How good are value-added measures of teacher performance? A review and empirical investigation using data from Turkey

Aslantas, Ismail

How good are value-added measures of teacher performance? A review and empirical investigation using data from Turkey

Ismail ASLANTAS

Thesis submitted in partial fulfilment of the qualification of
Doctor of Philosophy

School of Education
Durham University

October 2021

ABSTRACT

Education systems all over the world aim to provide good quality education for their citizens. This would require a good supply of quality teachers. The role of teachers is now more complex than ever before. Consequently, evaluating the quality of a teacher has also become more complex. While we may feel that we know intuitively what an effective teacher looks like, there is little consensus on how best to measure or capture the essence of a good teacher. Classroom observations protocols, interviews and surveys with teachers and pupils are commonly used to assess teachers. Increasingly, governments and schools are using standardised pupil test scores in teacher performance appraisal as a way of estimating how much difference teachers can make to student attainment by comparing the progress students make. This is seen as perhaps more objective or fair because students' test scores are considered objective measures. Such evaluation of teachers, also known as value-added models or VAMs, are increasingly used to measure teacher effectiveness for high-stake decisions, such as teachers' salaries and promotions. Teachers are rewarded or penalise based on these value-added measures.

VAMs have attracted considerable attention in recent years. Many researchers have raised concerns about their validity and reliability. There are also concerns about VAM's ability to predict the effectiveness of teachers consistently. Value-added measures of teachers are known to vary from year to year and from subject to subject. Different value-added models can also produce different estimates of teacher performance depending on the student achievement test scores used.

This study adds to the current debates by examining the stability of VAMs to see whether teacher effectiveness can be predicted consistently using different parameters, such as observable student, teacher/classroom and school characteristics, the number of student test scores obtained over time, and the data analysis methods used. Value-added measures can only be useful in estimating teacher effectiveness if they produce consistent results for the same teachers across time for different students.

This new research begins with a systematic review of the existing literature examining the stability of different VAMs as a measure of teacher performance. Of 1,439 results, 50 studies met the inclusion criteria to be included in the synthesis. Each of these studies was given a padlock rating in terms of the trustworthiness of its findings based on four criteria (such as

research design and threats to validity) using a bespoke assessment tool. Studies were rated from 0 (very weak) to 4 padlocks (the most secure that can be expected). Since the main research question (stability of estimates) is descriptive, correlational/comparative studies are appropriate in design. Most studies retrieved were correlational/comparative in design. Some of them rated the highest 4🔒, as they were large-scale, allowed random teacher-student allocations, and had low attrition. The majority of the studies in the review were rated 3 padlocks as they employed administrative/panel data where students are not randomly assigned to teachers in value-added estimates and/or were smaller or had higher attrition.

The strongest studies (4🔒) revealed that using one prior attainment score is sufficient to predict teacher performance. Using additional prior test scores does not increase the stability of value-added teacher effectiveness estimates consistently. Including student, school and teacher/classroom-level variables adds little to the predictive power of teacher performance assessment models. This suggests that these variables are not good predictors of teacher effectiveness. The systematic review found no evidence that any particular data analysis method is better in its ability to estimate teachers' effectiveness reliably.

Most studies in the review were conducted in the US using national administrative data. To see if the findings also apply in other contexts, longitudinal data of five teaching subjects (maths, Turkish science, history, and English) from one province in Turkey was then used to test the stability of value-added estimates. The data included 35,435 Grade 8 students (age 13-14, equivalent to Year 9 in the UK), matched to 1,027 teachers. To test how much progress in student academic achievement is available to be attributed to a teacher from one year to the next, a series of regression analyses were run. Models included contextual predictors at student-, school-, and teacher/classroom-level.

Consistent with the findings of the systematic review, the results show that the best predictor of students' later test scores is their prior attainment. Using additional years' test scores instead of a single prior-year attainment score contributed little to improving value-added teacher effectiveness estimates. Including other factors, such as student, teacher, and school characteristics in the model also explains very little in the variations in students' test scores once the prior attainment is taken into account (although the data on teacher characteristics was limited in the dataset). Correlation analyses suggested that there was no meaningful relationship between teacher effectiveness scores and the teacher/classroom-level variables.

Interestingly, teacher experience, regardless of whether it refers to their total experience or only that in their current schools, is negatively related to teacher effectiveness scores. In other words, more experienced teachers tend to have lower effectiveness scores on the value-added estimate. There was no evidence that teachers are more effective in smaller classes. Only a modest correlation was found between class size and teacher effectiveness. Intriguingly, students in large classes tend to have more “effective” teachers in value-added terms (except in history), although the difference is minimal.

The analysis also found that teachers’ previous effectiveness scores had little or no relationship with their current effectiveness scores, regardless of teaching subjects. Consistent with the literature in the review, this study also found that teacher effectiveness scores based on value-added estimates vary substantially across years. This means that the same teacher can be considered “effective” in one year and “ineffective” in other. This casts doubt on the reliability and meaningfulness of value-added measures.

As with previous studies in the systematic review, there is no evidence from the Turkish data that any single value-added approach is superior to any other approach regarding the ability to consistently estimate teachers’ effectiveness. There is no advantage in using more sophisticated statistical models.

The findings of this study suggest that regardless of the number of test scores, or variables used or data analysis methods, there is no consistent or reliable way of measuring teacher effectiveness. This highlights the danger of using value-added models in measuring teacher effectiveness. Studies suggested that some of the inconsistencies could be the result of measurement error and the timing of the test. There is, therefore, the risk of misclassifying teachers as “effective” or “ineffective”. Some teachers may be deemed 'effective' on one test but not another simply based on when the tests are scheduled. These findings have important implications for policy and practice. Value-added models should not be used to make high stake personnel decisions. They may have some value for research purposes or to provide formative feedback to headteachers about a class or a teacher as part of a larger set of evidence.

One major limitation of VAMs is that they measure teacher performance using tests designed to measure student performance. The assumption is that student performance is directly related to teacher quality. While there has been a lot of research on developing teacher quality, measuring teacher quality is itself problematic. The issue of measuring teachers performance has been one of the leading issues in education policies. A critical question that needs to be asked is not how effective teacher are, but what is the purpose of evaluating teacher performance? If such an exercise aims to differentiate “effective” from “non-effective” teachers since there is no reliable method or no methods that have been robustly tested and shown to work in identifying effective teachers, why are we still doing it? To improve teachers’ effectiveness and keep them updated with robustly tested and proven teaching approaches, it might be better to provide teachers with training, professional development to develop pedagogic skills, social and personal relationship skills, behavioural management, and subject knowledge. Assuming that classroom teachers have gone through teacher training and are certified, then they should be qualified to teach. If they are not deemed “effective”, it is perhaps the failure of the selection and training process more so than the quality of the individual teacher.

Another major limitation of VAMs is that they are comparative and zero-sum. For a teacher to be deemed effective, another must be deemed ineffective. Thus, if all teachers were actually effective (or ineffective), a VAM would still assess up to half of them to be ineffective (or effective). They are not fit for purpose.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

DEFINITION OF KEY TERMS

Some of the key terms used in this study are defined below.

Value-added Models (VAMs)
A variety of approaches based on student progress used to measure the effect of a teacher on the attainment of their students by isolating it from other factors beyond the teacher’s control.

Value-added teacher effectiveness
Here, it is defined as an estimation of the differences between expected and observed student test scores in a value-added model (Kersting et al., 2013).

Effectiveness
The causal contribution of a teacher in increasing student achievement.

Causality
The effect of a particular action that leads to specific, measurable results or effects (Stock & Watson, 2012).

Correlation
A statistical indication of a relationship that represents how two variables vary in accordance with each other.

Stability
Consistency of results across various specifications or models

Reliability
A statistical measure of the stability of results over repeated testing or modelling, also refers to the degree to which results obtained are free from random errors (APA, n.d.).

Validity
A statistical reflection of the ability of a measurement instrument to measure what it claims to measure

LIST OF ABBREVIATIONS

CVA	Contextual Value Added
HLM	Hierarchical Linear Modelling
IRR	Inter-Rater Reliability
MET	The Measures of Effective Teaching
MGP	Median Student Growth Percentile
NCLB	The American No Child Left Behind Act of 2001
OLS	Ordinary Least Squares
PM	Progress-Monitoring Model
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT	Randomized Control Trial
RG	Residual Gain Model
SBSA	The Step-by-Step Achievement
SGP	Student Growth Percentile Model
SEN	Special Education Needs
SES	Socioeconomic Status
TPES	Teacher Performance Evaluation Systems
TVAAS	The Tennessee Value-Added Assessment System
VAMs	Value-Added Models

DECLARATION

The work in this thesis is based on research carried out at the Department of Education, Durham University, England. No part of this thesis has been submitted elsewhere for any other degree or qualification.

STATEMENT OF COPYRIGHT

ACKNOWLEDGEMENTS

I am deeply grateful to Allah for the presence of many people in my life who have given me support, understanding, and encouragement on this journey. Without these individuals on my long and arduous journey, I would not have been able to achieve it.

First of all, I would like to remark that this dream would not be possible without funding from the Ministry of National Education, Turkey. Therefore, I would like to express I will always be grateful to the Republic of Turkey.

I would like to extend my sincere thanks to my supervisors Prof. Stephen Gorard and Prof. Beng Huat See for their invaluable advice, continuous support, and patience during my doctoral study. Their friendly guidance, valuable and expert feedback, and perhaps most importantly, their endless support has made an invaluable contribution to the development of my work. I thank them for their generous contribution to the success of this research project from the very beginning. I’ve always felt so lucky to have supervisors like you.

And finally, I need to express tremendous gratitude to my beloved wife, Sercan, and our kids, Emir Eymen and Kayra Yagiz, for their endless support, love, and tolerance during this entire journey. Whenever I needed you, you were always with me and always believed in me. My acknowledgements would be incomplete without thanking my parents, Zulfiye and Mevlut, you taught me to persevere and confidence in myself. These qualities were crucial in achieving my goal.

To my family

SECTION I

INTRODUCTION AND CONTEXT OF THE STUDY

Section I of this thesis consists of three chapters. Chapter 1 provides the rationale for the study and background to the study, explaining the aim and the research questions. Chapter 2 discusses the relevance and importance of teachers and why it is necessary to measure their performance. It also outlines conventional methods of evaluating teachers. Chapter 3 introduces the idea of growth models used in educational accountability systems, and in particular value-added models in teacher performance evaluation.

CHAPTER 1
THE RATIONALE OF THE RESEARCH

1.1 Background

For over almost three centuries, we have been struggling to find the best way to evaluate teachers. Historically, in the 17th to early 19th-century, teacher evaluations used to be simple inspections to see whether teachers were doing what was expected of them (Jewell, 2017). From the mid to late 19th century, more attention was paid to training and improving teacher practice. As the focus of schools shifted to social efficiency in the early 20th century, teachers were evaluated using observations and feedback. According to Jewell, it was between 1900 and 1920 that it is was proposed that teaching could be made more efficient using business productivity methods, and these models influenced the modern-day teacher evaluation model.

However, unlike businesses where there are clear and measurable outputs, teaching is a complex task, and the outputs of an ‘effective’ teacher go beyond simply test scores. Developing confident, socially well-adjusted individuals, preparing students for life in the future, enhancing their employability, and so are just as important as academic attainment. Because of the complex tasks that teachers perform, measuring their effectiveness is therefore challenging (Gorard, 2013; McCaffrey et al., 2003; Darling-Hammond, 2000).

Nevertheless, it is hard to disagree that highly qualified and effective teachers matter. As schools become more accountable for student outcomes, there is an increasing need to ensure that effective teachers are hired and retained in classrooms. Similarly, parents also desire their children to be enrolled in the best schools and be taught by the best teachers. This throws up questions such as how do we measure teacher quality and what constitutes quality teaching.

One measure of teacher outcome that is understood that can be objectively measured is student attainment. It is widely accepted that teachers are considered one of the most significant school-related factors in enhancing students’ academic achievement (Aaronson et al., 2007; Rivkin et al., 2005; Darling-Hammond, 2015; Sanders et al., 1997).

However, while it is accepted that evaluating teacher performance is beneficial in enhancing teacher development and student outcomes, there is no single agreed method to measure it. Teacher performance can be assessed in a range of ways, such as via classroom observation, survey, self-evaluation, portfolio, or student achievement growth analysis (Kane & Cantrell, 2010; Coe et al., 2014; Kane et al., 2015; See, 2020). In the past, teachers (and schools) were evaluated on students' attainment in a single test against a given benchmark score, however in recent years academics and decision-makers have focused on measures based on students' achievement growth, called value-added models (VAMs). At this point, it is helpful to highlight the difference between *achievement* and *attainment*, both of which provide different information but are often used interchangeably. While *attainment* is a snapshot measure of how well students are performing against a given standard, *achievement* refers to the learning progress students make over time. Value-added models are a statistical way of isolating and analysing a teacher's contribution to student learning.

Although VAMs are commonly used in the field of business and economics, they have also become popular among educational researchers. Consequently, these models are now increasingly employed in many countries to measure teacher performance. VAMs are based on the assumption that students' achievement, not attainment, can be attributed, at least partly, to their teacher quality (Darling-Hammond, 2015). The American No Child Left Behind Act of 2001 (NCLB), for example, was introduced based on the belief that teacher quality is the most significant of the school-related factors influencing students' academic achievement (Aaronson et al., 2007; Rivkin et al., 2005).

Such policies are effectively holding schools and teachers accountable for students' achievement. Stronge (2006) asserted that teacher performance evaluation is also vital for improving students' learning outcomes. Therefore, it is not surprising that many states and school districts in the US were compelled to modernise their own teacher performance systems based on student test scores or obtain developed ones from other states (U.S. Department of Education, 2009). Since how well students perform might be affected by their background characteristics, their prior learning, and other factors are beyond the teacher's control (Wei et al., 2012), VAMs are considered useful as they can isolate these background factors to see how much the progress made by the student can be attributed to the teacher net of these other factors.

In order to estimate the accountability or effectiveness of schools and teachers, various types of VAMs have been proposed and applied by many states and school districts in the US, for example. One of the first states in the US to develop a teacher performance measurement system based on student outcomes was Tennessee. The Tennessee Value-Added Assessment System (TVAAS), probably the earliest value-added model used routinely in education, was developed by William Sanders, former professor at the University of Tennessee, and his colleagues (Sanders et al., 1997; Sanders & Horn, 1994). Although several different VAMs have been developed since then to estimate an individual teacher's effects on student attainment, the fundamental idea of all VAMs involves determining the changes in the students' academic performance over the years using their prior and subsequent test scores. The value a specific teacher adds to their students' learning can be estimated by controlling for school-related factors (e.g., class size, school location), teacher-related factors (e.g., sex, teaching subject qualification, and experience), home-related factors (e.g., parents' attitude, parental socio-economic background), and student-related factors (e.g., prior attainment, attendance, having special needs).

In all VAMs, a particular teacher's performance is statistically estimated by using their students' test scores in a subject and grade. Changes in the students' performance in tests taken for at least two consecutive years are then attributed to the teacher's "effects". In the VAM concept, the word "effect" has a conceptual meaning that refers to the estimation of the differences between expected and observed student test scores (Sanders et al., 1997; Kersting et al., 2013). Conceptually, a teacher's effect on a student's achievement is the difference between the student's achievement under a particular teacher in their own class setting and the achievement that can be achieved by assuming that the same student is in another plausible class setting (McCaffrey et al., 2003). There is a common acceptance that an individual teacher performance based on VAMs reflects the contribution a teacher makes to a student's measured achievement gains (Darling-Hammond et al., 2012a). The value-added predictions are based on one or more previous test scores and often other students' characteristics (Ouma, 2014). These predictions are potentially beneficial in identifying the most and least effective teachers in the school, district or whole country (Schmitz, 2007).

In opposition to policy-makers' view that VAM-based accountability systems are fairer (Swanson, 2009), an increasing number of studies have questioned the reliability and validity of VAMs (Rothstein, 2007; Garrett and Steinberg, 2015; Stacy et al., 2018). The reliability and validity of such models are of paramount importance, especially when used for high-stakes personal decisions, such as promotion, dismissal, and wage increases for teachers. According to many researchers (Amrein-Beardsley & Geiger, 2020; Perry, 2016b; Newton et al., 2010; McCaffrey et al., 2003), the application of VAMs should be limited to providing formative feedback to teachers and principals and providing information on students' academic progress. The feedback provided by the teacher performance assessment could be used to determine the development needs of the teachers and also contribute to the enhancement of their knowledge and professional development.

Several different VAMs have been developed over the years, so the critical problem is that there is no single agreed-upon value-added model for measuring teacher and/or school effectiveness. Each has its own advantages and disadvantages. Therefore, the objective of this thesis is to examine the stability of VAMs estimates that use different contextual variables and analysis methods for teacher evaluation.

1.2 Significance of the Study

Despite there is a substantial body of research questioning the reliability and validity of VAMs (McCaffrey et al., 2004; Swanson, 2009), VAMs are still widely used in evaluating teacher and school effectiveness. In the UK, VAMs are used as a school effectiveness measure for allocation of funding and to identify schools that are placed in special measure. Schools are forced to close down based on such measures. In England, policy decision-makers believe that the imperfect system can be improved by using more complex models and including a greater number of contextualising variables (Kelly and Downey, 2010). Similarly, researchers have tried to improve VAMs used for teacher effectiveness by including more contextual factors and using more sophisticated statistical analyses. However, there have been no comprehensive studies that systematically review and evaluate the stability of VAMs under different conditions. Although some literature review studies have been completed (Goldhaber, 2015; Yeh, 2012; Everson, 2017; Koedel et al., 2015; Berliner, 2014; Darling-Hammond, 2015), there is only one study that systematically discusses a wide variety of issues raised by VAMs (McCaffrey et al., 2004), but this is now rather dated, and since then new and improved versions of VAMs have been developed.

This current study will update research evidence on the stability of value-added models used in teacher performance evaluation, specifically looking at the stability of VAMs that use a different number of students' prior test scores, different data analysis methods used, and different predictor variables, such as student/teacher/school characteristics. Unlike previous reviews, this review will assess the credibility or strength of evidence of each individual piece or research. As far as it is known, this is the only study that considers the design of the research in evaluating the evidence. Most previous narrative reviews have been conducted without weighting prior evidence in terms of its trustworthiness.

This study is also significant because it is the only study conducted in Turkey that evaluates the stability of VAM in evaluating teacher performance, bearing in mind that longitudinal administrative data are not routinely collected in the country. Much of the existing body of research on teacher performance evaluation based on VAMs has been conducted in the United States, partly because of the availability of longitudinal administrative datasets, which contain data on student test scores over time, their demographic characteristics (e.g., sex, ethnicity, socio-economic status and special education needs status). While such data may be routinely collected in the US, UK and many Western countries, they are not readily available in many countries. This new study will therefore contribute to the evidence on the stability of VAM from international school context. This is the aim of the second phase of the study, which involves a secondary analysis of longitudinal data of secondary school students from one province in Turkey to assess the stability of VAM as a teacher evaluation model in the Turkish context. The results of this secondary data analysis in combination with that of the systematic review, will allow for a comparison of the findings with those in the literature to see whether VAM is more stable in a different context. The findings will have important implications for research and policies on teacher evaluation practices using VAMs.

1.3 Purpose of the Study

The primary purpose of this study is to examine the stability of VAMs in estimating teacher effectiveness. It begins by reviewing previous research to assess the stability of VAMs as a measure of teacher performance under different conditions. In addition, the study also tests the stability of value-added estimates under various conditions in the Turkish context using longitudinal data from one province in Turkey.

The overarching aim of this study is to derive evidence-based recommendations to inform policy and practice. The findings of this study will provide guidance and inform policymakers and other stakeholders on the use of VAMs in teacher performance appraisal, especially for high-stakes purposes, such as decisions on dismissal and monetary reward.

1.4 Research Questions

To assess the stability of VAMs as a measure of teacher effectiveness, the main research question is:

How stable are teacher effectiveness estimates measured by VAMs?

The sub-questions are:

- How stable are teacher effectiveness measured by VAMs that consider student, school, and teacher-classroom characteristics?
- How stable are teacher value-added effectiveness estimates over a two-year period of time?
- How stable are teacher value-added effectiveness estimates when including an additional prior score (t-2)?
- Do different methods of analyses used in VAMs produce consistent teacher effectiveness estimates?

1.5 Overview of the Study Design and Methods

These sub-research questions are answered first by reviewing and synthesising existing research on the use of VAMs in estimating teacher effectiveness (except for the sub-research question about the stability of estimates over two years, as it was decided to be investigated after the systematic review study was completed), followed by primary research analysing based on real-life student attainment data from Turkey. Table 1.1 provides an overview of the data used in this study and the analysis method applied to answer each of the research questions.

Table 1.1 Research Questions, the Data Used, and the Analysis Method Applied in Each of the Research Questions

	Research Questions	Data	Data Analysis Methods
SYSTEMATIC REVIEW STUDY	How stable are teacher effectiveness estimates measured by VAMs?	50 eligible prior studies Studies were retrieved regarding: a) the predictors used in the estimations b) the number of test scores used c) the analysis methods applied	Followed the stages of the systematic review, clarified by Torgerson (2003).
PR IM	How stable are teacher effectiveness measured by	Student-level predictors	Multiple regression analysis using the forward

VAMs that consider student, school, and teacher-classroom characteristics?	School-level predictors Teacher/classroom-level predictors	selection method to retain the largest R- squared value by including as few predictors as possible
How stable are teacher value-added effectiveness estimates over a two-year period of time?	The prior year test score (e.g., *Grade 7) Two years prior test score (e.g., Grade 6)	(1) Multiple linear regression analysis (2) Pearson's/Spearman's correlation coefficients (3) Transaction matrix
How stable are teacher value-added effectiveness estimates when including an additional prior score (t-2)?	The prior year test score (e.g., Grade 7) Two years prior test score (e.g., Grade 6)	(1) Multiple linear regression analysis (2) Pearson's/Spearman's correlation coefficients (3) Transaction matrix
Do different methods of analyses used in VAMs produce consistent teacher effectiveness estimates?	The determined a) student-level predictors b) school-level predictors c) teacher/classroom-level predictors	(1) Multiple linear regression analysis (2) Residual gain model (3) Two-level HLM (4) Pearson's/Spearman's correlation coefficients (5) Transaction matrix (6) SD analysis

*The term "grade" is used with different meanings in different educational contexts (for example, used for exam, test, assessment outcome in the UK context, for statutory education in the educational context of Turkey and the USA), but it has been used to refer to the curriculumyear of study throughout this thesis.

A more detailed description of the methods used for data collection and analysis are presented in Chapter 4 for the systematic review study and in Chapter 5 for the primary research. Findings related to each research question are presented separately at first and then discussed synthetically.

1.6 The Scope of the Research and Limitations

Despite the complexities and challenges in conceptualising what an "effective" teacher is, many education systems have tried to use performance measurements to reward or penalize teachers. Such performance measurements are often based on assessing teachers' ability to improve student test scores, which has come to be used to mean teacher effectiveness.

However, if such performance measures cannot reliably estimate how effective a teacher is in improving students' learning outcomes, then their use in rating teachers cannot be warranted.

The study adds to existing research in this area by examining the stability of VAMs in estimating teacher performance under different conditions. This thesis begins with a systematic review by synthesizing prior research on the stability of VAMs, and the scope of this first part involves a total of 50 primary studies. The second part of the thesis examines the stability of VAM estimates using a longitudinal administrative data set extending over three school years, 2014-2017, from secondary schools in Samsun Province in Turkey. This part of the study involves a total of 1,027 teachers linked to 35,435 students.

Finally, while this thesis is limited to the inclusion of the relevant studies up to May 2019 in the systematic review, it is also limited to the use of a secondary dataset representing the date that the provincial education directorate possessed in May 2018. Missing data is a common problem when conducting research utilizing secondary longitudinal data and may influence the conclusion of this study. As missing data is not random in reality, I acknowledge that it has the potential to cause bias in my estimates.

1.7 Outline of The Thesis

The thesis is organised into five main sections.

Section 1 is made up of three chapters. Chapter 1 is an introduction to the study. It presents the background and rationale for the study, the main objectives and the research questions. Chapter 2 is a discussion of the role of teachers, what effective teachers and teaching look like, and how these qualities can be measured. Chapter 3 is a detailed discussion of growth models as a measure of teacher effectiveness. These are widely used in school accountability systems and forms the basis for this thesis.

Section 2 comprises two chapters (Chapters 4 and 5). This section describes the research design and methods of the study. Chapter 4 describes the process of conducting the systematic, from database searches to identifying, screening, quality assessment of the studies and synthesising the evidence. Chapter 5 deals with the research design and the methodology of the secondary data analysis for the primary research.

Section 3 presents the results of the systematic review. It contains four chapters. Chapter 6 describes the outcomes of the database search, the results of the quality assessment, and the characteristics of the included studies. Chapter 7 discusses the results of the systematic review looking at studies that evaluate the stability of VAMs that include student/teacher/school variables as predictors. Chapter 8 discusses the results of the systematic review looking at studies that evaluate the stability of VAMs that use one previous year's and additional years' test scores. Chapter 9 synthesises the results from reviewing studies that look at the consistency of VAMs using different data analysis methods.

Section 4 presents the results of the primary research, which analyses longitudinal data from Turkey to assess the consistency of VAMs. It consists of four chapters (Chapters 10, 11, 12 and 13). Chapter 10 describes the student attainment scores at Grade 8 (outcome variable used in the regression analyses) and the two previous years' test scores, which were used as predictors. Results of pre-analyses checks for normality, linearity, multicollinearity, and homoscedasticity are also presented. Chapter 11 analyses the stability of VAMs that consider student, school, and teacher/classroom-level characteristics in the analysis. Chapter 12 analyses the stability of value-added teacher effectiveness estimates over two-year periods and in terms of the inclusion of an additional prior year's test score. Chapter 13 examines the stability of VAMs that use different methods of analysis.

Section 5 is the concluding section, which has two chapters. Chapter 14 begins with a section that summarizing the main findings of the research, bringing together the results from both the systematic review and the primary research, and then the chapter addresses some of the limitations and challenges of this study. Chapter 15 discusses the implications of the findings for three groups of stakeholders: policymakers, researchers, and parents, and this chapter concludes with a sub-section by looking at what the findings of this study mean. It questions the purpose of teacher evaluation measures and what could we do instead?

CHAPTER 2
CONTEXT OF THE STUDY – DO TEACHERS MATTER?

This chapter begins with providing a general idea of teachers' materiality and effectiveness. In addition, the meaning of the term 'effectiveness' in the concept of value-added evaluation is also discussed.

2.1 The Policy Context - Why Teachers Are Important and Need A Performance Evaluation

In the 21st century, with the influence of globalization, radical developments and changes have occurred in every field. These rapid changes around the world have also directly affected the field of education. Developments in the field of information and technology, and the changing demands based on these developments, made it necessary for schools, which are social institutions, to pursue new developments, enhance their own innovations, and adapt to the increasingly competitive environment and the changing world. Nick Gibb, who has served as Schools Minister in England, stated in his speech on "The Purpose of Education" that the main purposes of schools in the changing world are to "ensure that more people have the knowledge and skills they need to succeed in a demanding economy", "resist attempts to divide culture from knowledge" and to "ensure that [young people] have the character and sense of moral purpose to succeed" in their adult lives (Gibb, 2015). Similarly, Secretary of State for Education Damian Hinds highlighted the importance of teachers at the annual conference for school and college leaders in 2018: "There can be no great schools without great teachers. To motivate children, to make knowledge meaningful, to inspire curiosity. The quality of teaching matters more than anything else, and it matters even more for disadvantaged pupils." (Hinds, 2018).

In the focus of both speeches, teachers have been assigned the important task of ensuring that students leave compulsory education with the knowledge and skills needed to meet the requirements of the demanding economy. This focal point is also among the 10 main goals set in the UNESCO 2030 Education agenda. UNESCO established its framework action plan for countries in 2015 to ensure that students finish school with "relevant skills for decent work" (UNESCO, 2016). For these reasons, teachers have begun to take huge responsibility on their shoulders to improve their students' academic achievements, instincts, willingness to take risks and accept challenges, character enhancements, mental and physical readiness, intellectual confidence; in short, almost all aspects of the student's educational needs. It can even be said that this responsibility has become unique by moving ahead of many educational components such as environment, parents, or even students. In other words, among the many personal-, family-, and school-related factors that have been considered as having an impact on student academic achievement, an effective teacher is regarded as one of the most important in school by many researchers (Sanders & Horn, 1998; Rockoff, 2004; Aaronson et al., 2007; Rivkin et al., 2005). Sanders and Horn (1998), for example, using multivariate longitudinal analyses, suggested that the teacher "effect" is the best predictor of student learning gain, rather than students' socio-economic status, class size, or heterogeneity.

Teachers, who have such an important role among educational components, constitute the agenda of policy makers in terms of increasing teachers' selection, training, and effectiveness. One of the major concerns of decision-makers is to ensure that effective teachers are recruited and retained in classrooms, and similarly, parents also desire their children to be enrolled in the best schools and taught by well-qualified teachers. Similarly, the UNESCO Sustainable Development Goal 4: Quality Education agenda has issued a call for member states to recruit 69 million qualified teachers to achieve universal primary and secondary education by 2030 through improved recruitment, retention, status, working conditions and motivation of teachers (target 4.c) (UNESCO Institute for Statistics, 2016). However, it seems that in most Western countries, teacher recruitment and retention has become increasingly difficult (Ovenden-Hope & Passy, 2020), and encouraging individuals to stay in teaching through financial incentives (a tpical practice of agencies) may only be a solution for a short time (Huat See et al., 2020). Some important factors such as self-efficacy and professional value that affect the retention of the teacher in classes or schools beyond financial reasons were mentioned (Ovenden-Hope et al., 2018). Therefore, for a long-term sustainable solution, a performance evaluation system that focuses on teacher development and training is needed to improve teachers' effectiveness or quality.

