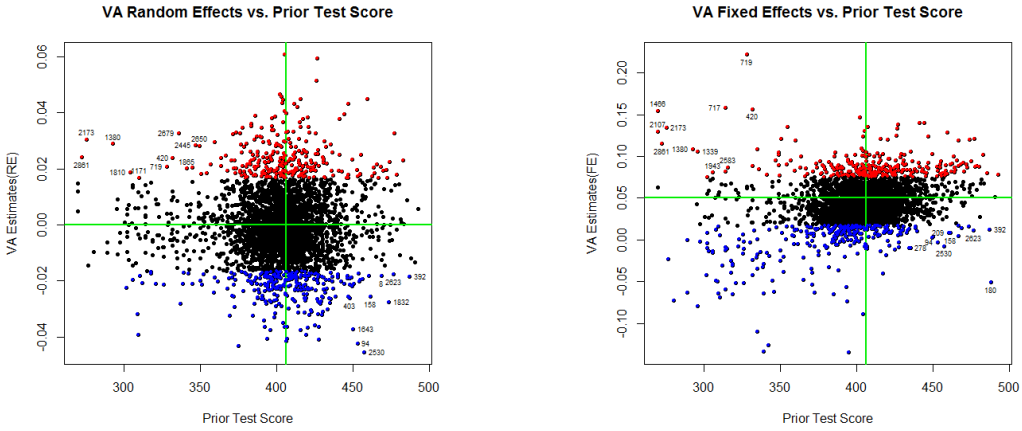


Table 10.6: Exemplary and non-exemplary teachers based on VA estimates and average prior performance

Exemplary and Non-Exemplary Teachers Based on VA and Initial Performance Level

Random Effects				Fixed Effects			
Exemplary		*Non-Exemplary*		*Exemplary*		*Non-Exemplary*	
37	1865	7	509	20	1943	8	1643
420	2093	8	842	294	2107	50	1832
470	2173	94	1643	420	2127	94	2038
560	2195	116	1742	657	2173	158	2095
719	2406	133	1832	717	2445	180	2505
1117	2445	158	1969	719	2502	209	2509
1171	2650	209	2336	1339	2583	278	2530
1380	2679	306	2530	1380	2650	392	2531
1469	2793	392	2531	1466	2679	1089	2551
1810	2861	403	2623	1716	2730	1265	2623
				1865	2861		

Note: The top 22 exemplary teachers for the fixed effects were considered take into account of several tied ranking and to accommodate few more matches

The quadrant analysis for the other student background variables (equity indicators) also illustrated good matching rates of the two models in identifying the exemplary and non-exemplary teachers. The exemplary teachers based on the percentage of White students and students' average prior test scores illustrated approximately 50% matching rate. The exemplary teacher based on the percentage of female students, on the other hand, illustrated relatively low matching rate of 20% but when the top 30 candidates were considered (as shown in the Appendix), the number of matches increased. And for the non-exemplary teachers, the average prior test score illustrated high match rate of 50% and the percentage of White students and female illustrated relatively low rate of match rate (25% and 10%, respectively). But when further candidates were considered

for the latter two variables, the match rates again increased significantly. In the following section, we now further investigate the exemplary and non-exemplary teachers identified in the different quadrant analysis to possibly learn some of the reasons behind their performances. But before we do so, the key message regarding the above quadrant analysis is that, it greatly extends and deepens the conventional VA analysis which only looks at the dimension of quality (the magnitude of the VA estimates) and not equity. Through incorporating equity related indicators (such as the student contextual background), the above analysis enabled us to identify a special group of effective teachers who achieved quality outcomes with high VA estimate and simultaneously contributed towards reducing the achievement gap inherent between different groups of students. Similarly, it also enabled us to identify a group of teachers who were unable to achieve the expected quality performance despite of teaching group students with high prior levels of achievement. The quadrant analysis therefore greatly improves the application of the VA estimates for policy relevant purposes. Through making the teachers of success and in need of improvement even more vivid, it assists policy makers in making even more effective targeted policies.

Identifying the Extra-Exemplary and Extra Non-Exemplary Case Teachers

In this section, further exploration and examination of the "exemplary" and "non-exemplary" teachers identified in the four quadrant analysis is conducted. Through comparing the findings from the quadrant analyses, this analysis further solidifies the validity and confidence in the findings. Specifically, teachers who achieved high VA despite teaching a group of students with multiple underachieving backgrounds will be identified. These teachers will be referred to as the "extra-exemplary" teachers. These teachers achieved quality outcomes and contributed towards closing more than one form of achievement gap simultaneously. On the other hand, teachers who could not achieve high VA despite teaching a group of students with multiple high performing backgrounds will also be identified. These teachers will be referred to as the "extra-non-exemplary" teachers. These teachers performed below what was expected. The underlying features of these two groups of teachers will then be closely examined. The "exemplary" and "non-exemplary" teachers identified by <u>both</u> fixed and random effects in the respective quadrant analysis are summarized again below.[215] Teachers who are identified as exemplary and non-exemplary in <u>more than one</u> quadrant analysis are then classified as "extra-exemplary" (highlighted in blue) and "extra- non-exemplary" (highlighted in green).

[215] The top 32 candidates for exemplary teachers based on the percentage of female students in considered due to low match rates in the top 20. The top 40 candidates for the non-exemplary teachers for all equity indicators are considered as the top 20 identified only 1 extra-non-exemplary teacher. These extended candidates and ranking are provided in the Appendix.

Table 10.7: Exemplary and non-exemplary teachers identified by both fixed and random effects

Exemplary and Non-Exemplary Identified By Both Random and Fixed Effects							
VA and % FRPL Students		**VA and % White Students**		**VA and % Female Students**		**VA and Avg Initial Perform**	
Exemplary	*Non-Exemplary*	*Exemplary*	*Non-Exemplary*	*Exemplary*	*Non-Exemplary*	*Exemplary*	*Non-Exemplary*
362	460	1117	656	719	259	420	8
719	692	1393	1070	1285	278	719	94
932	980	1776	1098	1380	656	1380	158
1285	1070	1855	1746	1847	793	1865	209
1380	1098	1881	2013	2173	1108	2173	278
1608	1432	2021	2551	2195	1639	2445	392
2001	1528	2430	2819	2332	1813	2650	1265
2085	2368	2650	2829		1818	2679	1643
2430	2511	2679			2074	2861	1832
2434	2551	2861			2439		2038
2569	2819				2589		2505
2861					2608		2530
					2760		2531
					2819		2551
							2623

Note: Due to the low matching rate in the top 20, the top 32 candidates for the exemplary teachers based on the percentage of White students is considered. The top 40 candidates for the non-exemplary teachers for each equity indicator is considered. These extended list of candidates are provided in the Appendix.

As you can see, there are number of "extra-exemplary" and "extra- non-exemplary" teachers being identified. To obtain a better understanding of these teachers, we now study their underlying characteristics using the available data.

Understanding the Extra-Exemplary Teachers

Looking first at the student peer effects (types of students these teachers teach) and the teacher background characteristics of the extra-exemplary teachers illustrate the following results. Further clarification of the variables is provided in the footnote.[216]

[216] St = students, T = teachers, S = schools. The percentage of FRPL students is the percentage of FRPL students in the respective teachers. This calculation applies for the rest of the student background characteristics. The teacher characteristics are defined specifically for each of the teachers. T size is the class size or number of students the respective teachers teaches. Finally, the overall teacher mean is the average of these variables across all the 5th grade math teachers in Washington.

Table 10.8: Teacher characteristics of extra-exemplary teachers

Extra-Exemplary Teachers Characteristics										
TeacherID	% FRPL St	% White St	% Female St	Avg Initial Perf	T Female	T White	T with MA	T Experience	T Size	School ID
719	1.000	1.000	0.000	328.0	1.00	1.00	1.00	21.9	1	317
1285	0.556	0.444	0.222	404.0	0.00	1.00	1.00	14.4	9	602
1380	0.500	0.500	0.000	292.8	0.00	1.00	1.00	11.4	4	416
2173	0.333	0.333	0.000	275.3	1.00	1.00	0.00	5.2	3	467
2430	0.500	0.000	0.500	396.4	0.00	1.00	0.00	2.0	8	345
2650	0.500	0.056	0.556	349.7	1.00	1.00	1.00	1.2	18	67
2679	0.267	0.000	0.600	335.9	1.00	1.00	0.00	1.0	15	752
2861	0.667	0.000	0.333	272.3	1.00	1.00	1.00	0.0	3	404
Mean	0.540	0.292	0.276	331.8	0.625	1.000	0.625	7.14	7.63	
SD	0.209	0.332	0.242	47.4	0.484	0.000	0.484	7.44	5.74	
Median	0.50	0.19	0.28	331.9	1.00	1.00	1.00	3.60	6.00	
Mini	0.27	0.00	0.00	272.3	0.00	1.00	0.00	0.0	1.0	
Max	1.00	1.00	0.60	404.0	1.00	1.00	1.00	21.9	18.0	
Overall T Mean	0.174	0.650	0.476	402.4	0.750	0.935	0.663	13.57	17.86	

Notes: St - students, T=teacher, S=School

As expected, the extra-exemplary teachers teach a more challenging group of students. They are teaching more FRPL, non-White, male, and low initial average performance students in comparison to the overall means across the entire 5th grade math teachers. Looking at the teacher characteristics, the mean percentage of female, White, and with MA degree among these teachers illustrate very similar values to the overall means. But a few noteworthy points are also evident with the teacher experience and teacher class size variables.[217] Looking first at the experience level, the first three teachers have a close to the average years of experience (slightly more for the first teacher) but three of the other teachers (highlighted in purple) have significantly fewer years of experience. They all have 2 or fewer years of experience and can be considered as novice teachers. But as it was shown earlier that teacher experience has a positive significant effect on student performance, these latter three teachers "beat the odds" of these teacher experience effects to achieve higher than expected performance (high VA) despite of teaching a group of challenging students. Something extraordinary might be characterizing these three teachers. We might expect them to be highly educated (possibly with a special MA degree pertaining to teaching) but as highlighted above, only one of these teachers have a MA degree. Moreover, these teachers all teach classes with decent class sizes. Their success is not explained by a few extreme or exceptional students. Instead, their class sizes provide us with the reliability in the findings.[218] In all, these three teachers may have special insights into the factors contributing to their success which is unmeasured in the data set.

[217] To re-clarify, the teacher class size variable is the number of students the teacher teaches in their math classes.

[218] For this reason, the other two teachers who also had very low levels of experience but only had three students were not emphasized in the text. Similarly, the reliability of the finding for the first teacher (teacher ID 719) with only 1 student is also questionable. As shown in the school level data, this teacher is also the only 5th grade math teacher in the school. Further investigation of these small class size teachers is therefore necessary before we interpret the findings.

Looking now at the characteristics of the schools in which the extra-exemplary teachers work illustrate the following results. These school level characteristics are defined only for the 5th grade students and 5th grade math teachers. Further clarification of the variables is provided in the footnote.[219]

Table 10.9: School characteristics (5th grade students and 5th grade math teachers) of extra-exemplary teachers

Extra-Exemplary Teachers' School Averages (of 5th Grade Students and 5th Grade Math Teachers)										
School ID	% FRPL St	% White St	% Female St	Avg Initial Perf	% Female T	% White T	% T with MA	Avg T Exp	Avg T Size	S Size
317	1.000	1.000	0.000	328.0	1.00	1.00	1.00	21.9	1.0	1
602	0.389	0.185	0.500	399.3	0.43	0.65	0.78	10.5	11.5	54
416	0.333	0.587	0.429	397.0	0.71	0.78	1.00	6.7	14.1	63
467	0.089	0.750	0.464	409.5	1.00	1.00	0.64	8.4	17.0	56
345	0.348	0.261	0.435	397.6	0.65	1.00	0.65	1.7	7.7	23
67	0.436	0.128	0.538	368.1	1.00	0.97	1.00	6.3	17.6	39
752	0.275	0.075	0.550	371.7	1.00	1.00	0.21	5.8	18.6	120
404	0.357	0.690	0.452	403.3	0.55	1.00	0.07	29.8	18.3	42
Mean	0.403	0.460	0.421	384.3	0.793	0.925	0.669	11.38	13.23	49.75
SD	0.246	0.320	0.165	25.4	0.221	0.127	0.337	8.90	5.84	32.47
Median	0.35	0.42	0.46	397.3	0.86	1.00	0.71	7.51	15.57	48.00
Mini	0.09	0.08	0.00	328.0	0.43	0.65	0.07	1.7	1.0	1.0
Max	1.00	1.00	0.55	409.5	1.00	1.00	1.00	29.8	18.6	120.0
Overall S Mean	0.178	0.656	0.490	404.6	0.742	0.932	0.669	13.75	19.33	54.7

Notes: St - students, T=teacher, S=School

It is evident that these teachers are working in less favorable school environments characterized with high percentage of academically challenging 5th grade students. The mean values all indicate less favorable conditions than the overall means across the schools in which the 5th grade math teachers work. It could be possible that these teachers are working in schools situated in less affluent areas. Looking at the types of 5th grade math teachers working in these schools, the percentages of female, White, and with MA degree teachers do not illustrate much difference from the overall school means. But as highlighted in purple, it is evident that the 5th grade math teachers' average experience level in the schools in which the three highlighted teachers mentioned previously work is very low. These three schools comprise a young group of 5th grade math teachers. And this finding is again not associated with the smallness of the school size (number of 5th grade students) or the smallness of the number of 5th grade math teachers working in the schools. In fact, the third school is a very large school with 120 5th grade students which is more than twice the overall school average.

To sum, the investigation of the extra-exemplary teachers with the available data indicated that the success of these teachers was not due to their background or the favorable school environment characterized with high concentration of high achieving 5th grade students and highly qualified 5th grade math teachers. In fact, three extra-exemplary teachers were highlighted as they achieved unexpected high performance despite having very little teaching experience and without a MA degree. These teachers therefore also beat the odds of the overall effects of teacher characteristics on student performance. These teachers also had average class sizes. These

[219] St = students, T = teachers, S = schools. The percentage of FRPL students is now the percentage of 5th grade students who receive FRPL in the respective school. This calculation applies for the rest of the student background characteristics. The percentage of female teachers is the percentage of 5th grade math teacher who is female in the respective school. This calculation applies for the rest of the teacher characteristics. Finally, school size is the number of 5th grade students in the particular school. It is not the number of students in the entire school (across all grades). The overall school means are the average of these variables across all the schools to which the 5th grade teachers belong.

findings hints us and policy makers that the success of these three teachers are most likely due to un-measurable factors beyond the available data such as their instruction methods, teaching philosophy, leadership, use of special technology, etc. Understanding of these factors can provide us with invaluable lessons as to how to achieve both quality outcomes in face of challenging group of students.

Understanding the Extra Non-Exemplary Teachers

Looking at the student peer effects variables and the teacher background characteristics of the extra-non-exemplary teachers illustrate the following results.

Table 10.10: Teacher characteristics of extra-non-exemplary teachers

Extra Non-Exemplary Teachers Characteristics										
TeacherID	% FRPL St	% White St	% Female St	Avg Initial Perf	T Female	T White	T with MA	T Experience	T Size	School ID
278	0.333	0.667	0.667	435.3	0.00	1.00	1.00	31.9	6	495
656	0.100	1.000	0.800	424.9	1.00	1.00	0.00	22.0	10	528
1070	0.000	1.000	0.000	339.0	1.00	1.00	1.00	12.3	1	897
1098	0.000	1.000	0.333	404.0	1.00	1.00	1.00	27.6	3	437
2551	0.000	1.000	0.500	434.3	0.00	1.00	1.00	1.4	4	14
2819	0.000	1.000	1.000	342.0	1.00	1.00	0.00	0.0	1	904
Mean	0.072	0.944	0.550	396.6	0.667	1.000	0.667	15.87	4.17	
SD	0.122	0.124	0.324	41.0	0.471	0.000	0.471	12.29	3.13	
Median	0.00	1.00	0.58	414.4	1.00	1.00	1.00	17.15	3.50	
Mini	0.00	0.67	0.00	339.0	0.00	1.00	0.00	0.0	1.0	
Max	0.33	1.00	1.00	435.3	1.00	1.00	1.00	31.9	10.0	
Overall T Mean	0.174	0.650	0.476	402.4	0.750	0.935	0.663	13.57	17.86	

Notes: St - students, T=teacher, S=School

As you can see, unlike the previous case, the extra-non-exemplary teachers teach much higher concentration of students with high achieving background. The percentage of FRPL, White, and female students also illustrate better values than the overall 5th grade math teachers means. The average initial performance is approximately at the the overall average but it is clearly larger than the extra-exemplary teachers of 339.6. Looking at the teacher characteristics, the percentage of female, White, and with MA teachers show close resemblance to the overall teacher mean values. But looking at the teacher experience and teacher class size variables illustrate interesting results. As highlighted in purple, some of these teachers have near to above average years of experience and at the same time possess a MA degree. This finding is unexpected as these characteristics were found to have significant positive effects on student performance. That is, these teachers could not achieve quality outcomes despite having a qualified background themselves. On the other hand, as highlighted in green, the last two teachers are beginning novice teachers with less than 2 years of teaching experience. These teachers may require immediate assistance as they are teaching a group of academically challenging students without the extensive and qualified backgrounds. But one finding which challenges the reliability of these findings is the extremely small class sizes for these teachers. All the teachers have less than or equal to 6 students in their class. Their VA estimates can be susceptible to the extreme and outlying observations problem described earlier e.g. few students just happen to perform horribly on the day of the test. This also opens the door to other alternative interpretations. For example, the qualified teachers highlighted in purple

may have been purposefully assigned to few students with particular needs and challenges e.g. attention problems, hearing problems, special difficulty in mathematics, etc. Similarly, the teachers highlighted in green may be young, promising, and highly acknowledged teachers who were purposefully assigned to teach the challenging group of selective students. For both of these (highlighted) groups of teachers, further data collection and/or field level investigation is now required to fully understand their circumstances.

Looking now at the school characteristics where these teachers work illustrate the following results.[220]

Table 10.11: School characteristics (5th grade students and 5th grade math teachers) of extra-non-exemplary teachers

Extra Non-Exemplary Teachers' School Averages (of 5th Grade Students and 5th Grade Math Teachers)										
SchoolID	% FRPL St	% White St	% Female St	Avg Initial Perf	% Female T	% White T	% T with MA	Avg T Exp	Avg T Size	S Size
495	0.099	0.846	0.560	425.7	0.648	1.000	0.868	21.3	12.0	91
528	0.045	0.909	0.500	407.7	0.455	1.000	0.000	26.4	11.6	22
897	0.162	0.757	0.446	402.0	0.392	1.000	0.703	6.5	20.7	74
437	0.077	1.000	0.513	415.3	0.667	1.000	0.667	20.4	11.4	39
14	0.000	1.000	0.500	434.3	0.000	1.000	1.000	1.4	4.0	4
904	0.014	0.972	0.535	419.7	1.000	1.000	0.986	15.0	23.1	71
Mean	0.066	0.914	0.509	417.4	0.527	1.000	0.704	15.16	13.80	50.17
SD	0.055	0.089	0.035	10.7	0.305	0.000	0.339	8.70	6.37	30.87
Median	0.06	0.94	0.51	417.5	0.55	1.00	0.79	17.71	11.81	55.00
Mini	0.00	0.76	0.45	402.0	0.00	1.00	0.00	1.4	4.0	4.0
Max	0.16	1.00	0.56	434.3	1.00	1.00	1.00	26.4	23.1	91.0
Overall S Mean	0.178	0.656	0.490	404.6	0.742	0.932	0.669	13.75	19.33	54.66

Notes: St - students, T=teacher, S=School

Unlike the extra-exemplary teachers, it is clearly evident that the extra-non-exemplary teachers are working in more favorable school environments characterized with higher concentration of 5th grade students with high achieving backgrounds. It seems like these teachers are working in schools situated in affluent areas. Looking at the types of 5th grade math teachers working in these schools, it is also evident that the percentages of teacher characteristics associated with higher student achievement are near or better than the overall means. This implies that the extra-non-exemplary teachers could not achieve quality outcomes despite being surrounded with highly qualified 5th grade math teachers as their colleagues. Looking now at the purple and green highlights, it is evident that the previously highlighted teachers work in schools which have a decent size average 5th grade math classes. Some of the schools have average class size which is larger than the overall school means (average of all the school means). This finding implies that the extremely small class size of the previously highlighted extra-non-exemplary teachers is very small even within their respective schools. And this may suggest the alternative interpretation pertaining to these teachers i.e. teachers who are assigned to few students with particular needs. Finally, the second to the last (previously highlighted as green) teacher was no longer highlighted in the above table. The finding above illustrate that this teacher (ID number 2551) is the only 5th grade math teacher in this school and there is only 4 5th grade students in this school. This schools is either extremely small school or potentially susceptible to measurement and coding error. Further investigation of this teacher and school is necessary to provide any reliable interpretation.

[220] As described in previous footnote, these school level characteristics are again defined only for the 5th grade students and 5th grade math teachers.

Summary and Conclusion

This chapter greatly extended the conventional application of VAMs for policy making purposes. It first conducted the conventional VA analysis to identify the "effective" and "non-effective" teachers based solely on the VA estimates. It then studied the underlying features of these two groups of teachers to find that effective teachers on average were surrounded with more favorable conditions to raise student performance. But unlike the majority of the VA analyses today, this study followed suit with the three innovative studies, to extend the analysis with an student background variables (equity related indicator) to identify the so called "exemplary" teachers who performed better than expected despite teaching a group of academically challenging students and the "non-exemplary" teachers who performed below what was expected despite teaching students with prior academic achievement. The former teachers achieved quality outcomes particularly for the underachieving students and therefore contributed towards reducing the achievement gap. This extended analysis was conducted using a variety of student background variables related to equity and the results were then compared to identify the "extra-exemplary" teachers who achieved quality despite facing multiple forms of underachieving students and the "extra non-exemplary" teachers who could not achieve quality despite teaching multiple forms of high achieving students. Studying these two groups of teachers also identified several promising teachers who were extra-exemplary without a higher educational degree or many years of experience. On the other hand, several problematic teachers who were extra non-exemplary despite having a higher education degree and above average years of experience were also identified. Further investigation and assistance can possibly be provided for these teachers. In all, this extensive exercise opened new doors and possibilities to further maximize the potential of VAMs. It enabled us to further narrow our search for teachers of true achievement and teachers in significant need of further improvement. And it thereby improved the effectiveness and efficiency of applying the results to inform targeted policies and interventions.

Yet, through the process of conducting the above analysis, a fundamental limitation of VAMs was also realized. That is, the identification of the true reason(s) for success or failure is beyond the scope of the available data used in conducting VA analyses. The un-measurable factors such as teachers' instructional methods, leadership, class management, teaching philosophy, teachers' cultural and ethnic understanding, application of innovative technologies, to school policies, professional development opportunities, teacher collaboration, and district leadership, district financing and resource, and more, all have important roles in determining the academic success of a student. What the VA analyses did provide us is an initial crucial step towards realizing these factors. It opened the first door towards better understanding of the spectrum of reasons determining educational success and failure. As emphasized in the following paragraph by Stecher et al. (2010) (in an evaluation report of the NCLB administered by the RAND Corporation) the ideal use of VAMs demands a two-step strategy where the first step comprised of thorough identification of the successful

and non-successful teachers and second step comprised of field level investigation to understand the educational and instructional practice of the identified teachers.

> A more flexible and effective system is needed to allow states and districts to identify and prioritize the teachers most in need and to design consequences to address their particular needs. A two-stage process might work better than the current set of automatic interventions [based on AYP]. In the first stage, states would use a test-based metric to identify teachers that were potentially having problems. In the second stage, states would gather additional information from these teachers about local deficiencies so that they could craft interventions that responded to these shortcomings. A number of states and districts use independent "inspectorates" to conduct field reviews of teachers to provide more complete and thorough information about local practices and improvement options. (Stecher et al. (2010), p. 59)

This study successfully completed and also strengthened the first step. It paves the path for the second step. This study hopes for an "exemplary" education system which achieves both quality and equity simultaneously for all students.

CHAPTER 11: CONCLUSION AND FUTURE REMARKS

A successful educational reform requires collective action. Collective action is indispensable particularly when the educational outcome of a child is in the hands of multiple stakeholders from the federal and state policy makers, district officials, school principals, teachers, community leaders, to the parents. As Hilary Clinton once said, "it takes the entire community to educate a child," it is critical for all stakeholders to come together as one to offer the best education system that students deserve. But collectivity and uniformity are challenged when confusion and chaos characterize the topic of interest such the VAMs. VAMs are studied by researchers from a broad range of academic disciplines who remain divided over the best methods in analyzing the models. Moreover, other stakeholders who play a pivotal role in the education system but without the extensive technical background in statistics – policy makers, district officials, school principals, teachers, and parents have been excluded from the technical debate of VAMs. The lack of transparency and unified understanding in the implementation of VAMs has inevitably devalued VAMs from becoming an effective tool to help transform the educational system.

Interestingly, history attests that chaos and confusion bring positive outcomes. During the post war era, Akio Morita, the then CEO of Sony electronics stated: "chaos is a window of opportunity for great innovation and success." Similarly, Albert Einstein professed that "in the middle of every difficulty lies opportunity." This study, therefore, conducted a comprehensive investigation of VAMs to cut through its Gordian knot so that all parties irrespective of their background are equally informed and can benefit from the model findings. This study, unlike the majority of existing VA research which conducts highly advanced statistical analyses, took a hands on pedagogical approach in describing the models in plain language. It targeted a wider range of audience particularly those without an extensive background in statistics – policy makers, district officials, school principals, teachers, and parents. In Part I of this study, the conceptual and technical dimensions characterizing the VAM were clarified. The definition and purpose of statistics were re-examined. Focused attention was then provided for the linear regression models which lie at the heart of the VAMs. And the overarching ideas and thought processes underlying the assumptions of the linear models to ensure the accurate and reliable BLUE estimates were thoroughly explained. The linear regression framework was then applied to the construction of VAMs by identifying the two major tasks underlying the models – to define the VA parameter and to take into account factors outside the control of teachers. For the first task, comparison of the two major modeling option of the VA parameter – fixed effects and random effects was provided. And for the second task, the importance of collaboration between our theories and empirical evidence in deciding which variables to take into account in the models was highlighted. In Part II of this study, the conceptual explanation was applied to conduct teacher VA analyses of the 5^{th} grade math teachers in Washington State. But unlike the majority of the VA studies which often jump straight into the conclusion and policy analyses, a thorough and hands on step by step analysis, diagnosis, revision, and validation of the estimated models in

conjunction with the conceptual explanation in Part I were conducted. Only when the accurate, reliable, and robust VA estimates were achieved, the findings were applied for policy relevant purposes. The study then demonstrated how the conventional policy analysis which focuses only on the VA estimates can be extended to incorporate a student background equity related indicator. This analysis enabled the identification of the teachers who achieved high VA particularly for the historically underachieving students. Such "exemplary" teachers achieved both quality and equity in student performance simultaneously. The teachers who performed lower than expected by not achieving a high VA even when teaching a group of high prior achievement were also identified. By identifying the teachers of extraordinary success whom we can learn from and the teachers in need of improvement whom we can provide assistance to, the analysis illustrated how the models can be used as an effective and efficient tool to better target policies and interventions.

The VA research continues to unfold and advance. There are a number of other challenges and topics which were not extensively considered in this study. Some of these topics include the highly advanced VAM models such as the cross-classified models and the Bayesian models; the psychometric (testing) designs such as the vertically linked scales, the equal interval scales, and the linking error problems to measure student growth; and the management of the longitudinal data systems to minimize attrition and missing data and to extend the data collection of teacher and school background information. These factors can all potentially affect the statistical properties of the VA estimates. But the studies covering these topics are limited and the specific implications and solutions for these features are not well understood. Furthermore, as many of these topics are specific to different academic disciplines, the interdisciplinary divide further impedes the understanding of these issues. For example, the conceptual and technical link (mathematical proof) among the psychometric properties, estimation procedures, and the statistical properties VA estimates not exactly known. The implications of the new statistical assumptions underlying the complex models to ensure BLUE are also not fully understood. And the diagnosis and revision of these models are also not extensively studied. Finally, in addition to the technical issues, improvement in the reporting, dissemination, presentation, and communication of the VA findings is critical to improve the stakeholders operation in practice. There are only few training programs and capacity building opportunities to date to spread the awareness of VAMs to the stakeholders without the extensive background in statistics. But on the positive side, the explanations provided in this study apply universally to all these challenges. This study does serve as the platform and basis for engaging in further dialogue regarding these issues. The cross-classified and Bayesian models are a natural extension of the random effects models.[221] The psychometric properties and the non-randomly missing data are forms of omitted systemic patterns which need to be properly removed from the residuals and taken into account in the

[221221] The cross-classified models simply incorporate more than one random effect teacher VA parameter to estimate the cumulative effect of the multiple teachers on students' academic progression. The Bayesian models stipulate additional prior distribution assumption on top of the sampling distribution (or the maximum likelihood) in order to estimate the distribution of the parameter instead of the point and the confidence interval estimates. This idea is illustrated as follow $f(\theta/y) = f(\theta) \cdot f(y/\theta)$ where the posterior distribution is the $f(\theta/y)$ product of prior distribution $f(\theta)$ and sampling/data distribution (likelihood function) $f(y/\theta)$.

model. And for the capacity building exercise, the user friendly and step by step explanation provided in this study can contribute to the existing training programs which often do not fully delineate the complexities underlying the models. With the explanation provided in this study and by engaging in inter-disciplinary collaboration among the researchers and exercising the leadership to further disseminate the work on VAMs, the above challenges can be resolved in the near future. As addressed by the NRC (2011), "the disciplinary tradition should not dictate the model choices … but rather we need start thinking about how the different disciplines can incorporate features that are missing in their own approaches." (p.58)

An anonymous person once said, "the whole is greater than the parts" and "team play is bigger than the collection of individual acts." This study hopes that the work provided serves as the common ground and bridge to connect all stakeholders irrespective of their background so that we can all help, assist, complement, and support one another to successfully implement VAMs. It hopes that all parties are now equally informed of the essentials of VAMs and can benefit from the model findings to improve their roles in educating a child. It hopes nothing but that the VAMs can become an effective tool in helping us achieve an extra-exemplary education system by which all students irrespective of their background can achieve high academic performance.