Evaluation can simply be defined as the process of analysing work performance based on predetermined criteria through different sources and mediums. However, since the speed and capacity of students to acquire their educational needs are not the same, in order to determine the performance of teachers in classes, tracking the changes in students' attainments can provide more reliable results in determining the performance of teachers in classes, rather than looking at students' learning at a single point in time. Therefore, when assessing teachers, it is important that the evaluation is comprehensive, fair, valid, reliable, provides incentive, and involves those evaluated in this interactive process. These factors also shape the success of the performance evaluation systems. It is believed that providing feedback to teachers through objectively measuring their performance can contribute to their development by providing opportunities for them to improve. This is reiterated in the OECD Teaching and Learning survey, which states:

“The appraisal and feedback they [teachers] receive are beneficial, fair and helpful for their development as teachers.” (OECD, 2009, p.139).

As often emphasized in the literature, training qualified teachers and ensuring the continuity of their professional development within the process are important factors in increasing the quality of the education system. From this point of view, new teacher performance evaluation systems, which bring a new dimension to inspection in education, have taken their place in many education systems across the world. However, not all of them can be successful in their intended purpose. For instance, Robertson-Kraf (2014) investigated the impact of a new teacher evaluation system that combined student achievement growth and observation in a school district in the US and found that the assessment system had a negative relationship with teachers’ expectations and even had no significant impact on teachers’ effectiveness and their decision not to quit teaching.

Turkey, like other countries, has also produced some policies on improving the teacher evaluation process. The first work on supervision, known as the Regulation on the Duties of the Primary Education Inspectors, came into effect in 1923 during the Republican period (Korkmaz & Ozdogan, 2005), and since then, it has undergone many changes. With the increase in school enrolment, fundamental changes were made to the Legislation Decree on the Organization and Duties of the Ministry of Education in 2014. Until this change, there was a dual structure in supervision whereby the Ministry of National Education organised one part, and the provincial directorates of national education the other (Ergen & Esiyok, 2017). One of the most important of these changes was the removal of this dual inspection structure, the aim of which was to centralize the inspection system. The other important change was the transfer of the responsibility of inspection by inspectors to the school principals.

Following this legislation, in 2015, the Ministry of Education introduced a teacher performance evaluation model just for candidate teachers (MoNE, 2015). To be a teacher in Turkey, graduates must pass a nationwide, annual exam after obtaining a professional teaching degree.

Then, the Ministry of Education appoints graduates as candidate teachers if they pass the exam. However, this evaluation model brought new conditions to the criteria of being a teacher that has been applied for many years. Since 2015, after working as a teacher candidate for a year, teachers are subjected to the teacher evaluation process in order to be tenured. Performance appraisal is carried out through the evaluation forms filled out by school principals. However, this teacher performance form has been criticised by researchers, educators, union officials and candidate teachers as being without objective criteria (Education and Science Workers' Union, 2016; Educators Trade Union, 2016; Turkish Education Union, 2017). Despite all these criticisms, the evaluation form was implemented for all teachers (not just for candidate teachers) in Turkey in June 2016. As a result, debates have flared up even more.

After a short time, the authorities were unable to remain indifferent to the outcry, and the existing teacher assessment system was revoked in 2017 after one year of use. However, while some studies to examine the multiple evaluation measures in the teacher performance evaluation system continued, no official performance evaluation work was carried out. Thus, this situation caused the teacher performance evaluation issue to become an open wound in Turkey. For this reason, new and comprehensive teacher performance evaluation models are urgently needed, such as those which are conducted in education systems in developed countries.

2.2 Teacher Effectiveness

Teachers' contribution to their students' learning differs significantly, and improving students' learning is a widely accepted characteristic of an effective teacher. There are several theories about what makes an effective teacher (e.g., Creemers 1994; Scheerens 1992). Researchers have tried to come up with a comprehensive list of what constitutes effective teaching and teachers. However, identifying what does so is challenging. While there is some consensus as to what such characteristics are (e.g., classroom climate and pedagogical skills), they are difficult to encapsulate because they are often also related to the kind of students in the class and its composition. For instance, the Hay McBer report (DfEE, 2000) suggests that teacher characteristics, teaching skills and classroom climate contribute as much as 30% of the variance in pupil progress. Others suggest behavioural management and pedagogical skills, such as making clear the learning objectives and making links explicit and assessment for learning as features of an effective teacher (Siraj-Blatchford & Taggart, 2014; Siraj-Blatchford et al., 2011). Coe et al. (2014) added content knowledge, teacher expectations, and professional behaviour to the list. The ability to inspire, motivate and enthuse are among other characteristics of an effective teacher (DfE, 2013; Sammons et al., 2014). In order to robustly measure teacher effectiveness, a comprehensive perspective on teacher effectiveness and constitution is needed. To meet this need, there are a number of methods that have been used throughout the past to measure teacher quality. Apart from the teacher performance assessment method based on the progress students make over years, which is the subject of this study and will be discussed later, a few of the most common methods, such as classroom observation, student evaluation, certification, and self-reporting, will be discussed in this section to lead to the definition of an effective teacher.

Classroom observation

Classroom observation is one of the most widely used methods of teacher performance evaluation and can provide more accurate signs of classroom practices when conducted by well-trained evaluators or observers. Various competencies and personal characteristics of the teacher are attempted to be evaluated through observation. In general, teachers are observed and evaluated in terms of their behaviour, lesson plans and preparations, teaching techniques applied in class, subject knowledge, teacher-student interaction, the ability to encourage student participation in lessons, considering individual differences, effective communication, classroom management, and other professional competencies. However, it is very important to be able to use valid and suitable tools to measure teacher performance through classroom observation in order to make an accurate assessment (Little et al., 2009).

Many studies conducted for this purpose have reached similar conclusions about the effectiveness and limitations of the observation method. For example, in a study involving 375 schools in the United States, which aimed to collect data on teacher evaluation practices, information about the tools used in the teacher evaluation process was collected from different regions. The research results showed that the observation method was the most used method in all schools; however, classroom visits take place once a year or less, and the study also revealed that there are school-to-school differences in evaluation forms used and procedures applied (Kowalski, 1978). In many countries, teacher evaluation work is determined by regulations, but explanations of observation conditions and procedures are inadequate. This evaluation method can be performed in different ways, such as live or video recording, while observations can be carried out by the school principal or an external evaluator such as an inspector. In addition, these evaluators may conduct classroom observations once or several times a year, depending on cost and context, either formally planned and announced or without prior notice. However, there is no sufficient explanation about the observation period of these visits, and what and how to observe. Assessment guidebooks often assume that teaching skills and other attributes are visible and evaluable in any teaching situation.

In many countries and also in Turkey, observations are the most important and widely used teacher evaluation methods in classrooms. However, it is a fact that evaluating teachers through observation has important limitations. For example, effective teaching cannot be defined independently of environmental conditions (Ko et al., 2016). While making classroom observations to evaluate the teachers, the constantly functioning structure of the teaching environment, such as daily routines, and the factors affecting the environment should also be taken into account. Current practices focusing on the classroom observation method should be followed in order to cope with the problem of continuity and endurance in the teaching environment. Additionally, the observed situation may not fully represent the teacher's teaching ability and classroom practices. Therefore, assuming that principals and inspectors, as evaluators, observe typical behaviours of teachers in the classroom may be misleading.

Moreover, classroom observations may also bias measures of teacher effectiveness since teachers are rarely randomly sorted into classes (Rothstein, 2009, 2010). To address this, Steinberg & Garrett (2016) randomised teachers to classes, and they found that teachers' performance, assessed using a teacher observation protocol, was strongly and positively related to students' prior attainment. They found that the quality of students in the class (indicated by their prior attainment) also strongly influenced teachers' interactions with them. Some studies (e.g., Stecher et al. 2018) suggest that observation instruments could positively predict student achievement gains (more so for maths than English). But this was when multiple observations by multiple observers were used. There were issues with reliability when scores were rated by one rater as would be the case when a teacher is assessed by the principal. There was high volatility in scores between observers and between lessons. To obtain a reliability of around 0.65 would require four observations, each by a different observer. Observers also need to be trained to score accurately, had no relationship with the teachers (hence not unconsciously bias), and observations were done via digital videos rather than in the actual classroom. These are all very controlled conditions that are rarely achieved in real-life situations. It is also possible that school leaders may be reluctant to give adverse reports to their teachers. In the Bill & Melinda Gates multimillion-dollar initiative to measure teacher effectiveness, school leaders were reluctant to give teachers a low rating, so few teachers were rated ineffective (Stecher et al., 2018).

All this suggests that caution has to be taken when implementing teacher evaluation based on classroom observations, and an assessment that reflects the teacher's true competence and effectiveness cannot be determined with a small number of observations (Garrett and Steinberg, 2015). Promoting, punishing, or rewarding the teacher based on such observations would not be fair, as they are not able to fully reflect the teacher's effectiveness.

Student evaluation

Another measure most commonly used in teacher performance assessment is student evaluation in the form of a questionnaire that asks students to rate teachers. In this form of measurement, students are direct sources of information about the classroom environment, such as the extent to which teaching activities are understandable/beneficial for them, the ability of teachers to motivate/encourage their students to learn, and the degree of communication between teachers and students. However, while it is considered that the information obtained by students' ratings of classroom experience can be valuable, student feedback is sometimes not seen as a reliable source of information due to the difficulties in determining the direction of causality. For example, students who do well in particular subjects and those who have a good rapport with their teachers are more likely to rate their teachers highly. In fact, due to the students' lack of knowledge of the teaching context, there is a great concern that students may evaluate teachers based on relationships with their teachers or their teachers' personality rather than the quality of the teaching activities. Worrell and Kuterbach (2001) found that student assessment results are generally reliable, but students at different grade levels were interested in different aspects of teaching. For instance, students in lower grades were affected more by the teacher-student relationship in teacher performance ratings, while students in higher grades tended to evaluate their teachers' performance regarding their own learning. For these reasons, the existing literature, in general, suggests that student ratings should never be used as a single measure in teacher performance evaluation; instead, such ratings can be combined with other measurement tools (Little et al., 2009).

Certification

Another measure utilised in measuring teacher performance is related to teacher training, specifically, certification. Based on the view that there is a relationship between teachers' qualifications and their classroom performance, the NCLB Act required all teachers in the

United States to be highly qualified (2002). According to this act, one of the ways to improve teacher quality is certification programmes, where teachers are provided with the necessary competencies. In other words, it is aimed for teachers to improve themselves by obtaining more education and increasing their performance.

However, there are some problems regarding the content of the certificate programmes that teachers attended. For example, it is not known whether such programmes specialised in a certain field or were designed to improve teachers' ability by increasing the quality of the teaching process in class. The notion of expecting too many certificates from teachers may lead them to attend short-term and accelerated programmes. However, since temporary and emergency licences are only valid for one or two years (and alternative certificates often turn into standard certificates within two years), it is unclear how much the efficiency of teachers can be increased during this period (Goldhaber and Brewer, 2000). For example, there are many certification programmes that teaching assistants must attend in order to work in a school, and as a result, these people are expected to be well equipped with teaching methods and techniques. However, studies show that many assistants need in-service training after completing their certification programmes and do not consider themselves fully qualified (Blatchford et al., 2007). However, of course, this does not show that such programmes are implemented completely randomly and are ineffective in improving the quality of teaching. The examples given emphasize that teacher performance cannot be evaluated solely on the number of certificates they have.

This research acknowledges that one of the ways to improve teacher quality is to participate in certification programmes. Thus, steps are taken to ensure that teachers have the competencies required for teaching. In fact, as a result of these methods, which are used in many countries to evaluate the performance of teachers, teachers who are considered competent are rewarded with an appropriate increase in their salaries. Paying teachers more may affect their motivation and likely improve students' learning. Within the scope of teacher incentive programmes based on this perspective, certificate programmes lead to salary increases for teachers, as in the United States (National Board for Professional Teaching Standards (NBPTS), 1987). Although there is a perception that this situation will increase teacher performance and effectiveness, according to Loeb and Page (2000), there is no definite relationship between increases in the salaries of teachers and improvements in the academic success of students. However, a relationship can be established between increased salary and teacher motivation, which may indirectly increase teachers' desire for more certification through more training. Although Hawk et al. (1985) stated that student achievement was positively affected by teachers with certificates, it is unclear whether the teachers' salary led to an improvement in student attainment (Loeb & Page 2000). Therefore, it is unclear whether the teacher certification program accompanied by a salary increase is a valuable endeavour.

As a result, all the limitations explained above show that the certificates alone are insufficient in evaluating teacher performance, so additional assessment methods are needed.

Self-reporting

Another method that researchers have focused on in the evaluation of teacher performance is self-reporting. This is an evaluation method where teachers report what they do in the classroom through large-scale surveys, instructional logs, or interview reports. In this method, teachers have the opportunity to evaluate their own performances. However, like the observation method, this may cause the performance of the teacher to be viewed from a very broad perspective or, in contrast, to miss the whole picture of teaching by focusing on certain subjects. In particular, the fact that instructional logs and some interviews are highly structured may cause teachers to evaluate their performance very strictly, which also results in some important aspects of teacher performance being overlooked. On the other hand, open-ended questions give teachers the opportunity to explain their teaching activities with why and how questions (Ball & Rowan, 2004). Although self-reporting is a method that may be preferred in terms of its ease of use, providing teachers with a wide range of possibilities to express their feelings, thoughts, knowledge, and beliefs, it has some limitations in terms of validity, reliability, and bias. The fact that teachers misreport their activities to exclude their shortcomings may limit the quality of the evaluation process. For this reason, researchers recommend using surveys, instructional logs, and interview methods together in cases where self-reporting is used (Camburn & Barnes, 2004; Moorman & Podsakoff, 1992). Thus, while this method enables the evaluation of teacher performance in multiple dimensions, it also enables teachers to evaluate their performances with a critical perspective.

As can be understood from the limitations of all the methods mentioned above and the concerns of the researchers, while it is accepted that evaluating teacher performance is beneficial in enhancing teacher development and student outcomes, it is a complex process, and there is no perfect measure (See, 2020). Academic achievement is one of the primary goals of education; however, student test scores alone, as an indicator of academic success, do not provide comprehensive information on teachers' classroom activities and how students perceive these activities. Therefore, combining multiple measures can provide a wide range of information about teacher performance in the classrooms.

2.2.1 Defining Effective Teacher

In the Race to the Top competitive grant programme (2009), an effective teacher was described as a teacher whose students showed at least one grade level improvement in student growth throughout an academic year. In this respect, teacher effectiveness is generally defined as a teacher's ability to make improvement in students' learning, typically measured by achievement tests (Burgess, 2019; Little et al., 2009; Goe, 2007). This definition is also the concept behind value-added models (VAMs), where "effectiveness" refers to the estimation of the differences between expected and observed student test scores (Sanders et al., 1997; Kersting et al., 2013), as will be explained in the next section.

Although this definition refers to an important role of an effective teacher, it is quite a narrow one because teachers play many roles in school, from planning lessons to classroom management, from motivating students to inspiring and encouraging critical thinking; supporting student learning is just one of them. For this reason, associating teacher effectiveness only with their contribution to test scores remains a very shallow measure compared to the scope of actual teacher effectiveness. Therefore, it is not easy to create a single definition of an effective teacher. However, based on all these criticisms and assumptions, by an effective teacher, I mean one who has the ability to plan the teaching process in line with educational goals, has pedagogical content knowledge, communicates with students about educational goals and student expectations, is able to enhance student learning outcomes by providing additional supports in accordance with the needs of the students, and plays a guiding role in this whole process. The focus in this definition is that the student and the teacher play a role in the education process together because a teacher can only be 'effective' if students are willing or able to learn. Although teacher effectiveness is associated with the characteristics and qualities of teachers (Walker, 2008; Stronge, 2018), the most widely used measure of an effective teacher is student academic growth because it is tangible and quantifiable, which is the focus of the rest of the thesis.

2.3 Conceptualising "Teacher Effectiveness" in VAMs

Teachers' contribution to the academic achievements of their students is an undeniable fact. One of the ways to show teacher contribution given in the literature is the analysis of student achievement data. With the proliferation of longitudinal data at the student level, researchers have focused on identifying teachers who were successful in making an above-average contribution to student achievement. Student achievement test scores are used to evaluate teacher contribution based on the level of student achievement growth over at least two consecutive years, called value-added measures.

Value-added models (VAMs) attempt to measure a teacher's effect on his or her students' achievement. This involves using a variety of measures to predict each student's test score and then comparing these predicted scores to how the teacher's students actually scored on the test.

In this measure, how effective a teacher is in improving student learning is estimated by predicting how their students would have done by controlling their previous attainments and some characteristics, such as sex and socioeconomic status (SES), and comparing these predictions with how they actually performed. More specifically, the predicted score obtained by controlling for the student attainment to some degree is subtracted from the actual score, and then student-level differences derived are aggregated at the teacher level. The means of the differences at the teacher level are then attributed to a teacher's value-added effectiveness scores. In the VAM concept, this difference between the predicted and actual performance of the same students is called "teacher effectiveness". However, since "effectiveness" inherently refers to causality and the design of this study is not suitable to reveal this causality, instead of using the term "teacher effectiveness", I prefer to use "effectiveness score" or "value-added score" throughout the thesis.

In VAM estimates, teachers whose students' actual performance is better than expected are considered more effective than those whose students score lower than expected in tests. However, the use of pupils' test scores is not without problems, as measures are not perfect. As Gorard (2018a) pointed out, a very high proportion of the pupil gain scores in England is the result of error propagation due to missing data, measurement errors, and representation errors. Consequently, VAM estimates can be volatile and highly sensitive to the kinds of data used and the level of aggregation.

When looking at the VAM estimates from another perspective, the tautology element stands out. Accordingly, teachers whose students do well are considered effective teachers, and effective teachers are those whose students perform well. In other words, the concept of VAMs seems to contain within itself a circularity. Moreover, the link between a good student and an effective teacher also raises some concerns arising from the use of student data in VAM estimates. On the other hand, in addition to Gorard's (2018a) emphasising the concerns about errors arising from student achievement scores, Haertel (2013) also stated that students' test results do not fully reflect the contribution teachers make to students' learning. However, student achievement scores are not the only concern with VAM estimates. For instance, some concerns such as bias, reliability, and validity arising from estimations other than student data have also been discussed by researchers (detailed discussion on these concerns will be provided in the next chapter).

VAMs are currently widely used approaches in accountability systems to evaluate teacher effectiveness, which take into account other contributing factors, such as student backgrounds or prior performance. As these approaches allow such factors beyond the teacher's control to be controlled, any progress made by the students can be attributed to the teacher, and this makes them useful in the evaluation of teachers. Prior to VAMs, teachers were evaluated based on how their students performed by simply measuring how much progress students made. These simple models are known as growth models. The next chapter will provide information about the different types of growth models, including VAMs.

CHAPTER 3
THE CONCEPT OF GROWTH MODELS

The chapter provides a summary of growth models used in educational accountability systems. There are different types of growth models, and the most commons are discussed in this chapter. Value-added models (VAMs) are a kind of growth model and also considered an adjusted form of growth models. General information about VAMs, such as the fundamental principle underlying them, working principles, and essential characteristics, is given in this chapter. In the last section, some ongoing concerns arising from the use of VAMs to measure teacher performance are also discussed.

3.1 Growth Models in Educational Accountability Systems

Assigning effective teachers to classrooms is one of the most important educational issues for policymakers. Since the close relationship between teacher effectiveness and student achievement is a striking issue in education, the issue of determining whether the teachers assigned to classrooms are effective or not has gained importance. With longitudinal student achievement data being easily reachable, researchers have started to use approaches based on the use of students' academic achievement growth – growth models – as an indicator of teacher effectiveness. Before describing such models as used in educational accountability systems, it is essential to explain the expression 'status' in terms of improving conceptual integrity. Status can be defined as a single snapshot of student academic performance (Castellano & Ho, 2013a). This means that a student's academic attainment is based on their performance at a single point in time. One main limitation of a status model is that it is unable to determine the progress made over time by the student. Nevertheless, it is the most widely used method in measuring student attainment by educators, as it is simple to implement, and no sophisticated analytical skills are required in interpreting the results.

This shortcoming of the status model has led researchers to focus on alternative measures, such as growth models. Growth models are an improvement over status models as they look at the change in the academic performance of a student or group of students by measuring the same student academic attainments over two or more different time points.

It is important when describing 'growth' models to distinguish between 'growth' and 'improvement'. 'Growth' measures the change in performance of the same individual over time, while 'improvement' assesses the difference in the performance of different students or groups of students over time. The cross-cohort model, for example, is technically an ‘improvement’ model (Blackorby et al., 2016), although it is often referred to as a type of growth model by some researchers (O’Malley et al., 2011) as it only registers the growth or change in performance of one cohort in relation to another, and the cohorts do not necessarily contain the same individuals. Cross-cohort models are widely used in school accountability systems; for example, in comparing the percentage of those reaching the proficiency level in Grade 7 in 2014 and the percentage of those reaching the proficiency level in Grade 7 in 2017.

Table 3.1 An Illustration of a Cross-Cohort Model

		Grade 5*	**Grade 6***	**Grade 7***	**Grade 8***
Year	**2014**	65	75	74	69
	2015	69	50	70	65
	2016	55	61	55	70
	2017	74	69	60	60

*Average score of the relevant grade level out of 100

Table 3.1 illustrates what a cross-cohort model looks like in practice. The vertically highlighted cells represent the performance of the same grade over the years. The improvement chart can also be used horizontally as within-year improvement across grades. The improvement models are generally used for school accountability purposes. The main shortcoming of the cross-cohort model is that it is highly affected by the student profile of the school, as it does not allow the achievement of the same students to be tracked. Changes in student intake, therefore, can have a noteworthy link to the school’s performance. For instance, if the school enrolment rate shifts from high-performing students to low-performing students the following year, it will look like the school has not made any progress. As the model does not control for student background factors, such as socio-economic status (SES) and parental involvement, any change in the school performance may be due to these factors, which are beyond the control of the school.

Growth models, on the other hand, take into account the growth made by the same student or group of students over time. This requires longitudinal data to allow the same student’s achievements to be tracked over a number of years. Table 3.2 is an illustration of what a growth model looks like.

Table 3.2 Example of Growth Over Years

		Grade 5*	Grade 6*	Grade 7*	Grade 8*
Year	**2014**	65	75	74	69
	2015	69	50	70	65
	2016	55	61	55	70
	2017	74	69	60	60

*Average score of the relevant grade level out of 100

The shaded cells across the diagonal in Table 3.2 represent the performance of the same group of students and their class averages from 2014 to 2017. They show changes in the achievement of the same students over the years.

This kind of model is seen as fairer, unlike status and improvement models, which compare the current performance of a student and/or group of students at a single point in time to a threshold value for the proficiency level determined by the school, district, or state, growth models consider how the same student/group of students perform over time. Since students in different cohorts may differ in terms of their characteristics, comparing different cohorts over time against a pre-determined criterion of performance may appear to be 'equitable' for everyone in the same grade, but it is not necessarily a 'fair' system as all individuals begin from a different starting point. Figure 3.1 illustrates the difference between equality and equity (justice, fairness).

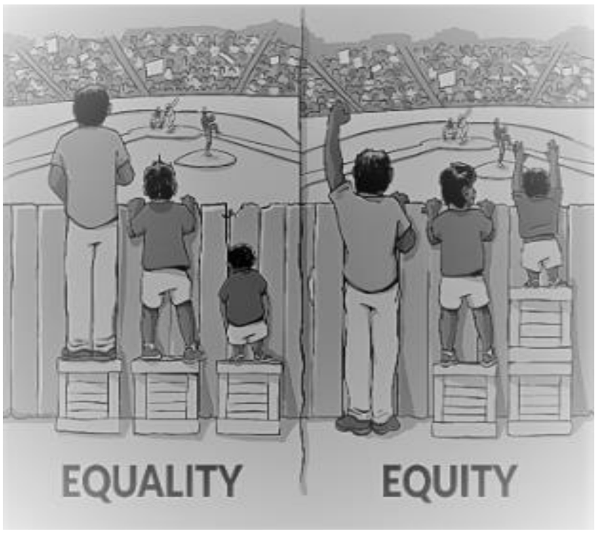


Figure 3.1 Equality and equity

Source 1 : https://artplusmarketing.com/equality-equity-freedom-55a1d675b5d8

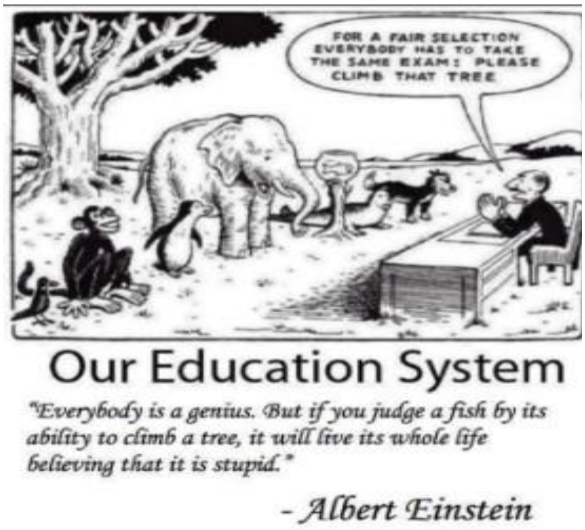

FOR A FAIR SELECTION EVERYBODY HAS TO TAKE THE SAME EXAM: PLEASE CLIMB THAT TREE
Our Education System
"Everybody is a genius. But if you judge a fish by its ability to climb a tree, it will live its whole life believing that it is stupid."
- Albert Einstein

Figure 3.2 Our education system

Source 2 : https://i.pinimg.com/originals/19/59/85/195985660924652b3c007a764e78ce81.jpg

Equality is not fair in each circumstance. Figure 3.2 illustrates that all of the animals are expected to compete by climbing the tree, although the fish and elephant are not naturally equipped to do so, while the monkey is naturally able to climb a tree very easily; so, although the task is the same for all of the animals, the test is obviously unfair as they do not have an equal chance of passing it. In order to bring equity to the education system, instead of expecting the same success from everyone, each individual should be evaluated individually based on the goals they can achieve. For instance, while a monkey is expected to be able to climb a tree within a certain period of time, the fish is expected to swim from one place to another instead of climbing a tree.

To assign personal goals for each individual, first of all, a tracking system should be constituted that allows the academic progress of students to be monitored. The monitoring process can be carried out through growth models, which no educators oppose, but the main issue is that the educators cannot come to an agreement with each other over how to assign an appropriate benchmark for each individual. The answer to this issue is also important for this current research because a teacher's effectiveness is related to the number of students who meet their proficiency level in his/her classroom. The different approaches currently in use for teacher effectiveness estimates will be discussed in more detail in the following sections.

Before explaining growth models used in educational accountability systems, the confusions in terms used in the literature will be discussed in order to clarify the context of this study. Having a clear understanding of what terms are in use is one of the prerequisites for research in a new field. Conducting research in a field where many different terms are used, but most of which have the same/similar meaning increases the difficulty of carrying out research. The field of research related to growth models, especially VAMs, is one where there is such confusion because different names given to the same models by some researcher. For instance, Castellano and Ho (2013a) mentioned seven different terms used in the literature as aliases, variants, and close extensions of the gain model, such as *growth relative to self, raw gain, simple gain, slope, average* gain, *gains/slopes-as-outcomes*, and *trajectory model*. There are some comprehensive studies in the literature in which the answer to this labelling confusion can be found (Ligon, 2008; Castellano & Ho, 2013a).

The other confusion mentioned in this section is about the grouping of growth models; researchers group the models with regard to their own criteria. For instance, Ligon (2008) stated in the second part of the study called Growth Model Series, "There are only a few really different approaches to growth models, but many different formulas for calculating them. If we understand which question each model answers, then making a selection among them will be easier," and grouped the models according to their capacities to answer the prospective questions. The author also specified which models are used by different researchers to find out the answers to similar questions. In another grouping study, O'Malley et al. (2011) investigated fifteen states' accountability systems and classified all growth models into three groups: *growth to proficiency,* in which students' previous performance is taken into account to provide a yearly growth target for the student; *value/transition,* in which the changes in student performance is evaluated based on performance categories over two years; and *projection,* in which students' future performance can be predicted by using the students' previous and current performance and that of prior cohorts who have had similar performance scores in the past. Perhaps the most comprehensive study in the literature was carried out by Castellano and Ho (2013a). As can be understood from its name, the study, '*A Practitioner's Guide to Growth Models',* takes on a guiding role for those who work in this field. To create categories for the models, the authors considered a variety of criteria, such as models' *primary interpretation* (for growth description, growth prediction or value-added) and their *statistical foundation* (gain-based models, conditional status models, multivariate models), etc. While mainly benefiting from Castellano and Ho (2013a) in the grouping of growth methods, a new grouping was created, taking into account the common aspects of existing growth models. In the following section, some common growth models will be explained in detail.

3.1.1 Progress-Monitoring Model

The first model presented here to measure students' academic growth is the progress-monitoring model (PM). PM embodies successive data collection procedures to determine the extent to which the curriculum and instructions applied in the classroom serve to achieve educational goals (AASEP, n.d.). The use of PM is not usually preferred in educational accountability systems (Blackorby et al., 2016) as it is used for monitoring or diagnostics rather than evaluation. It is useful in determining which student(s) need additional help at an individual student level and to what extent the teaching activities used in the classroom benefit the students at the classroom level. It is commonly used in response to intervention (RTI) approach, which is a method used by educators to determine those students who are not responsive to the teaching procedures used and thus require more intensive support (King et al., 2012).

As the primary use of PM is for diagnostics or monitoring, data about the student(s) is collected regularly (weekly, bi-weekly, monthly). Although other growth models compare students' test scores over at least two school years, PM's data collection procedure is completed within the same school year. This allows issues to be identified early so that appropriate interventions can be taken.

The primary advantage of PM over other growth models is that as data is collected in the same school year, the change in a student's or group of students' performance can be detected immediately, and then if there is a situation that requires taking action, PM allows it to be done more quickly. Therefore, students who need help in the classroom can be determined without delay, and after the appropriate additional interventions are applied to them, their progress can be evaluated regarding the determined year-end targets (Jenkins et al., 2013). The other advantage of PMs is that as the data is collected frequently, such as once a week, any change in a student's performance can be attributed to their teacher's action. For this reason, PMs are often considered sensitive models that are highly influenced by current instructional interventions, whereby teachers can discover whether their recent interventions work well enough to make the desired changes in students' success (National Center on Response to Intervention, 2010).