References:

Aaronson, D., Barrow, L., & Sander, W. (January 01, 2007). Teachers and Student Achievement in the Chicago Public High Schools. *Journal of Labor Economics, 25* (1), 95-135.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole Publications.

Allison, P. D. (2001). *Missing data*. Thousand Oaks, CA: Sage Publications.

Andrejko, L. (2004). Value-Added Assessment: A View from a Practitioner. *Journal of Educational and Behavioral Statistics*, *29*(1), 7-9.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Arellano, M. (2003). *Panel data econometrics*. Oxford: Oxford University Press.

Atkinson, A. C., & Riani, M. (2000). *Robust diagnostic regression analysis*. New York: Springer.

Ballou, D. (2005). Value-Added Assessment: Lessons from Tennessee. In R.W. Lissitz (Ed.), *Value-added models in education: Theory and application* (pp. 272-297). Maple Grove, MN: JAM Press.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for Student Background in Value-Added Assessment of Teachers. *Journal of Educational and Behavioral Statistics, 29* (1), 37-65.

Bartholomew, D. J., Steele, F., Moustaki, I. & Galbraith, J.I. (2008). *Analysis of multivariate social science data*. Boca Raton, FL: CRC Press.

Bell, P. (2008*). Learning science in informal environments: People, places and pursuits*. National Research Council. Washington D.C.: National Academy Press.

Bell, B. A., Ferron, J. M., & Kromrey, J. D. (2008). *Cluster Size in Multilevel Models: The Impact of Sparse Data Structures on Point and Interval Estimates in Two-Level Models*. Retrieved November 15, 2012 from http://www.amstat.org/sections/srms/proceedings/y2008/Files/300933.pdf

Bickel, R. (2007). *Multilevel analysis for applied research: It's just regression!*. New York: Guilford Press.

Bill & Melinda Gates Foundation (2010). *Learning About Teaching: Initial Findings from the Measures of Effective Teaching Project.* Retrieved October 17, 2010, from http://www.metproject.org/downloads/Preliminary_Findings-Research_Paper.pdf

Boardman, E.E. & Murnane, R.J. (1979). Using Panel Data to Improve Estimates of the Determinants of Educational Achievement. *Sociology of Education*, 52, 113-121.

Bock, R. D., University of California, Los Angeles., & NORC (Organization). (1989). *Multilevel analysis of educational data*. San Diego: Academic Press.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. Hoboken, N.J: Wiley-Interscience.

Boudett, K. P., City, E. A., & Murnane, R. J. (2005). *Data wise: A step-by-step guide to using assessment results to improve teaching and learning*. Cambridge, MA: Harvard Education Press.

Bransford, J., National Research Council (U.S.), & National Research Council (U.S.). (2000). *How people learn: Brain, mind, experience, and school*. Washington, D.C: National Academy Press.

Braun, H. I. (2005). *Using student progress to evaluate teachers: A primer on value-added models*. Princeton, NJ: Education Testing Service.

Braun, H. I. (2005). Value-Added Modelling: What Does Due Diligence Require? In R. Lissitz, *Value Added Models in Education: Theory and Applications* (pp. 19-39). Maple Grove, Minnesota: JAM Press.

Braun, H. I. (2006). Empirical Bayes. In J. G. (eds.), *Complementary methods for research in education.* Washington, DC.: American Educational Research Association.

Braun, H. I., Y. Qu & C. S. Trapani. (2008). *Robustness of value-added analysis of school effectiveness.* ETS RR-08-22. Princeton, NJ: Educational Testing Service.

Briggs, D.C., Weeks, J.P., & Wiley, E. (2008). *The sensitivity of value-added modeling to the creation of a vertical score scale*. Paper presented at the National Conference on Value- Added Modeling, University of Wisconsin-Madison, April 22-24.

Briggs, D. C. & Domingue, B. (2011). *Due Diligence and the Evaluation of Teachers*. Retrieved May 12, 2013 from http://nepc.colorado.edu/thinktank/review-learning-about-teaching

Brown, S. R., & Melamed, L. E. (1990). *Experimental design and analysis*. Newbury Park, Calif: Sage Publications.

Bryk, A., Y. Thum, J. Easton & S. Luppescu. (1998). *Academic productivity of Chicago public elementary schools, technical report.* Chicago, Il.: The Consortium on Chicago School Research.

Bryk, A. S. (2010). *Organizing schools for improvement: Lessons from Chicago*. Chicago: University of Chicago Press.

Callender, J. (2004). Value-Added Student Assessment. *Journal of Educational and Behavioral Statistics*, *29*(1), 5-5.

Casella, G., & Berger, R. L. (2002). *Statistical inference*. Australia: Thomson Learning.

Chatterjee, S., & Price, B. (1977). *Regression analysis by example*. New York: Wiley.

Christensen, R. (1991). *Linear models for multivariate, time series, and spatial data*. New York, N.Y: Springer-Verlag Inc.

Clarke, P., & Wheaton, B. (2007). Addressing Data Sparseness in Contextual Population Research. *Sociological Methods & Research, 35* (3), 311-351

Clotfelter, C.T., Ladd, H.F. & Vigdor, J.L. (2006). Teacher-Student Matching and the Assessment of Teacher Effectiveness. *Journal of Human Resources, XLI*(4), 778-820.

Clotfelter, C.T., Ladd, H.F., & Vigdor, J.L. (2007). Teacher credentials and student achievement in high school: A cross-subject analysis with student fixed effects. Working paper 13617. Cambridge, MA: *National Bureau of Economic Research.*

Cohen, J., & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, N.J: Lawrence Erlbaum Associates.

Coleman, J. S., Azrael, J. R., & Social Science Research Council (U.S.). (1965).*Education and political development*. Princeton, N.J: Princeton University Press.

Coleman, J. S., & Hoffer, T. (1987). *Public and private high schools: The impact of communities*. New York: Basic Books.

Cowart, B. & Myton, D. (1997). The Oregon's Teacher Effectiveness Work Sample Methodology: Rationale and Background. In J. M. (Ed.), *Grading Teachers, Grading Schools: Is Student Achievement a Valid Evolutional Measure?* (pp. 11-14). Thousand Oaks, CA: Corwin Press, Inc.

Cox, D. R., & Solomon, P. J. (2003). *Components of variance*. Boca Raton, FL: Chapman & Hall/CRC.

Cunningham, E. P., & Henderson, C. R. (1968). An Iterative Procedure for Estimating Fixed Effects and Variance Components in Mixed Model Situations.*Biometrics, 24* (1), 13-25.

David, H.A. & Edwards, A.W.F. (2001). Annotated readings in the history of statistics. York, N.Y: Springer-Verlag Inc.

Diggle, P., & Diggle, P. (2002). *Analysis of longitudinal data*. Oxford: Oxford University Press.

Doran, H. & J.Cohen. (2005). The Confounding Effects of Linking Bias on Gains Estimated from Value-Added Models. In R. Lissitz, *Value-Added Models in Education: Theory and Applications* (pp.80-110). Maple Grove, MN: JAM Press.

Doran, H. C., & Lockwood, J. R. (2006). Fitting Value-Added Models in R. *Journal of Educational and Behavioral Statistics, 31* (2), 205-230.

Doran, H. & T. Jiang. (2006). The Impact of Linking Error in Longitudinal Analysis: An Empirical Demonstration. In R. Lissitz, *Longitudinal and Value-Added Models of Student performance* (pp. 210-229). Maple Grove, MN: JAM Press.

Doran, H. C., & Lockwood, J. R. (January 01, 2006). Fitting Value-Added Models in R. *Journal of Educational and Behavioral Statistics, 31*, 2, 205-230.

Dougherty, C. (2007). *Introduction to econometrics*. Oxford: Oxford University Press.

Economic Policy Institute. (2010, August). *The Problems with the Use of Student Test Scores to Evaluate Teachers.* (Briefing Paper No. 278). Retrieved May 12, 2013 from http://www.epi.org/page/-/pdf/bp278.pdf

Edwards, A.L. (1985). *Multiple regression and the analysis of variance and covariance*. New York, N.Y.: W.H. Freeman and Company.

Eliason, S. R. (1993). *Maximum likelihood estimation: Logic and practice*. Newbury Park, CA: Sage.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, N.J: L. Erlbaum Associates.

Engineering Statistics Handbook (2012). Retrieved May 12, 2013 from http://www.itl.nist.gov/div898/handbook/

Ferrão, M. (2007). Sensitivity of VAM Specifications: Measuring Socio- Economic Status*: A Background Report for the OECD Project on the Development of Value-added Models in Education Systems*. Warsaw.

Finkel, S. E. (1995). *Causal analysis with panel data*. Thousand Oaks, Calif: Sage Publications.

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

Fisher, R. A. (1935). *The design of experiments*. Edinburgh: Oliver and Boyd.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004).*Applied longitudinal analysis*. Hoboken, N.J: Wiley-Interscience.

Fox, J. (2004). Robust Diagnostic Regression Analysis (Book). *Sociological Methods & Research, 32*,4.

Fox, J. (2008). *Applied regression analysis and generalized linear models*. Los Angeles: Sage.

Frees, E. W. (2004). *Longitudinal and panel data: Analysis and applications in the social sciences*. Cambridge, UK: Cambridge University Press.

Galton, F. (1962). *Hereditary genius: An inquiry into its laws and consequences*. Cleveland: Meridian Books.

Gamoran, A. (2007). *Standards-based reform and the poverty gap: Lessons for No Child Left Behind*. Washington, D.C: Brookings Institution Press.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Goe, L. (2007). *The Link Between Teacher Quality and Student Outcomes: A Research Synthesis*. Retrieved October 17, 2010 from http://www.tqsource.org/publications/LinkBetweenTQandStudentOutcomes.pdf

Goldhaber, D. & Brewer, D. (2000). Does teacher certification matter? High school teacher certification status and Student Achievement. *Educational Evaluation and Policy Analysis,* 22 (2), 129-145.

Goldhaber, D. (2007). Everyone's Doing It, but What Does Teacher Testing Tell Us about Teacher Effectiveness? Working Paper 9. Washington, D.C.*: National Center for the Analysis of Longitudinal Data in Education Research, Urban Institute*.

Goldhaber, D., & Hansen, M. (2010). Using Performance on the Job to Inform Teacher Tenure Decisions. *The American Economic Review, 100* (2), 250-255.

Goldhaber, D. & Hansen, M. (2010). *Assessing the Potential of Using Value-Added Estimates for Teacher Job Performance for Making Tenure Decisions*. Retrieved October 3, 2012 from http://www.urban.org/uploadedpdf/1001369_assessing_the_potential.pdf

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

Goldstein, H. (1987). *Multilevel models in educational and social research*. London: C. Griffin.

Goldstein, H. (1997). Methods in School Effectiveness Research. *School Effectiveness and School Improvement*, 8, 369-95.\

Green, J.L. (2010). Estimating Teacher Effects Using Value-Added Models. Dissertations and Thesis. Department of Statistics, University of Nebraska at Lincoln

Greene, W. H. (2003). *Econometric analysis*. Upper Saddle River, N.J: Prentice Hall.

Grubb, W. N. (2009). *The money myth: School resources, outcomes, and equity*. New York: Russell Sage Foundation.

Gujarati, D. N. (2008). *Basic econometrics*. New York, NY: McGraw-Hill.

Hamilton, J. D. (1994). *Time series analysis*. Princeton, N.J: Princeton University Press

Hamilton, L., Koretz, D. M. & McCaffrey, D. F. (2006). Validating Achievement Gains in Cohort-to-Cohort and Growth-Based Modeling Contexts. . In R. Lissitz, *Longitudinal and Value-Added Models of Student Performance* (pp. 407- 435). Mapple Grove, MN: JAM Press.

Hanushek, E. a, & Rivkin, S. G. (2010). Generalizations about Using Value-Added Measures of Teacher Quality. *American Economic Review*, *100*(2), 267-271.

Hanushek, E. A., & Rivkin, S.G. (2010). Using Value-Added Measures of Teacher Quality. *CALDER Working Paper* No. 9. May 2010.

Harris, D. N. (2009). Would Accountability Based on Teacher Value Added Be Smart Policy? An Examination of the Statistical Properties and Policy Alternatives. *Education Finance and Policy*, *4*(4), 319-350.

Harris, D. N. (2011). *Value-Added measures in education: What every educators needs to know. Cambridge*, MA: Harvard Education Press

Harris, D.N., & Sass, T. (2005). *Value-added models and the measurement of teacher quality.* Paper presented at the annual conference of the American Education Finance Association, Louisville, KY, March 17-19.

Harris, D. N., & Sass, T. R. (2009). The effects of NBPTS-certified teachers on student achievement. *Journal of Policy Analysis and Management, 28* (1), 55-80.

Hausman, J. A. (1978). Specification Tests in Econometrics. *Econometrica 46* (6), 1251–1271.

Henderson, C. R., Kempthorne, O., Searle, S. R., & von, K. C. M. (1959). The Estimation of Environmental and Genetic Trends from Records Subject to Culling.*Biometrics, 15* (2), 192-218.

Henderson, C. R. (1975). Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics, 31* (2), 423-447.

Henderson, C. R. (1977). Best Linear Unbiased Prediction of Breeding Values Not in the Model for Records. *Journal of Dairy Science, 60* (5), 783-787.

Hershberg, T. & Robertson-Fraft, C. (2009). *A grand bargain for education reform: New rewards and supports for new accountability*. Cambridge, MA: Harvard Education Press.

Hilden-Minton, J. A. (1995). *Multilevel diagnostics for mixed and hierarchial linear models*.

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*. New York: Springer.

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, N.J: Lawrence Erlbaum Associates.

Ishii, J., & Rivkin, S. G. (2009). Impediments to the Estimation of Teacher Value Added. *Education Finance and Policy*, *4*(4), 520-536.

Jakubowski, M. (2007). Volatility of Value-Added Estimates of School Effectiveness: A Comparative Study of Poland and Slovenia. *Paper presented to the Robert Shurman Centre for Advanced Studies, European University.* Florence.

Johnston, J. (1997). *Econometric methods*. New York: McGraw-Hill.