Lastly, with increasing computer-based commercial publishers, such as DIBELS, AIMSWEb, EasyCBM, or FAST, the data can be collected online very quickly and be instantly visualized, which makes interpretation easier even by non-technical individuals, such as teachers. On the other hand, besides the advantages of PM, there are also some disadvantages. PM requires a set of data collection procedures in the same school year. The frequency of data collection can sometimes be weekly. This data collection frequency makes the implementation of PM difficult, even when the data is collected online, so teachers might tend to track the learning growth of only selected students who need additional help instead of the whole class. Secondly, a shortage of PM is related to implementing the content area. PM can be implemented in a very limited teaching field, such as reading (McCardle et al., 2001; Fuchs & Fuchs, 2011; January et al., 2018) for students with special needs (Legere and Conca, 2010; Tichá et al., 2009; Denton, 2012). Although the majority of research has been done in the field of reading, there has been done some research in the teaching field of algebra (Foegen, 2008; Foegen & Morrison, 2010).

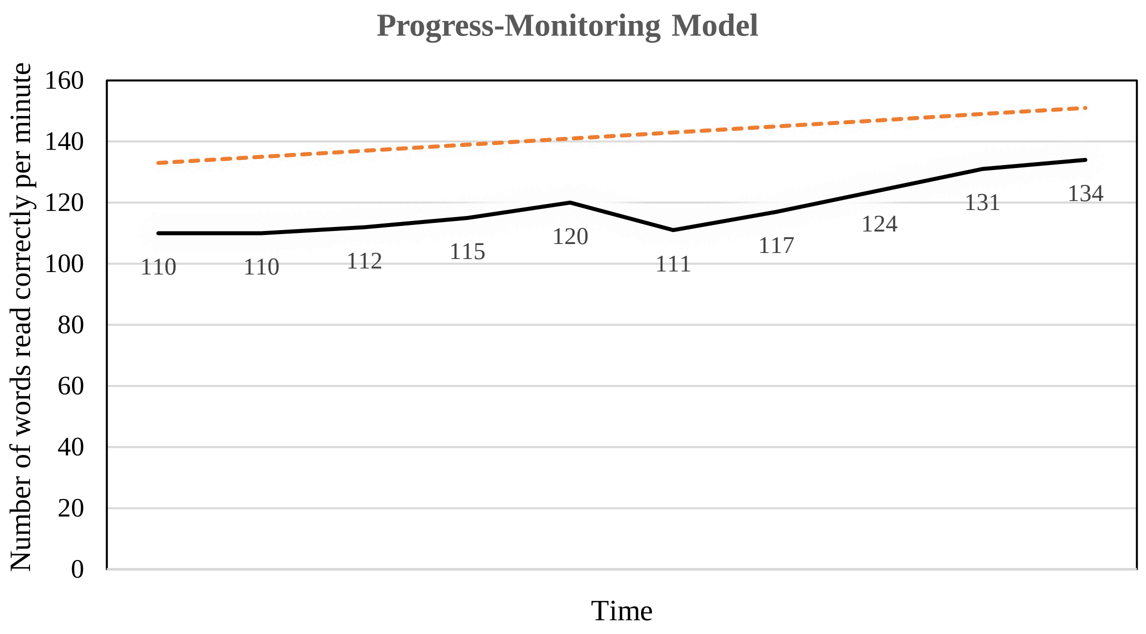


Figure 3.3. The progress-monitoring for the correct number of words read per minute

Figure 3.3 shows how an individual student's reading performance can be monitored over time. Similar progress-monitoring graphs can be generated at the classroom and school level also. The dotted line shows the target or predetermined benchmark for each measurement point, and the solid line shows the student's actual scores relative to that target. This comparison allows educators to determine whether the student is meeting the target and if more intensive intervention is required.

3.1.2 Simple Gain and Trajectory Model

The simple gain and trajectory model is also known in the literature as *growth relative to self, raw gain, gain score, gain/slope-as-outcomes, growth-to-standards* – terms that have nuanced differences from each other. This model has a very simple approach to determine the growth in the achievement of the student or group of students. As the name suggests, the simple gain model is a growth model that requires simple mathematical operations. As the simple gain can be computed as the difference between two time points, past and current performance, the growth calculated serves the descriptive purpose (Castellano & Ho, 2013a). The simple gain model requires vertically scaled test scores where the data are collected from two grade levels. If we assume that the rate of growth of a student is linear, that is, the student improves at the same rate now as in the future, we can predict his performance three years later (the trajectory model). If the trajectory model is considered a two-step ladder, the simple gain model represents the first step of the ladder. For instance, if the student gains a score of +2 in maths (he achieved 63 in maths last year and 65 this year), the student's trajectory score three years later can be estimated.

The biggest advantage of using simple gain and trajectory models over other growth models is that they are simple and require only basic mathematical skills; the results are easy to understand by non-technical educators. The direction and magnitude of the student's gain score indicate whether the student is making progress or not and by how much. In the trajectory model, as the student's future gains can be estimated, the educators can clearly understand which students are on their trajectory line. Another advantage of this type of model is that models do not require a comprehensive dataset, such as students' background characteristic data and teacher-level data; one only needs two test scores collected at different times for the same students. But this simplicity is also a limitation. The requirement of the desired scores on a common scale can be considered as another disadvantage in the use of these models. They are not ideal for accountability purposes because higher and lower scoring students are likely to make different rates of progress, so they cannot be used to evaluate teacher effectiveness. Other student contextual factors must be taken into account to explain the different rates of growth for each student. Another shortcoming of this model is the assumption that growth over time is linear. This is an oversimplistic assumption, unlikely to happen in real life.

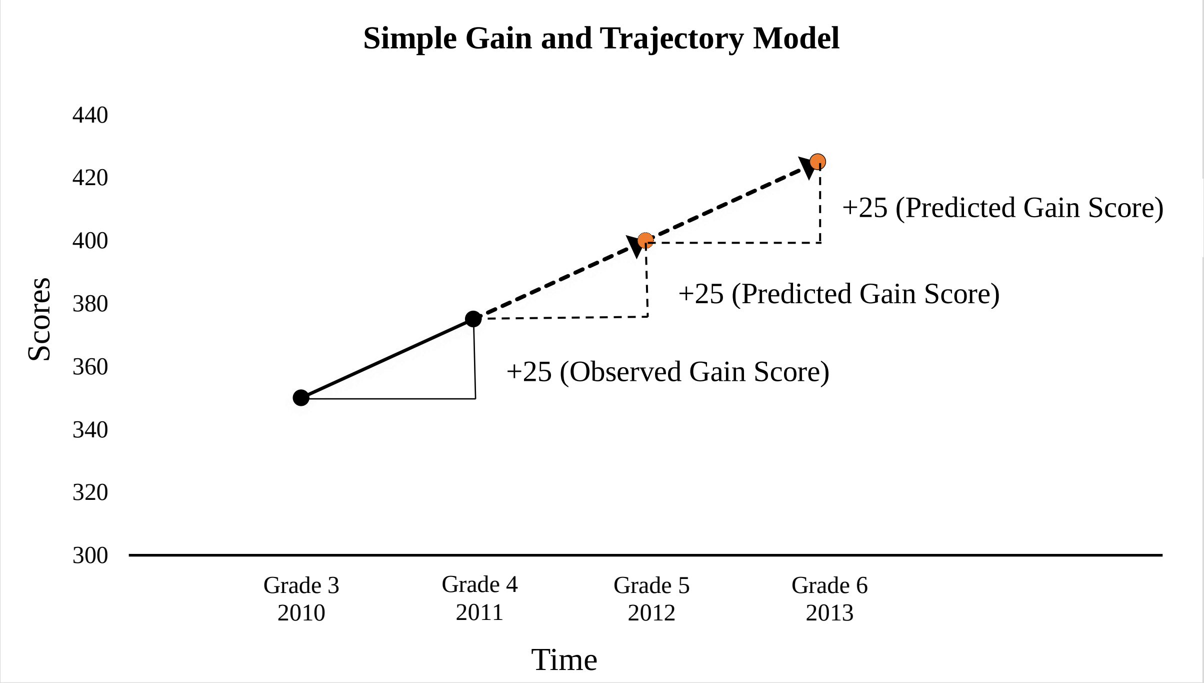


Figure 3.4 The simple gain and trajectory model for prediction of future student maths performance

(Source: A Practitioner's Guide to Growth Models by Castellano & Ho, 2013a)

Figure 3.4 illustrates the simple gain model and trajectory model's basic calculation using the maths scores of a student in Grade 3 through to Grade 6. While the horizontal axis represents the grades of the student over time, on the vertical axis, the student's maths scores throughout these grades have been shown. As required by the single gain and trajectory models, the scores in the different grades must be converted to a common scale. The solid black dots represent the student's maths test scores in Grade 3 and Grade 4. The actual (observed) gain score is calculated by the vertical change in the student's test scores, which is 375−350= +25. Assuming linear growth, the student can be expected to make a +25 gain in their score each year and thus, their performance two years later can be predicted.

3.1.3 Categorical Model

The categorical model can also be considered as a kind of simple growth model that shows the transition of students from one category to another by using a transition table consisting of columns and rows over two successive years. These columns and rows contain information about the position of students in a certain year who were in a particular proficiency category level last year. It is similar to the simple gain or trajectory model in that it assumes that students' progress is linear. The only difference is that instead of discrete gain scores, progress is measured in categories. The transition tables show the percentage or the total number of students who remained at the same performance level for two consecutive years or who moved one or more performance level(s) up or down. The model, instead of tracking individual student growth, is generally applied to determine the effectiveness of teaching programmes or educational organizations, such as school performance evaluations in an accountability system (Blackorby et al., 2016). The categorical model has recently been used for teacher performance evaluation, taking into account the average value points earned by the students in a particular teacher's classroom (Dwyer, 2016).

The categorical model, in practice, covers two very similar approaches: the transition (matrix) model and the value table. Although these approaches are used interchangeably in the literature, the value table is actually an extended form of the transition model. The main advantage of the categorical model over some growth models is that it does not require complex statistical estimates and can be easily understood by non-technical educators. On the other hand, the categorical model also has shortcomings. As the vertical scores are grouped into categories, some information that could be crucial will be lost. For instance, two students were in the same category last year, but although one's score was just over the bottom threshold of the category, the other's score was just below its top cut score; these two students made progress and moved one category up. In this example, just looking at the result of both students moving one category up would be misleading. Although the two students made positive progress over two years, the amount of progress is different, as the progress of the student who was just above the bottom threshold is much greater than the other. Therefore, although the scores are not necessarily required to be on the same vertical scale, the vertical scale is needed in the interpretation. To address this problem, sometimes sub-categories are created so that more realistic interpretations can be made. Another way is to assign a cut-off score and value points for each category. However, this relies on human judgement (Buzick and Laitusis, 2010) and would require careful work by a group of experts who know the educational system well and the characteristics of the data to be used.

Table 3.3 is used as an illustration of how this works. For instance, 22 per cent of students (n= 19) were in the "developing" level last year, but by moving up two levels are in the "distinguished" level this year, and the value-added score of the teacher can be estimated by averaging the multiplications of the number of students and the value score at each achievement level.

Table 3.3 An Example of a Categorical Model

			Post Test Levels			
			Basic	Developing	Proficient	Distinguished
Pre-Test Levels	Basic	N	16	20	15	9
		Percentage	27	33	25	15
		Value Score	**10**	**50**	**90**	**130**
	Developing	N	21	25	20	19
		Percentage	25	29	24	22
		Value Score	**-50**	**20**	**90**	**140**
	Proficient	N	10	22	29	20
		Percentage	12	27	36	25
		Value Score	**-100**	**-30**	**40**	**110**
	Distinguished	N	5	11	21	25
		Percentage	8	18	34	40
		Value Score	**-150**	**-50**	**50**	**150**

3.1.4 Residual Gain Model

While the simple gain model and the trajectory model estimate the absolute amount of gain, the residual gain model looks at the relative gain in achievement. Using the linear regression method, this model estimates the degree of resemblance between the observed performance in a given year and the predicted performance based on the previous outcome(s) (Blackorby et al., 2016). The residual is the difference between the observed and expected values. The sign (positive or negative) and magnitude of the residual indicate whether the student's current performance is above, similar to, or below the expected performance based on their previous score(s). This model describes a student's current growth status by regressing his/her current score on previous score(s), not by predicting forthcoming growth (Castellano & Ho, 2013a). Although the linear regression method allows for the consideration of the influence of other demographic predictors, such as sex of student, prior performance, and teacher characteristics, the residual gain model, generally, includes only one or two lagged test scores. Based on the assumption that there is a strong and linear relationship between past and current scores, linear regression gives expected values of each baseline score by creating a linear regression line (trend-line) with the smallest vertical distance (Castellano & Ho, 2013a).

Instead of comparing a student with all other students in the dataset, by controlling previous test score(s), the residual gain model allows the student to be compared with students whose previous scores are similar so that a more realistic trajectory can be created for that student. Together with being difficult to understand by non-technical people, the model is not useful for estimating overall gains made by the entire group, as the mean of the residuals in the dataset is always equal to zero.

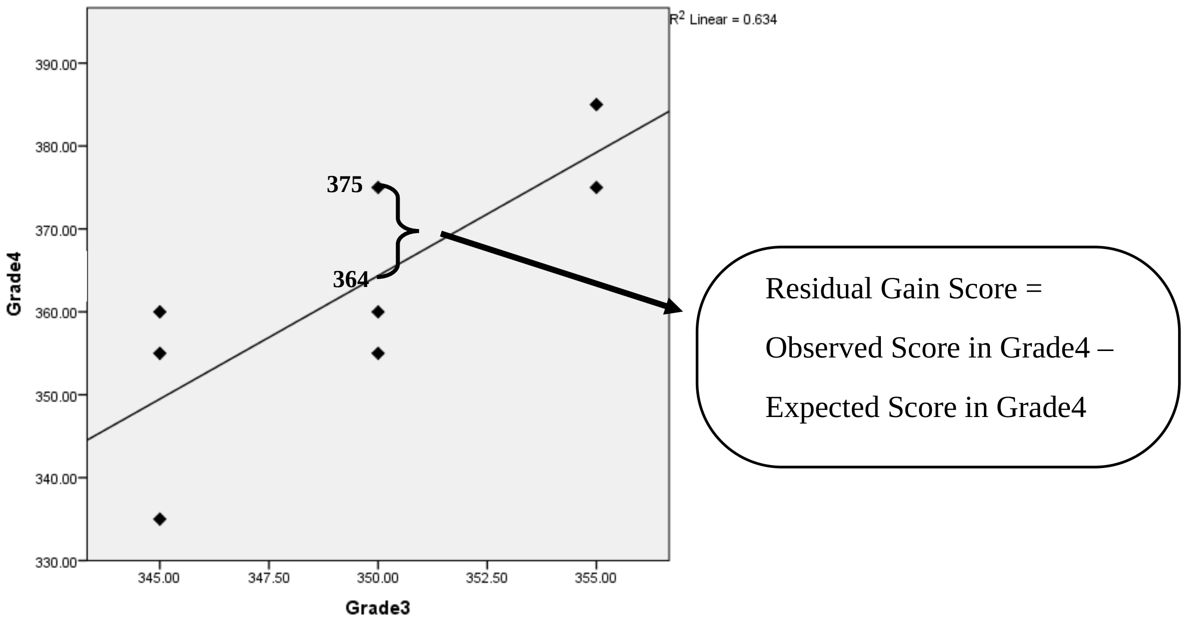


Figure 3.5 A visualisation of the residual gain model

Source: A Practitioner's Guide to Growth Models (Castellano & Ho, 2013a)

Figure 3.5 illustrates the residual gain model with the horizontal axis indicating test scores in Grade 3, and the vertical axis the student's test scores in Grade 4. The diamond shapes on the chart illustrate students' test scores obtained in consecutive grades. In addition, the linear regression line – the solid line shown on the graph, also called the trendline – also represents the line that best fits all the students' observed test scores in Grades 3 and 4. For instance, according to the trend line, all students whose scores were 350 in third grade in the data set are expected to earn a score of 364 in fourth grade, so the residual gain score of the student chosen on the chart can be found by subtracting the expected test score of 364 in fourth grade from his/her observed test score of 375 in the same grade. Here, that is +11, which means that the student performed 11 points above expectation. Most importantly, the trendline also depicts what the students' overall performance might be by regressing test scores in Grade 4 on test scores in Grade 3 in the whole dataset. It is worth noting that growth models are simple, unadjusted versions of VAMs; the residual gain model used for value-added purposes, such as for teacher effectiveness, is called the covariate-adjusted model.

3.1.5 Projection Model

The projection model, also called the prediction model, has a similar approach to the residual gain model. As the name suggests, student's future performance can be predicted by conducting linear regression in the model. While the residual gain model describes a student's current growth, the projection model predicts the student's future growth. To make a regression line, longitudinal achievement data from a past cohort of students who have already completed the target grade is used in the projection model, and then the equation of this regression line is used for the performance data of a current cohort of students (Castellano & Ho, 2013a). The projection model is similar to the trajectory model in terms of extrapolating a student's future score; however, although the trajectory model predicts a student's future score by taking into account their own past performance, the projection model estimates students' future performance by an equation of the regression line obtained using a past cohort of students' longitudinal performance data (Blackorby et al., 2016).

To improve the accuracy of the regression equation, along with the additional years of test scores, background characteristics of past student cohorts, such as gender, ethnicity, age, etc., can be added to the equation (Blackorby et al., 2016). Moreover, as the model is used to predict current students' future scores, the production results may also be used for the purpose of taking precautionary measures (Goldschmidt et al., 2005). It can be predicted from about a year in advance which students will need more additional support, thereby helping them reach the proficiency level. On the other hand, as the projection model is based on the assumption that consecutive cohorts have similar school experiences in the course of time, if the population of the school or the content of teaching tested has dramatically changed over time, the prediction may no longer be accurate (Blackorby et al., 2016).

To visualize the operational procedure of a prediction model, the same equation of the regression line estimated in the residual gain model is accommodated in the projection model in Figure 3.6.

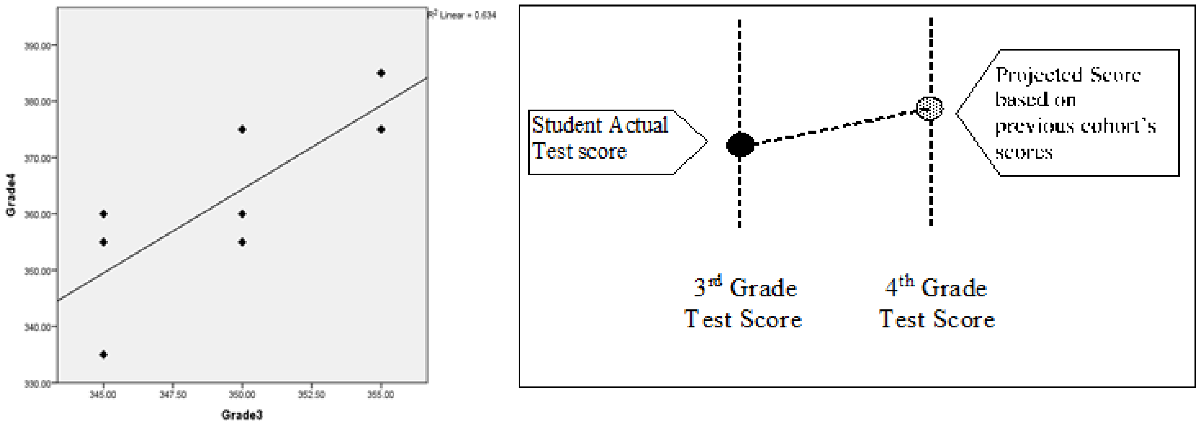


Figure 3.6 A visualisation of the projection model

Source: A Practitioner's Guide to Growth Models (Castellano & Ho, 2013a)

The scatter chart on the left was created using the test scores of a group of students who had already completed Grade 4 in the previous section. Thanks to the regression line obtained, any possible scores in Grade 3 can be input into the equation to predict the same cohort of students' scores in Grade 4 (Castellano & Ho, 2013a). The figure on the right-hand side of Figure 3.6 represents information of a student from the current cohort. The solid bubble on the line above the third-grade test score symbolizes the student's actual score earned in the current year. The dashed bubble on the line to the right represents the student's projected score for next year by using the same equation of the regression line on the left. Any third-grade scores of the current cohort can be inserted into the equation to predict future test scores in fourth grade. If the predicted score of each student or any other cut-off point is assigned as the threshold score, a teacher's effectiveness can be considered in parallel with the number of their students who met or exceeded the target points. If the problem of missing data can be overcome, to increase the accuracy of the predictions, along with the inclusion of more previous years' test scores, other predictors that represent students' language learner status and special education status, as in the Delaware Student Growth Model in the USA, can be included in the regression equation.

3.1.6 Student Growth Percentile Model

The student growth percentile (SGP) is also known as the Colorado growth model in the literature. SGP is a model that is based on the normative measure of student growth. It determines the growth a student has made in a year by comparing the student to their academic peers who have had a similar achievement history (National Center for Education Statistics, 2012). While the attainment score provides information about whether the student is meeting expectations at a single point in time, SGP reflects the academic achievement growth of the student from year to year. Like a Paediatric Growth Preference Chart commonly used by doctors to inform parents about their child's current weight and height in percentiles by comparing them to other children of the same age, academic growth is also expressed as a percentile in SGP. For instance, a mean of 75 SGP means that the student's performance is the same or better than 75 per cent of their academic peers. SGP has been used for many purposes in educational accountability systems, such as school accountability, teacher effectiveness, and instructional improvement. Along with the similarities to the residual gain model, which are to predict students' current scores by using their previous scores as predictors and be used for the description of the students' current growth, there is also a fundamental difference. This is that rather than drawing a single best fit line, SGP fits 99 lines, one for each percentile from 1 to 99, by applying the quantile regression model (Castellano & Ho, 2013a). Moreover, apart from its use to explain current growth, SGP is also conducted to predict growth by combining aspects of the trajectory model and the projection model (Castellano & Ho, 2013a). SGP assumes that the growth will continue at the same rate in the future as in the trajectory model and uses the regression equation created using the scores of a cohort that already has future scores, as in the projection model.

The remarkably dominant characteristics of SGP over some growth models are because the results are easier to understand for non-technical people (Blackorby et al., 2016). As the results are expressed with a readily explainable metric in SGP, the interpretation of the results is relatively simpler; an SGP score of 60 means that a student's performance is better than 60 per cent of their cohort's performances (National Center for Education Statistics, 2012). In contrast to the other regression-based models – which are very strict in order to meet the assumption of there being a linear relationship between the predictors and outcome, and which have equal variance in current scores across initial scores – even though SGP involves more complicated regression analysis, it also has a more flexible statistical structure that embraces these requirements (Castellano & Ho, 2013a). On the other hand, the large sample size requirement stands out as a remarkable disadvantage of conducting SGP. According to Castellano and Ho (2013b), although the sample size requirement depends on model-supported inferences, the general guideline for the minimum sample size for SGP estimation is 5000. It is not a big challenge for states to estimate SGP scores for teachers or schools, but a large sample size requirement might not be comfortable for the researchers. Recently, in research by Culbertson (2016) to investigate the accuracy of SGP estimates, when the SGP estimates are based on a small sample size, the researcher reached the conclusion that the SGP scores of high- and low-achieving students are more affected by small sample size than students with average success.

Figure 3.7 depicts the heuristic approach of SGP in the educational accountability system. As students at different academic levels are in the same class, to compare a student only to other students in the same class would not make sense in the accountability system, so the student should be compared to other students such as those in the school district, state, or whole country who have similar previous test scores in the area of interest. That is also the common principle for all regression-based growth models.

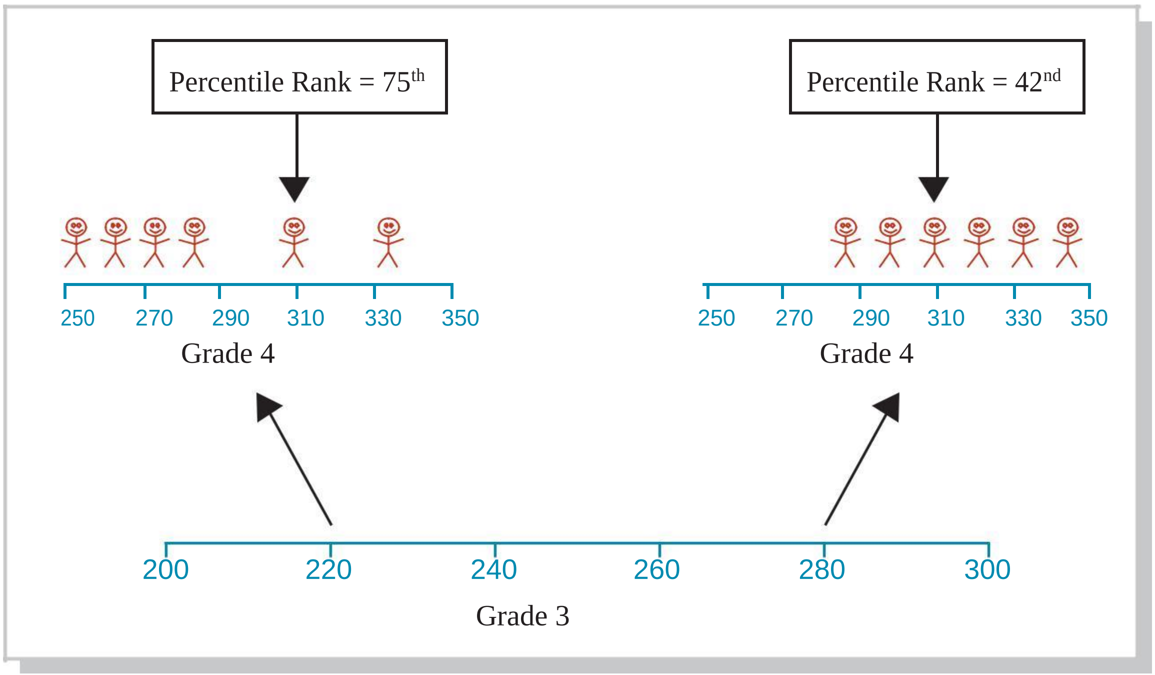


Figure 3.7 An illustration of the SGP model

Source: A Practitioner’s Guide to Growth Models (Castellano & Ho, 2013a)

The score line in the lower part of Figure 3.7 represents students’ prior test results in Grade 3. Each student in the class has a test score in a specific teaching subject from the previous year. Imagine selecting two students in the class who earned scores of 220 and 280 in Grade 3. Now we need the other students, who also had the same scores in third grade, their “academic peers”, or a “comparison group/cohort” of the student to be selected. The academic peers’ scores will be compared to the selected students’ test scores in the current grade. The students in the comparison group are ranked based on their test scores in fourth grade. The position of a student in the comparison cohort represents the student’s SGP score. While the first student received a student growth percentile of 75, which means that the student’s academic attainment in the current year is better than 75% of his/her academic peers, the second student earned an SGP score of 42, which means that 58% of students in his/her comparison group did better in the fourth-grade test than the second selected student. To characterise the performance of the teacher based on the individual student SGP scores, after determining SGP scores of all students in the classroom, the scores are ranked from lowest to highest, then the middle percentile score is determined. This middle percentile score gives the group’s median number that refers to the median student growth percentile (MGP), which is used to determine the performance of the teacher. The median growth percentile method is also used for school accountability in the educational accountability system in various states in the USA, such as Colorado, Massachusetts, and Washington.

3.1.7 Value-Added Models

VAMs are listed among growth models as they are statistical techniques based on regression analysis used to measure the academic growth of students over time. While both VAMs and other growth models are based on changes in students' test scores over time, VAMs are specifically used to determine the extent to which changes in student academic performance are attributable to a particular teacher/school. VAMs can also be defined as extended forms of some growth models that can be used for value-added purposes so that these models associate students' academic growth with a particular teacher and/or school, allowing inference to be made about the cause of this growth. Hence, VAMs can be simply defined as adjusted growth models (Ligon, 2008).

The most important characteristic that distinguishes VAMs from other growth models is that VAMs control the impact of selected factors, such as students' SES and/or interventions such as the programme, teacher, school, etc., on the student's current academic performance. Since VAMs make it possible to take into account student achievement gains after adjusting for some background characteristics, these approaches provide fairer estimates than judgements based on students' test scores at a single point, as in the status model, or on comparing different students at the same point at different times, as in cohort models. A growth score is usually calculated simply as the difference between the student's current and prior attainment, while a VAM score is statistically more complex, as it is obtained by separating non-educational factors such as SES from the student's academic achievement. Then, the student's isolated achievement growth can be associated with the educational practices of the school and teacher (McCaffrey et al., 2003). Since the main purpose of VAMs is to determine the impact of teachers and the school on student achievement, taking into account non-educational factors for student achievement, VAMs often deal with outcomes at the teacher and school level, not individual student growth as in growth models.

A student's academic achievement might be affected by various factors such as language learner status, family SES, etc., and a growing number of studies have reached a consensus that among school-related resources, the most crucial is the quality of teachers (Aaronson et al., 2007; Opper, 2019; Rivkin et al., 2005; Wright et al., 1997). Thus, since the teacher is the greatest contributor to a student's achievement, the presence of effective teachers in the classroom is one of the most important educational issues for policymakers. For this reason, policymakers want to make sure that their classrooms are staffed with effective teachers to enhance students' academic achievement. The issue of determining whether class teachers are effective or not, which is closely related to student achievement, gains importance here. Consequently, how to evaluate teacher effectiveness has been an ongoing issue of debate for researchers and policy makers.

VAMs, statistical methods adapted from economics, are designed as a set of approaches based on student academic achievement growth to be used in teacher accountability. To estimate a teacher's impact on student achievement, most VAMs take account of students' prior attainments and some demographic characteristics. Then, the teacher's value-added score is usually calculated by averaging the difference between the actual scores of all students in the teacher's class and their predicted scores based on prior attainment and some demographic characteristics. The difference between the actual and estimated scores are also conceptually considered to be the teacher effectiveness. Since their first use in teacher evaluation, VAMs are the models most studied by researchers and consequently the most debated. VAMs, theoretically, isolate the effectiveness of a particular teacher on the achievement of their students from other factors that contribute to student achievement outside the teacher's control, such as family, peers, prior attainment (McCaffrey et al., 2004; Darling-Hammond et al., 2012a). From another perspective, one of the uses of VAMs in the educational accountability system is to distinguish effective teachers from ineffective ones, and they do this by treating, on average, teachers whose students perform better than expected as more effective than those whose students do not meet the expected performance.