Kane, T.J. & D.O. Staiger. (2002). Volatility in School Test Scores: Implications for Test-Based Accountability Systems. In D. R. (Ed.), *Brookings Papers on Education Policy* (pp. 235-269). Washington, DC: Brookings Institution.

Kingston, N. & Reidy, E. (1997). Kentucky's Accountability and Assessment Systems. In J. M. (Ed.), *Grading Teachers, Grading Schools: Is Student Achievement a Valid Evolutional Measure?* (pp. 191-209). Thousand Oaks, CA: Corwin Press, Inc

Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences*. Monterey, CA: Brooks/Cole Pub. Co.

Klockars, A. J. (2005). *Basic education statistics*. Course material for Educational Psychology, Statistics 490 at the University of Washington.

Kline, R. B. (2005). *Principles and practice of structural equation modeling*. New York: Guilford Press.

Knapp, M. S., Copland, M. A., Honig, M. I., Plecki, M. L., & Portin, B. S. (2010). Urban Renewal: The Urban School Leader Takes on a New Role. *Journal of Staff Development, 31* (2), 24-29.

Koedel, C. & Betts, J. (2009). Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique. Working Paper. Columbia, Mo.: University of Missouri-Columbia.

Koedel, C., & Betts, J. (2009). *Value-added to what? How a ceiling in the testing instrument influences value-added estimation.* Available: http://economics.missouri.edu/workingpapers/ koedelWP.shtml [accessed September 2009].

Koedel, C., & Betts, J. (2010). Value Added to What? How a Ceiling in the Testing Instrument Influences Value-Added Estimation. *Education Finance and Policy, 5*(1), 54-81.

Kolen, M. & R. Brennan. (2004). *Test equating, scaling and linking: Methods and practices.* New York, NY: Springer Science and Business Media.

Kolen, M.J. & Tong, Y. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education, 20* (2), 227-253.

Ladd, H. F., Hansen, J. S., & National Research Council (U.S.). (1999). *Making money matter: Financing America's schools.* Washington, D.C: National Academy Press.

Larsen, R. J. & M. L. Marx. (2005). *Introduction to Mathematical Statistics and Its Applications.* Upper Saddle River, NJ: Prentice Hall.

Lindsey, J. K. (1995). *Introductory statistics: A modelling approach*. Oxford: Clarendon Press.

Linn, R.L. (2008). *Measurement issues associated with value-added models.* Paper prepared for the workshop of the Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Educational Accountability, National Research Council, Washington, DC, November 13-14. Available: http://www7.nationalacademies.org/ bota/VAM_Workshop_Agenda.html.

Lipscomb, S., Teh, B., Gill, B., Chiang, H., Owens, A., & Mathematica Policy Research, Inc. (2010). *Teacher and Principal Value-Added: Research Findings and Implementation Practices. Final Report*. Retrieved May 12, 2013 from http://www.mathematica-mpr.com

Lissitz, R., H. Doran, W. Schafer & J.Willhoft. (2006). Growth Modelling, Value-Added Modelling and Linking: An Introduction. In R. Lissitz, *Longitudinal and Value-Added Models of Student Performance* (pp. 1- 46). Mapple Grove, MN: JAM Press.

Little, R. J. A. & D. B. Rubin. (1987). *Statistical analysis with missing data.* New York, NY: Wiley.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V.N., & Martinez, J. F. (2007). The Sensitivity of Value-Added Teacher Effect Estimates to Different Mathematics Achievement Measures. *Journal of Educational Measurement, 44*(1), 47-67.

Lockwood, J. R., McCaffrey, D. F., Mariano, L. T., & Setodji, C. (2007). Bayesian Methods for Scalable Multivariate Value-Added Assessment. *Journal of Educational and Behavioral Statistics, 32*(2), 125-150.

Lockwood, J.R., & McCaffrey, Daniel F. (2007). Controlling for Individual Heterogeneity in Longitudinal Models, with Applications to Student Achievement. *Electronic Journal of Statistics, 1*, 223-252.

Lockwood, J. R., & McCaffrey, Daniel F. (2009). Exploring Student-Teacher Interactions in Longitudinal Achievement Data. *Education Finance and Policy, 4*(4), 439-467.

Loehlin, J. C. (2004). *Latent variable models: An introduction to factor, path, and structural equation analysis*. Mahwah, N.J: L. Erlbaum Associates.

Lomax, R. G. (1998). *Statistical concepts: A second course for education and the behavioral sciences*. Mahwah, N.J: L. Erlbaum Associates.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage Publications.

Loveless, T., & TIMSS International Study Center. (2007).*Lessons learned: What international assessments tell us about math achievement*. Washington, D.C: Brookings Institution Press

Lunneborg, C. E., & Abbott, R. D. (1983). *Elementary multivariate analysis for the behavioral sciences: Applications of basic structure*. New York: North-Holland

Maas, C. J. M., & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica, 58* (2), 127-137.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient Sample Sizes for Multilevel Modeling. *Methodology, 1* (3), 85-91.

Mariano, L. T., McCaffrey, D. F., & Lockwood, J. R. (2010). A Model for Teacher Effects From Longitudinal Data Without Assuming Vertical Scaling. *Journal of Educational and Behavioral Statistics*, *35*(3), 253-279.

Martineau, B. J. A. (2010). The Validity of Value-Added Models An Allegory. *Journal of Personnel Evaluation in Education*, (April), 64-68.

McCaffrey, D., Lockwood, J.R., Koretz, D.M., & Hamilton, L.S. (2003). *Evaluating value-added models for teacher accountability.* Santa Monica, CA: RAND Corporation.

McCaffrey, D., Koretz, D.M., & Hamilton, L.S. (2004). The Promise and Peril of Using Value-Added Modeling to Measure Teacher Effectiveness.. Santa Monica, CA: RAND Corporation.

McCaffrey, Daniel F, Lockwood, J R, Koretz, Daniel, Louis, T. a, & Hamilton, Laura. (2004). Models for Value-Added Modeling of Teacher Effects. *Journal of Educational and Behavioral Statistics*, *29*(1), 67-101.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A, & Hamilton, L. (2004). Let's See More Empirical Studies on Value-Added Modeling of Teacher Effects: A Reply to Raudenbush, Rubin, Stuart and Zanutto, and Reckase. *Journal of Educational and Behavioral Statistics*, *29*(1), 139-143.

McCaffrey, D. F., Lockwood, J. R., Mariano, L. T. & C. Setodji, (2005). Challenges for Value-Added Assessment of Teacher Effects. In R. Lissitz (Ed.) *Value added models in education: Theory and practice* (pp. 111-144). Maple Grove, MN: JAM Press.

McCaffrey, D. F., & Hamilton, L. S. (2007). *Value-added assessment in practice: Lessons from the Pennsylvania Value-Added Assessment System Pilot Project* (Technical Report TR-506). Santa Monica, CA: RAND Corporation. May 9, 2008,

Mccaffrey, D. F., & Lockwood, J R. (2011). Missing Data in Value-Added Modeling of Teacher Effects. *Education*, 1-31

McCaffrey, Daniel F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The Intertemporal Variability of Teacher Effect Estimates. *Education Finance and Policy*, *4*(4), 572-606.

McCulloch, C. E., & Searle, S. R. (2001). *Generalized, linear, and mixed models*. New York: John Wiley & Sons.

Miller, I., Miller, M., Freund, J. E., & Miller, I. (2004). *John E. Freund's mathematical statistics with applications*. Upper Saddle River, NJ: Prentice Hall.

Monahan, J. F. (2008). *A primer on linear models*. Boca Raton: Chapman & Hall/CRC.

Mood, A. M. F., & Graybill, F. A. (1963). *Introduction to the theory of statistics*. New York: McGraw-Hill.

Morgan, S. & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research*. New York, N.Y: Cambridge University Press.

Muijs, D. (2004). *Doing quantitative research in education with SPSS*. London: Sage Publications.

National Association of Standards Based Education. (2005). *Evaluating value-added: Findings and recommendations from the NASBE study group on value-added assessments.* Alexandria, VA: National Association of State Boards of Education.

National Research Council, National Academy of Education, Braun, H., Chudowsky, N., & Koenig, J., (2010). *Getting value out of value-added*. Washington D.C.: The National Academies Press.

Neter, J., & Wasserman, W. (1990). *Applied linear statistical models: Regression, analysis of variance, and experimental designs*. Homewood, Ill: R.D. Irwin.

Neter, J., Kutner, M., Nachtsheim, C. & Li W. (2004). *Applied linear statistical models*. Homewood, Ill: R.D. Irwin.

Noell, G., & Burns, J. (2006). Value-Added Assessment of Teacher Preparation. *Journal of Teacher Education, 57* (1), 37-50.

Organization of Economic Cooperation and Development. (2005). *Teachers matter: Attracting, developing and retaining effective teachers.* Paris: OECD Publications.

Organization of Economic Cooperation and Development. (2007). *PISA 2006: Science competencies for tomorrow's world*. Paris: OECD Publications.

Organization of Economic Cooperation and Development (2008). *Measuring improvements in learning outcomes: Best practices to assess the value-added of schools*. Paris: OECD Publications.

Pearl, J. (2009). *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.

Plecki, M. & Castaneda, T. (2009). Whether and how money matters in K-12 education. In Sykes, G. & Plank, D.(Eds.). *American Education Research Association Handbook of Education Policy Research*.

Plecki, M. L., Elfers, A. M., & Nakamura, Y. (2012). Using Evidence for Teacher Education Program Improvement and Accountability: An Illustrative Case of the Role of Value-Added Measures. *Journal of Teacher Education, 63* (5), 318-334.

Ponisciak, S. & Bryk, A. (2005). Value-Added Analysis of the Chicago Public Schools: An Application of Hierarchical Models. In R. Lissitz, *Value-Added Models in Education: Theory and Applications* (pp. 40-79). Maple Grove, MN: JAM Press.

Public Impact & Thomas B. Fordham Institute. (2008). *Ohio value-added primer: A user's guide*. Washington, DC: Thomas B. Fordham Institute.

Rabe-Hesketh, S. & Skrondal, A. (2008). *Multilevel and longitudinal modeling using Stata*. College Station, Texas: A Stata Press Publication.

Raftery, A. E. (1999). Bayes Factors and BIC. *Sociological Methods & Research, 27* (3).

Raftery, A. E., Gneiting, T., Balabdaoui, F., & Polakowski, M. (2005). Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Monthly Weather Review,133* (5), 1155-1174.

Ravitch, D. (2011, January 18). The Pitfalls of Putting Economists in Charge of Education. The Education Week. Retrieved from http://www.edweek.com/

Raudenbush, S. W. (2004). What Are Value-Added Models Estimating and What Does This Imply for Statistical Practice? *Journal of Educational and Behavioral Statistics, 29*(1), 121-129.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks: Sage Publications

Raudenbush, Stephen W. (1993). A Crossed Random Effects Model for Unbalanced Data with Applications in Cross-Sectional and Longitudinal Research. *Journal of Educational Statistics, 18*(4), 321.

Raudenbush, Stephen W. (2009). Adaptive Centering with Random Effects: An Alternative to the Fixed Effects Model for Studying Time-Varying Treatments in School Settings. *Education Finance and Policy, 4*(4), 468-491.

Ray, A., McCormack, T., & Evans, H. (2009). Value Added in English Schools. *Education Finance and Policy, 4*(4), 415-438.

Reardon, S. F., & Raudenbush, Stephen W. (2009). Assumptions of Value-Added Models for Estimating School Effects. *Education Finance and Policy, 4*(4), 492-519.

Reckase, M. D. (2004). The Real World is More Complicated than We Would Like. *Journal of Educational and Behavioral Statistics, 29*(1), 117-120.

Rivkin, S.G. (2007). Value-Added Analysis and Education Policy. *CALDER Working Paper* No. 1. November 2007.

Rivkin, Steven G., Hanushek, A.E. & Kain. J.F. (2005). Teachers, Schools and Academic Achievement. *Econometrica* 73(2): 417–58.

Rockoff, J. E. (2004). The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data. *American Economic Review, 94*(2), 247-252.

Rothstein, R., Jacobsen, R., & Wilder, T. (2008). *Grading education: Getting accountability right*. Washington, DC: Economic Policy Institute

Rothstein, J. (2009a). Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables. *Education Finance and Policy, 4*(4), 537-571.

Rothstein, J. (2011). *Learning About Teaching: Initial Findings from the Measures of Effective Teaching Project*. Retrieved May 12, 2013 from http://nepc.colorado.edu/thinktank/review-learning-about-teaching.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, 63, 581-592.

Rubin, D. B., Stuart, E. a, & Zanutto, E. L. (2004). A Potential Outcomes View of Value-Added Assessment in Education. *Journal of Educational and Behavioral Statistics*, *29*(1), 103-116.

Rutherford, A. (2001). *Introducing ANOVA and ANCOVA: A GLM approach*. London: SAGE.

Sanders, W., & Rivers, J. (1996). *Cumulative and residual effects of teachers on future student academic achievement.* Retrieved June, 2009 from http://www.cgp.upenn.edu/pdf/Sanders_Rivers-TVASS_teacher%20effects.pdf.

Sanders, W. L., Rivers, J. C., & Hall, M. (1997).*Graphical Summary of Educational Findings from the Tennessee Value-Added Assessment System*. . Retrieved September 30, 2010 from http://www.shearonforschools.com/summary/GRAPH-SUM.HTML

Sanders, W.L., Saxton, A.M. & Horn S.P. (1997). The Tennessee Value-Added Assessment System A Quantitative Outcomes-Based Approach to Educational Assessment. In J. M. (Ed.), *Grading Teachers, Grading Schools: Is Student Achievement a Valid Evolutional Measure?* (pp. 137-162). Thousand Oaks, CA: Corwin Press, Inc.

Sanders, W., & Horn, S. (1998). Research findings from the Tennessee value-added assessment system (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education*, *12* (3), 247-256.

Sanders, W. L. (2000). Value-Added Assessment from Student Achievement Data: Opportunities and Hurdles. *Journal of Personnel Evaluation in Education*, 329-339.

Sanders, W. L. (2006). Comparisons Among Various Educational Assessment Value-Added Models. Presented at The Power of Two – National Value-Added Conference, Columbus Ohio.

Sanders, W., Wright, S. W. & Rivers J.C. (2006). Measurement of Academic Growth of Individual Students Toward Variable and Meaningful Academic Standards. . In R. Lissitz, *Longitudinal and Value-Added Models of Student Performance* (pp. 385- 406). Mapple Grove, MN: JAM Press

Sass, T. (2008).The Stability of Value-Added Measures of Teacher Quality and Implications for Teacher Compensation Policy. *CALDER Working Paper* No. 4. November 2008.

Scheerens, J., & Bosker, R.J. (1997). *The foundations of educational effectiveness.* Oxford: Elsevier.

Schleicher, A. & OECD (2008). *Education at a glance*. OECD, Paris

Searle, S. R. (1971). *Linear models*. New York: Wiley.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. New York: Wiley.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford: Oxford University Press

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage Publications.