As stated before, VAMs do not refer to one single approach; they consist of various techniques from simple models, such as the covariate-adjusted model, to complex regression models, such as multi-level modelling that are used in estimates of teacher effectiveness, based on the growth in the average academic achievement of students in the classroom over time (usually over a few years) (Rubin et al., 2004). Although a number of different VAMs have been developed to predict teachers' impact on student learning, each has its own advantages and disadvantages. All VAMs, in general, are based on the logic that students' academic achievement reflects their teachers' performance, so teachers should be held accountable for the changes in students' academic attainment (Shaw, 2012). On the other hand, VAMs differ from each other in terms of which student, classroom, and school background characteristics are taken into account and how they are controlled, or whether teachers are compared within or across schools, or whether prior teachers' influence is to be considered as diminished or undiminished into the future. For instance, the Tennessee Value-Added Measurement System (TVAAS) (Sanders et al., 1997) is also called a layered model and assumes that teacher influence will continue undiminished into the future. The model suggests that the effectiveness of the classroom teacher, which affects student performance, will continue unabated in later grades. Therefore, the estimated teacher effectiveness in the higher grades is shared equally between the current teacher and the former teacher(s). Alternatively, it is possible to adjust the former teachers' effect in the future by using the "persistence model" (McCaffrey et al., 2003). The differentiation underlying VAMs, unfortunately, causes different value-added scores to be estimated for the same teacher, especially in teacher quality rankings (Goldhaber & Theobald, 2012). The results on different teacher performance, estimated based on the chosen model, reveal the potential bias in using VAMs alone in the evaluation of teacher performance, thus the doubts about the reliability of these models, especially in high-stakes personal decisions. Such decisions include improper promotion or demotion, unjust pay rises or cuts in salary, permanent appointment (tenure) or dismissal. On the other hand, it may be appropriate to use value-added measures to identify teachers who need assistance (Murphy, 2012). Alternatively, it is also suggested that these measures may be instruments to improve practices by providing valuable information about the deficiencies and strengths of the curriculum, teaching methods, and other teaching practices applied at school (Hong, 2010).

Finally, this thesis focuses on VAMs for evaluating teacher performance, but similar models have been used in the UK to evaluate school performance based on pupil contextual background and prior attainments, such as contextual value added (CVA) and Progress 8. However, the ongoing concerns that will be discussed in the next section, which are common ones arising from the use of VAMs to measure teacher or school performance.

3.2 Ongoing Concerns Related to VAM Estimates

VAMs are intended to measure a teacher's performance in a more objective way by revealing how much value that teacher added to their students' learning. As mentioned in the previous section, researchers have created several different models as a result of their efforts to solve various technical problems that have arisen in measuring teachers' performance. However, none of them has completely overcome all the problems mentioned. Some of the problems discussed in this section also led the researcher to conduct this research.

VAM estimates are made by statistically measuring changes in students' academic performance from the previous year to the next. However, besides the fact that the test scores do not fully reflect the learning of the students, VAMs are based on the assumption that the contents of the previous and next tests are equivalent. For example, suppose students took an exam that was predominantly about geometric shapes, reflecting the previous year's curriculum, but what if they are subjected to an algebra-based test the following year? Subtracting the scores from these two exams may lead to a false judgement about teacher performance. Therefore, VAM teacher performance evaluators should take into account the scope of grades' curricula before making a decision.

All value-added measures use pupils prior scores and gains from that score as an indication of teacher effectiveness. But these test scores are not perfect. As Gorard noted (2018a), in reality, a highly important proportion of the pupil gain scores in England is due to error propagation as they have missing records, measurement errors, and errors in representation. A very good example of how unreliable such value-added models can be is the Tennessee Value-Added Assessment System (TVAAS), where (by accident) a school that lies on the county line was given two VA measures in the TVAAS. The same school was given two completely different scores (Glass, 2004).

Perhaps one of the most agreed issues is that the actual dynamics of schools are ignored in VAM estimates. More specifically, VAM assumes that students are randomly assigned to schools or teachers. However, in reality, teachers who are considered to be successful according to their students' test scores in previous years tend to choose their own classes, or school principals tend to assign successful students to high achieving teachers. Therefore, successful and highly motivated students are more likely to have effective teachers in school, contrary to what VAMs assume. Without random assignment, it is difficult to ascertain whether students' high levels of achievement were caused by their teachers or by the motivations of the students themselves or something else. Therefore, it is inevitable that one obtains biased results in the evaluation of teachers based on VAM estimates (Rothstein, 2009; Paufler & Amrein-Beardsley, 2014). Even if the random assignment issue is overcome, it remains a mystery whether improving teacher quality can also improve student outcomes. For instance, although the first large-scale study - the Bill & Melinda Gates initiative (Kane et al., 2013) – tackled a number of challenges, it could not resolve this mystery. First, random assignment was subverted in a number of ways in that although teachers were randomly allocated to classes, students were taught by more than one teacher, and some students swapped classes in the same school. Some teachers also left teaching or taught a different subject or grade. There were also students who were assigned to one teacher but ended up in another teachers' classroom, and some schools simply ignored the randomization. In the end, many students ended up with a teacher different to the one assigned to them. As a result of this multimillion-dollar project is that students in the intervention group did not do much better than those who did not.

On the other hand, in a very recent study by Bacher-Hicks et al. (2019) using random assignment of teachers, value-added, student surveys and classroom observations were compared, and it was concluded that the value-added measures are unbiased predictors of teacher performance. However, this study is based on a very small number of teachers who were actually randomised (N = 66). Only one-year test scores were used, and over 30% of students remained in their randomised classrooms. All this reduces the trustworthiness of the findings.

The uncertainty about which variables should be used in models in teacher value-added performance assessments is another issue discussed in the literature. It is acknowledged that there are many personal, family, and school-related factors that are thought to have an impact on the academic success of the student apart from teachers. There is a great consensus in the existing literature that students' previous performance play an important role in their current attainment (Hu, 2015; Kersting et al., 2013), but unfortunately, similar consensuses have not been reached on other factors that may have an impact on student achievement. In line with the view that students' academic achievement is greatly influenced by their families' well-being, Gorard and See (2009) found that SES (socio-economic status) is associated with student attainment. However, no clear relationship has been revealed between the growth in students' achievement as measured by VAM and the well-being of their families (Muñoz et al., 2011; Ehlert et al., 2014; Rothstein, 2009). In short, the lack of knowledge on the impact of student, teacher, and school characteristics on student achievement, and hence the question of whether these predictors should be included in the model in teacher value-added performance estimates, is an important topic of discussion in VAM.

Another of the most discussed topics, possibly the most problematic one, by researchers is that VAMs produce unstable results concerning teachers' performance. There are many VAM studies that have determined that teachers who were defined as highly effective one year were ineffective in the following year or vice versa, or very talented teachers might be identified as "ineffective" (Schmitz, 2007; Sloat et al., 2018; Berry, 2010; Goldhaber et al., 2014). Sometimes, volatile estimates are caused by predictors that are included or not in particular models, sometimes due to the additional previous test scores added or the applied model. In other words, this instability is caused by many factors beyond the teacher's control. Therefore, VAM results should not be used as the sole or primary evaluation tool for making high-risk decisions about individual teachers (Goldhaber, 2010; McCaffrey et al., 2003).

The last concern discussing in this section is related to models' complexity. Since the use of VAMs in teacher performance evaluation, many models have been developed by researchers, and accordingly, many studies have attempted to determine how different value-added scores are produced by simpler and more complex models. Some researchers have developed complex models, considering that it may be beneficial to take into account important factors that may affect teacher effectiveness, in parallel with the complex and stratified structure of school and education, and have advocated the use of these models in performance evaluation (Sanders et al., 1997 [layered model]; Raudenbush & Bryk, 2002 [cross-classified model]; McCaffrey et al., 2004 [general model]). On the other hand, some researchers point out that if complex models make little difference in practice, simpler models may intuitively be preferred due to their practicality (Cunningham, 2014; Schmitz, 2007). For example, students interact with many teachers in their education, and the contribution of the previous teachers they had on student achievement is undeniable. Therefore, in modelling teacher contributions to student achievement, the contributions of previous teachers should also be taken into account; that is, the longitudinal achievement data to be used should allow students to regroup with different teachers in different classes. However, these are not all the concerns about such models; many other concerns are still awaiting answers from researchers, such as how to treat previous teachers' contributions (persist undiminished into the future, decrease gradually), or what other stakeholders should be taken into account, as individual scores are influenced by multiple stakeholders, such as teaching assistants and librarians.

Many VAMs in teacher performance assessment have been designed in response to these and similar problems, but as can be seen, none of the models has been able to overcome all the problems. For this reason, this research aims to contribute to some of the concerns discussed in the literature and mentioned in this section with the answers to the following main research question:

How stable are teacher effectiveness estimates measured by VAMs?

The sub-questions are:

- How stable are teacher effectiveness measured by VAMs that consider student, school, and teacher-classroom characteristics?
- How stable are teacher value-added effectiveness estimates over a two-year period of time?
- How stable are teacher value-added effectiveness estimates when including an additional prior score (t-2)?
- Do different methods of analyses used in VAMs produce consistent teacher effectiveness estimates?

SECTION II

RESEARCH DESIGN AND METHODS

The research design and methods section contains two chapters. In Chapter 4, the stages of conducting the systematic review study are discussed in detail. The purpose of the review, the searching strategies used, how the relevant studies are screened, the tool used in data extraction and quality appraisal, and data analysis methods used in this chapter are explained. In addition to this, Chapter 5 also deals with the research design and the methodology of the secondary data analysis study. This chapter discusses the study population and types of data used in analyses, and the data collection procedures. In this methodology chapter, the data analysis methods employed in each sub research question are also presented in detail.

CHAPTER 4
SYSTEMATIC REVIEW – DESIGN AND METHODS

This chapter describes the method used to identify, review and synthesise studies from a systematic search of the usual educational, sociological and psychological databases. The systematic review synthesises research on VAMs to determine the stability of teacher effectiveness estimates with regards to the number of contextual predictors used, previous test scores used, and data analysis methods employed.

Value-added models or VAMs have been widely used in teacher performance appraisal for high-stake purposes, such as decisions on dismissal and monetary reward. The findings of this review will provide evidence to justify the use of such models in making judgments about essential decisions concerning teachers' careers and to make recommendations to policymakers about using VAMs in policy and practice for high-stakes personnel decisions.

This review will synthesize the results of previous relevant studies that analyse the contribution of contextual predictors, such as student, school and teacher/classroom characteristics, the number of student test scores over time and data analysis methods to teacher performance evaluation.

4.1 Purpose of the Review and Research Question

The aim of a systematic literature review is to identify, select and critically appraise relevant research to answer a clearly formulated question(s) in a systematic and explicit manner (Torgerson, 2003). This is to prevent bias in selecting only particular types of studies, such as those that report positive effects or those that use one method of discovery.

Specifically, this review aims to determine how stable estimates of teacher performance evaluation are based on VAMs that use contextual predictors, the number of previous test scores, and data analysis methods. Understanding the contribution of predictors, the number of lagged test scores, and data analysis methods to the stability of VAM estimates is critical for education policy-makers, school administrators, and practitioners as they are commonly used in high-stake decision-making purposes for teachers and schools in many countries of the world, especially in the USA.

The main research question for this review is:

Teacher effectiveness is operationally defined by VAM as the estimation of the differences between expected and observed student test scores (Kersting et al., 2013). When the literature is examined, it is understood that stability studies are carried out from three main perspectives; therefore, in this systematic review, the stability of the estimates refers to the stableness of the estimates due to (a) *the predictors used in the estimations*, (b) *the number of test scores used*, and (c) *the analysis methods applied*. Existing literature on the stability of VAMs estimates will be retrieved from these three perspectives.

4.2 The Search Strategies

To ensure that the search process is systematic and organised, the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram is used. PRISMA is a diagram published in 2009 by the PRISMA group (Moher et al., 2009) to help researchers to map out the number of studies identified, included and excluded based on the criteria established. The PRISMA involves a series of steps presented as a flow chart. To give more details about the comprehensive search process, a modified form of PRISMA flow diagram is used in this research (see Figure 4.1). In the modified flow diagram, information on the databases and their providers are shown as separate rectangles at the top with a separate rectangle for the alerted results added later. Search alerts were set up in each database. When new records regarding the search criteria were available, the researcher was informed by the databases' alert system. The search was completed on1st of May 2019.

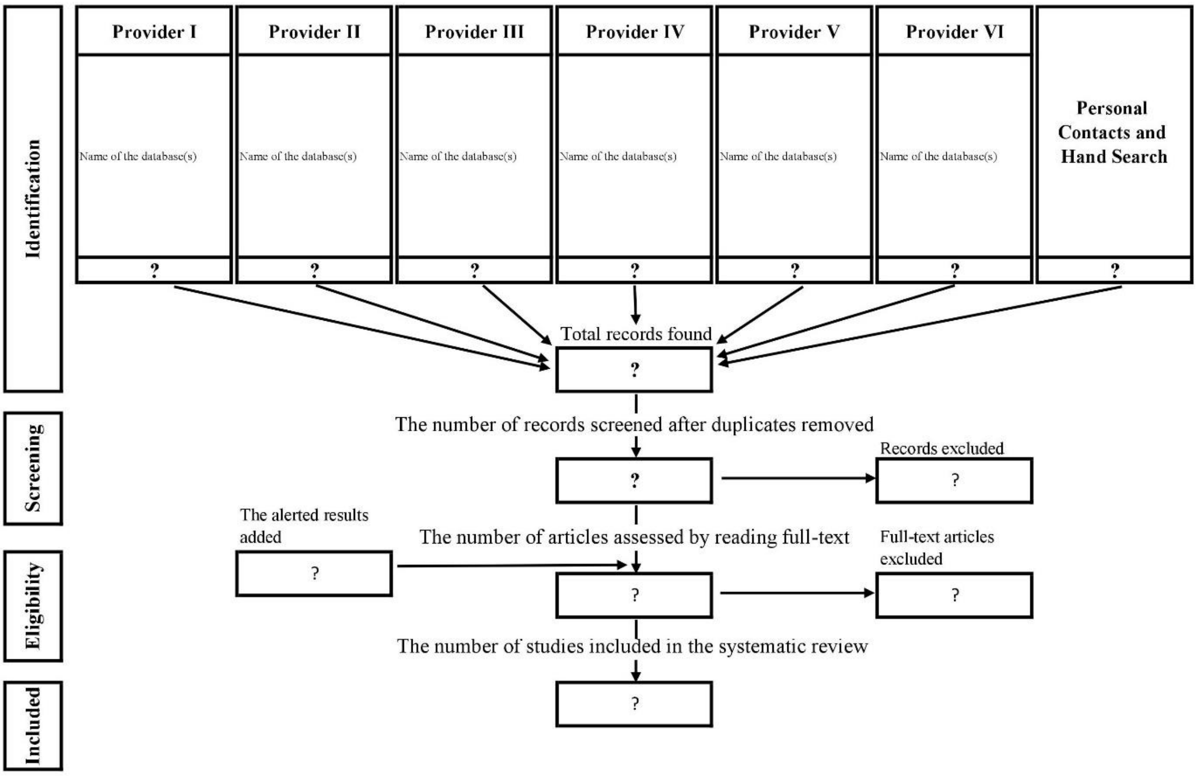


Figure 4.1 The modified PRISMA flow diagram

4.2.1 Databases

A total of 17 electronic databases from 6 major providers (see Table 4.1) were accessed via the Durham University Online Library system. The appropriate databases for this research were recommended by experts in this field and suggestions from personnel at the university library.

Table 4.1 Databases and Their Providers

	Provider	Database
1	ProQuest	ProQuest Dissertations & Theses Global: Social Sciences
		Education Database
		ERIC
		International Bibliography of the Social Sciences (IBSS)
		Social Science Database
		Applied Social Sciences Index & Abstracts (ASSIA)

2	EBSCOhost	OpenDissertations
		British Education Index
		Business Source Premier
		Education Abstracts
		Educational Administration abstracts
		PsycINFO
3	Web of Science	Web of Science Core Collection
		Current Contents Connect
4	Elsevier	SCOPUS
5	SAGE Research Methods Core	SAGE Journals
6	Taylor & Francis Online	Educational Research Abstracts Online

Besides these databases, the relevant sources of this systematic review were obtained from personal contact with research centres, foundations, and researchers who have worked on teacher performance evaluation based on VAMs. To ensure that both published and unpublished studies are included in this study, websites of relevant research centres and foundations were also searched. Some relevant websites included Nber.org, Caldercenter.org, gatesfoundation.org, and nepc.colorado.edu.

Contacts were also made with well-known researchers in this area via e-mails to identify isolated published and/or unpublished studies related to the review topic. Further search for experts in the field was done through “ResearchGate.net”. Hand searches of reference lists in journal articles and well-known studies were also made to identify studies that may not have been picked up in the electronic databases using a snowballing approach. A search of the Google search engine and Google Scholar were also made to look for grey literature. To be sure that the review is comprehensive, both published and unpublished materials were included. For this reason, the ProQuest database was also searched to look for PhD or master’s theses and dissertations that may not be published.

4.2.2 Keywords and Search Strings

To facilitate the search, appropriate search strings were formulated that are relevant to the research question, which is: How stable are teacher effectiveness estimates measured by VAMs? The search terms included “teacher performance”, “student performance”, “value-added model”, and “stability”. In order to identify the related search terms, alternative or synonym terms used in studies that were already known to me were determined through the Durham University Online Library search system and Google Scholar. (shown in Table 4.2).

Table 4.2 Search Keywords

Search Terms	Related Terms	
Teacher performance	Teacher effect* Teacher evaluation Teacher performance evaluation Teacher appraisal Teacher performance appraisal Teacher quality Teacher assessment Teacher performance assessment Teacher accountability Teacher proficiency	Teacher proficiency-rank Teacher judgment Educational effectiveness Educator performance appraisal Educator performance Educator evaluation Educator quality Teaching effect* Measuring teach* Evaluating teach*
Student performance	Academic achievement Academic gains Student test score Student test performance Student test-score	Achievement Achievement measure* Outcome* Outcome measure*
Value Added Model	Value added modelling Value-added model* Teacher value-added	VAM* Value added estimate* Value-added estimat*
Stability	Concord* Robust Sensitivity Instability Precision	Imprecision Variat* Fluctua* Persistence Shrink*

These search terms were then applied to databases and providers shown in Table 4.1. Most of these databases allow the use of advanced search using Boolean and truncation operators. This enables search terms to be expanded using the OR command. The search can also be more focused using the AND and NOT commands. Truncation, also known as wildcard, lets the researchers search without using all the alternative spellings of a search term by removing a letter(s) at the end of the search term where the spelling differences start and attach an asterisk (*) at the end of that word.

The search strings used in this review were:

> (((teacher OR educator) AND (effect* OR evaluat* OR quality OR perform* OR appraisal OR assess* OR accountability)) OR ((teacher OR educator) AND performance AND (evaluation OR appraisal OR assessment)) OR "teacher proficiency-rank" OR "teacher judg*" OR "educational effectiveness" OR "teaching effect*" OR "measuring teach*" OR "evaluating teach*") AND ((academic AND (achievement OR gain*)) OR ((student AND test) AND (score OR performance)) OR achievement OR outcome* OR ((achievement OR outcome*) AND measur*)) AND (VAM* OR (Value-added AND (model* OR estimat*)) OR ((value AND added) AND (model* OR estimat*)) OR (teacher AND value-added) OR (teacher AND value AND added)) AND (stabil* OR concord* OR robust OR sensitiv* OR instabil* OR precis* OR imprecise* OR variat* OR fluctuat* OR persistence OR shrink*).

The search strings were adapted to the idiosyncrasies of the different databases. Applying these search strings to the different databases identified 1,103 articles.

As Google Scholar does not have an advanced search function, the search strings were modified using the following keywords: teacher effectiveness estimated by value-added model stability OR concordance OR robust OR sensitivity OR unstable OR precise OR imprecise OR variation OR fluctuation OR persistence OR Shrinkage (see Appendix A). Using these keywords in the Google Scholar search revealed 26,8000 results. However, when the results were sorted by relevance, only 260 sources in the first 16 pages were identified to be substantively the most relevant to this systematic review’s purpose.

Personal contacts with research centres, foundations, and well-known researchers in the field identified a further 71 studies. Altogether a total number of 1,439 studies were found, including the five results that were reached by hand search in "Google" and "ResearchGate".

4.2.3 Cleaning the datasets

Following the database searches, all the results obtained from each provider were merged in an excel spreadsheet and exported to EndNote X8, a reference management software. As the search involved a number of databases, it is not surprising to find many duplicate versions of some of the studies. The dataset will, therefore, need to be cleaned to remove these duplicates. EndNote has a function that helps to identify duplicate cases. However, there were some duplicated results that were not flagged up by the software. This can happen because different databases sometimes record the same study differently; for example, one may be recorded as a report and another as a journal article. These were detected and eventually removed during the screening process when the title, abstract, and full text of the articles were read. Details about the search strings used, the results found in each database and the duplicated cases are in Appendix A.

4.3 Screening the Relevant Studies

The next stage in the review process was to screen the studies to remove irrelevant ones by applying a set of inclusion and exclusion criteria. At Phase I, titles and abstracts of the studies in the list were screened and included if they met the following criteria: a) written in English, (b) relevant to education, (c) took place in K-12 settings, (d) primary or empirical research, (e) related to teacher evaluation (see Phase I Screening Checklist in Appendix B). All irrelevant cases were filtered out from the review list, and prospective relevant studies remained for the second step of the screening. In order to avoid inadvertently eliminating relevant studies at the title-abstract screening process, an option of "not sure, yet" was placed for each criterion in the Phase I checklist. In this way, any studies where it was not clear from the title and abstract if they were relevant due to incomplete or vague information were kept. These studies would be rejected at Phase II screening when the full text is read if they were found not to have met the inclusion criteria. In the second stage of the screening process, the full texts of the prospective articles remaining were read, and it was determined whether they could be discarded or retained. Each study was assessed on 10 criteria (see Phase II Full-text Screening Checklist in Appendix C). The first five checklist points were specifically to assess those studies in Phase I where it was not clear from the titles and abstracts alone if they were relevant. These studies were revisited and reviewed. The other screening criteria relate to whether they include: (a) statement of stability of estimates, (b) student test or gains scores as dependent variables, (c) observable student, teacher and school characteristics (e.g., sex, ethnicity, teaching experience, school type) as predictors, (d) the contribution of predictors to estimates, and (e) the number of test scores used. As soon as it was clear that the studies did not meet the first few criteria on the list, they were immediately removed from the review list, and the reading process stopped. For example, if the dependent variables of a study are not about teacher effectiveness or student outcomes, the study was excluded at this point.

To ensure that relevant pieces were not mistakenly removed, a second reviewer was engaged to review 10% of the literature (Torgerson, 2003). For this systematic review, the second reviewer was the researcher's second supervisor, who reviewed a random sample of around 10% of the literature in the review list. To check that both reviewers were in agreement regarding the relevance of the literature, inter-rater reliability (IRR) was calculated. Along with the percentage agreement between the raters, as the agreement variable was coded as nominal, Cohen's kappa statistics (κ) was also computed for assessing inter-rater reliability, which is one of the most commonly applied statistical methods for estimating IRR in systematic review studies (Hallgren, 2012). The range of Cohen's kappa coefficients must be between 0 and +1. The closest coefficient value to +1 indicates high agreement exists between raters.

The following inclusion and exclusion criteria were established to effectively determine the research that examined the contribution of the contextual predictors and/or data analysis methods used to the stability of teacher performance evaluation estimates based on VAMs estimates.

4.3.1 Inclusion Criteria

The included studies in this systematic review met all the criteria listed in Table 4.3.

Table 4.3 Inclusion Criteria

Inclusion Criteria	
Criteria	**Description**
The population of this study is teachers	Only studies focused on teacher performance evaluation based on student test scores will be included in this systematic review. Where studies that evaluated the performances of multiple subjects, such as curricula, teachers and schools in a single study, only studies will be included in this systematic review if one of its interest areas is teacher effectiveness.
The issue of the study is the stability of the estimates	The operational definitions of the term of stability in this systematic review refer to the stability of the estimates based on; (a) the number of student test scores employed (b) the predictors used (c) the analysis methods applied Studies are included if they use any one of the above measures of stability.
Only empirical studies are reviewed for this study	Empirical studies refer to primary research as opposed to secondary research, such as reviews and government reports, but individual studies from the systematic review will be added to the review list. However, studies analyzing secondary data, such as panel and administrative data, are considered primary research.
The study setting of the research interest is K-12	All studies conducted from kindergarten (age 5-6, equivalent to Year 1 in the UK) to the 12th-grade (age 17-18, equivalent to Year 13 or 6th form in the UK) setting are included in this systematic review.
Published in English	Studies reported in English

This systematic review includes the studies that met all the described eligibility criteria in order to narrow the review to focus on the substantive research questions.

4.3.2 Exclusion Criteria

Studies are excluded if they were:

- Not reported or published in English
- Not primary research
- Not about education
- Not within K-12 (e.g., higher education, reception year or nursery)
- Not about the evaluation of teacher effectiveness
- About the use of value-added measures of teachers to predict teacher attrition
- The outcome is not student test scores or gains (e.g., children's behaviour or attendance)
- Using measures of teacher effectiveness to predict outcomes
- Just about school effectiveness or school improvement (but if the studies focused on both school (principal) and teacher effectiveness, they would be included in the review, but only their findings of teacher effectiveness will be used)
- About teacher effectiveness in non-mainstream school
- Just about pupils with special educational needs (SEN)
- About theories and policies, opinion pieces, discussion pieces
- Instructional manual or promotional literature about how to measure teacher effectiveness
- Literature about the characteristics of effective teachers

4.4 Data Extraction and Quality Appraisal

Unlike most systematic reviews, which use complex technical checklists for quality appraisal, this review evaluates each study using a set of robust appraisal criteria based on the research design and threats to validity. As El Soufi and See (2019) asserted, an essential feature of a systematic review is the quality appraisal. The reason for doing this is to ensure that the findings reported in the studies are trustworthy and thus represent the best evidence. Quality appraisal refers to the internal validity of the research conducted, which is related to how far the studies are methodologically free from biases (Petticrew and Roberts, 2006). The quality appraisal is crucial for the systematic review studies because it helps to distinguish the relationship between the differences in the strength of evidence of the research and the differences in the results of these studies. This is necessary because bundling weak and robust evidence in the same pot with equal weighting can lead to invalid or misleading conclusions (Gorard, 2014a). It also helps for the interpretation of the findings (Bettany-Saltikov and Mcsherry, 2016) so that more weight is given to research rated higher on the strength of evidence than those rated lower. The appraisal tool used in this systematic review is the “sieve” (Gorard, 2014a). This is preferred over the complicated technical checklists used in some literature because it is a practical way for evaluating the quality of individual studies, which takes into account the research design and factors that affect the validity of the study (e.g., sample quality and size, attrition). It is a quality appraisal framework originally designed for active designs to address causal research questions, but it can be used for other research designs. To judge the trustworthiness of the findings identified through the screening processes, a simplified extract form of the sieve was applied (see in Table 4.4).

Table 4.4 A ‘Sieve’ to Assist in the Estimation of Trustworthiness of Any Research Study

Design	Scale	Completeness of data	Data quality	Rating
Strong design for research question	Large number of cases per comparison group	Minimal missing data, no evidence of impact on findings	Standardised, independent, pre-specified, accurate	4🔒
Good design for research question	Medium number of cases per comparison group	Some missing data, possible impact on findings	Standardised, independent, not pre-specified, some errors	3🔒
Weak design for research question	Small number of cases per comparison group	Moderate missing data, likely impact on findings	Not standardised, independent, or pre-specified, some errors	2🔒
Very weak design for research question	Very small number of cases per comparison group	High level of missing data, clear impact on findings	Weak measures, high level of error, too many outcomes	1🔒
No consideration of design	A trivial scale of the study, or number is unclear	Huge amount of missing data, or not reported	Very weak measures, or accuracy not addressed	0🔒

Each included study was judged according to four criteria: the research design (e.g., whether the research design is an RCT with random allocations of the population), scale (sample size per comparison group), level of attrition (the incompleteness of data) and data quality (how outcomes are measured) and given a padlock or security rating from 0🔒 (no evidence) to 4🔒 (most trustworthy). Studies rated four padlocks can be considered as the most trustworthy or secure.

Each criterion in the columns and rows has a hierarchical structure within itself. The sieve is to be read from left to right, starting from the strongest design. Since the systematic review is to determine the contribution of contextual factors and data analysis methods to the stability of VAM models in estimating teacher effectiveness, a descriptive study design is considered appropriate to fulfil these objectives. Therefore, studies that are large scale correlational/comparative studies with low attrition and also allow random student-teacher allocations are considered to have a strong design (4🔒). However, if the trial concerns only a small number of cases, then it drops a padlock or two and moves to row 2 or 3, depending on how small the sample is. For example, if a small number of cases for a comparison group (e.g., 50 teachers) are involved in the estimates, then the quality of the study will be rated as 2 padlocks as these sample size would be insufficient to demonstrate variations between groups. Moving to the third column, if the correlational study involves a large sample, it will start with 4 padlocks, but if it loses a large proportion of the cases, then it may drop two padlocks. And if the measure of outcome is not reliable, for example, based on teachers' or pupils' self-report of pupils' performance, then the results will be rendered invalid, and the trial will drop a further padlock. So, the final rating for the study would be one padlock.

Another critical issue needing to be explained here is that the criteria in the subsequent columns cannot compensate for a deficiency of the previous criterion. This means that the padlock rating can never move up. To give an example, if a study's design is determined as a "weak design for research question" (2🔒), the rating can stay in the same row or move down, but the study cannot move up to 3🔒 in the subsequent columns.