Springer, M. G. (2009). *Performance incentives: Their growing impact on American K-12 education*. Washington, D.C: Brookings Institution Press

Stecher, B. M. & Vernez, G. (2010) *Reauthorizing No Child Left Behind: Facts and recommendations*. Santa Monica, CA: RAND Corpoaration

Stevens, J. & Zvoch K. (2006). Issues in the Implementation of Longitudinal Growth Models for Student Achievement. In R. Lissitz, *Longitudinal and Value-Added Models of Student Performance* (pp. 170-209). Mapple Grove, MN: JAM Press.

Tabachnick, B. G., & Fidell, L. S. (2000). *Using multivariate statistics*. Boston, MA: Allyn and Bacon.

Tate, R. L. (2004). A Cautionary Note on Shrinkage Estimates of School and Teacher Effects. *Florida Journal of Educational Resaerch*, 42, 1-2

Tekwe, C., R. Carter, C. Ma, J. Algina, M. Lucas & J. Roth. (2004). An Empirical Comparison of Statistical Models for Value-Added Assessment of School Performance. *Journal of Educational and Behavioral Statistics*, 29 (1), 11-36.

Theall, K.P., Scribner, R., Lynch, S., Simonsen, N., Schonlau, M., Carlin, B., & Cohen, D. (2008*). Impact of Small Group Size on Neighborhood Influences in Multilevel Models.* Retrieved May 12, 2013 from http://mpra.ub.uni-muenchen.de/11648/1/MPRA_paper_11648.pdf

The Achievement Gap Initiative (2009). How High Schools Become Exemplary – Ways that Leadership Raises Achievement and Narrows Gaps By Improving Instruction in 15 Public High Schools. Retrieved May 12, 2013 from http://www.agi.harvard.edu/

Thomas, S. & Mortimore, P. (1996). Comparison of Value-Added Models for Secondary School Effectiveness. *Research Papers in Education*, 11 (1), 5-33.

Thum Y. M. (2006). Measuring and Comparing Academic Progress Towards a Standard Using Bayesian Performance Profiles. In R. Lissitz, *Longitudinal and Value-Added Models of Student Performance* (pp. 436- 479). Mapple Grove, MN: JAM Press.

Thum, Y.M. & Bryk, A.S. (1997). Value-Added Productivity Indicators. In J. M. (Ed.), *Grading Teachers, Grading Schools: Is Student Achievement a Valid Evolutional Measure?* (pp. 100-109). Thousand Oaks, CA: Corwin Press, Inc.

Todd, P. E., & Wolpin, K.I. (2003). On the Specification and Estimation of the Production Function for Cognitive Achievement." *Economic Journal* 113(485): F3–33.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, Mass: Addison-Wesley Pub. Co.

Value-Added Research Center (2013). *Quadrant Analysis.* Retrieved May 12, 2013 from http://varc.wceruw.org/tutorials/

Vanneman, A., Hamilton, L., Baldwin Anderson, J., & Rahman, T. (2009). Achievement Gaps: How Black and White Students in Public Schools Perform in Mathematics and Reading on the National Assessment of Educational Progress, (NCES 2009-455). *National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education*. Washington, DC.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York: Springer.

Wainer, H. (2004). Introduction to a Special Issue of the Journal of Educational and Behavioral Statistics on Value-Added Assessment. *Journal of Educational and Behavioral Statistics*, *29*(1), 1-3.

Webster, W. J. (2005). The Dallas School-Level Accountability Model: The Marriage of Status and Value-Added Approaches. In R. L. (ed.), *Value added models in education: Theory and Applications* (pp. 233-271). Maple Grove, MN: JAM Press.

Webster, W.J. & Mendro, R.L. (1997). The Dallas Value-Added Accountability System. In J. M. (Ed.), *Grading Teachers, Grading Schools: Is Student Achievement a Valid Evolutional Measure?* (pp. 81-99). Thousand Oaks, CA: Corwin Press, Inc.

Weisberg, S. (2005). *Applied linear regression*. New York, NY: Wiley

Willms, J.,& Raudenbush, S. (1989). A Longitudinal Hierarchical Linear Model for Estimating School Effects and Their Stability. *Journal of Educational Measurement*, 209-232.

Willms, J.D. (2008). *Seven Key Issues for Assessing "Value-Added" in Education.* Paper prepared for the workshop of the Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Educational Accountability, National Research Council, Washington, DC, November 13-14. Available: http://www7.national academies.org/bota/VAM_Workshop_Agenda.html.

Winerip, M. (2011, March 6). Evaluating the New York Teachers, Perhaps the Numbers Do Lie. *The New York Times*. Retrieved from http://www.nytimes.com/

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.

Wooldridge, J. M. (2003). *Introductory econometrics: A modern approach*. Australia: South-Western College Pub.

Wright, S. P., Horn, S. P., & Sanders, W. L. (1997). Teacher and Classroom Context Effects on Student Achievement: Implications for Teacher Evaluation. *Journal of Personnel Evaluation in Education*, 57-67.

Appendix:

Chapter 4:

Three Different Derivations of the Empirical Bayes Estimates:

1. Regression Derivation:

The long historical work and techniques of the conditional expectation (which defines the regression function) can be applied to the study of predicting unobservable random effects in light of relevant evidence. To do so, we assume the normal distribution of the random variables a_j and y_{ij} (or e_{ij}) in $y_{ij} = \mu + a_j + e_{ij}$. Using the properties of normal distributions, this leads to a_j and $\bar{y}_{.j}$ being jointly distributed with a bivariate normal density having mean and variance of

$$E\begin{bmatrix} a_j \\ \bar{y}_{.j} \end{bmatrix} = \begin{bmatrix} 0 \\ \mu \end{bmatrix} \quad and \quad Var\begin{bmatrix} a_j \\ \bar{y}_{.j} \end{bmatrix} = \begin{bmatrix} \sigma_a^2 & \sigma_a^2 \\ \sigma_a^2 & \sigma_a^2 + \sigma_e^2/n_j \end{bmatrix}$$

And then by applying the following theorem base on the seminal work of Francis Galton who founded the concept of regression, we can obtain the closed form equation of $E(a_j|\bar{y}_{.j})$[222]

> Theorem: If X and Y have a bivariate normal distribution, the conditional density of Y given $X = x$ (also known as the regression function) is a normal distribution with the mean
>
> $$E(Y|X = x) = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x) = \mu_y + \frac{\sigma_{yx}}{\sigma_x^2}(x - \mu_x)$$
>
> and variance
>
> $$\sigma_{Y/x}^2 = \sigma_x^2(1 - \rho^2)$$

And applying this theorem to the problem of predicting random effects, we obtain

$$E(a_j|\bar{y}_{.j}) = E(a_j) + \frac{cov(a_j, \bar{y}_{.j})}{var(\bar{y}_{.j})}[\bar{y}_{.j} - E(\bar{y}_{.j})]$$

and substituting the corresponding parameters,

$$E(a_j|\bar{y}_{.j}) = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2/n_j}[\bar{y}_{.j} - \mu]$$

And this can be naturally extended to the case of $\mu_j = \mu + a_j$ by simply adding μ as follows

$$E(\mu_j|\bar{y}_{.j}) = \mu + \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2/n_j}[\bar{y}_{.j} - \mu]$$

Keeping these equations in mind, we move to the next derivation using the Bayesian framework.

[222] Proof is provided on Miller (2002). ρ is the correlation coefficient equal to the ratio of the covariance and the product of standard deviation $\sigma_{XY}/\sigma_x\,\sigma_y$. The same functional form applies for the case of $E(X/y)$ by simply switching the place of Y and X.

2. Bayesian Derivation:

The study of conditional expectation is also the raison de entre for the study of Bayesian inference. Bayesian inference is defined as the process of inductive learning via the Bayes rule which is defined as follows

$$p(A|B) = \frac{p(A \cap B)}{p(B)}$$

where probability of event A given B is the ratio of the probability of event A and B and probability of event B. A standard interpretation of probability is that it is a numerical measure to express our beliefs or uncertainty about unknown quantities, outcomes or statements.[223] Bayesian inference is a rational method of updating our beliefs of the unknown population parameter in light of new information perceived by the observed data. Quantifying this chance in probability/uncertainty is the purpose of Bayesian inference. More formally, we have some prior belief/information p(theta) which describes our belief that theta represents the true population characteristics and we also have some belief of the data $p(y|\theta)$ which describes our belief that y would be the outcome of our study if we knew theta to be true. y is also the subset/sample of the population distribution. Once we observe the data y, we update our prior beliefs about theta to obtain the posterior distribution/belief $p(\theta|y)$ which describes our belief that theta is the true value, having observed the dataset y. The posterior distribution is obtained from combining the prior distribution and the sampling/data distribution in accordance to the Baye's rule

$$p(\theta|y) = \frac{p(\theta \cap y)}{p(y)} = \frac{p(y|\theta)p(\theta)}{p(y)}$$

The denominator is a constant that does not depend on θ. In Bayesian estimation, we can focus primarily on the numerator. It is important to note that Bayes' rule tells us how our belief, knowledge, uncertainty, confidence, probability or distribution changes after seeing new information. It provides a rational method of learning and updating our beliefs in light of new information.[224] And as described earlier, it is directly on the point with the task of predicting random effects where our objective is to determine the form of $E(a_j|\bar{y}_{.j})$.

Now, applying the Bayesian methods to the random effects model, we can mathematically obtain the posterior distribution $p(a_j|\bar{y}_{.j})$ and subsequently the expression of the posterior expectation of $E(a_j|\bar{y}_{.j})$. Given the model specification of $y_{ij} = \mu + a_j + e_{ij}$, we know that yij (the sampling distribution) is normally distributed with mean zero and variance σ^2 and $\mu_j = \mu + a_j$ (the prior distribution) is normally distributed with mean mu and variance τ^2. Using the properties of normal distribution and combining the distributions in accordance to the Baye's rule, we can obtain the posterior distribution as follows. Note that we drop the constant terms of the normal distribution which do not dependent on the parameters of interest to simply the mathematical process.

[223] Hoff (2009)
[224] Hoff (2009)

The prior distribution is

$$p(\mu_j|\,\mu\,,\tau^2) \sim N(\mu, \tau^2,)\ \propto exp\left\{-\tfrac{1}{2\tau^2}\left(\mu_j-\mu\right)^2\right\}$$

The sampling distribution is

$$p(y_{ij}\big|\,\mu_j,\sigma^2) \sim N(\mu_j,\sigma^2) \propto exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n_j}(y_{ij}-\mu_j)^2\right\}$$

And the posterior distribution is

$$p\left(\mu_j\big|\,\mu\,,\tau^2,\sigma^2\,,y_{ij}\right)\sim N\left(E(\mu_j|y_{ij})\,,\Sigma^2\,\right)\ \propto\ p(\mu_j|\,\mu\,,\tau^2)\prod_{i=1}^{n_j}\prod p\left(y_{ij}\big|\,\mu_j,\sigma^2\right)$$

Our task is to find the explicit form of $E(\mu_j|y_{ij})$. To do so, we simply the posterior distributional form as follows. Adding the terms in the exponents and ignoring the -1/2 for the moment, we have

$$\frac{1}{2\tau^2}\left(\mu_j^2-2\mu_j\mu+\mu^2\right)+\frac{1}{\sigma^2}\left(\sum y_{ij}^2-2\mu_j\sum y_{ij}+n\mu_j^2\right)=a\mu_j^2-2b\mu_j+c$$

$$a=\frac{1}{\tau^2}+\frac{n_j}{\sigma^2}\,,\qquad b=\frac{\mu}{\tau^2}+\frac{\sum y_{ij}}{\sigma^2}\ \ and\ c=c(\mu\,,\tau^2,\sigma^2,y_{ij})$$

Now,

$$p\left(\mu_j\big|\,\mu\,,\tau^2,\sigma^2\,,y_{ij}\right)\propto\exp\left\{-\frac{1}{2}(a\mu_j^2-2b\mu_j)\right\}$$

$$=exp\left\{-\frac{1}{2}a(\mu_j^2-2b\mu_j/a+b^2/a^2)+\frac{1}{2}b^2/a\right\}$$

$$\propto\exp\left\{-\frac{1}{2}a(\mu_j-b/a)^2\right\}$$

$$=exp\left\{-\frac{1}{2}\left(\frac{\mu_j-b/a}{1/\sqrt{a}}\right)^2\right\}$$

This function as exactly the same shape as the normal density curve, with $1/\sqrt{a}$ playing the role of the standard deviation and b/a playing the role of the mean. Since probability distributions are determined by their shape, this means that $p\left(\mu_j\big|\,\mu\,,\tau^2,\sigma^2\,,y_{ij}\right)$ is indeed a normal density with the following mean and variance,

$$E\left(\mu_j|\bar{y}_{.j}\right)=\frac{b}{a}=\frac{\frac{1}{\tau^2}\mu+\frac{n_j}{\sigma^2}\bar{y}_{.j}}{\frac{1}{\tau^2}+\frac{n_j}{\sigma^2}}$$

$$\tau_j^2=\frac{1}{a}=\frac{1}{\frac{1}{\tau^2}+\frac{n_j}{\sigma^2}}$$

Now, comparing the final form of the conditional expectation of the random effects in the regression and Bayesian methods shown below we find that the results are identical

Regression Derivation	**Bayesian Derivation**
$E(\mu_j \vert \bar{y}_{.j}) = \mu + \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2/n_j}[\bar{y}_{.j} - \mu]$	$E(\mu_j \vert \bar{y}_{.j}) = \frac{\frac{1}{\tau^2}\mu + \frac{n_j}{\sigma^2}\bar{y}_{.j}}{\frac{1}{\tau^2} + \frac{n_j}{\sigma^2}}$

By expanding the equation of the regression derivation, isolating the parameter μ and $\bar{y}_{.j}$ and finally by multiplying $\frac{1}{\sigma_a^2 \cdot \sigma_e^2}$ the regression derivation is equivalent to the Bayesian derivation.

3. Linear Mixed Model Derivation:

The above equations are also equivalent to the solution to the linear mixed model equations $y = X\beta + Zu + e$ with the distribution assumptions of $\begin{bmatrix} u \\ y \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ X\beta \end{bmatrix}, \begin{bmatrix} 0 & DZ' \\ ZD & V \end{bmatrix}\right)$ as shown below.[225]

$$\tilde{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y$$

$$\tilde{u} = DZ'V^{-1}(y - X\tilde{\beta})$$

To illustrate the equivalence, the random effects $\tilde{u}$ can be rewritten as $\frac{cov(u,y)}{var(y)}(y - X\tilde{\beta})$ and in the case of one way random effects model ($y = \mu I_N + Z\alpha + e$), $\tilde{u} = \frac{cov(\alpha,y)}{var(y)}(y - \hat{\mu}I_N)$ where $\hat{\mu}$ is the GLS estimate of $(I'V^{-1}I)^{-1}I'V^{-1}y$ which simplifies to the weighted grand mean (weighted by each treatment's sample size). The final form of $E(\alpha_j \vert \bar{y}_{.j})$ and $E(\mu_j \vert \bar{y}_{.j})$ reduces to[226]

$$E(\alpha_j \vert \bar{y}_{.j}) = \frac{n_j\sigma_\alpha^2}{n_j\sigma_\alpha^2 + \sigma_e^2}[\bar{y}_{.j} - \mu]$$

$$E(\mu_j \vert \bar{y}_{.j}) = \mu + \frac{n_j\sigma_\alpha^2}{n_j\sigma_\alpha^2 + \sigma_e^2}[\bar{y}_{.j} - \mu]$$

And this formula is equivalent to the solution derived under the regression method. By dividing the numerator and denominator of the above equation by n_j show the equivalence.

Thus the three methods yield the same expression of the prediction of the random effects. This sheds light on intricate complexity and the different avenues in deriving the same solution in statistics. The above derivations can be extended to include covariates and other random effects e.g. random slope. For the latter case, the final

[225] X are fixed explanatory variables and u are random variables just like the random effects aj. The random effects model presented in the text is nothing but a simplified version of the linear mixed model with only one random effect for the intercept. If random slopes with different effects of the explanatory variables across the groups were to be modeled, then this leads to another random variable into the u matrix. The Z or design matrix monitors the respective model specifications.

[226] Detailed proof is provided in Searle et al. (2002)

reduced form resembles the random intercept model where the random slope is a weighted balance of grand or pooled slope (calculated from the regression without taking into account the units) and un-pooled or group specific slope (calculated from the dummy variable fixed effects regression).[227]

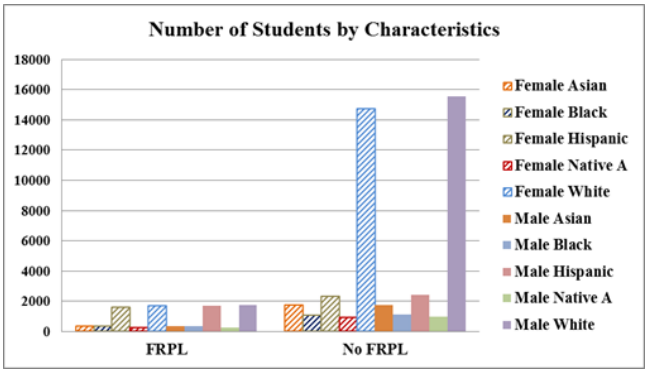
Chapter 6

Multivariate Descriptive Analysis of the 5th Grade Student Background Variables

The student background/contextual variables themselves can also be interrelated. Without taking into account its underlying relations, it can confound its own effects on the outcome variable. To explore these interconnections, the multivariate descriptive analysis conducted in Chapter 6 is repeated to show the following results.

Number of Students by Student Characteristics

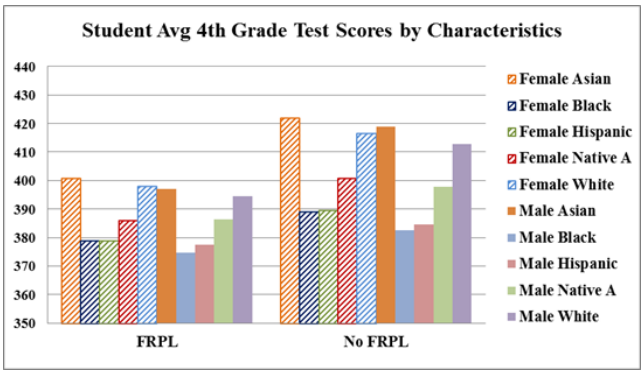
		FRPL	
Gender	*Ethnicity*	Yes	No
Female	Asian	333	1754
	Black	320	1056
	Hispanic	1580	2317
	Native American	248	939
	White	1664	14739
Male	Asian	355	1766
	Black	352	1105
	Hispanic	1716	2398
	Native American	270	969
	White	1723	15557



The most frequent groups of students characterizing the 5th grade students are white male and white female students who do not receive FRPL. Each of these two groups of students each comprises approximately 30% of the total students. There is no clear gender difference in counts with the FRPL status and the White ethnicity factor. Balanced gender distribution is also evident across other combinations of ethnicity and FRPL status e.g. there are very similar number of Asian, Black, Hispanic and Native American female and male students who receive and not receive FRPL. But there is clearly a direct association of White ethnicity and FRPL status where White students by a large margin do not receive FRPL. The proportion of students within each ethnic group who receive FRPL is led by Hispanic students of 23.7% who is followed by Black students of 23.7% and Native American students of 21.4%. White and Asian students have the lowest proportion with 17.2% and 16.3%, respectively. This finding is in line to the existing research which identifies higher concentration of economically disadvantaged students amongst the minority ethnic groups. Looking now at the average prior achievement levels across the different groups of students show the following results.

[227] Gelman and Hill (2002) provides extensive explanation of the random slope.

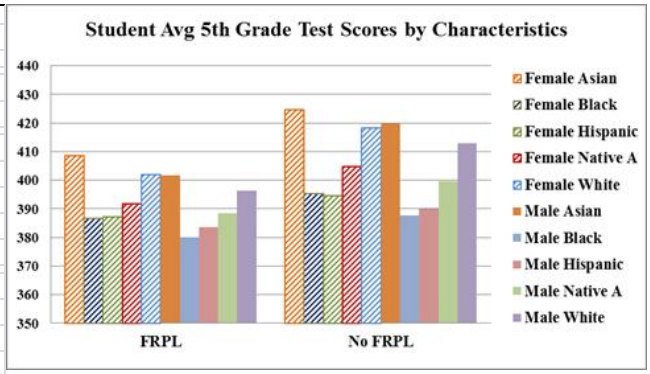
Student Avg 4th Grade Test Scores by Student Characteristics			
		FRPL	
Gender	*Ethnicity*	Yes	No
Female	Asian	400.95	422.01
	Black	379.08	389.11
	Hispanic	379.04	389.64
	Native American	386.26	400.91
	White	398.16	416.53
Male	Asian	396.90	418.89
	Black	374.54	382.55
	Hispanic	377.45	384.64
	Native American	386.34	397.77
	White	394.57	412.65



As you can see, White and Asian female students who do not receive any FRPL have the highest initial performance. On the other hand, Male Black and male Hispanic students who receive FRPL have the lowest initial performance. These findings are in line to the individual univariate summaries with the current performance shown above. Combination of different contextual variables together show clear association with the students' initial performance. The former group of students in away brings high initial level of academic challenge while the latter group brings a lot of academic challenge and difficulty to the 5th grade teachers.

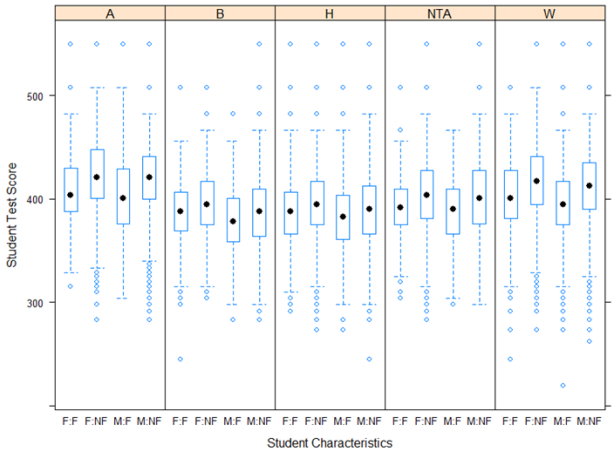
And looking now at the current 5th grade performance by the different combination of contextual variables shows the following result.

Student Avg 5th Grade Test Scores by Student Characteristics			
		FRPL	
Gender	*Ethnicity*	Yes	No
Female	Asian	408.59	424.74
	Black	386.63	395.46
	Hispanic	387.17	394.76
	Native American	391.90	404.74
	White	402.04	418.32
Male	Asian	401.72	419.86
	Black	380.02	387.61
	Hispanic	383.50	389.88
	Native American	388.30	399.69
	White	396.27	412.94



As you can see, the identical trend to the prior test score results is evident with the current test scores. White and Asian female students who do not receive any FRPL again have the highest initial performance. On the other hand, Male Black and male Hispanic students who receive FRPL again the lowest initial performance. Looking at the entire (conditional) distribution of students in each combination of student variables is illustrated below. Despite different sample sizes, each distribution shows very similar pattern indicated with similar variation around the conditional mean values as indicated with the similar size/shape of the boxplot and the pattern of the outlying observations.[228]

[228] The black dots are the conditional mean values summarized in the table above.



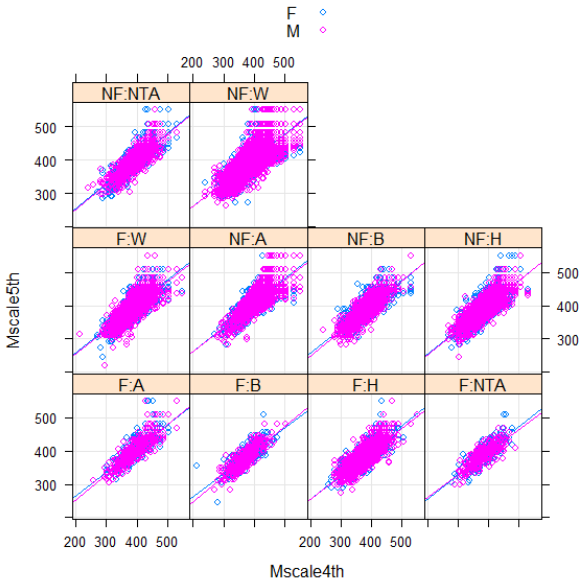
The key finding of the above result is that the combination of different contextual variables portrays clear association with the students' initial performance. The findings show the same/consistent direction/sign of the association each variable separately has on the outcome variable (as shown in the previous univariate/bivariate analysis). But unlike the previous section, the key difference of the multivariate analysis is that it conditions on multiple student contextual factors to calculate the variable's association on the outcome variable. That is, it controls for various confounding effects inherent between the variables to provide even more direct/precise unique estimate of the variable's association on the outcome. In essence, it dissects the univariate finding to illustrate further student performance patterns. The approximation of confounding effects is illustrated by comparing the adjusted/conditional mean values to the unadjusted mean values (univariate analysis) re-summarized below.

Mean Test Scores (Un-conditional) By Student Characteristics			
Gender:		*Ethnicity:*	
Student is Female	411.2	Student is White	413.9
Student is Male	405.8	Studentis Native A	399.5
		Student is Black	389.5
FRPL Status:		Student is Hispanic	389.4
Student with FRPL	392.5	Student is Asian	419.5
Student without FRPL	411.7		

Comparing the mean values to the conditional mean values above, it is evident that the values differ widely depending on the variables conditioned. [229] Thus depending on the variables that is taken into account/controlled the effects of each variable on the outcome variable alternates. If other variables which are shown to have clear association with the variables are excluded from the model, its effects will be picked up/miss-attributed to the variables included in the model. That is, the parameter estimates will be biased.

[229] To obtain conditional mean values after conditioning on two variables (instead of three) you simply sum across the conditional mean values across the third variable.

Now to estimate the association of prior test score on the current test score after conditioning on other student variables, we construct the conditional scatterplot and calculate the conditional correlation coefficients as shown below. As you can see, for any group of students, a clear positive association of the two variables is illustrated. This is confirmed with a strong positive correlation (at east 0.78) as shown in the right hand table. The similarities (lack of variation) of this finding across the different groups provide hints the robustness and "un-confounded" relation between the two variables. That is, the effects of prior test scores on current test score is very strong/firm for all groups of students and does not change even if the contextual factors are taken into account.



Correlation of 4th and 5th Scores by St Characteristics

		FRPL	
Gender	*Ethnicity*	Yes	No
Female	Asian	0.781	0.790
	Black	0.817	0.817
	Hispanic	0.835	0.825
	Native A	0.823	0.816
	White	0.821	0.786
Male	Asian	0.846	0.805
	Black	0.834	0.830
	Hispanic	0.817	0.834
	Native A	0.805	0.830
	White	0.794	0.789

To sum, the findings above clearly illustrate the interconnection and variability inherent between the different student background variables defined at the teacher level. Given the significant individual effects on the outcome variable (shown in the text), unless we separate out this web of associations, the effect of each variable on the outcome variable will be confounded. This therefore provides further support of the VAMs which can simultaneously take into account all these variables in a single model to unravel these interconnections and provide the reliable unique estimate of each variable on the outcome variable.

5Th Grade Math Teacher Characteristics – Nexus of Associations and Rough Approximation of Confounding Effects on Teachers

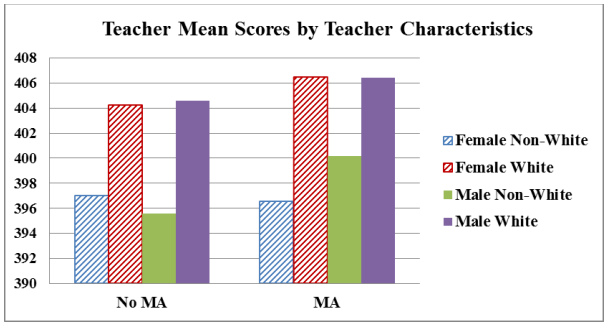
The interaction between the teacher characteristics variables of the 5th grade math teachers are illustrated as follow.

Number of Students by Teacher Characteristics			
		Higher Degree	
Gender	*Ethnicity*	No MA	MA
Female	Non-White	51	81
	White	679	1337
Male	Non-White	20	33
	White	214	449

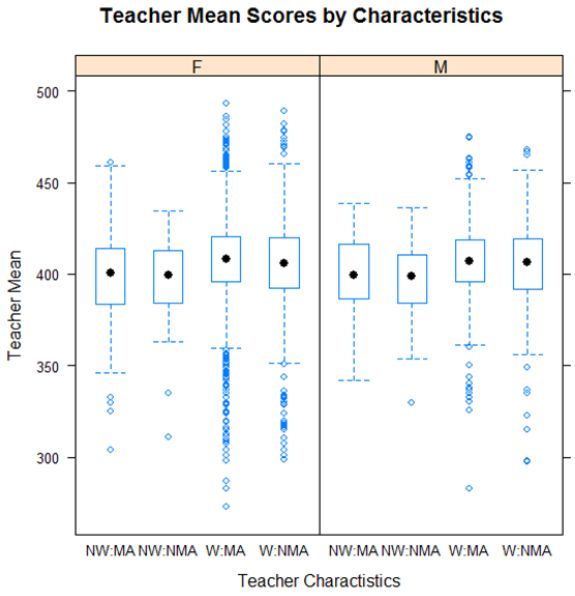
Avg Teacher Experience by Teacher Characteristics			
		Higher Degree	
Gender	*Ethnicity*	No MA	MA
Female	Non-White	9.07	12.76
	White	11.59	14.49
Male	Non-White	12.08	12.14
	White	13.33	14.75

As you can see from the left hand table, there is a clear variation in teacher characteristics across the 5th grade math teachers i.e. not all the teachers have the same characteristics. Teachers who are female, White with MA degree comprises 46.7% of the teacher labor force with 1337 teachers. These teachers comprise the largest group and surpass other groups of teachers by a large margin. Furthermore, the proportion (not absolute amount but percentage within the ethnic group) teachers who are non-White and do not have a MA degree is 38.4% which is larger than the White teachers with 33.3%. That is, White teachers not only have more MA teachers in absolute amount but also in relative/proportionally in comparison to non-White teachers. The teachers' gender balance across ethnicity and MA degree on the other hand shows relatively equal distribution. Looking now at the right hand table, it is evident that teachers who are White have a higher on average teaching experience than non-White teachers with or without MA degree. Thus there is a clear association between the White ethnicity, experience and MA degree teacher characteristics and not so much with respect to the gender of the teachers. To examine how these different teacher characteristics is associated with student performance, the following cross tabulated table is constructed.