Moreover, without interfering with the essence of the appraisal tool, for the convenience of making decisions about the scale (sample size per comparison group) of the studies with the "sieve", three scale categories were determined for the comparison group. Comparison group size refers to the size of the smallest group, whether comparison or not. Consequently, below 1000 students or 50 teachers sample cases for a comparison (or smallest) group in the estimations were determined as "small number of cases", while a comparative (or smallest) group sample of more than 2000 students and 100 teachers was determined as a "large number of cases". The comparison group sample size between these two groups (between 1000 and 2000 students, or 50 and 100 teachers) was identified as a "medium number of cases". Moreover, the attrition rates were clustered into five groups. Minimal missing data is missing up to 19%, some missing data is missing between 20% and 39%, moderate missing is between

40% and 59%, high level of missing data is missing between 60% and 79%, and lastly, a huge amount of missing data refers to more than 80% of data is missing or not reported.

The most challenging criterion for this systematic review study was the completeness of data used in the estimation. Because almost all the included studies used longitudinal panel or administrative data, where data loss is inevitable, I was more flexible or lenient in terms of missing data. In this review, a balance had to be struck between the number of cases and the attrition. For instance, Goldhaber and Hansen (2010) could only link 3% of the teacher data with student data, losing 97% of the data. But the scale was still large, with 609 teachers linked with 26,280 students. So, I rated this as 1 padlock rather than 0 padlock. The same strategy was applied to studies that do not report missing cases in the estimations. Instead of giving 0 padlock and discarding them from the synthesis process, it was treated as having a high attrition rate; its rating was dropped to the lowest value (1 padlock). In this way, the bias likely to occur in the synthesis of findings was intended to be minimised.

Once the quality appraisal process has been completed for the articles retained after the screening stages, the key data from the relevant studies are extracted and recorded in an excel spreadsheet. The sheet includes the key information about the individual studies in accordance with the purpose of the review, the research question and the determined review perspectives.

The findings from the articles retained were classified under the three perspectives clarified in the following section and merged within an excel spreadsheet. The data extracted for each article includes the following information: the author(s) names, date of the publication and title, type of research design used, the country of data collected, the number of participants, method of assigning teachers to students, scale, completeness of data used, study setting, dependent variable(s), data analysis method(s), stability of estimates due to the number of test scores, stability of estimates due to the predictors, and stability of estimates due to analysis method(s) (see in Appendix D).

4.5 Data Analysis

After data extraction, the studies are then analysed. To facilitate analysis, all the included studies were classified from three main perspectives:

- the predictors used in the estimates,
- the number of previous test scores employed in the estimates, and

- the data analysis methods applied for the estimates.

To avoid the loss of data, the data extraction sheet was stored in a password protected cloud storage and file synchronization system.

4.6 Summary

The purpose of this systematic review study was to investigate the contribution of the number of previous test scores, the contextual predictors, and data analysis methods to the stability of teacher performance evaluation estimates based on VAMs. The data were collected from the included studies by following the stages of the systematic review clarified by Torgerson, 2003. Inclusion and exclusion criteria were applied to eliminate irrelevant studies. Relevant data in each study were extracted, which facilitated the evaluation of the quality of the evidence and synthesis of findings. To ensure that the evidence from the review was valid and trustworthy, each included study was quality appraised using the “sieve”. Then, relevant information that answers the review question was extracted and recorded in an excel spreadsheet. The extracted data were synthesised based on the number of previous test scores employed in the estimates, the predictors used in the estimates, and the data analysis methods applied for the estimates.

CHAPTER 5
THE SECONDARY DATA ANALYSIS - DESIGN AND METHODS

This chapter presents the methodology used to explore the stability of teacher value-added effectiveness estimates using secondary data analysis. This chapter is divided into five main sections. The research design is introduced in the first section. The next section explains the study population, which includes students, teachers, and schools, as well as how they were selected. Then, the data employed in the analyses and the data collection procedures are introduced. The following section discusses the missing data issue and how to address and analyse them. The last section explains the data analyses methods utilised in each sub research question in details.

5.1 Study Design

The nature of this research design is a retrospective study intended to estimate the contribution of contextual predictors at student, school, and teacher/classroom-level, students' prior test score(s), and the choice of data analyses method to teacher effectiveness estimates in five subjects (mathematics, Turkish, science, history, and English) by employing longitudinal data from an administrative data set extending over three school years, 2014-2017, from secondary schools in the Samsun Province in Turkey. Since the retrospective study design allows the relationship between independent and dependent variables to be examined without any manipulation of the variables, the chosen research design is appropriate for this study.

Along with the nature of the overall study that is retrospective, the specific design of this current study is the longitudinal correlation. To assess the stability of teacher value-added effectiveness estimates, the main research question is:

How stable are teacher effectiveness estimates measured by VAMs?

To answer this main research question, four sub research questions were formulated. See Table 5.1 for a summary of analysis methods to be utilised for each of the sub research questions. More detailed explanations of the analysis methods conducted are also provided in each related research question in the result chapter.

Table 5.1 Summary of Research Designs and Data Analysis Methods

Research Questions	Data Analysis Methods
How stable are teacher effectiveness measured by VAMs that consider student, school, and teacher/classroom characteristics?	Multiple regression analysis using the forward selection method - having the largest R-squared by including as few predictors as possible
How stable are teacher value-added effectiveness estimates over a two-year period of time?	(1) Multiple linear regression analysis (2) Pearson's/ Spearman's correlation coefficients (3) Transaction matrix
How stable are teacher value-added effectiveness estimates when including an additional prior score (t-2)?	(1) Multiple linear regression analysis (2) Pearson's/ Spearman's correlation coefficients (3) Transaction matrix
Do different methods of analyses used in VAMs produce consistent teacher effectiveness estimates?	(1) Multiple linear regression analysis (2) Residual gain model (3) Two-level HLM (4) Pearson's/ Spearman's correlation coefficients (5) Transaction matrix (6) SD analysis

5.2 The Population of the Study

The target population of this study is all teachers who taught Turkish, mathematics, science, history, and English (as a foreign language) in 8th grade (age 13-14, equivalent to Year 9 in the UK) during the 2016-2017 school year in the Samsun Provincial Directory of National Education, Turkey. During the 2016-17 academic year, the provincial directory enrolled 272,261 students (from kindergarten through grade 12) and employed 17,965 teachers in 1,129 schools (MoNE, 2017a). As the value-added estimate requires at least one previous year's test score along with the outcome score of the same student, it was decided to conduct this study in secondary schools where data are available.

The target student population was those who can be tracked academically from Grade 6 through to 8 (Key stage 3 – Years 7 to 9). However, although a total number of 18,986 students enrolled in 8th grade in 315 secondary schools in the 2016/17 school year (MoNE, 2017a), due to the fact that not all secondary schools in Samsun participated in the Step-by-Step Achievement project (explained in the following section), and the mobility of students taking the test, the total number student population in this study is around 16,000 in each teaching subject. As a part of the purpose of the study is to examine the contribution of schools' characteristics to the estimates of secondary school teachers' effectiveness, this research involved all secondary schools without any discrimination on the school type in Samsun province. Because of the same limitation reason explained above (e.g., data availability), only 282 secondary schools could be included in the study for each teaching subject.

In order to assess the stability of value-added teacher performance estimates, it is essential that student data can be linked to teacher data longitudinally (Clotfelter et al., 2006; Goldhaber, 2007). Since the information about the teachers was not available in the administrative data set that the researcher obtained, the teacher information was requested from the schools where they work through the Samsun National Education Directorate (more details about the data collection procedure are given in the next section). Unfortunately, not many school directorates were willing to share the requested teacher information with the researcher (none of the private schools shared their teacher information); therefore, more than half of the data available in the administrative dataset could not be used in the analyses of this study. The number of teachers in each teaching subject involved in the research ranged from 173 to 232 and were those who are linked to a total of 35,435 students in 8th grade.

Table 5.2 displays the sample sizes of this study. On the left-hand side of the table, unrestricted sample sizes (original total available cases in administrative dataset) are given, while on the right-hand side, the restricted number of participants (number of available cases to be used in analyses) involved in the estimates are displayed.

Table 5.2 The Population of the Study

		N of Unrestricted Sample	N of Restricted Sample
(Student)			
	Mathematics	16,444	7,543 (21%)
	Turkish	16,827	7,594 (21%)
	Science	16,419	7,116 (20%)
	History	16,410	6,638 (19%)
	English	16,376	6,544 (19%)
	Total	**82,476**	**35,435**
(Teacher)			
	Mathematics		230 (22%)

Turkish		232 (23%)
Science		204 (20%)
History		174 (17%)
English		187 (18%)
Total		**1,027**
(School)		
Mathematics	282	145 (21%)
Turkish	282	150 (21%)
Science	282	137 (20%)
History	282	131 (19%)
English	282	132 (19%)
Total	**1,410**	**695**

5.3 Data Available for Analyses and Collection Procedure

Data used in the study included longitudinal students' achievement data spanning three consecutive school years, students' characteristics, teacher/classroom background information, and school information.

Several challenges were encountered in accessing student data due to changes in the school exam. In the Turkish education system, students take the first national exam at the end of Grade 8, which is the last grade of secondary school. This national exam was compulsory for 8th graders but was discontinued in 2017. Although there is another national exam taken at the end of high school, which is Grade 12, it is challenging to establish a link between the exam scores of students in Grade 8 and Grade 12. Even were it to be possible, it would not be fair to attribute the changes in student performance over 4 years to a single teacher. Therefore, other alternative data sources were searched. It was found that the Samsun Provincial Directory of National Education had been running a project named "Step by Step Achievement" since 2015 (Samsun Provincial Directorate of National Education, 2014), and within the scope of this project, every year since then, secondary and high school students have taken low-stakes exams in various subjects at the same time throughout the province. Unlike high-stakes tests, although low-stake tests apparently may be better at predicting student achievements attributed to teacher effectiveness (Goldhaber et al., 2013), it should be kept in mind that students might not try their best because they underestimate these tests, so the low-stakes tests might not be a very good measure of students' real learning (Koretz, 2008).

Therefore, in the absence of the high-stakes test, which was scrapped, the Step by Step Achievement (SBSA) exam scores were used in the analyses of this research. Under this project, all the SBSA exam scores over the years, including students' background information, their school and class information (i.e., teacher's name) were collected and stored electronically in the provincial directorate's own electronic systems. To obtain the data to be used in this study, the provincial directorate was contacted to request permission to access the data. After approval, the researcher was given a username and password to log into the electronic data storage system. In the electronic system, there are data about students' test scores over the years and some basic information such as name, sex, and language learner status, as well as school and classroom names. Although the names of the teachers associated with the students were available in the system, as the researcher was given restricted access to the system, the teacher names were not accessible at this stage. However, this information was then provided via email in an excel spreadsheet.

Once student data was accessed, the longitudinal test scores of all students registered on the system were downloaded, and this data was then merged with other student-level data, including their names, their unique school number, classroom, sex, and their language learner status.

Students' names and their unique school number information prevented the same students from being recorded twice. All duplicated records in the system were deleted from the datasheet during the merging process. The next step was to link the student data with the teacher data, whose names were provided in the excel spreadsheet. Since the electronic storage system did not contain teacher-level data, this data had to be obtained from each of the schools separately via the provincial directorate. The schools' directorates provided the following information about teachers: sex, number of years of teaching experience, number of years teaching in the current school, teaching appointment field, teachers' major degree subject, their highest level of qualification and field. After obtaining this information, teachers' background information was merged with the student-level variables in another excel spreadsheet.

Last, school-level data were obtained from the Ministry of National Education's official website (MoNE, 2017b). A list of all secondary schools was first downloaded from the website of Samsun Provincial Directory of National Education (Samsun Provincial Directory of National Education, 2017). Only schools that were included in the project were retained. School-level data included school type (private or state-funded), school categories (general, regional boarding or vocational secondary school), location of school (urban, suburban or rural), and school service scores.

After all the three data files were merged, to maintain confidentiality, participants' identities were removed from the data set, and identification numbers were assigned to each student, teacher, school, and school location. Data on the excel spreadsheet was then saved in SPSS (Statistical Package for the Social Sciences) for analyses.

Table 5.3 summarises the outcome variables and independent student, teacher/classroom and school-level variables included in this study.

Table 5.3 Summary of Variables Included in the Study

Outcome Variables		The number of correct answers in maths, Turkish, science, history, and English at Grade 8 (2017)
Independent Variables	Student Characteristics	The number of correct answers in maths, Turkish, science, history, and English at Grade 7 (t-1) (2016)
		The number of correct answers in maths, Turkish, science, history, and English at Grade 6 (t-2) (2015)
		Sex (1= female, 0= male)
		Language Learner ID (1= yes, 0= no)
	Teacher/ Classroom Characteristics	Sex (1= female, 0= male)
		Class size
		Percentage of female students in the classroom
		Classroom-level average students' test scores at Grade 7
		Classroom-level average students' test scores at Grade 6
		Number of years of teaching experience (overall)
		Number of years of teaching experience in the current school
		Assignment field (1= if the current teaching field is the same as his/her assignment field, 0= otherwise)
		Graduation field (1= if the teacher's major degree subject is the same as her/his current teaching field, 0= otherwise)
		Terminal degree (1= if the teacher has a master's or higher degree, 0= otherwise)

		Field of the terminal degree (the teacher's field of highest-level qualification is related to her/his current teaching field, not related and unknown)
	School Characteristics	School type (1= state-funded, 0= private)
		School categories (general, regional boarding, and vocational secondary school)
		School locations (rural, suburban and urban)
		School's service scores (1 to 6)
		School-level average students' test scores at Grade 7
		School-level average students' test scores at Grade 6

Before any analysis could be conducted, it is necessary to ensure that the same teachers taught the same students in previous years. However, as is the case with most longitudinal data, adding each prior test score results in a loss in the number of cases that can be used in the estimates. This might be due to students not taking the test or moving out of the province during the testing period. The average loss rate from adding a previous one-year test score in cases where 8th-grade test scores are available in this study is approximately five per cent.

Besides meeting a minimum necessary number of previous years' student test scores in order to estimate value-added performance scores, as one of the aims of this study is to examine the stability of the value-added estimates over the years and in terms of using additional previous test scores, two-lagged test scores (t-2, Grade 6) were also employed in the estimates. In order for the test scores from two years ago (t-2) to be included in the stability estimates, these scores had to be associated with the current teachers; however, the classroom rosters of the two lagged years (2015) are not available in the dataset. All teachers, therefore, could not be directly linked to the whole sample of students in the sixth grade. To identify teachers responsible for changes in students' exam scores in the relevant course since Grade 6, selection criteria, including the schools having only one teacher in a specific teaching subject and the teachers who have at least two years of experiences in those schools, were applied in the stability estimates. The selection criteria caused the total number of teachers to drop to a sub-sample of 151 who were linked to 2,526 students.

The selection criteria also caused the disappearance of sub-categories of some categorical variables that were to be used in the eventual models for each course. For instance, the language learner identity variable, suggested to be included in the eventual models created for science and history teachers, was excluded from the stability estimates, which contained the second lagged test scores, since none of the data to be used in the estimates belongs to students with language learner identity. It is also the same for school categories and terminal degree variables, so these variables could not be included in the models in sub-research questions about the stability of estimates over a two-year period and in terms of using additional previous test scores. Moreover, since two lagged test scores of 7th-grade students were not available in the dataset (e.g., Grade 5), a similar exclusion was applied to teachers' previous effectiveness estimates, where 7th-grade test scores were used as the response variable and an average classroom/school scores in two-year prior (e.g., Grade 5) was requested.

5.3.1 Student Longitudinal Data and Demographic Characteristics

Students in secondary and high schools in Samsun are required to take province-wide exams in order to increase their academic achievement. Students are tested in mathematics, Turkish (language arts), English (as the foreign language), education of religion and ethics, science, and history, starting in Grade 6 until graduating from high school in Grade 12. Due to the lack of obtaining information about the education of religion and ethics teachers (just one school shared the information), test scores in the other five content areas are used in this study. Instead of using only maths and reading exam scores as in most of the studies in this area, the use of students' test scores in various content areas is preferred in this study in order to examine whether contextual predictors have a similar contribution to teacher effectiveness estimates in a variety of teaching subjects.

The mathematics, Turkish, English, science, and history test scores were available, spanning three consecutive years (2015, 2016, and 2017) in this study. The study focused on 8th-grade secondary school students because of the availability of at least one previous year's test scores. It is important to note that the number of correct/wrong answers out of 20 questions in each teaching subject test is available for each student separately, along with their sex and language learner status information. Therefore, the outcome scores in each teaching subject, the number of correct answers in Grade 8 in 2017, were used as the dependent variables of the study. As the value-added models are approaches based on statistically measuring students achievement growth from one year to the next, in addition to outcome test scores, most of the models require at least one previous year’s test scores and other contextual student-, teacher/classroom-, and school-level variables if any (Amrein-Beardsley & Holloway, 2019; Newton et al., 2010; Wei et al., 2012).

Dummy variables were created for the categorical predictors in student-level: sex and language learner status. Boys were grouped into 0, and girls were coded as 1; therefore, boys are the reference variables in sex. Similarly, students who are Turkish language learners were also measured as a binary variable by coding the language learners with 1. Tables 5.4 and 5.5 summarise the student-level variables employed in this study.

Table 5.4 Student Longitudinal Data Used in the Equations

Subject	Grade	Number of students	Mean	Standard Deviation (SD)
Mathematics				
	8	7,543	9.18	4.34
	7	7,230	9.52	4.99
	6	7,186	8.73	4.24
Turkish				
	8	7,594	12.90	4.23
	7	7,353	14.17	4.35
	6	7,228	11.45	4.07
Science				
	8	7,116	12.32	4.88
	7	6,815	12.14	4.33
	6	6,741	9.88	3.96
History				
	8	6,638	12.91	4.98
	7	6,364	10.93	4.44
	6	6,275	10.51	4.73
English				
	8	6,544	10.60	5.03
	7	6,275	9.94	5.05
	6	6,221	10.34	5.45

Table 5.5 Student Demographic Characteristics Used in the Equations

	Number of students	Percentage
Sex (Female)	17,110	48.3
Language Learner Status (LLS)	73	0.2

5.3.2 Teacher/Classroom Characteristics and Average Attainments

Since the teacher-level data file obtained from the school directorates also contained information about the classroom and the school where the teachers work, the data set enabled the students to be fully connected to their teachers in five teaching subjects. The teacher-level data set indicates the teachers' demographic, educational and teaching background information. Table 5.6 and 5.7 summarise the teacher-level independent variables analysed in this study.

Table 5.6 Teacher/Classroom Characteristics Used in the Equations

Teacher/Classroom Characteristics	Mean ($n_{Teacher}$ = 1,027) *($n_{Classroom}$= 1,728)	Standard Deviation (SD)
Class size	20.11	5.96
Percentage of female students in the classroom	0.48	0.17
Total year of teaching experience	10.70	6.35
Experience in the current school	3.50	2.33

** A total of 510 teachers were assigned into multiple classrooms*

Table 5.7 Other Teacher/Classroom Characteristics in Percentage

Teacher/Classroom Characteristics	Number of teachers	Percentage
Sex (Female)	575	56
Teaching assignment subject		
Related to the teaching field	1,022	99.5
Field of bachelor's degree		
Related to the teaching field	994	96.8
Having a master's degree	34	3.3
Field of the terminal degree (out of 34 teachers)		
Related to the teaching field	11	32.3
Not related	8	23.5
Unspecified	15	44.2

The sex of teacher variable is coded into a dichotomous variable; while female teachers were coded as 1, male teachers were coded as 0. Therefore, the male sex was assigned as the reference variable in the estimates. Slightly over half of the teachers were female (56%). "Class size" is a continuous variable and indicates the actual number of students in the class to which the teachers were assigned. Class sizes of the sample ranged from 5 as the minimum class size to 35 as the maximum. It is worth noting that as a total of 510 teachers in five teaching subjects were assigned to multiple classrooms, therefore the number of the classrooms is higher than the number of teachers involved in the study ($n_{Classroom}$= 1,728). Thanks to the availability of individual students' sex and class roster information, the percentages of female students in each classroom were also calculated, and it was revealed that, on average, almost half of the classes consist of female students.

Two types of teacher-level variables about experience were collected from school directorates; while one indicates the number of years teachers have taught in their current school, the other variable refers to the total number of years of teaching experience teachers have in their professional career. For a teacher to be accountable for the change in student attainment, the teacher must affect the students' learning experiences. The teachers should, therefore, be given the opportunity to spend adequate time with their students. Therefore, the information for the teachers who have been teaching for at least one year in their current schools was only requested from school directorates.

It is known that in some periods of the Turkish education system, teacher candidates were assigned to irrelevant teaching fields, regardless of which higher education programs they graduated from. Therefore, such teacher level variable was also available in the data set for estimations. The teaching appointment subject variable was defined as a dummy variable by grouping the teaching appointment areas based on the relationship with their teaching fields according to the national education board (MoNE, 2018). The teaching appointment subjects related to the teaching field received a code of 1, whereas non-related appointment subjects were given a value of zero. Here follows an example to explain creating a dummy variable in maths. There are three different teaching appointment subjects in the study sample among the mathematics teachers: mathematics teaching, elementary mathematics teaching, and primary school teaching. According to the schedule for teaching fields published by the National Education Board (MoNE, 2018), as the appointment subjects of mathematics teaching and elementary mathematics teaching are related to the teaching subject of mathematics, these variables were grouped and coded as one, whereas the appointment subject of primary school teaching was given a value of zero. A similar coding strategy in creating a dummy variable was applied for the variable about fields of bachelor's degree.

Teachers' highest level of qualifications were also denoted by a dummy variable which indicates whether the teachers have a master's or higher degree. In the teacher-level data set, teachers' education background ranged four levels: bachelor's degree, master's degree (non-thesis), master's degree, and PhD (only one teacher has a PhD degree). Teachers having a bachelor's degree were given a value of zero, whereas master or higher education degrees were grouped and coded as one. To indicate whether the fields of the highest level of qualification the teachers had are related to their teaching subjects, three dummy variables were created: "related", "non-related", and "unknown". As there are missing cases in the variable of the field of highest-level of qualification, a new variable named "unknown" was created (details for treatments of missing cases were provided in the related section). The "related" cases were coded as one, whereas others received a code of zero. Likewise, "non-related" cases received a code of one, and others were coded as zero. The variable of "unknown" was appointed as the reference variable in the estimates.

Along with the teacher/classroom characteristics obtained from the school directorates, classroom-level students' average prior test scores in each teaching subject were also calculated (see Table 5.8). Comparison of classroom-level students' average prior test scores shows that while students in Turkish classrooms had the highest score on average, the maths averages of the classes were the lowest in both previous grades. On the other hand, the most varied course in class-level average students' attainemment was English (SD=2.72 and 2.93 in Grade 7 and 6, respectively).

Table 5.8 Students' Classroom-level Average Attainments

Classroom-level average test scores		Mean	Standard Deviation (SD)
Mathematics	Grade 7	9.48	1.89
	Grade 6	8.65	2.31
Turkish	Grade 7	13.96	2.02
	Grade 6	11.36	1.95
Science	Grade 7	12.10	2.15
	Grade 6	9.82	2.08
History	Grade 7	10.90	2.17
	Grade 6	10.43	2.39
English	Grade 7	9.89	2.72
	Grade 6	10.26	2.93

5.3.3 School Demographic Characteristics and Average Attainments

The last set of variables contains information on the demographic characteristics of the students' schools: the name of school districts, school names, school type, school category, service score, location and school-level students' average test scores. All secondary schools' names in the province were downloaded from the official website of Samsun Provincial Directory of National Education (Samsun Provincial Directory of National Education, 2017) and compared with the list of the schools involved in the project. The names of the schools not involved in this project were deleted from the data set. After this elimination, a total of 1,410 schools remained in the dataset. Unfortunately, because of the lack of teacher characteristics in the administrative dataset and the reluctance of many school directorates to share the related information with the researcher, the total number of schools involved in this project dropped to 695. The demographic characteristics of the 695 schools were then obtained from the Ministry of National Education's official website (MoNE, 2017b) (see Table 5.9). At the next stage, all levels of variables were merged in an excel spreadsheet and transformed into SPSS format. After completing the merging process for the data at all levels, all information about the participants' identity in this study were deleted from the data set. To maintain confidentiality, identification numbers were assigned to each student, teacher, school, and school districts.

Table 5.9 School Characteristics Used in the Equations

School Characteristics		Number of schools	Percentage
School Type			
	State-funded	695	100
School Category			
	General	619	85.7
	Regional Boarding	21	2.8
	Vocational	55	11.5
Service Score			
	1 (highest)	244	35.1
	2	91	13.1
	3	69	9.9
	4	220	31.7
	5	57	8.2
	6 (lowest)	14	2.0
School Location			
	Urban	273	39.3
	Sub-urban	175	25.2
	Rural	247	35.5

The name of the district and of the school were collected just to prevent the double cases in the dataset and to link schools to teachers and students precisely. Therefore, this information was not used in the estimations. The school type was intended to be coded as a dummy variable, but as none of the private schools shared the requested information about their teachers, the school type variable was revoked in the estimates. School categorisation was denoted by three dummy variables that indicate the categories of the schools: general, regional boarding, and vocational secondary schools. When regional boarding schools received a code of 1, other school categories were coded as zero. Similarly, vocational secondary schools were coded as one, whereas other categories were coded as zero. The variable of general secondary school was chosen as the reference variable in the estimates.

Three service regions have been constituted by the Ministry of National Education by grouping the provinces that are similar in terms of the number of teachers needed, geographical location, economic and social development level, transportation conditions, and meeting the service requirements (MoNE, 2017c). According to this schedule, the province of Samsun is located in the first service zone. In addition to these service regions, similar schools are also grouped into six service areas and given a service score range from 1 (highest score) to 6 (lowest score) in terms of their difficulties in the appointment and employment of teachers and the facilities they have.

The school location variable contains three options: urban, suburban, and rural areas. Therefore, the variable was coded into three dichotomous variables. When the schools in the suburban area were coded as one, other school locations received a code of zero. Likewise, rural school locations were given a value of one, whereas schools in urban and suburban areas were coded as zero. The variable of "urban" was assigned as the reference variable for school location in the estimates.

Lastly, school-level students' prior average test scores in each teaching subject were also computed in order to be employed as school-level predictors in the analyses (see in Table 5.10). Comparison of school-level students' average prior test scores shows that students' school-level scores in Turkish were the highest, while the maths averages of the school were the lowest in both previous grades. On the other hand, the most varied course in school-level average students' attainment was again English (SD=2.13 and 2.37 in Grade 7 and 6, respectively).

Table 5.10 Students' School-level Average Attainments

School-level average test scores		Mean	Standard Deviation (SD)
Mathematics	Grade 7	9.48	2.20
	Grade 6	8.66	1.84
Turkish	Grade 7	13.99	1.66
	Grade 6	11.36	1.66
Science	Grade 7	12.09	1.68
	Grade 6	9.81	1.63
History	Grade 7	10.90	1.74
	Grade 6	10.44	1.88
English	Grade 7	9.92	2.13
	Grade 6	10.20	2.37

5.4 Treatment of Missing Values

The longitudinal dataset contains 35,435 students linked to 1,027 teachers in 695 secondary schools. After eliminating the cases that could not be linked to teachers in the dataset, the remaining dataset still contained missing data. As in other studies, the issue of missing data is important for this research. Instead of using the listwise deletion method, as the missing cases cause a high attrition rate and have the potential for bias in the estimates (Gorard, 2016, 2015 and 2014b), missing data was manipulated by applying different methods based on missing data types in order to use as much data as possible in the estimates.

Although there were no missing cases in the variables of student's sex and language learner status, missing values increased in the records of the students' previous years. The missing values of real numbers, which are the test scores in Grades 7 and 6, were replaced with the overall mean score in each teaching subject in the related year (Gorard, 2020). Table 5.11 illustrates the comparison of the samples used in the analyses in each teaching subject and grade. Comparison of means of the unrestricted and the restricted sample shows that there are very small differences between them; the average test scores in all teaching subjects are slightly below in each grade from the overall means of the unrestricted sample.

Table 5.11 Comparison of the Samples Used in the Analyses

	Number of cases (Restricted Sample) Mathematics Turkish Science History English	**Std. Deviation (Restricted Sample)** Mathematics Turkish Science History English	**Mean (Restricted Sample)** Mathematics Turkish Science History English	**Overall Mean (Unrestricted Sample)** Mathematics Turkish Science History English
Grade8	7,543	4.35	9.18	9.33
	7,594	4.23	12.90	12.70
	7,116	4.88	12.32	12.29
	6,638	4.99	12.91	13.01
	6,544	5.03	10.60	10.81
Grade7	7,230	4.99	9.52	9.77
	7,353	4.35	14.17	13.79
	6,815	4.34	12.14	12.14
	6,364	4.44	10.93	10.95
	6,275	5.05	9.94	10.28
Grade6	7,186	4.24	8.72	8.89
	7,228	4.07	11.45	11.63
	6,741	3.96	9.88	10.02
	6,275	4.73	10.51	10.51
	6,221	5.45	10.34	10.53

As all school information was downloaded from the Ministry of National Education's official website, there is no missing data in the school-level data set. Most of the cases in the teacher-level dataset had even no missing data. Since there were some missing data in the variable of the field of highest-level qualification, a further category was needed to indicate the missing cases in that variable (Gorard, 2020). Although 1,194 cases were associated with teachers who had a master/higher degree in the dataset, only 636 cases indicated that the fields of master/higher education degree their teacher earned are associated with their teaching fields or not. All missing values were coded as "unknown".

Lastly, although the p-value and confidence interval of the statistical significance tests, such as the t-test or chi-square etc., is still widely reported in the social sciences, as this current study involves a non-random sample from the study population and the sample also has missing data, the main assumption of reporting the p-value is not met (Figueiredo Filho et al., 2013; Gorard 2018b); therefore, p-values of the significance tests used were not reported in this study. The issue of what the p-value and confidence interval of the significance testing actually tell us was widely discussed by Gorard (2019, 2016, 2014b), Greenland et al. (2016), Cohen (1994) and White & Gorard (2017).

5.5 Data Analysis

5.5.1 Data Analysis for Sub-Research Question 1

How stable are teacher effectiveness measured by VAMs that consider student, school, and teacher-classroom characteristics?

An answer to this sub-research question will be sought in three steps hierarchically. First, the contribution of using student characteristics in models to predictions of teacher effectiveness will be examined, then school characteristics, and finally teacher/classroom characteristics.