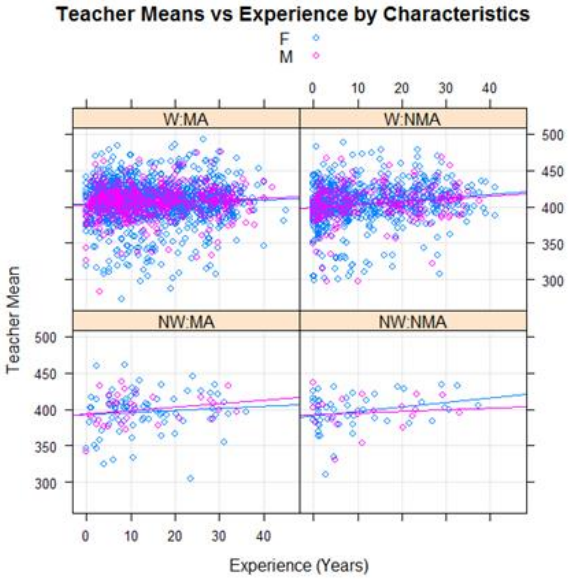
Teacher Mean Scores by Teacher Characteristics			
		Higher Degree	
Gender	*Ethnicity*	No MA	MA
Female	Non-White	397.01	396.57
	White	404.21	406.44
Male	Non-White	395.58	400.16
	White	404.58	406.39



As you can see, the combination of White ethnic background and the possession of MA degree have the highest mean performance. And the gender difference in performance is not as distinct. Looking at the entire (conditional) distribution for each group most of the groups have similar distribution pattern e.g. similar boxplot. White female teachers with MA degree indicate relatively large amount of outlying values but this can be (partially) attributed to the very large sample size



Finally, looking at the association of teacher experience for each group (combination of categorical teacher characteristics) shows the folloing results. As you can see from the diagram, despite the difference in sample size, the slightly positive association of the teacher experience on teacher mean performance is clearly evident in each group. The correlation coefficient for each group is provided in the right hand table which again indicate a low positive value for majority of the groups. Yet, there are some gender and MA degree possession differences in the experience effects for non-White teachers. The effects of experience on student performance is relatively high for non-White female students with out MA degree and non-White male students with MA degree compared to other groups of teachers. Thus the teacher experience variable is interrelated with the other teacher characterstics variables.



Correlation of Teacher Means and Experience

		MA Degree	
Gender	*Ethnicity*	Yes	No
Female	Non-White	0.091	0.252
	White	0.050	0.165
Male	Non-White	0.232	0.104
	White	0.098	0.174

To sum, the above findings illustrated the nexus of associations inherent between the teacher characteristics variables. The triangulation of these variables also illustrated different mechanims in affecting the student test score outcome. There findings therefore further support the need of VAMs to separate these web of associations in order to capture the unique effect of each variables.

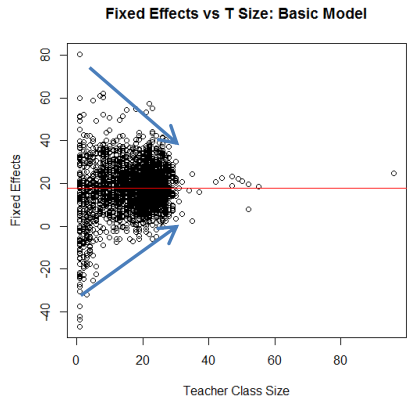
Chapter 7:

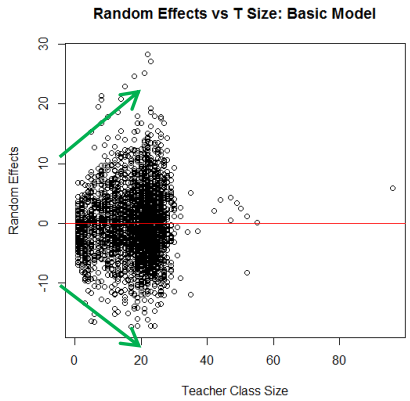
Shrinkage Property and Borrowing Effect – Comparing Fixed Effects and Random Effects

There is a clear difference between fixed effects and random effects (empirical Bayes) VA estimates of the basic VAMs. Looking at the caterpillar plots in Chapter 7 again, the magnitude of fixed effects is clearly larger and standard error estimates (confidence interval band) are not only longer but clearly more unstable than the random effects values. Yet, reflecting on the earlier explanation of these two models in Chapter 4, this finding is nothing surprising or new. It illustrates the important "shrinkage property" (or borrowing effect) underlying the random effects empirical Bayes predictions.[230] As described in Chapter 4, the random effects shrinks fixed effect VA estimates (adjusted teacher means) toward the adjusted overall grand mean in accordance to the amount of information provided by the teachers (teachers' class size). As the teacher sample size (teacher class size) fall, such teachers have less and less information/data to provide estimates of their adjusted teacher mean values. Their estimates therefore become more unreliable and unstable. Random effects acknowledges this problem by shrinking the fixed effects teacher adjusted means toward the overall adjusted mean (which is estimated with all the students in the data thus with more data and higher reliability) in accordance to amount of information provided by each teacher. For this reason, the shrinkage property is also referred to as the "borrowing effect" as teacher with small sample size borrow information from the rest of the teachers by gravitation toward the adjusted grand mean. In essence, the random effects adjust and compensate the unreliability of the individual teacher fixed effects estimates.

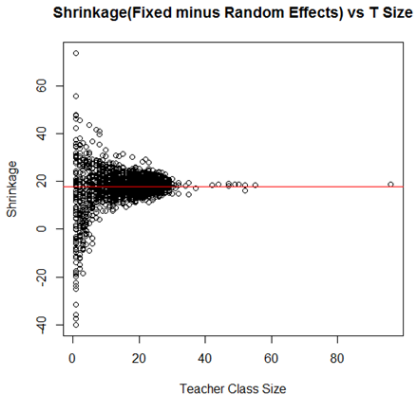
Following suit with the analysis of Gelman and Hill (2007) and Hoff (2011), who all work extensively on empirical Bayes and Bayesian methods, we investigate this shrinkage property through plotting the fixed and random effects by the teacher class size as shown below.

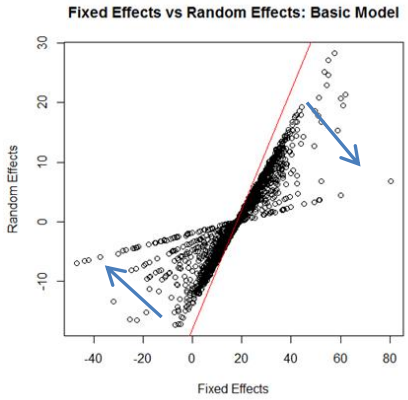
[230] From here on, random effects estimates are used to imply the empirical Bayes predictions.





The two diagrams illustrate opposite pattern. Fixed effects show more scatter and variation for teachers with small sample size. As shown in the blue arrows, as the teacher sample size increases, fixed effects increasingly gravitate/cluster between the values of 0 to 40.[231] With more data, it shows more stability in the estimates. The random effects, on the other hand, show much less variation/scatter for low values of teacher sample size. As shown in the green arrows, as the sample size increases, the random effects increasingly illustrate more scattering of the data points. In fact, as the sample sizes increase, the random effects are much closer to fixed effects values on the left. These trends all highlight the shrinkage property. For teachers with small sample size, the random effects are shrunk more towards the adjusted grand mean of 0 while for teacher with large sample size such adjustment is minimized and unique teacher adjusted means (fixed effects) prevail. To further illustrate the shrinkage property, the following diagrams are constructed.

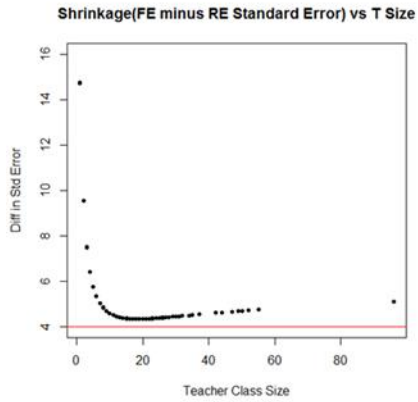


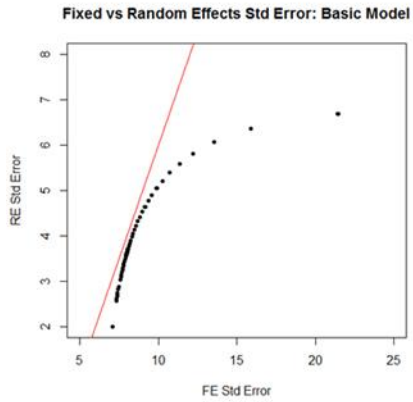


The diagram on the left plots the shrinkage value (fixed effects minus random effects) and the teacher sample size. As you can see, the two estimates show the most difference for small teacher sample sizes. The difference decrease as the sample size increases. It gravitates toward the value 17 which is the difference in the adjusted grand mean of the random effects and the adjusted mean of the reference/based group (teacher ID = 1) of the fixed effects. This finding is identical to the analysis conducted by Hoff (2011). Hoff (2011) also

[231] For technical reasons with the statistical software R, the dummy coding (with reference group teacher ID = 1) was used instead of the effects coding in providing the fixed effects VA estimates. Both coding methods provide the same information and results (i.e. the relative ranking of the VA estimates).

plotted the fixed and random effects to illustrate that the largest difference (due to small sample sizes) were evident particularly with the extreme values of fixed effects. And the author therefore (indirectly) linked that extreme fixed values were due to small sample sizes. To investigate this point, the two VA estimates are plotted as shown in the above diagram on the right. For moderate fixed effects values (the majority of the data points concentrated between 0 and 30), random effects also show a similar pattern as it lies in line on the straight line with slope of 1 as indicated by the red line.[232] But for extreme fixed effects, the random effect deviates more from the straight line leaning more towards the value of 0. This is illustrated with the observations rotating clockwise (as shown with the blue arrows). This confirms the Hoff (2011) point that extreme fixed values due to the small sample sizes and most shrinkage is evident for these extreme fixed effects values. Finally, comparing of the standard error estimates (used to construct the confidence interval band) shows the following trend.





As you can see, the difference in the standard error consistently falls as the teacher sample size increases. This again portrays how fixed effects with very small sample sizes are very unstable and imprecise. Plotting the two estimates (as shown on the right), it is clearly evident that all the random effects values are smaller than fixed effects values as the plot lies under the 45 degree line (line with slope 1 which signify the equivalence of the two estimates). For low to moderate fixed effects standard error values (with large sample sizes), random effects also show similar pattern as the plot lies very close to the 45 degree line. But again for large fixed effects standard error values (with small sample sizes), the random effect values deviates more from the 45 degree line.

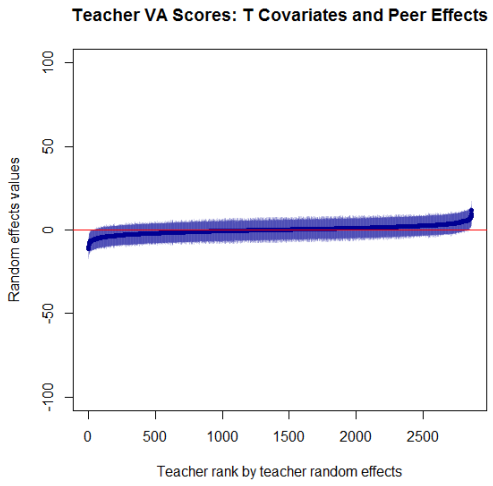
[232] The red line adjusts for the difference of 17 between the adjusted grand mean of random effects and adjusted mean of the reference group of fixed effects. In essence, it indicates the equivalence of the two estimates.

Extended VAMs with Teacher Characteristics and Student Peer Effects:

Random Effects VAM with Teacher Characteristics and Peer Effects			
	Estimate	*SE*	*t value*
Intercept	264.05	2.946	89.63
Student Characteristics:			
Student is Female	2.453	0.184	13.32
Student is Native American (base)	-	-	-
Student is Asian	3.608	0.536	6.73
Student is Black	-1.118	0.582	-1.92
Student is Hispanic	0.569	0.496	1.15
Student is White	1.841	0.444	4.15
Student is on FRPL	-1.998	0.272	-7.36
Students' Prior Test Score*	0.681	0.003	266.22
Teacher Level Variables:			
Teacher is Female	0.253	0.273	0.93
Teacher is White	0.533	0.494	1.08
Teacher has MA Degree	0.079	0.254	0.31
Teacher Experience	-0.012	0.012	-1.01
Teacher/Class Avg Test Score	0.350	0.007	48.94
Percentage of Female Students	-0.949	1.124	-0.84
Percentage of FRPL Students	9.117	0.995	9.17
Percentage of White Students	-3.294	0.623	-5.29
Teacher Class Size	-0.021	0.019	-1.14
School Fixed Effects	No		
Note: * Prior test scores are grand mean centered			

Variance Components:	
Between Teacher Variance	14.92
Within Teacher Variance	414.33
Intraclass Correlation	0.035
AIC	454935
BIC	455103
Number of Students	51161
Number of Teachers	2864
Degrees of Freedom	51143

The above estimates illustrate very little difference from the extended VAM which took into account only the student peer effects (as explained in the main text). The estimates of the teacher characteristics variables are very small and insignificant. These findings shed light on the profound and robust effect of the student peer effects variables on student performance and teacher VA estimates. And the teacher VA estimates as shown below virtually illustrated the same finding from the estimates with only the student peer effects. The mean of VA estimates is 2.16539e-12 with standard error of 2.35408 and the mean of standard error of the VA estimates is 3.04992 with standard error of 0.286439.