5.1.1.1 Stability of teacher value-added estimates using student characteristics

To determine the contribution of student characteristics to teachers' value-added estimates, multiple linear regression analysis with the *forward* method of entry was conducted. By using the *forward* method, whether adding new predictor(s) causes the noticeable improvement in the model fit compared to the model created in the previous step can be tested. Instead of including all predictors at the same time in the basic model to obtain a model with the highest predictive ability (having the largest R-squared value), the aim is to create the best-fit regression model with the largest R-squared by including as few variables as possible (which have predictive power on the value-added estimates).

Before revealing the contribution of students' characteristics to VAM estimates, whether there is any relationship between these characteristics and their current test score was checked. The relationship with students' current test scores was revealed by using Pearson's r coefficients for prior attainment (Grade 7) and by using Cohen's effect sizes for sex and language learner status variables (which was calculated by dividing the difference between the averages for each category of these variables by their overall standard deviation).

After determining the relationship between student characteristics and their current test scores, a basic model was, first, created using only a prior attainment score (t-1) to find out how much improvement will be achieved on the model fit when employing the student characteristics. Then, to find the highest R-squared value that a model can have by employing all the predictors, both sex and language learner identity variables were added to the basic models for each teaching subject with the *enter* method; thus, the highest R-squared value that can be obtained in this sub research question was determined. Finally, the same student characteristics variables were again included in the basic models by the *forward* method. When a proposed model reached the largest R-squared value that can be achieved, the model(s) proposed in the next step(s), if any, was not considered in order to keep only variables that have predictive power on the estimates in the best-fit regression model. The variables that were excluded as they have no predictive ability or are too small to be considered from the estimates in this analysis will not be included in the next analyses.

5.1.1.2 Stability of teacher value-added estimates that include school characteristics

The individual student residual scores obtained from the final models created in the previous section were aggregated at teacher level, and the class averages of these teacher-level residuals were tentatively attributed to teachers' individual value-added effectiveness scores. To examine the relationship between school-level variables and the teachers VAM scores, Pearson's correlation coefficients were calculated for school service scores, school average test scores in Grades 7 and 6, and Cohen's effect size were calculated for each sub-category of the variables of school categories and locations.

After revealing the relationship between school characteristics and teacher VAM scores, to determine which school characteristics make a notable contribution to teachers' value-added estimates, the models obtained for each teaching subject in the previous section were used as the baseline models in the analyses of this section. Next, all five school-level variables, school categories, service scores, locations, and school-level average test scores in Grades 7 and 6, were included in the baseline models with the *enter* method, whereby the highest R-squared values can be obtained at this stage were revealed. Finally, the same school-level independent variables were again included in the baseline models by the *forward* method. As the school-level dataset contained two categorical variables that have three sub-categories, three dummy variables were created for these two categorical variables, which are the school locations and the school categories variables. While the variable of general secondary school was chosen as the reference variable among the school category variables, the location variable of urban was assigned as the reference variable for school location in the regression estimates.

Like the selection strategy applied in the previous section, in order to keep only variables that have predictive power on the estimates in the best-fit regression model, once a model with the largest R-squared value was obtained by the *forward* method of entry, the model(s) suggested in the next step(s) (if any) was ignored. The excluded variables, as they have not contributed or are too small to be considered on the estimates in this current section, will also not be included in further analyses.

5.1.1.3 Stability of VAMs in teacher effectiveness estimates that include teacher/classroom characteristics

To determine the contribution of teacher/classroom characteristics to teachers' value-added estimates, multiple linear regression analysis was carried out with the *forward* method of entry. Again, before conducting the regression analyses, whether there is any relationship between the teacher/classroom characteristics and their value-added effectiveness scores was checked. Individual student residual scores (the difference between predicted and actual attainment level) obtained through the final model proposed in the previous section were aggregated at the teacher level. The mean of the residuals at teacher level was tentatively attributed to a teacher's individual value-added effectiveness score. Then the effectiveness scores were correlated with the teacher/classroom characteristics. Pearson's r coefficients were calculated by correlating real-number variables with the teachers' effectiveness scores, and Cohen's effect sizes for categorical variables were also calculated by dividing the difference between the averages for each category of these variables by their overall standard deviation.

After conducting the correlation statistics, the full regression models for each subject were developed by adding all teacher/classroom level variables with the *enter* method over and above the student and school characteristics identified in the previous section (baseline model), and thus, the largest R-squared value that a model can have at teacher/classroom-level was determined. To create a best-fit regression model with the largest R-squared that can be achieved employing as few variables as possible, the same teacher/classroom-level independent variables were included in the baseline models (revealed in the previous section) with the *forward* method. Once a model with the largest R-squared value was obtained, in order to include only variables with predictive power in the final model, the model(s) proposed in the next step(s) was ignored, if any. As the excluded teacher/classroom characteristics because of not contribute (or are too small to be considered) to the estimates in this current section, they will also not be included in the next analyses.

5.5.2 Data Analysis for Sub-Research Question 2

How stable are teacher value-added effectiveness estimates over a two-year period of time?

To determine to what extent teacher value-added effectiveness estimates is stable over two years, current and previous value-added effectiveness scores of the same teacher were estimated by conducting multiple linear regression analyses. For current effectiveness estimates, the 8th-grade student test scores in the related teaching subject were regressed on the same students' prior attainment scores in Grade 7 and other predictors determined for each teaching subject in sub-research question 1. Student level residuals obtained from the estimates were saved for use in value-added teacher effectiveness estimates. Similarly, for the previous effectiveness estimates of the same teachers, the 7th-grade student test scores in the related teaching subject were regressed on the prior attainment scores in Grade 6 and the same predictors determined for each course in the previous sub research question, and again residuals at student level were also saved. In the next step, the individual student residuals obtained through the final models for current and previous effectiveness estimates were aggregated at the teacher level. The means of the residuals at teacher level were tentatively attributed to a teacher's current and previous value-added effectiveness scores. Finally, Pearson's r coefficients were calculated by correlating the teachers' current and previous effectiveness scores to find out how consistent teacher value-added effectiveness estimates is over a two-year period. The correlation results were presented for each teaching subject separately.

5.5.3 Data Analysis for Sub-Research Question 3

How stable are teacher value-added effectiveness estimates when including an additional prior score (t-2)?

To find out how consistent teacher value-added effectiveness estimates can be achieved by adding additional prior attainment scores to the final models stated for each teaching subject in Table 11.17, teacher effectiveness estimates from the eventual models for each course (determined in the last section of sub-RQ1) were correlated with the corresponding ones derived by adding two lagged test scores (t-2). 8th-grade students' test scores in the related teaching subject were regressed on the same students' one lagged test scores (t-1, Grade 7) and other independent variables identified for each course in sub-RQ1, and then student-level residuals were aggregated at the teacher level. The means of the residuals at teacher level were tentatively attributed to a teacher's value-added effectiveness scores. To obtain the corresponding teacher value-added scores, multiple regression estimates were carried out by adding additional lagged test scores (t-2, Grade 6) to the same predictors for each teaching subject. Finally, correlation coefficients were calculated separately for each teaching subject by correlating the teachers' actual and the corresponding effectiveness scores. The effectiveness scores were also grouped into quartiles to reveal how consistently teachers remained in the effectiveness categories that were assigned in each estimate.

5.5.4 Data Analysis for Sub-Research Question 4

Do different methods of analyses used in VAMs produce consistent teacher effectiveness estimates?

This sub-research question was formulated to test whether the choice of a specific modelling approach influences the teachers' value-added effectiveness estimates. More specifically, in order to investigate whether similar effectiveness estimates can be obtained across different models, the Ordinary Least Square (OLS)-based multiple regression model, which used in all previous sub-research questions, was compared with a more simplistic model, the residual-gain model, and a more sophisticated model, two-level HLM (Hierarchical Linear Model). There are concerns that as the multiple regression method ignores the multi-level data structure, it might produce misleading results; therefore, the most common statistical approach, the OLS-based model, was compared to a complicated multi-level statistical approach. On the other hand, there is another debate about if a simpler model might produce a similar result for teachers, why more complex models are chosen. The sub-RQ4 has the purpose of investigating the concerns raised on both sides.

This sub-research question used a cohort of eighth-grade maths, Turkish, science, history, and English language teachers to compare the consistency of their value-added estimates derived from three common statistical approaches. The first statistical approach compared is the residual gain model (hereafter to be expressed as RG), also called the covariate-adjusted model in some research. RG is a linear regression-based model that takes into account students' prior attainments when predicting their current attainments. While in some research, students' background characteristics are included in the RG as covariates, in this study, in order to make a better comparison with OLS based-multiple regression model, only the prior test score was included in the RG model. The RG model can be formulated as follows:

$$y_{it} = \beta_0 + \beta_1 y_{it-1} + \varepsilon_{it} \qquad (1)$$

Where y_{it} is the student's current year test score, y_{it-1} is their prior year's score, β_0 is an intercept, β_1 is a coefficient of prior attainment predictor, and ε_{it} is the residual score for the ith student. In this model, prior test scores were only included as a covariate to predict students current test scores. After estimating residuals for each individual student, these individual residual scores were aggregated at teacher-level. The classroom averages of the teacher-level aggregated residuals were tentatively attributed to teachers' individual value-added effectiveness score.

$$T_j = \frac{\sum_{i=1}^{n} \varepsilon_i}{n} \tag{2}$$

Where T_j is the teacher j's value-added estimate, ε_i is the difference between observed and predicted scores of individual students belonging to teacher j, and n is the number of students in the classroom to which the teacher is assigned.

The second statistical approach is OLS-based multiple regression (hereafter to be expressed as OLS). In the OLS model, 8th-grade students' performance are predicted by controlling their prior attainment and the contextual characteristics depicted in Tables 13.1 and 13.2 for each teaching subject. The OLS model is the main statistical approach used in this study, and teacher effectiveness estimates generated from OLS were compared with the estimates derived from the other two approaches. The OLS-based multiple regression model can be specified as follows:

$$y_{it} = \beta_0 + \beta_1 y_{it-1} + \beta_2 X_{it} + \beta_3 T_{ijmt} + \beta_4 S_{imt} + \varepsilon_{it} \tag{3}$$

Where y_{it} is the student's current year test score, y_{it-1} is their prior attainment, β_0 is an intercept, β_1 ... β_4 are estimated parameters of each variable, X_{it}, T_{ijmt} and S_{imt} are vectors of student, teacher and school characteristics, and ε_{it} is the individual student residual score. The subscripts are used to indicate students (i), teachers (j), schools (m), and time (t). Student-level residuals were calculated by subtracting the actual test score (y_{it}) from the predicted test scores $(\hat{y}_{it})$, which were estimated by employing the contextual variables at the student-, school-, and teacher/classroom level as shown in Tables 13.1 and 13.2 for each teaching subject, along with using students' prior attainments. Finally, the teachers' individual value-added effectiveness scores were calculated using the same aggregation method as in equation 2.

The last statistical approach is two-level HLM (hereafter to be expressed as HLM). Although the hierarchical structure of the data set allows for conducting three-level analyses, since most of the school level variables were not included in models created for some teaching subjects, these limited number of school-level variables were situated to the models at the same level as the second level variables. Therefore, in this sub-research question, the lowest level of hierarchical data (level-1), where the student-related variables, such as prior attainment and sex are employed, were nested within level-2 teacher/classroom (and school) related variables. The student-level (level-1) model is formulated as:

$$y_{ij} = \beta_{0j} + \beta_1 X_{1ij} + \cdots + \beta_{\mathrm{p}} X_{\mathrm{p}ij} + \varepsilon_{ij} \tag{4}$$

Where y_{ij} is the student's current year test score for student i, within teacher j, $X_{1ij} + \cdots + X_{\mathrm{p}ij}$ are p predictors at student-level, including prior attainment. β_0 is a level-1 intercept and $\beta_1 + \cdots + \beta_{\mathrm{p}}$ are the regression coefficients of student-level predictors, and ε_{it} is the residuals in the level-1 equation. The teacher-level (level-2) model is specified as

$$\beta_{oj} = \gamma_{00} + \gamma_{01} T_{1j} + \ldots + \gamma_{0\mathrm{p}} T_{\mathrm{p}j} + u_{0j} \tag{5}$$

$$\beta_{1j} = \gamma_{10}$$

$$\cdots$$

$$\beta_{pj} = \gamma_{p0}$$

Where γ_{00} is the intercept for the level-2 equation, $T_{1ij} + \cdots + T_{\mathrm{p}ij}$ are p predictors at teacher/classroom-level, $\gamma_{01} + \cdots + \gamma_{0\mathrm{p}}$ are the regression coefficients of teacher/classroom-level predictors, $\gamma_{10} + \cdots + \gamma_{\mathrm{p}0}$ are the constant values, and u_{0j} is the random error component at the level-2 equation. Again, the individual level residuals were aggregated at teacher-level, and the teacher's individual value-added effectiveness scores were calculated using the same aggregation method as in equation 2.

To investigate the consistency of teacher value-added effectiveness estimates generated from the three statistical approaches, the estimation result obtained from one approach was compared to the corresponding estimation result obtained from another approach by using Pearson's correlation analysis. The strength and direction of the relationship between the results of paired models indicated the degree of concordance (or discordance) of the approaches. In addition to Pearson's correlation analyses between VAM raw estimates, it was also investigated whether teacher rankings and effectiveness classifications are consistent between the statistical approaches. As one of the main purposes of many education departments in implementing VAMs as a measure of the accountability for changes in student academic attainment is to rank and/or classify teachers, the statistical approaches were also compared in terms of their ranking and classification capabilities. The rankings of individual teachers related to their effectiveness raw scores generated from each approach were compared by Spearman rank correlation analysis. In addition, the effectiveness scores were also grouped into four effectiveness categories by dividing them into quartiles: highly effective, effective, partially effective, and ineffective. The consistency in the categorical assignment of the teachers was also examined by a transition matrix, which determines the percentage of teachers that remained or changed in their effectiveness categories from one approach to another one.

Finally, in order to investigate to what extent each model is intrinsically consistent, the value-added effectiveness scores of teachers assigned to more than one class at the same school were compared with each other. For this analysis, the individual student level residuals were aggregated at classroom level, but only for teachers who taught multiple classrooms. In cases where a teacher taught the same subject in different classrooms in the same school year, it is expected that the teacher is expected to have similar effectiveness scores in each classroom if the model used is able to isolate the teacher’s effect on students’ achievement from other factors outside the teacher's control (which is argued that VAMs can achieve this). For this intrinsic consistency analysis between classrooms, it was determined that there is a total of 510 teachers who taught in multiple classrooms. The dataset involved some teachers who were assigned to more than two classrooms; therefore, in order to make a comparison between pairs of classrooms, all data were converted as pairs, for instance, where a teacher has three classes, three pairs were created, class A and B, class A and C, and class B and C. The consistency analyses were examined by another transaction matrix, which determines the percentage of teachers that remained or changed in their effectiveness categories from one class to another. Moreover, to find out the normal range of results for the same teacher for different classes, the standard deviation (SD) was calculated for all classroom pairs.

SECTION III
RESULTS OF THE SYSTEMATIC REVIEW

This section consists of four chapters. Chapter 6 presents the initial search outcomes, phase I and II screening results, quality appraisal, and the general characteristics of the included studies. Chapter 7 presents the results of the systematic review of studies to determine the consistency of VAMs that include student/teacher and school characteristics. Chapter 8 reviews studies that look at the consistency of value-added teacher effectiveness estimates based on the number of previous years' test scores employed. Chapter 9 considers those studies in the review that examine the consistency of VAMs using different data analysis methods.

CHAPTER 6
RESULTS FROM THE DATABASE SEARCH

6.1 Search Outcomes

A total of 17 databases/search engines, including Google Scholar, were searched. In addition, some studies were identified through personal contacts and hand-searching through references of known studies. Thus, a total of 1,439 articles were initially identified. All results obtained from each source were exported to EndNote, a reference manager software. Of these, 492 cases were flagged as duplicates by the system and were thus deleted. A further 175 duplicated cases were identified by the researcher during the screening process. These were also removed. In total, 667 cases were duplicates, retaining 772 cases.

After the duplicated cases were removed from the list, 309 and 423 studies were excluded from the review list in phase I and II screening processes, respectively. The final number of studies included in this review for analysis was 50. These studies focused on the stability of teacher effectiveness measurement estimates by VAMs.

The PRISMA flow diagram (Figure 6.1) tracks the number of identified, included, and excluded studies (or records) at the different phases of the systematic review.

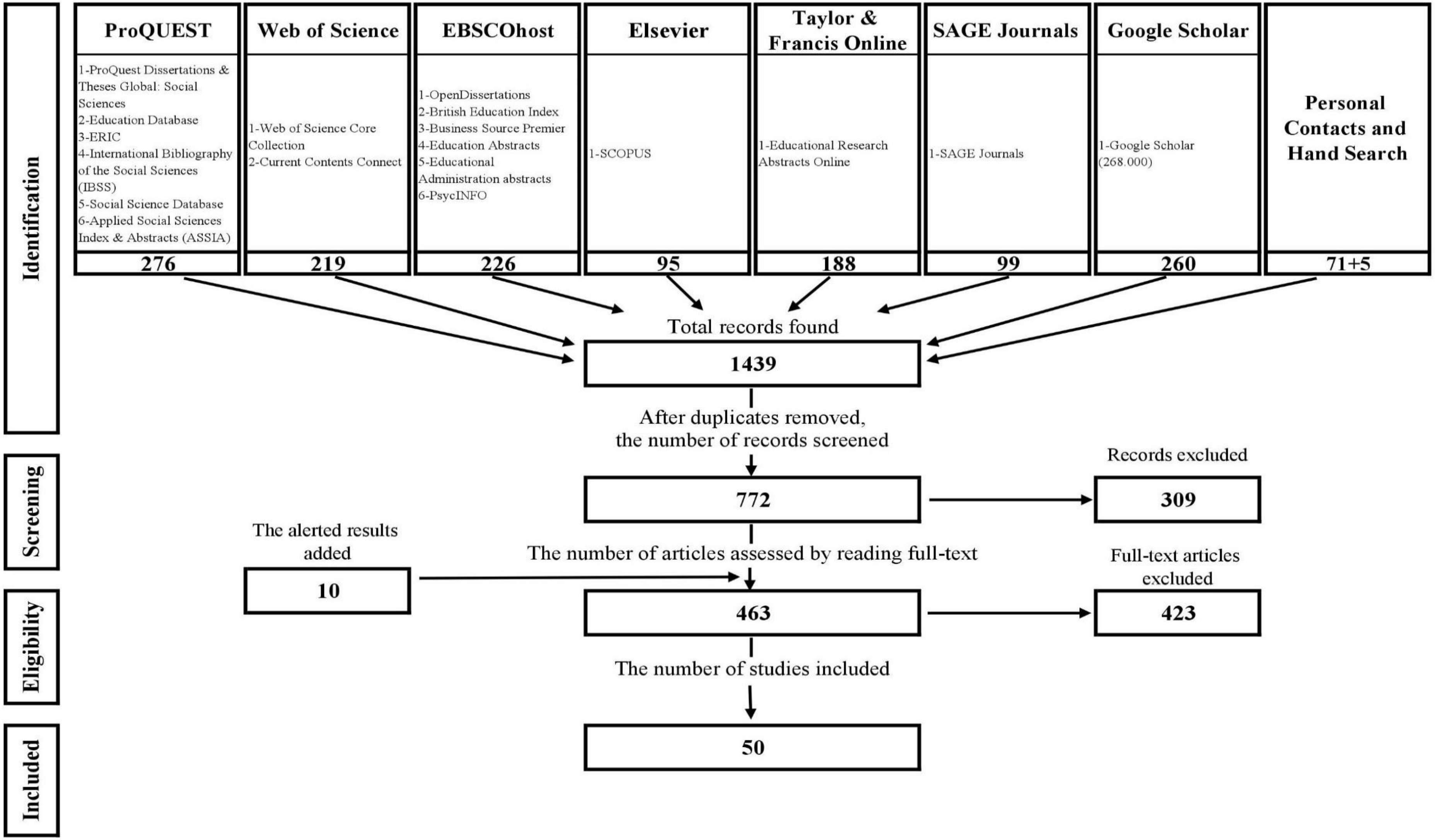

Identification
Screening
Eligibility
Included
ProQUEST
1-ProQuest Dissertations & Theses Global: Social Sciences
2-Education Database
3-ERIC
4-International Bibliography of the Social Sciences (IBSS)
5-Social Science Database
6-Applied Social Sciences Index & Abstracts (ASSIA)
276
Web of Science
1-Web of Science Core Collection
2-Current Contents Connect
219
EBSCOhost
1-OpenDissertations
2-British Education Index
3-Business Source Premier
4-Education Abstracts
5-Educational Administration abstracts
6-PsycINFO
226
Elsevier
1-SCOPUS
95
Taylor & Francis Online
1-Educational Research Abstracts Online
188
SAGE Journals
1-SAGE Journals
99
Google Scholar
1-Google Scholar (268.000)
260
Personal Contacts and Hand Search
71+5
Total records found
1439
After duplicates removed, the number of records screened
772
Records excluded
309
The alerted results added
10
The number of articles assessed by reading full-text
463
Full-text articles excluded
423
The number of studies included
50

Figure 6.1 The PRISMA flow diagram summarising the review process

6.2 The Results of Phase I Screening

In Phase I, the 772 studies that were retained after duplicates were removed were screened by titles and abstracts for relevance to the research question and whether they met the inclusion criteria.

Although a filter was applied in the search engines to restrict the search to contain studies only written in English, there were three studies that were not in English. These three were removed. Another 62 cases (20%) were removed because they did not relate to education. Most of these were related to health studies. A further 40 studies (13%) were removed as they were researched in higher education contexts and thus did not meet the review criteria that was about the K-12 school setting. Thirty-eight (12%) studies were excluded as these were not deemed primary or empirical research. A large majority (166 studies or 54%) were eliminated because they were not about teacher effectiveness. Most of these were about school effectiveness, program/curriculum effectiveness, and principal effectiveness. All in all, a total of 309 study reports were eliminated in this first phase just from screening the titles and abstracts, retaining 463 that proceeded to Phase II. Figure 6.2 is a summary of the number of results eliminated at Phase I screening and their reasons for elimination.

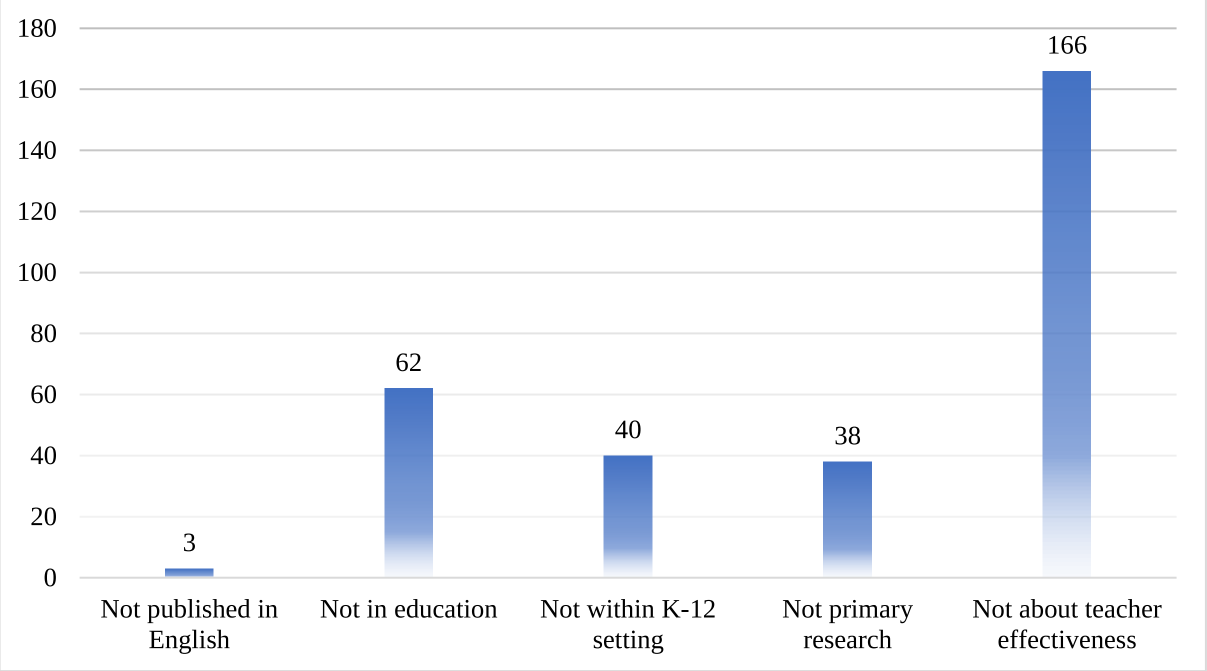


Figure 6.2 The number of results eliminated at Phase I screening

6.3 The Results of Phase II Screening

After the title and abstract screening process was completed, the full paper of the 463 retained studies was read at Phase II. Each study was assessed on 10 criteria (see Appendix C for details of the criteria checklist).

As explained in the method chapter, beginning of the search process, search alerts were set up in each database until May 1, 2019; therefore, a further ten studies were identified from the alerts and included in the total screened studies at Phase II. These were also put through the screening process. Of these, 61 (or 14%) were excluded because it was clear from the full text that they were about theories or policies, not primary research. Three studies (1%) were excluded because they were not conducted in mainstream schools. These three studies focused on students with special educational needs and their teachers. An additional 10 (2%) studies were removed when it was clear from the full text that they were about school effectiveness rather than teacher effectiveness. A large number (n= 151 or 36%) were eliminated because they did not use students' test or gain scores in teacher performance evaluation. Examples of such studies included those that used classroom observations and principal ratings to predict student test scores. Around 44% or 187 research reports on teacher effectiveness were excluded as they were not about the stability of teacher effectiveness estimates. Another 11 (3%) studies were excluded from the review list because their stability estimates were not based on the number of previous years' test scores employed. For instance, Bessolo (2013) investigated the stability of teacher effects on student math and reading achievement by correlating the students' results in one year with their results the following year. A total of 423 were removed at Phase 2, retaining 50 for synthesis. A summary of the number of results eliminated at Phase II is shown in Figure 6.3.

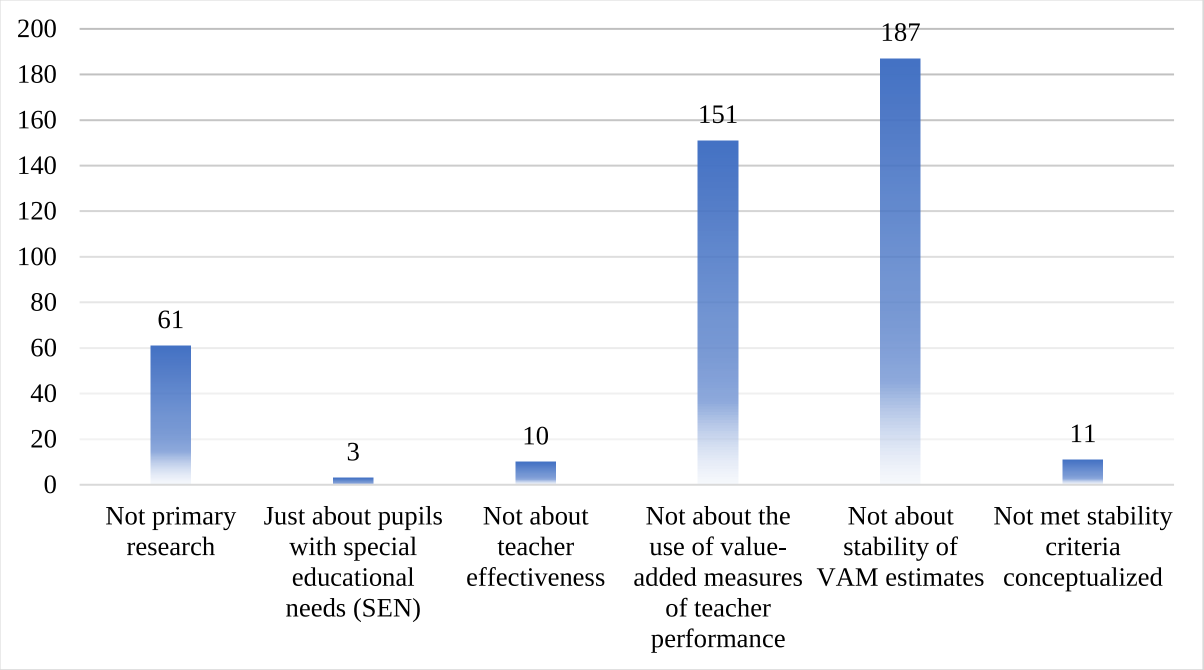


Figure 6.3 The number of results eliminated at Phase II screening

6.4 Inter-rater Reliability

To ensure that the screening processes were undertaken without prejudice by the researcher and to minimise the number of potentially relevant articles being discarded, 70 papers (roughly 10% of the 772 studies) were randomly selected and screened by a second independent reviewer, the second supervisor of the researcher. The selected sample references were used to estimate how much agreement was reached between the reviewers. This was to establish inter-rater reliability (IRR). IRR was estimated using Cohen's Kappa statistics (k). The values of Cohen's kappa range between 0 and +1. The closest value to +1 indicates high agreement between raters.

Table 6.1 Crosstabulation for Inter-rater Reliability Assessment

Inter-rater Reliability Assessment				
		The Second Reviewer		**Total**
		Include	Exclude	
The Researcher	Include	5	1	**6**
	Exclude	2	62	**64**
Total		**7**	**63**	**70**

Of the 70 articles reviewed, the reviewers reached an agreement on 67 references (5+62; see Table 6.1). This means that the two reviewers were in 95 per cent agreement ($\frac{67}{70}x100$).

Cohen's kappa coefficient estimated via SPSS for inter-rater reliability agreement was 0.75. According to the guideless suggested by McHugh (2012), the value of Kappa between 0.80 and 0.90 denotes strong agreement between the raters. One possible reason for the level of disagreement between raters might be that the inclusion-exclusion criteria were not clearly defined enough for the second reviewer. A discussion between the two reviewers was carried out to clarify the inclusion-exclusion criteria, and a consensus was reached.

6.5 Quality Appraisal

The final number of studies included in this review for analysis was 50. These studies focused on the stability of teacher effectiveness estimates based on VAMs. Before synthesising the existing evidence, each study was individually scrutinised to rate its weight of evidence. As stated in the methodology chapter, the quality appraisal of the papers was assessed using the "sieve" designed by Gorard (2014a).