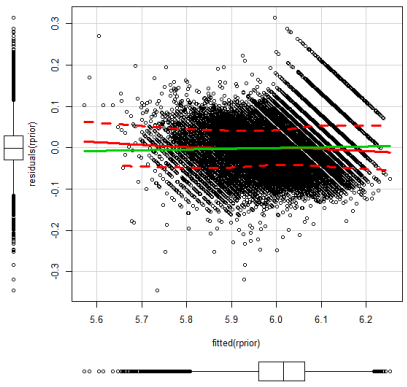


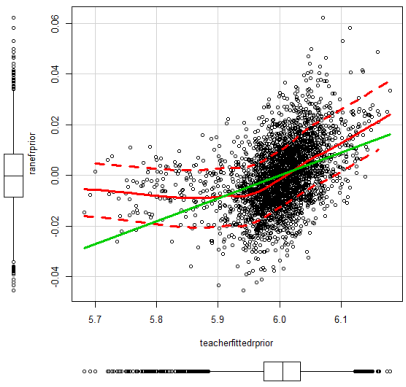
Chapter 8:

Diagnosis of the Random Slope and School Fixed Effects Extended VAMs

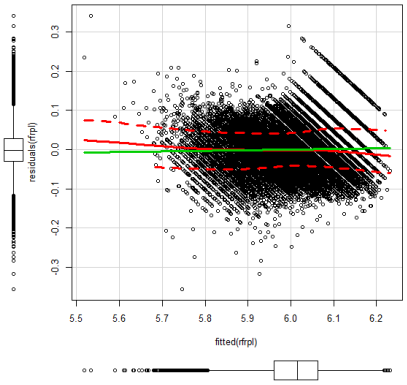
In the preliminary VA analyses in Chapter 7, the random slope models (for the FRPL and prior test scores variables) and the school fixed effects model were estimated. These modeling options could also potentially cure the un-modeled systemic pattern evident in the level 2 residuals shown above. Therefore the key diagnosis plots of these models using the log transformation with squared prior test score for the level 1 specification is provided below starting with the random slope model with prior test scores, followed by random slope with FRPL and finally the school fixed effects model.

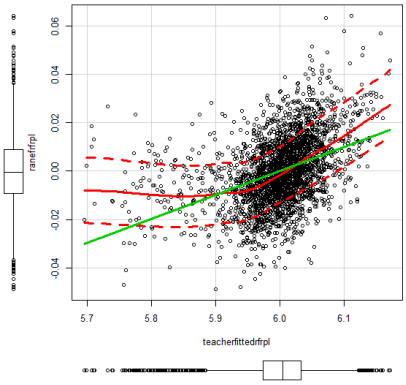
Level-1 and level-2 residual scatterplots for random slope model with prior test scores:



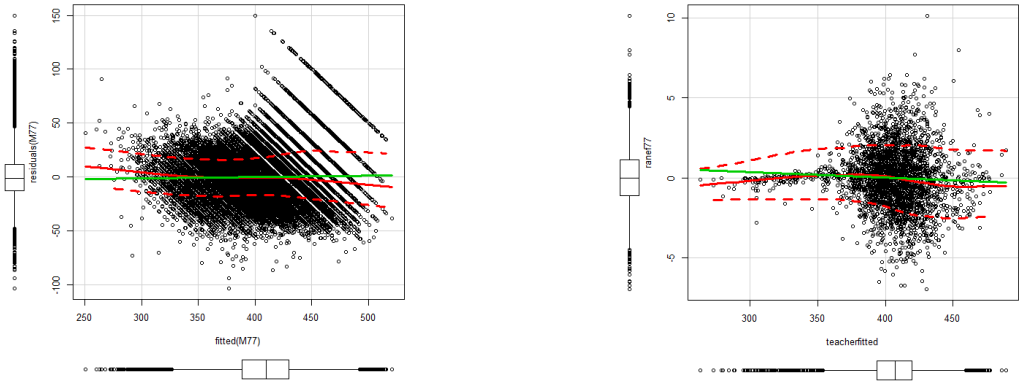


Level-1 and level-2 residual scatterplots for random slope model with FRPL:





Level-1 and level-2 residual scatterplots for school fixed effects model:



As you can see, all the models illustrate good model performance for the level 1 residuals (shown on the left hand column). The normal distribution was well approximated with skewness and kurtosis values all in the 0.3 range and 4.0 range. The randomly distributed pattern represented with a flat regression curves (indicating no omitted systemic patterns) and horizontal band shape is also clearly illustrated. But looking now at the level 2 residuals pattern (on the right hand column), although the normally distributed residual pattern is retained in all of the models,[233] the flat horizontal banded shaped randomly distributed pattern is clearly not achieved. Both random slope models fail to take into account the upward systemic pattern. The exogeneity condition is again violated and the bias is engendered in the model. Further diagnosis of the residual pattern to identify the source of systemic pattern is necessary.[234] The school fixed effects model, on the other hand, does a better job of taking this systemic pattern into account but it introduces new form of violation with a clear non-constant variance that consistently (and non-linearly) expands as the fitted values increase. Inefficiency in the model estimated is engendered. The school fixed effects model may be imposing too many restrictions with the 936 dummy school variables. The model could be too ambitious for the data to accommodate this model i.e. there is likely to be not enough information to provide reliable estimate for both the school fixed effects and the teacher VA effects. And this contributes to the increase in model misfit/residuals which then increases the variance of the residuals. Thus both the random slope models and the school fixed effect model demand further work and revision to provide convincing evidence that the necessary assumptions to ensure BLUE estimates are jointly met.

[233] Skewness and kurotsis values together with histogram and qqplots were examined for all models to illustrate close resemblance of the normal distribution.

[234] Moreover, random slope models introduce new set of assumptions. Just like the random intercept (random effects VA model), the random slope parameter (across the teachers) must also be randomly distributed normal with mean of zero and constant variance. This condition can be diagnosed by applying the empirical Bayes formula to the random slope parameter. Further diagnosis and revision process is therefore necessary to justify the BLUE estimates for the random slope models.

Chapter 10:

Top 32 Exemplary Teachers Based on VA and % of Female Students

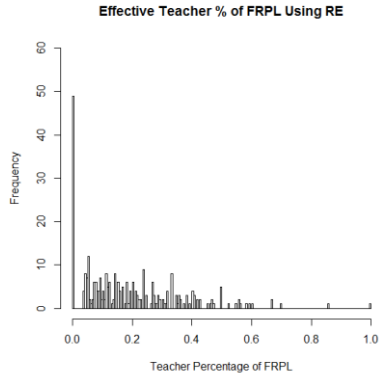
Top 32 Exemplary Teachers Based on VA and % of Female Students					
Random Effects			Fixed Effects		
204	1270	1977	20	1285	2107
463	1285	2085	249	1339	2127
560	1350	2173	459	1380	2173
620	1380	2175	495	1466	2195
719	1393	2182	657	1539	2332
854	1655	2195	668	1716	2415
932	1810	2332	717	1772	2502
1043	1847	2338	719	1774	2577
1171	1881	2434	814	1847	2606
1221	1976	2445	1200	2027	2630
		2571			2721
		2861			2730

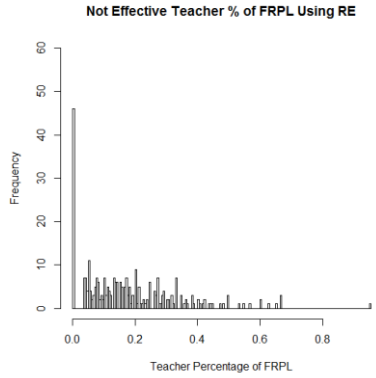
Bottom 40 Non-exemplary Teachers Based on Each Equity Indicator

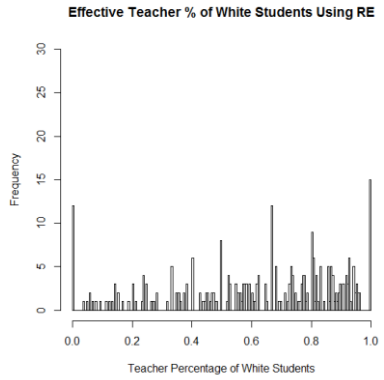
Bottom 40 Non-Exemplary Teachers																
% FRPL Students								% White Students								
Random Effect				*Fixed Effect*				*Random Effect*				*Fixed Effect*				
8	460	1528	2464	28	863	1544	2366	123	476	1276	1898	50	1070	1841	2366	
19	692	1672	2511	180	894	1631	2368	143	509	1361	2013	108	1098	1886	2408	
76	723	1683	2551	330	980	1637	2433	177	656	1457	2262	180	1156	1995	2433	
133	928	1700	2573	460	986	1660	2511	256	697	1465	2551	330	1304	2013	2520	
143	980	1771	2590	544	1027	1822	2520	306	784	1481	2573	594	1448	2100	2551	
158	1037	1891	2596	557	1070	1876	2551	371	910	1580	2608	623	1544	2168	2566	
209	1070	1893	2608	673	1098	2100	2566	387	928	1672	2670	656	1631	2177	2627	
259	1098	1969	2623	692	1304	2168	2652	400	1070	1742	2701	673	1660	2180	2819	
296	1217	2080	2670	748	1432	2180	2683	425	1098	1746	2819	927	1746	2210	2829	
387	1361	2342	2701	821	1528	2230	2819	447	1257	1830	2829	1067	1822	2245	2832	
392	1432	2368	2723													
		2456	2819													
Note: the top 46 for the random effects for the % FRPL was considered to accommodate few more extra-non-exemplary teachers																

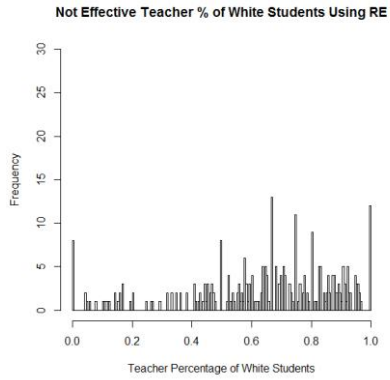
Bottom 40 Non-Exemplary Teachers																
% Female Students									% White Students							
Random Effect				*Fixed Effect*					*Random Effect*				*Fixed Effect*			
116	509	1108	2074	259	1089	1813	2418		7	296	1564	2386	8	444	1643	2509
143	617	1482	2086	278	1108	1818	2439		8	306	1643	2505	50	656	1683	2530
152	619	1524	2342	623	1296	1876	2520		19	392	1672	2530	94	910	1771	2531
173	656	1572	2439	656	1304	2063	2589		94	403	1742	2531	143	1089	1804	2551
202	682	1607	2507	673	1511	2074	2608		116	434	1832	2551	158	1217	1832	2608
259	793	1639	2589	793	1631	2100	2615		123	509	1969	2573	180	1236	2038	2623
278	803	1813	2608	821	1637	2125	2652		133	807	2038	2623	202	1265	2080	2692
353	924	1818	2760	863	1639	2168	2760		158	842	2263	2638	209	1268	2095	2701
419	989	1898	2805	894	1673	2177	2819		209	1037	2325	2723	278	1445	2120	2704
434	1037	1906	2819	927	1748	2245	2827		278	1265	2336	2835	392	1465	2505	2827

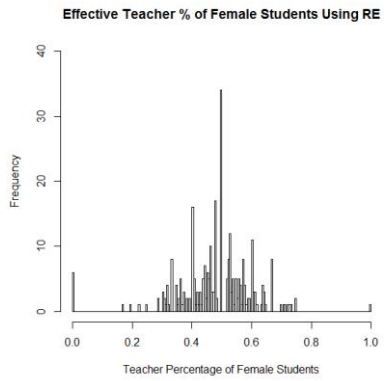
Descriptive Summary Distribution of Effective and Non-Effective Teachers Based on Random Effects Estimates

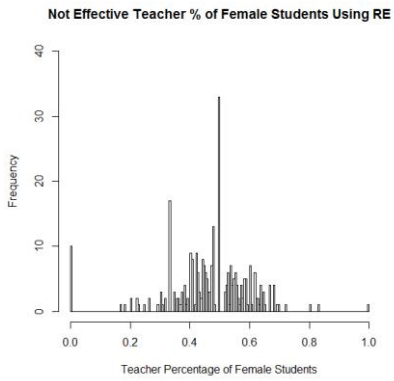


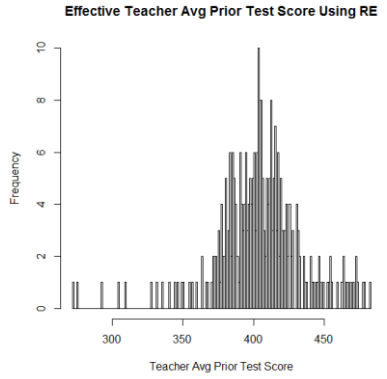


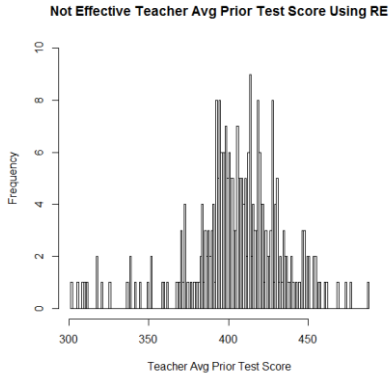


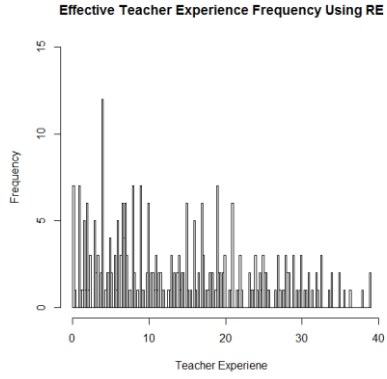


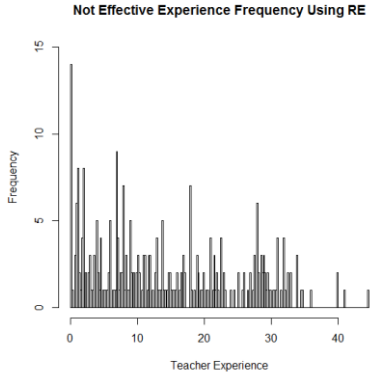

Effective Teacher % of Female Students Using RE
Frequency
0
10
20
30
40
0.0
0.2
0.4
0.6
0.8
1.0
Teacher Percentage of Female Students


Not Effective Teacher % of Female Students Using RE
Frequency
0
10
20
30
40
0.0
0.2
0.4
0.6
0.8
1.0
Teacher Percentage of Female Students


Effective Teacher Avg Prior Test Score Using RE
Frequency
0
2
4
6
8
10
300
350
400
450
Teacher Avg Prior Test Score


Not Effective Teacher Avg Prior Test Score Using RE
Frequency
0
2
4
6
8
10
300
350
400
450
Teacher Avg Prior Test Score


Effective Teacher Experience Frequency Using RE
Frequency
0
5
10
15
0
10
20
30
40
Teacher Experiene


Not Effective Experience Frequency Using RE
Frequency
0
5
10
15
0
10
20
30
40
Teacher Experience