A padlock rating system representing the security of the evidence is used to rate each study. In this review, studies that were comparative or correlational in design, large scale with low attrition, and allowed random student-teacher allocation were rated with a 4-padlock rating which represents the most secure evidence or most trustworthy finding (see in Table 6.2).

Table 6.2 A Summary of Study Ratings

Rating	Number of studies
4	13
3	33
2	2
1	2
Total	**50**

Since almost all of the studies retrieved were large scale comparative/correlational study with low attrition, most of these studies were rated 3🔒 out of 4, as they used administrative/panel data where students were not randomly assigned to teachers in value-added teacher effectiveness estimates

To ensure inter-rater reliability of the appraisal quality process, a randomly selected eight studies were also rated by a second-rater. First, four studies were sent to be evaluated by the second-rater, and any disparities in rating were discussed to reach a consensus. After reaching an agreement with the second-rater, all studies were revised with regards to the second rater's feedback (see Appendix E for quality appraisal of all studies).

6.6 Characteristics of the Included Studies

Studies included in the review were published over a range of 20 years, between 1999 and 2019. Only one study, which was a doctorate thesis, was dated 1999. Value-added models as teacher performance appraisal tool for school accountability systems were developed in the 1900s by American statistician William Sanders. Their implementations expanded throughout all states in the US with the No Child Left Behind Act (NCLB) in 2002, which required states to test all students in third to eighth-grade levels to receive federal school funding. Therefore, school districts had a growing amount of longitudinal test scores, which are basic requirements for value-added analysis. That might be a reason for the increase in the number of studies conducted after 2000. Around thirty per cent (n= 14) of the studies were conducted between 2000 and 2009.

Half of the studies were published between 2010 and 2015. This is the period where developments in the school accountability system intensified, a consequence of the grant scheme, Race to the Top (RTT), introduced by the US former president Barack Obama in 2009. Therefore, it is not surprising that 52% of the studies were conducted in this period. Nine of the most recent papers published in 2019 came from alerts set on search engine providers.

It is clear that most of the research in this field is dominated by US researchers. These constituted 94% of all the studies included in this review. Only three studies were conducted outside of the United States. One was conducted in India, one in Australia and one in the United Kingdom. Again, a possible main reason for such a small number of studies done outside of the United States may be that VAMs require extensive longitudinal data on individuals and that such data are very difficult to obtain.

Table 6.3 Descriptive Statistics of Eligible Study Characteristics

	Number of Studies	Percentage (%)
`Publication Year		
Before 2000	1	2.0
Between 2000 and 2009	14	28.0
Between 2010 and 2015	26	52.0
After 2015	9	18.0
Country of Study Location		
United States	47	94.0
United Kingdom	1	2.0
Australia	1	2.0
India	1	2.0
Type of Publication		
Journal Article	28	56.0
Dissertation or Thesis	15	30.0
Working Paper or Report	7	14.0
Search Databases		
Education Database (ProQuest)	2	4.0
ERIC (ProQuest)	5	10.0
Social Science Database (ProQuest)	2	4.0
Applied Social Science Index &Abstracts (ProQuest)	2	4.0
ProQuest Dissertations & Theses A&I (ProQuest)	15	30.0
Web of Science Core Collection (Web of Science)	1	2.0
Educational Research Abstracts Online (Taylor&Francis)	5	10.0
Sage Journals (SAGE)	5	10.0
Scopus (Elsevier)	5	10.0
Google Scholar (Google)	7	14.0
CALDER (Hand Search)	1	2.0
Type of Research		
Longitudinal study	25	50.0
Longitudinal comparison study	23	46.0
Causal comparative research (with using longitudinal data)	1	2.0
Mixed factorial design (with using longitudinal data)	1	2.0

Sample Size (total)		
Limited Sampling Size < 2000 students or 200 teachers	4	8.0
Sufficient Sampling Size ≥ 2000 students or 200 teachers	46	92.0
Sample Size (comparison group)		
Small Number of Cases (Below 1000 students or 50 teachers)	2	4.0
Medium Number of Cases (Between 1000 and 2000 students, or 50 and 100 teachers)	8	16.0
Large Number of Cases (Over 2000 students or 100 teachers)	40	80.0
Study Setting		
Elementary School	21	42.0
Elementary and Middle Schools	17	34.0
Middle School	3	6.0
High School	5	10.0
Middle and High Schools	1	2.0
Elementary, Middle and High Schools	1	2.0
Not Reported (Simulated Data)	2	4.0
The Subject Area of Outcome		
Only mathematics test scores	13	26.0
Only reading test scores	4	8.0
Mathematics and readings test scores	16	32.0
Mathematics and English language art (ELA) test scores	6	12.0
Mathematics and communication art test scores	1	2.0
At least three subject areas' test scores	7	14.0
Not Reported (Simulated Data)	3	6.0
Attrition Rates		
Minimal missing data (Up to 19%)	26	52.0
Some missing data (Between 20% and 39%)	14	28.0

Moderate missing data (Between 40% and 59%)	6	12.0
High level of missing data (Between 60% and 79%)	2	4.0
Huge amount of missing data (Over 80% or not reported)	2	4.0

Of the 50 eligible studies, slightly more than half (56%, n= 28) were journal articles. All the journal articles would apparently have been peer-reviewed by experts in the subject area. Table 6.4 shows the journals and the databases with the number of included studies. Most of the studies included in this review were published in economics of education and finance journals, e.g., Economics of Education Review, Education Finance and Policy and Statistics and Public Policy. Thirty per cent (n= 15) were doctoral thesis. Three were reports by research organisations in the USA, and one report came from India (Goel and Barooah, 2018). The others were working papers.

Table 6.4 The Name of Journals Included in the Systematic Review

	The Number of Studies
Journal of Educational and Behavioral Statistics (Sage Journals)	3
American Economic Review (Google Scholar)	1
Economica (Web of Science Core Collection)	1
Economics of Education Review (SCOPUS)	4
Education Finance and Policy (ERIC)	4
Education Policy Analysis Archives (Google Scholar)	2
Educational Assessment, Evaluation and Accountability (ERIC)	1
Educational Evaluation and Policy Analysis (Sage Journals)	2
Journal of Labor Economics (Google Scholar)	1
Journal of Personnel Evaluation in Education (Google Scholar)	1
Journal of Research on Educational Effectiveness (Education Research Abstracts Online)	1
Planning and Changing (ERIC)	1
Practical Assessment, Research & Evaluation (ERIC)	1
Statistics and Public Policy (Education Research Abstracts Online)	4
Oxford Bulletin of Economics and Statistics (Education Database)	1
Total	28

Half (50%, n= 25) of the fifty included studies were longitudinal studies. Twenty-three studies (46%) made comparisons across various value-added models/data analysis methods using a longitudinal comparative design. One study (2%) examined the relationship between teacher effectiveness and their characteristics and attitudes, employing a causal-comparative research design. Another study systematically compared several methods for integrating multiple measures of student performance into traditional value-added methods using a mixed factorial design.

For ease of judgement, I defined the sample sizes as “sufficient” or “limited” (see Table 6.3). This is a subjective judgement. After identifying the sample sizes (for total and comparison groups) of all studies, the cluster intervals of these sample sizes were then determined, and finally, threshold values are assigned for each category based on these cluster intervals. A “sufficient sample size” here refers to at least 2000 students or 200 teachers. Of the 50 eligible papers, 46 (92%) were judged to have an adequate or reasonable number of cases. Four studies (8%) were judged to have a “limited” sample size (that is under 2000). The majority of the eligible studies (80%, n= 40) were conducted with a comparative (or smallest) group sample of more than 2000 students or 100 teachers, which is determined as a “large number of cases”. On the other hand, two studies (4%) had 754 and 534 student cases in the comparison group (small number of cases). According to Table 6.3, eight studies (16%) had a medium comparison (or smallest) group sample size of between 1000 and 2000 students or 50 and 100 teachers.

As all public schools in the USA were mandated to replace their assessments with a state-wide standardised test under the No Child Left Behind Act of 2001, it is therefore not surprising to see the prevalence of standardised tests throughout school districts or whole states, such as the Florida Comprehensive Assessment Test (FCAT), California Standards Tests (CSTs) and the California Achievement (CAT). The availability of such standardised tests explains why a large number of educational research done at elementary and middle school levels to evaluate school accountability systems, such as the teacher effectiveness, is based on students’ test results (see Table 6.3). Most of the studies on teacher effectiveness using value-added measures were conducted at the elementary school level (42%). Two studies (4%) did not include study setting information because they used simulated data.

The longitudinal design of these studies makes attrition at the student level a problem, particularly when two or more previous test scores were used in the estimation. Except in studies that used simulated data - there were three studies in this review that used simulated data – attrition is to be expected in longitudinal studies. Although the rate of attrition in the data used in the estimations and the reasons for these attritions are not mentioned in many studies, authors of some studies commonly reported that attrition was due to the following issues: a) missing observations, b) students' mobility (moving, withdrawing from school, etc.), and c) inability to link students to teachers.

Of the 50 included studies, slightly more than half (52%, n= 26) had minimal data loss with up to 19% of their overall sample cases. Two studies (4%) have more than 80% data loss - one of which falls into this category because it did not report attrition rates in their data set. Two other studies were missing between 60% and 79% of cases, one of which was Guarino et al. (2015b). Although their dataset includes 1,488,253 total students, only 482,031 students' test scores were used in the estimations; the attrition rate is around 68%.

CHAPTER 7

REVIEW OF STUDIES THAT EXAMINE THE STABILITY OF VAMS USING STUDENT/TEACHER/SCHOOL CHARACTERISTICS

In this review, value-added teacher performance estimates were reviewed from three perspectives: a) the predictors used, b) the number of previous test scores employed, and c) the data analysis methods applied. This review is perhaps the most comprehensive study of its kind that synthesised the results of single studies on teacher effectiveness estimates. The 50 studies were synthesised according to these three measures. As some of the studies cover two or all three measures, the background information of such studies is given only in one of the summaries, while the results are explained in each relevant section.

This chapter is focused on studies that look at the consistency of VAM estimates that include student-, teacher/classroom-, and school-level variables.

7.2 Stability of Value-added Estimates that Include Student, School and Teacher/Classroom Characteristics

Twenty-five studies in this review considered the use of different levels of predictors to test the contribution of the predictors. Out of 25 studies, four were rated 4 🔒, eighteen rated 3🔒, two were rated 2🔒, and one was rated 1🔒. The findings are presented starting with the study having the highest quality score. The quality appraisal of the studies in this section is shown in Table 7.1.

Table 7.1 Quality Appraisal of the Studies: The Predictors Used

Author(s) and Year	Design	Smallest Cell	Allocation	Attrition (roughly)	Quality
Aaronson et al. (2007)	Longitudinal Study	25.299 Students	Random	48%	4🔒
Kane and Staiger (2008)	Longitudinal Study	1.925 Teachers	Random	36%	4 🔒
Rothstein (2009)	Longitudinal Study	2.733 Teachers	Random	4%	4 🔒
Nye et al. (2004)	Longitudinal Study	5.766 Students	Random	5%	4 🔒

Alban (2002)	Longitudinal Comp. Study	5.487 Students	Non-Random	32%	3 🔒
Ballou et al. (2004)	Longitudinal Study	120.646 Students	Non-Random	9-14%	3 🔒
Buddin (2011)	Longitudinal Study	36.484 Students	Non-Random	1%	3 🔒
Chetty et al. (2014)	Longitudinal Comp. Study	3.5M Students	Non-Random	34%	3 🔒
Cunningham (2014)	Longitudinal Comp. Study	1.001 Students	Non-Random	20%	3 🔒
Ehlert et al. (2014)	Longitudinal Comp. Study	289 Teachers	Non-Random	9%	3 🔒
Gagnon (2014)	Longitudinal Study	10.657 Students	Non-Random	3%	3 🔒
Gallagher (2002)	Longitudinal Study	532 Students	Non-Random	Complete data	3 🔒
Goel and Barooah (2018)	Longitudinal Study	144 Teachers	Non-Random	24%	3 🔒
Heistad (1999)	Longitudinal Study	182 Teachers	Non-Random	22%	3 🔒
Hu (2015)	Longitudinal Study	1.210 Teachers	Non-Random	5%	3 🔒
Johnson et al. (2015)	Longitudinal Comp. Study	2.778 Teachers	Non-Random	Complete data	3 🔒
Kersting et al. (2013)	Longitudinal Study	38.503 Students	Non-Random	22%	3 🔒
Kukla-Acevedo (2009)	Longitudinal Study	754 Students	Non-Random	35%	3 🔒
Leigh (2010)	Longitudinal Study	59.612 Students	Non-Random	45%	3 🔒

Slater et al. (2012)	Longitudinal Study	7.095 Students	Non-Random	24%	3 🔒
Tobe (2008)	Causal Comparative	6.263 Students	Non-Random	19%	3 🔒
Munoz et al. (2011)	Longitudinal Study	235 Teachers	Non-Random	17%	3 🔒
Germuth (2003)	Longitudinal Comp. Study	258 Teachers	Non-Random	66%	2 🔒
Munoz and Chang (2007)	Longitudinal Study	56 Teachers	Non-Random	3%	2 🔒
Goldhaber and Hansen (2010)	Longitudinal Study	7.732 Teachers	Random	97%	1 🔒

One of the studies had the highest quality rate (4🔒), **Aaronson et al. (2007)**, involved roughly 53,000 ninth-grade students linked with 1132 maths teachers. Although the authors used an administrative dataset from Chicago public high schools, they also estimated the teacher effectiveness from simulated classroom settings. As observable characteristics, the teacher-level predictors such as race, sex, experience, tenure status, advanced degree, undergraduate major, university ranking attended, and teaching certifications were employed to estimates teacher effectiveness on a ninth-grade math test score. Based on the analysis of the changes in R^2, the authors concluded that none-of-the observable teacher characteristics added have a noteworthy contribution to the explanation of the variance in estimated teacher quality. Specifically, the predictors of the advanced degree, tenure, and undergraduate major explained at 5‰ of the total variation, and R^2 never exceeded 0.08 in all cases. This finding is similar to the result of **Rothstein (2009)** that other highest-ranking research in this section. As it was conducted in longitudinal research design with adequate sampling size, and it made the random allocation possible, the study was rated 4🔒. This study used a longitudinal administrative data set for students from Grades 3 to 5 in North Carolina. The sample consists of 49,453 students for whom all four scores (pre-test in Grade 3, end of grade tests in Grades 3,4 and 5) were available and linked with 2,844 reading teachers in 838 elementary schools. The researcher created various assignments strategy by controlling statistically observable and unobservable predictors, including random student-teacher allocation; however, the allocation was not purely random like in an RCT. Excluding the 111 reading teachers who have less than ten students from the data set, the attrition rate was 4%. Although the reason for this exclusion is not mentioned by the researcher, the concerns discussed in the literature regarding the impact of small sample sizes on the error in value-added estimates may be the reason why teachers with less than 10 students were excluded from the analysis. The researcher investigated the contribution of predictors to the estimates through changes in the R^2 of the models. The researcher added twenty-eight predictors such as race, gender, free/reduced lunch status, parental education etc. in a model where the nearest prior year test score was used as a unique predictor, and in another model where along with the nearest prior year data, two lagged test scores also employed. The inclusion of twenty-eight predictors, however, resulted in an increase of 0.05 in the R^2 of the first model and 0.01 in the second model. Therefore, the study suggests that the use of twenty-eight contextual variables had little contribution to the teacher performance evaluation estimates once the prior attainment is included in the estimates.

The other study, with 4🔒 rating score, **Kane and Staiger (2008)**, evaluated various specification of the model used in estimating teacher effectiveness based on student achievement growth. This study involved 47,320 elementary school students linked to a total of 1,925 math and reading teachers from experimental schools (the sample size of the experimental teacher group is 140 and the non-experimental group is 1,785), and 273,525 students linked to 11,352 teachers from non-experimental schools. The researchers used the data from a random-assignment experiment in the Los Angeles Unified School District to test various non-experimental methods used in estimating teacher effectiveness. Along with students' Stanford 9 math and reading/language art achievement test scores grades from 2 to 5, the administrative data also contained other demographic characteristics including race/ethnicity, grade repetition status, the status of eligible for free/reduced lunch status, gifted, talented, special education status. Ordinary least square (OLS) method was preferred to estimate teacher value-added effectiveness by employing empirical Bayes' techniques. The end-of-year math and reading test scores were used as a dependent variable in the estimates. As the simplest specification, none of the control variables was included in the estimates, and the estimates were mainly based on the average achievement for each classroom. As the second specification, along with including students' prior test scores in maths and reading, the students' demographic variables and the averages of these variables at the classroom level were included. As the following specification, the school fixed effect was added into the second specification, and as the last specification, all specifications (one, two and three) were repeated by replacing the dependent variable with students’ test score gains (the difference between their end-of-year test scores and prior test scores). As a result of this study, the researchers reported that the teacher effectiveness estimates performed best by controlling for students’ previous test scores and mean peer characteristics. The standard deviation in teacher effect was dropped to 0.231 in maths and 0.184 in English language art from 0.448 and 0.453 (estimated by not controlling any student or classroom level covariates) in the same teaching subjects, respectively. Including the school fixed effect also contributed to a decrease in the standard deviation in teacher effect, but the contribution was limited.

The last study rated 4🔒, **Nye et al. (2004)**, examined the contribution of teacher characteristics to teacher effectiveness estimates using a hierarchical linear model. The study involved the data from a four-year experiment, the project STAR, also known as The Tennessee Class Size Experiment. The four-year experiment involved over 7,000 students from kindergarten through grade 3 in 79 schools were randomly assigned to one of three treatment conditions: small class (13 to 17 students per teacher), regular class (22 to 25 students per teacher), and regular-with-aide class (22 to 25 students with a full-time teacher's aide). Teachers were also randomly assigned to one of these classrooms (Achilles et al., 2008). The authors carried out two sets of HLM analyses, one of which estimated teacher effectiveness based on student achievement gains, and the other employed achievement status. Two of the teacher characteristics, teacher experience and education, were investigated whether the variance of teacher effectiveness changes by controlling them. The authors found that neither teacher experience nor teacher education had a notable contribution to teacher effectiveness estimates. The explained variance in each case never exceeded 5%.

The first study rated with 3🔒 reported that prior test scores are essential in value-added estimation for teacher effectiveness. This doctorate thesis based on the longitudinal comparison **(Alban, 2002)** involved 17,559 eight grade students linked with a total of 911 English, mathematics, science, and social science teachers from two school systems. The researcher compared hierarchical linear models and multiple regression models using a variety of student, teacher and school-level predictors such as gender, race, prior attainment, English as a second language (ESL) status, the number of year teaching, current degree status, socioeconomic level of the school, per cent of students receiving special service, etc. In order to determine the significant predictors in the estimates, the researcher preferred to look at F statistics along with the standardised regression coefficients in multiple regression analyses. For instance, in the first round of multiple regression in school system A, the researcher run twenty-one analyses in each teaching subject, and it was found that the prior test score is the only significant variable in each estimation, and per cent of students receiving special service in school and gender followed it. It was also found that the coefficients of the same variable and their signs differed significantly in each teaching subject. Therefore, based on these findings, the researcher concluded that educational researchers might need to take into account which variables are the most important to their models related to teaching subjects.

Gallagher (2002) examined the relationship between teacher effectiveness estimates and student achievement in a charter elementary school by employing hierarchical linear modelling. The sample of the study consisted of thirty-four reading, maths, and language arts teachers and roughly their 1,700 students in Grades 2 through 5, whose two years of achievement data were available. Unlike other related studies, the researcher used the term classroom effects instead of teacher effects since the group level residuals were calculated by hierarchical linear modelling after controlling individual and group characteristics for a given classroom. Students' prior year test scores and a variety of student-level characteristics such as English proficiency status, special education status, attendance, etc., were controlled to be estimated the classroom effects. To determine the significant predictors in the models, the author took into account the coefficient value of the predictors. Not surprisingly, the student individual prior attainment score was found the strongest predictor for current academic performance by controlling other student-level predictors. For instance, each point increase in the student's prior reading test scores caused an increase of 0.59 points on their current test scores.

Another doctorate thesis **(Gagnon, 2014)**, consisting of three essays and rated with 3🔒, was interested in the contribution of various student-level predictors, including student's achievement data, lower-income, race, gender, English language learner status, disability status, attendance, suspension to the value-added estimates for teacher effectiveness. For the estimates, a total of 53,411 students with at least a 75% attendance rate, not changing school between Grades 7 and 8, and having three consecutive years of achievement data in mathematics and English language arts were involved. The researcher claimed that the prior test scores are the most powerful predictors of future achievement, with an example that 75.8 per cent of the variation in Grade 8 maths scores may be explained by the prior test scores in Grades 3 through 7, and similarly, 65.6 per cent of the variation in ELA test scores in Grade 8 may be explained by the prior test scores in Grades 3 through 7. The author used alpha level criteria (p< 0.10, p< 0.05, p< 0.01, and p< 0.001) in the results. Opposite of the other studies in this section, the research also claimed that almost all students’ characteristics used remain important in determining teacher effectiveness even though prior test scores were controlled. For instance, in the 8th-grade mathematics model, all student-level controls, except the student race, reached the alpha level with 0.10.

The following six studies, rated with 3🔒, reached similar findings related to the contribution of contextual predictors to value-added effectiveness estimates. The first study in this group **(Cunningham, 2014)** examined the contribution of student-level variables in the value-added teacher performance estimates by comparing the correlation coefficients between the models. Although the design of the study was not explicitly stated by the researcher, it was determined as a longitudinal comparison study regarding have longitudinal data set of the same students for three successive years to determine the contribution of variables by comparing model specifications. The study population consisted of three cohorts of students; cohort 1 included 1,001, cohort 2 included 1,060 and cohort 3 included 1,094 students, giving a total of 3,155 students in Grades 3, 4, and 5. Because of similar concerns about the magnitude of error in estimates resulting from employing limited participants, teachers who have less than fifteen students were excluded, which was roughly 40% of the teachers. This also meant that around 20% of the whole student population were dropped from the estimates. The author employed two covariate adjustment models, one of which used student’s prior test scores only, and the other used the student characteristics as well as the previous test scores. The high degree of consistency between covariate adjustment models (0.97 in a single year and 0.96 in multiple years analysis) suggested that the use of the student characteristics –free/reduced lunch, special education status, and English language learner status– in the value-added estimates were unnecessary.

The other study **(Ehlert et al., 2014)** estimated the VAM scores for school and teachers using up to 3 years of lagged test scores, eligibility of free/reduced lunch status, language learner status, special-education status, race, and gender. Forty-two thousand middle and junior high school students linked to 289 maths and 390 communication art teachers were involved in this study. Students with incomplete score histories were excluded from the estimations, and this resulted in a loss of 9% of the study population. The authors run various estimations by including and/or excluding the control variable(s) each time and found that the correlation between the estimations in the various models is very high with at or above 0.90. Although the authors found a high correlation across the estimates, which means that whether inclusion or exclusion of the variable at student-level had very limited contribution on the teacher effectiveness estimates, but they also double-checked this result by ranking the teachers related to their effectiveness estimates in the quartiles. They concluded that despite the high correlation found across the estimates, the teacher rankings are meaningfully influenced by the selection of predictors to include in the models and how to include them.

Another longitudinal study with 3🔒 **(Hu, 2015)** estimated teacher effects by using up to three years of student test scores in maths and reading across five grades (from Grade 4 to 8), student background characteristics including gender, ethnicity, LEP (limited English proficiency) status, gifted status, and student with disability status and only class size as a classroom-level in the hierarchical linear models (HLMs). Weighted Least Squares (WLS) analysis, logistic regression, and point-biserial analysis were employed to examine the relationships among teachers' VAM scores. The study involved 1,210 maths and 1,239 reading teachers with less than 5 per cent missing data. In order to explore the contribution of contextual variables to the teacher effectiveness estimates, the researcher added the covariates one by one into models on each occasion. Compared with the amount of variance explained from the models with student prior scores only, even models that include all predictors explained slightly much variance in the students' current attainment, up to 2%. The other longitudinal study **(Kersting et al., 2013)** explored the contribution of contextual variables to value-added teacher effectiveness estimates. The study population consists of 208,137 students linked with 3,878 fifth-grade mathematics teachers in 474 schools. Data of students with complete three successive year data in Grades 3, 4, and 5 and all their relevant characteristics information were linked with the data of teachers who stayed in the school during the year of the study. Since not all students could be successfully linked to their students, there was a 22% loss in the student data. Therefore, the final sample size of the study was 161,811 students, linked to 3,651 math teachers from 469 schools. The authors compared the relative impact of statistical control variables which are previous years data, student gifted and special education status, free/reduced lunch status, and ethnicity, across four models. Although in Model 1, one previous year data was used as a unique controlled variable, additional prior year test scores were controlled in Model 2. Moreover, in Model 3, the authors included the student background information in Model 1, and in Model 4, the student background information was added in Model 2. The findings were mainly revealed that the statistical controls in the teacher value-added estimates have a little contribution. The pairwise correlation coefficients across these four models were very high with ranging from 0.97 to 0.99. It was also found that although 68% of the variance in the students' current scores was explained by controlling for only one previous year data (Model 1), the percentage of explained variance increased only 1% by adding all student background information in the estimate (Model 3). In other words, missing student characteristics would have little contribution to teacher value-added estimates.

The next study **(Heistad, 1999)** was designed to explore the predictive power of student characteristics employed in estimates. The study sample consisted of three cohorts of teachers (first cohort from 1993/94, second from 1994/95, and the third one from 1995/96). The sample included 585 class teachers and 3,237 students. Multiple regression was conducted in the estimates where students' second-grade reading test score was regressed on the same student's previous academic performance, and characteristics including free/reduced-price lunch, limited English proficient, special education, race, parent or guardian "resides with" status. To determine the power of the predictors on the estimates, the researcher examined the changes in R^2 by adding each variable one-by-one on each occasion. Mainly, very minimal changes in R^2 were found across the model specifications. In the first specification, which only included in previous test score in reading and, R^2s were found 0.632, 0.656, 0.560 in each school year. It means that the previous tests score alone can explain 56% to 66% of the variance in a current test score. Adding race caused an increase of 0.011 in R^2 in the year 1 and 2 groups and 0.027 in the year 3 group. Similarly, gender caused an increase of 0.002, 0.001 and 0.004 in R^2s in the same groups, respectively. Family compositions and poverty resulted in increases of R^2s between 0.009 and 0.018, while special education status raised R^2s by 0.008 in each group.

Johnson et al. (2015) also focused on the sensitivity of teacher value-added estimates regarding student and peer background characteristics used. The authors employed multiple regression models adding and excluding the contextual predictors in each model. The longitudinal comparison study involved a total number of 9,269 maths and 9,944 reading teachers from elementary and middle school in an unknown school district in the USA. The authors estimated the VAM score for all teachers in their study population who have at least ten students with no missing data. The findings mainly suggested that the control variables have little contribution to teacher VAM performance estimates. The authors compared the baseline model, which included one prior year data and student and peer characteristics, with the restricted models that excluded predictors. The correlation coefficients were found to be above 0.9 in each comparison. However, the authors also warned that the correlation coefficients above 0.9 do not prevent teachers from being classified incorrectly across performance categories. For example, 26 per cent of teachers were placed in the lowest quintile in the baseline model but rated in higher performance categories in an alternative model.

Kukla-Acevedo (2009), rated 3🔒, examined the relationship between teacher characteristics and the variance explained in students' current performance. This study involved 3,812 5th grade students linked with 120 maths teachers from 46 schools. In addition to using mathematics test scores of fifth-grade students as an outcome variable, student's prior attainment in reading and demographics (gender, race, subsidized lunch status), teacher characteristics (college coursework, GPAs, mathematical contents and number of hours of mathematics education during pre-service training, experience, gender and race), and school-level characteristics (percentages of subsidized lunch status and race) were included in the estimates. Although the study had various predictors at different levels, primarily it focused on the relationship between teacher qualification and student achievement. To determine the relationship, the author preferred to look at regression coefficients (β) and ρ-values of F-tests in the fixed effect models. Mainly, the study found that only overall math teachers' undergraduate performance (GPA) among the other teacher characteristics consistently positive link to student math achievement and suggested being included in models to predict maths teachers performance.

The following three studies, rated with 3🔒, reporting very similar results that prior attainment is only the strongest predictors in the value-added estimates. **Slater et al. (2012)** estimated the effectiveness of individual teachers on student achievement by using the longitudinal dataset from the UK. In the estimations, while student the GCSE score (as known as Key stage 4) was used the dependent variable, various predictors in teacher and school level such as gender, age, experience, degree, percentage of free school meals, percentage of the ethnic minority, school population, etc. were also used as the explanatory variables. The study included 7,305 students linked to 740 mathematics, science, and English teachers from 33 secondary schools. The authors suggested that subject-specific previous attainment is the only significant variable in the student fixed effect regression and also reported that none of the observable teacher and school characteristics is significant predictors in explaining teacher effectiveness. Last, the authors concluded their study by giving an answer to the question that do teachers matter? that is, “having a one-standard-deviation better teacher raises the test score by 27% of a standard deviation”.

The other study **(Chetty et al., 2014)** investigated which control variables in the estimates are most important by comparing several commonly used VA specifications. The dataset contained approximately 1.8 million test scores of students in Grades 3 through 8 in ELA and maths, the student information such as ethnicity, gender, age, receipt of special education services, and limited English proficiency, and the parental characteristics information obtained from US federal income tax returns. As a result of the analyses, it was found that the estimates controlled by student’s prior attainment provided more stable results of teacher effect on the student test score. The authors concluded that in a model where the student's own previous test scores controlled, prediction bias is around 5 per cent, correlation with baseline VA estimates where prior scores at student-, class-, and school-level and demographics employed is 0.96.

Budding (2011) focused on the teacher value-added effectiveness estimates using elementary school students’ achievement growth. The study used the panel data that involved 412,825 individual students in Grades 2 through 5 linked with 11,462 math and ELA teachers from 473 schools. The researcher examined whether teacher effectiveness estimates vary by controlling student and peer characteristics including prior attainment, gender, English language learner (ELL) status, eligibility of free/reduced lunch status, race, the proportion of ELL, race, gender, eligibility of free/reduced lunch status, etc. in four different specifications. While in the basic specification, student prior test scores in maths and ELA and class size were controlled, in the second specification, student characteristics were added in the first one. Similarly, in the next specification, the student peers’ characteristics added into the previous specification, and finally, in the last specification, students' average prior test scores added into the third specification. In parallel with other studies, this study also reported that prior test scores have a strong link to student’s current performance. In addition to the coefficient report, the study also calculated correlation coefficients between models, and it was found at above 0.90 in each comparison. Last, the R^2 indicated that the teacher characteristics -experience, degree, race/ethnicity, and gender- were able to explain little variance in value-added teacher effectiveness with less than 1.5 per cent.

Goel and Barooah (2018) examined the contribution of teachers in enhancing student test scores at the higher secondary level in public schools in Delhi, India. The administrative data set, obtained from the Directorate of Education (DOE), included 18,552 student-subject-

(grade)-teacher observations (1,733 students) in sixteen subjects such as English, Hindi, political science, economics, history, etc. in Grade 12, and student and teacher characteristics including age, gender, religion, castle, parental education, income, number of siblings, teacher gender, marital status, religion, castle, degree, training, tenure status, etc. The study sample (2,207 students) were dropped to 1.733 as their prior and current test scores are not available (14.7% for Grade 10), their section information is missing (1.0% for Grade 12 and/or Grade11), and there is a recording problem such as same data belonging to the same students from the different classroom (8.1%). After estimating teacher fixed effect by using students test scores in Grade 12 as a dependent variable, the result was used as a dependent variable in the other estimates to examine the role of teacher information on the teacher effectiveness by comparing three specifications; a) included only the teacher characteristics, b) included only personality dimensions obtained from teacher survey and c) included both of them. The result indicated that the teacher effectiveness estimates were positively affected by the predictor of only being permanent (tenured), on the other hand, consistently with a large body of existing literature, other characteristics such as gender, educational qualification, training, experience, etc. have no or very little predictive power. Moreover, the personality dimensions (agreeableness, conscientiousness, extraversion, neuroticism, and openness) were found to be insignificant predictors in teacher effectiveness estimates. Last, the study also reported that having a one-standard-deviation better teacher raises test scores by 0.37 standard deviation. This result revealed that a better teacher contributed to the test scores about 10 per cent more than Slater et al. (2012)'s study.

The other doctorate thesis **(Tobe, 2008)**, ranked 3🔒, examined the relationship between teacher effectiveness estimates and student and teacher characteristics by involving 223 math teachers who were linked to12,369 students in Grades 5 through 8 in the two-level HLM model. Along with using student's maths test scores (shown as a percentage of correct answers and percentile rank) as the outcome variable, students-level variables including prior attainment, grade, gender, ethnicity, special education status, gifted status, poverty and were employed in the first level and teacher-level variables such as grade, gender, experience, ethnicity, certification, degree were controlled in the second level. To determine the significant predictors, the researcher examined the differences in the amount of variance explained in each specification by adding or/and removing predictors. The models indicated that over 47% of the variability in student current math score could be explained by student's prior test score alone; on the other hand, the inclusion of all other student-level variables in the estimate contributed to the increase in the variability explained by only extra two to four per cent. Moreover, the study reported that apart from being certificated by the state, none of the other teacher characteristics and attitudes had a noteworthy link to student achievement gains.

Another study with 3🔒 **(Munoz et al., 2011)** focused on the contribution of student and teacher characteristics to teacher effectiveness estimates. The study carried out a multilevel model controlling the student's prior academic attainment and characteristics, including gender, race, socio-economic status, parents’ education, attendance, age, special needs status, English language learner status, or gifted/talented status at level 1, and teacher/classroom characteristics including teacher’s years of experience, educational level, ranking, class size, and the aggregated data of the student-level variables at level 2. Fourth-grade end-of-year reading test scores were employed as the outcome variable for a total of 17,206 students linked to a total of 712 reading teachers. Consistently with a large body of existing literature and most of the studies in this section, the study determined that among all student characteristics examined, student’s previous performance in reading test was the strongest predictor in the teacher effectiveness estimates. On the other hand, the authors also determined that teacher experience is another valuable predictor of teacher effectiveness estimates at the teacher/classroom level.

Ballou et al. (2004) focused on modifying the TVAAS, Tennessee Value-Added Assessment System, by controlling student characteristics. The original TVAAS did not employ any student-level predictors rather than student prior test scores to estimate teacher and school effectiveness. William Sanders and his associates, who developed the TVAAS, explained the reason for not adding additional predictors to the model was that the student characteristics influence on the post-test is already reflected in the pre-test score, so no need to add them again. However, the approach has been criticised for lack of being controlled enough in some studies (Linn, 2001; Kupermintz, 2002). Therefore, the authors needed to conduct this study to check whether the criticisms are justified. The effectiveness of over 5,000 reading, language arts, mathematics teachers linked to over 120,000 students in Grades 3 to 8 spanning over five years in each subject were estimated. To investigate the predictive power of student background variables in value-added estimates of teachers, the author compared the results obtain in unmodified TVAAS and modified ones. Eligibility of free/reduced-price lunch status, ethnicity and class and school levels percentage of ethnicity were controlled in modified TVAAS. The researchers found that the inclusion of additional contextual variables has little contribution to the estimations. The correlation coefficients between teacher effectiveness estimates generated from both models exceed 0.90. The researchers also investigated the concordance between models regarding identifying the teachers who above or below the average and claimed that the agreement in reading is 2.7 times more likely than disagreement, and these proportions are 3.5 and 8.5 times more likely in language arts and mathematics, respectively.

The last longitudinal study rated 3🔒 **(Leigh, 2010)** examined the contribution of teacher characteristics to the teacher effectiveness estimated by the value-added model. Along with using students' test scores, teacher's gender, age, experience and DETA rating (The Queensland Department of Education, Training and the Arts) were included in the estimates. The research used data from state primary schools in the state of Queensland, Australia. The Queensland education department also provided the rating scores for about two-thirds of sample teachers. The teacher or principal candidates make mini-presentations about their experience and ability of teaching/managing, and then the interviewee answers the questions. Based on this process, the teachers receive their "suitability rating" in four-points scales. The study sample consisted of over 90,000 students in Grades 3 through 8 between 2001 and 2004 linked to over 10,000 literacy and numeracy teachers. To determine the contribution of predictors employed in the models, the researcher added each teacher-level predictor in the models one by one and included all in the last model and checked the changes in R^2s on each occasion. As an overall result of the estimates where students' achievements gains were regressed on the teacher characteristics, the teacher-level predictors did not explain almost any variation among teachers. The explained variance in teacher performance by the combination of the characteristics (gender, age, experience, and the DETA rating) never exceed 1%.

Another doctorate thesis **(Germuth, 2003)** investigated the contribution of student-, teacher-, and school-level variables to maths teachers' effectiveness estimates by employing data from middle schools. The study sample consisted of 21,634 students linked to 258 mathematics teachers across 26 middle schools. The study compared two value-added models' (OLS and HLM) specifications to determine the most powerful predictors in the teacher effectiveness estimates. In the models where student maths test results in 2002 were used as the outcome variable, and following predictors were used as independent variables: days in attendance, special education status, free/reduced lunch status, gender, ethnicity, and prior attainments at student-level; highest degree, experience, teaching in the graduated field, and per cent time teaching instruction at teacher-level; and student stability in the school, crowding, ESL percentage, special education percentage, school minority percentage, and school free lunch percentage at school-level. As adding two previous year test scores (in 2000) in the estimates caused to lose of 66% of the data, the study was rated with 2🔒. The researcher ranked the teachers based on the residuals obtained across the models' specifications, then Kendall's Coefficients of Concordance for Ordinal Data (Kendall's W) were used to compare the ranking concordance. Kendall's Ws were found as very similar across model comparisons (0.97 to0.99), suggesting that the contextual variables used caused very little variability in teacher rankings among the models. In general, over 90% of the variance in teacher effectiveness estimates was related to students-level predictors, and the student's prior attainment alone was the most critical predictor in all estimates with the capability to explain the variance explained between 76%-80%. The study also claimed that teacher- and school-level variables had a very little link to student achievement.

Another study **(Munoz and Chang, 2007)** focused on the contribution of teacher characteristics to student achievement gains using a multi-level growth model in an urban school district in Kentucky. Along with the selected teacher characteristics, including education level, years of teaching experience, and race, the researchers used student's test score obtained from the Predictive Assessment Series (PAS - The PAS was administered three times a school year; in September, December, and February-) in ninth grade as an outcome variable in the estimates. The study employed the data for 58 reading teachers in Grade 9 and 1,487 their students whose three data wave points are available in the two-level HLM. Comparing with the other studies, as the study sample is very small (n= 58), the study, therefore, was rated with 2🔒. The variable of "time" (indicates when the test was administered) was used at level 1 as a student-level variable in all models, while each teacher characteristics (years of experience, education level and ethnicity) was added at level 2 as one-by-one in each model. Each model was compared to the baseline model, where "time" was only employed to investigate the power of each variable to explain variance. The authors revealed that none of the teacher characteristics is not significantly related to student reading achievement growth. The findings of this research indicated that the teacher characteristics have little contribution to student achievement gains.

The final study retrieved in this section was also ranked the lowest ranking with 1🔒 is a working paper **(Goldhaber and Hansen, 2010)**. Although it is a longitudinal study with having adequate sampling size, as the study had a tremendous attrition rate in the estimates with over

95%, its rating was dropped to 1🔒. Its findings, therefore, have to be treated with caution when reporting. It started with a total of 19,586 maths and reading teachers, but for analysis, only teachers whose data could be linked to their students and who were assigned to only 4th and 5th grades classes with more than ten and less than twenty-nine students were included (using data that best represents typical class situations). The researchers created various student-teacher matching strategy by controlling statistically observable student characteristics (gender, race, etc.), including random student-teacher allocation; however, the allocation was not purely random like in an RCT. Teachers also needed to have a minimum of two years of teaching experience in a district before obtaining tenure and at least one year after it. Because of these requirements, the final number of teachers that were included in the analysis was 556, resulting in attrition of over 95%. The study examined the predictive power of teacher characteristics and earlier career performance estimates on using value-added models in high-stakes decisions like tenure and dismissal. Along with employing student's test scores in reading and maths and their characteristics, such as teacher's licensure status, experience, degree levels, college selectivity, average licensure scores and past year VAM estimate score were included in the estimates. The result of the comparisons across the models (all included student-level predictors) -one of which included only teacher characteristics, the second one included only teacher past performance, and the last one used both teacher variables and previous performance score- showed that if the previous performance estimates included in the models, the teacher quality variables are no longer significant predictors on the student's current test score. By adding only teacher characteristics in the estimates, the model's R^2 (strength of the relationship) was 0.69 in reading and 0.72 in maths, and similarly, if only prior performance of teachers (prior VAM scores) was included in the estimates, the R^2 was also 0.69 in reading and 0.73 in maths. Interestingly, in the estimates where both teacher characteristics and prior performance estimate were used, the teacher characteristics lost their predictive power, and the model's R^2 remained the same as the model used only prior performance score. Consequently, the researchers claimed that prior year teacher performance in the same teaching subject is the strongest predictor in the estimates.

Table 7.2 summarises only those studies that identify the key predictors for each of the variables. Of the 25 studies, 22 identified at least one of the student, school, or teacher/classroom variables as crucial predictors. Three studies showed that none of the teacher level variables was important. These were therefore not shown in the table.

Table 7.2 A Summary of Key Predictors Identified by Studies Reviewed

Key predictor identified	Number of studies			
	4🔒	3🔒	2🔒	1🔒
Prior attainment	1	10	1	
Sex (student)		2		
Lower-income, English language learner status, disability status, attendance, suspension		1		
Mean peer characteristics (classroom level)	1			
Teacher GPA score		1		
Being permanent		1		
Being certified by the state		1		
Experience		1		
Prior performance (teacher)				1
Percentage of students receiving special service at school-level		1		

*2 studies rated 4🔒, and 1 study in 2🔒 suggested that none of teacher-level variables were key predictors; therefore, these studies are not excluded in this table.

In summary, the majority of studies (see Table 7.2) identify students' prior attainment as the best predictor of teacher effectiveness estimates. Studies that focus on student variables (e.g. sex, socio-economic status, English language status, disability status, attendance and suspension) show mixed results. Most of these studies suggest that these other student variables contribute little to the prediction of teacher effectiveness. Of the 19 studies, only three identify student variables as the key predictor. Teacher-level and classroom predictors (e.g. permanent status, experience, qualification, prior performance, teacher GPA and peer characteristics) were also found to be not important predictors of teacher effectiveness. Of the 15 studies, only six considered these factors as key predictors. Most of these studies were rate 3, while one was rated 1. Of the 6 studies focused on the contribution of variables at school-level, only one study rated 3 suggested an advantage of including variable of per cent of students receiving special service. The very disparate results suggest that student-, school-, and teacher/classroom-level variables are not consistent measures of teacher effectiveness.

The strongest studies (rated 4🔒) show that students' previous academic performance is the best predictor of teacher effectiveness, and the inclusion of variables at the student and teacher level adds little to the predictive power of teacher performance assessment models. More studies rated with 4🔒 may be needed to confirm these results, but at the moment, there is some degree of robust evidence that using contextual variables except prior attainment is not useful.

CHAPTER 8

REVIEW OF STUDIES THAT EXAMINE THE STABILITY OF VAMS USING PREVIOUS YEARS' TEST SCORES

This chapter synthesises the findings of previous studies on the stability of VAM teacher effectiveness estimates with regards to the number of previous years' test scores used in the estimates. Since some of the studies examined in this chapter has been discussed in detail in other chapters, only the findings of these studies and not the background information are summarised here.

8.1 Stability of Estimates Using the Number of Previous Years' Test Scores Fifteen studies were retrieved regarding the contribution of using additional prior attainment in teacher value-added effectiveness estimates. Out of 15 studies, one was rated with 4🔒, thirteen had 3🔒, and one was rated 1🔒. The key findings of the studies are presented, starting with the study having the highest appraisal score. Quality appraisal of the studies in this section is depicted in Table 8.1

Table 8.1 Quality Appraisal of the Studies: The Number of Previous Test Scores

Author(s) and Year	Design	Smallest Cell	Allocation	Attrition (roughly)	Quality
Rothstein (2009)	Longitudinal Study	2,733 Teachers	Random	4%	4🔒
Cunningham (2014)	Longitudinal Comp. Study	1,001 Students	Non-Random	20%	3🔒
Ehlert et al. (2014)	Longitudinal Comp. Study	289 Teachers	Non-Random	9%	3🔒
Hu (2015)	Longitudinal Study	1,210 Teachers	Non-Random	5%	3🔒
Johnson et al. (2015)	Longitudinal Comp. Study	2,778 Teachers	Non-Random	Complete data	3🔒
Kersting et al. (2013)	Longitudinal Study	38,503 Students	Non-Random	22%	3🔒

Schafer et al. (2012)	Longitudinal Comp. Study	5,536 Students	Non-Random	1%	3🔒
Goldhaber and Hansen (2013)	Longitudinal Study	9,961 Teachers	Non-Random	47%	3🔒
Heistad (1999)	Longitudinal Study	182 Teachers	Non-Random	22%	3🔒
Koedel and Betts (2011)	Longitudinal Study	471 Teachers	Non-Random	49%	3🔒
Lash et al. (2016)	Longitudinal Study	390 Teachers	Non-Random	40-46%	3🔒
McCaffrey et al. (2009)	Longitudinal Study	2,070 Students	Non-Random	29%	3🔒
Potamites et al. (2009)	Longitudinal Study	103 Teachers	Non-Random	19%	3🔒
Stacy et al. (2018)	Longitudinal Study	5,283 Teachers	Non-Random	35%	3🔒
Goldhaber and Hansen (2010)	Longitudinal Study	7,732 Teachers	Random	97%	1🔒

The strongest study **Rothstein (2009),** rated 4🔒, suggests that the use of additional scores and contextual variables had a limited contribution to the stability of teacher performance evaluation estimation. The study showed that by adding the nearest previous year's test score, which is a 4th-grade score, in the model, the models' R^2 increased from 0.13 to 0.68. Including the two previous years' scores which are pre-test and end of year test scores in Grade 3 to the model in which prior attainment (test score in Grade 4) is already existing, the change on R^2 is 0.039 (R^2 raised from 0.68 to 0.719), suggesting that the additional prior test scores explained the variance for the test score in Grade 5 by an additional 3.9 per cent. In addition, the author also suggested that the estimates obtained from even the best value-added models used in teacher performance evaluation might be biased depending on the number of variables used.

Six other studies also reported limited contribution of using additional years' test scores. All were rated 3🔒. **Cunningham (2014)** examined the contribution of adding additional lagged test scores in the value-added teacher performance estimates. The study used up to three successive years' student data to estimate teacher effects from five value-added models. Teacher rank orderings obtained in the five value-added models using either one or three successive previous years' test scores together showed a high correlation with each other. The correlation among the five models exceeded 0.90 when using single-year data and 0.80 when using multiple previous years of data. The use of one year's test scores instead of three years resulted in a slight increase in correlation between the models and a slight decrease in teacher movement between quarters. **Ehlert et al. (2014)** also found a limited advantage in using multiple-year lagged test scores instead of a single lagged test score in the VAM estimates. The authors estimated the VAM scores for school and teachers by using up to 3 years of lagged scores and the following control variables; eligibility of free/reduced lunch status, language learner status, special education status, race, and gender. The authors compared the full model, which included three years of lagged test scores along with all control variables stated above with various restricted models. The findings of comparing the full model with counterpart models which containing the same demographic features, but only one-year test score, showed that removing additional lagged scores made little difference to the estimates. The correlation analyses were obtained over 0.90 among the VAMs in each teaching subject.

Another longitudinal study with 3🔒 ratings reported a noteworthy contribution of using students' nearest prior test scores, but adding additional previous years test scores to VAM estimates was of limited help. In this doctoral thesis, **Hu (2015**) estimated teacher effectiveness using up to three years of student academic outcomes in maths and reading across five grades (from 4th to 8th), student background characteristics (gender, ethnicity, English proficiency status, gifted status, and student with disability status) and class size. In order to explore the contribution of adding additional previous attainment of students into the teacher effectiveness estimates, the longitudinal students' data (one to three previous year test scores depend on grade and year) was used for creating hierarchical linear models to estimate the value-added scores for teachers. The study found that students' previous years' test scores (up to three years) explained a large proportion of the variance in their current performance. The nearest prior-year's test score alone accounted for an average of 63% of the variance in students' current attainments in both maths and reading. On the other hand, additional previous years' test scores did not contribute much to the variance explained in the current test score once the nearest prior attainment is included in estimates. For instance, 55% of the total variance in students' maths achievement in Grade 7 was explained by the students' maths scores in Grade 6, while 67% of the total variance in students' mathematics achievement in Grade 7 were explained by their attainments in Grades 6 and 5 (the contribution of test score in Grade 5 was around 12%).

Johnson et al. (2015) studied the sensitivity of value-added estimates for teacher effectiveness using different model specifications, including student and peer background characteristics and students' lagged test scores, by conducting multiple regression analyses. The correlation coefficient between a baseline model using one previous year test score, student and peer background variables, and an alternative model that added a second lagged score in the baseline model ranged between 0.958 and 0.988. In order to examine the stability of the teacher VAM estimates, the authors also checked the changes in the average standard errors of teacher VAM estimates and the percentage of teachers who significantly differ from the average. By employing the second lagged score, the average standard error decreased only 0.002 in both maths and reading and the percentage of the teachers who significantly differ from average is increased by 3.2 points in maths and 1.1 points in reading. The findings, therefore, suggest that the additional lagged scores have little contribution to the stability of the teacher VAM estimates.

Another study **(Kersting et al., 2013)** reported little advantage in using additional year lagged scores in the VAM estimates. The authors investigated the contribution of student-level predictors, including gifted status, special education status, free/reduced lunch status, ethnicity, and additional previous year's data to value-added teacher effectiveness estimates. Overall, the correlation coefficient between the model using only one previous year's test score and the corresponding model, which added additional lagged scores was above 0.97. This means that very similar value-added results were obtained in both models, suggesting that additional year's data did not add much to the variance explained. It was also found that 68% of the variance in the students' current scores was explained by controlling for only one previous year's data. Adding a second lagged data in the estimate increased the percentage of variance explained by only 2%. In other words, missing an additional lagged score would have little contribution to teacher value-added estimates.

Schafer et al. (2012) compared six student growth models for teacher effectiveness estimates, including quantile regression (QReg), ordinary least square (OLS), growth score difference (year two minus year one), and three different transition models. The sample consisted of 306 maths and 291 reading teachers from 107 elementary and 28 middle schools, and it was limited only to students with three years of achievement data in Grades 3 through 8 in mathematics and reading. Students with more than one maths and reading teachers in any one school year and those in classes with less than five students or in schools with less than twenty-five students were excluded in the estimates (The reason for the restriction on class size was not specified). To find out the contribution of additional previous years data in the VAMs estimates, two models (QReg, and OLS) were extended using two lagged data. The findings of the study suggest that using additional lagged scores in quantile regression (QReg) and ordinary least square (OLS) models has little advantage in value-added estimates across the four cohorts for maths and reading. The correlations between scores changed (calculated from year 1 to year 2 and from year 2 to year 3) in maths and reading for cohort 1 students were 0.19 using only one prior year's test scores and 0.18 using two prior years' test scores. For cohort 3, the correlation was the same whether using one year's prior test scores or two prior years' test scores. Similar findings were also found for the OLS models.

The other eight studies reported positive results in teacher effectiveness estimates using additional years' test scores. **Lash et al. (2016)**, rated 3🔒, involved students in Grades 4-8 from Washoe County School District, Nevada USA. Administrative data from the school district contained students' test scores and their growth percentile scores associated with their maths and reading teachers for three school years beginning in 2009/10. As the administrative data from the school district were used in this study, naturally, the students were not randomly allocated to teachers, and teachers were not allocated to schools. Since only teachers who were teaching particular subjects in Grades 4-8 and remained in the district within the three school years were included in the estimations, the sample size differed from the actual number of teachers working in the district. For instance, although 696 English language arts teachers worked in the 2010-11 school year, only 375 had test scores for all three years of scores. Similarly, only 369 of the 677 maths teachers had test scores for all three years. This meant that between 40-46% of the data were lost. In light of this, the study was rated as 3🔒. The study investigated the stability of teacher growth percentile scores generated using a number of prior attainment of students; so that the authors compared the reliability of coefficients of the estimations. They found that the stability of performance scores increased from 0.50 to 0.67 when the results were obtained by averaging over two prior years' test scores in maths and to 0.75 by averaging three prior years. Similarly, in reading, the stability of estimates increased from 0.41 to 0.58 by averaging two prior years' performance and to 0.68 by averaging three prior years. However, the researchers cautioned that the level of stability provided was not enough to justify its use in making high-stakes decisions about the teachers (the authors accepted the level of 0.85 or more reliability coefficient for high-stakes decision stated in Haertel, 2013 and Wasserman & Bracken, 2003).

Another study also claimed a substantial improvement obtained in the stability of VAM estimates using additional years of observations **(Goldhaber and Hansen, 2013)**, but the evidence for this study is weaker because 47% of the data could not be used. The authors used up to ten years of the longitudinal dataset, including information on standardised tests in maths and reading in Grades 3 to 5. The number of sample observations obtained (students) was 1,029,259 from school years 1994-5 to 2005-6, as they were restricted to include only Grade 5 teachers with a minimum of 10 students (for a reasonable level of inference) and a maximum of 29 students (it is the maximum class size in elementary level in the state where the data obtained), and students who have current and at least two years of prior test scores in both maths and reading. Consequently, the restricted sample is 541,552 students-year observations and 28,931 teacher-year observations spanning ten years (representing 9,961 unique teachers). The researchers conducted a series of estimates and concluded that using additional prior year test scores in VAM resulted in a noteworthy improvement from 0.29 to 0.52 in the reliability of the estimates in maths.

A doctorate thesis **(Heistad, 1999)** explored the contribution of the number of test scores to the stability of teacher effectiveness estimates in reading by utilising data for up to four years. Along with the student’s reading test score in Grade 2, other predictors included in the model were free/reduced-price lunch, limited English proficiency, special education status, race, parent or guardian “resides with” status, and students’ previous test scores in reading. Pearson product-moment correlations among the value-added coefficients estimated in the three study years were employed for checking the stability of teacher effectiveness as the first stage of the estimates, then the generalizability coefficient (G-studies) was calculated by including only value-added coefficients of teachers with at least seven students over the three study years. Cronbach’s Alpha was used as the generalizability coefficient of this study by the researcher. The analysis showed that the stability of the value-added estimates remarkably increased by using multi-years data. The Cronbach's Alpha increased from 0.48 using single year data to 0.65 for two years of data. The alpha using three years of data was 0.74, and this increased to 0.78 using four years of data. The researcher recommended that in the accountability systems, at least two years of completed value-added data should be used for the high-stakes decisions on teachers.

Another study, **Koedel and Betts (2011)**, rated 3🔒, was also a longitudinal study where 30,354 fourth grade students assigned non-randomly to 595 maths teachers. The study examined the reliability of value-added models in teacher performance evaluation by extending Rothstein's (2010) analysis by employing multiple years of data instead of a single year data. This study used a longitudinal administrative dataset for four cohorts of fourth-grade students in San Diego. Excluding those teachers who have less than twenty students from the dataset (because of concerns about sampling variation, the results are not sensitive to reasonable adjustments at the twenty-student threshold) and including only students who have three adjacent years' test scores resulted in a loss of sample. Only 15,592 students were linked to 389 teachers. This represented attrition of 49% for student records and 35% for teachers. The study compared three model specifications and concluded that the teacher effectiveness estimates based on their students' test scores might contain bias. However, the use of multiple lagged scores resulted in a decrease in the bias of value-added estimates. This finding also concurs with that of Rothstein.

McCaffrey et al. (2009) examined the year-to-year variability in value-added estimates by using a longitudinal data set spanning up to five years for students in Grades 3-8 from five large Florida school districts. Since only teachers who have at least fifteen students (as effectiveness estimates are very inaccurate for the teacher with a small number of students), and only students who have at least two achievement test scores with no missing data were included in the estimates, the number of observations dropped to 24,232 observations (29% attrition). Due to the lack of random allocation of students to teachers, the study was rated with 3🔒. The authors mainly focused on the with-in teacher variance measured by three VAMs; covariate (complete), covariate (partial), and student-fixed model. The researcher presented the comparison of the predictive power of single and two-year average data in teacher effectiveness estimates by county, grade level, and model type. Using one-year data to estimate a teacher's performance reduced uncertainty of their permanent contributions by about 20 to 60 per cent. On the other hand, using two-year average data in estimates resulted in an additional reduction in uncertainty by about 10 to 20 per cent. The authors also argued that these estimates could be expanded using three or more years' data, and the average of three-year data would increase the stability in value-added estimates for elementary school teachers by about 23% and for middle school teachers by about18%.

Potamites et al. (2009) studied the contribution of school and teacher on student achievement by controlling students' prior performances and other factors that beyond the control of school and teacher. A total of 908 teachers who were linked to at least fifteen students (the reason not given) with no-missing current and lagged test scores in mathematics and reading in each class were included, 572 of whom from 63 elementary school, 233 teachers from 30 middle school, and 103 teachers from 21 high schools. In addition to students' test scores in a particular subject, contextual predictors which may link to the students' achievement were also included. These predictors were eligibility for free/reduced lunch, limited English proficiency, special education status, gender, and ethnicity. Using two previous year's test scores resulted in decreasing the estimated mean of the standard error to 0.132 from 0.160, which obtained using one previous year data. Moreover, using additional years of data in VAMs estimates was reported to result in a marked improvement in reliability (1-SSE/σ^2) from 0.387 using one-year data to 0.439 with using two-years' data.

Stacy et al. (2018) also reported a noteworthy contribution of using additional prior years' test scores to the teacher value-added estimates. The study involved a total of 2,985,208 student observations in mathematics and ELA (English Language Arts) in Grades 3 through 6 in an anonymous state in the US. Teacher value-added effectiveness estimates were analysed using students' previous years' test scores and other contextual predictors such as special education status, socioeconomic status, demographic characteristics, and their teacher's experience information. To identify the parameters of the study interest accurately, students taught by more than one teacher in the same subject in the same school year and were in the classroom with fewer than twelve students were dropped from the analysis. The authors specified the proportion of the dropped cases in details, 6.3% of the total observations were excluded because of lack of a link between students and a teacher or being linked with multiple teachers, an addition 2.5% were lost due to missing observations, and finally 26.5% of observations were dropped due to fewer than twelve students' records being available for the estimates of teacher effectiveness. The teachers were simply clustered into three groups: teachers with students whose prior-year test scores in the bottom 25%, the middle 50%, and the top 25%. Although the magnitudes of improvements in the stability of estimates varied across the subgroups, the researchers suggested that the more stable value-added estimates could be obtained by increasing the number of previous years of observations used. The researchers, for instance, found that the reliability coefficient raises from 0.350 to 0.547 by just adding additional prior year data for the fourth-grade maths teachers with students in the bottom quartile, from 0.529

to 0.733 for in middle quartiles, and from 0.521 to 0.725 for in top quartile. Moreover, the findings also distinguished that the estimates for teachers with lower-achieving students were less precise than for other teachers.

The final study, **Goldhaber and Hansen (2010)**, was ranked 1🔒, which is the lowest rank in this systematic review. The study investigated the stability of value-added teacher effectiveness estimates regarding using multiple years data. Given the number of years of data employed, the reliability of VAM estimates was computed and mainly founded that using multiple years of data increases the estimates of reliability. More specifically, although the reliability of VAM estimates was 0.597 based on single-year data, the coefficient increased to 0.784 by adding two-year data and to 0.717 by adding three-year data in reading. Similarly, it was estimated 0.784 based on one-year data in maths and increased to 0.858 with two-year data and to 0.883 with three-year data.

In summary, although there is some evidence that using additional prior year's data increases the stability of value-added teacher effectiveness estimates, the evidence is weak. Only one 4🔒 studies was found. The strength of evidence for each of the 15 studies is shown in Table 8.2.

Table 8.2 The Strength of Evidence: Including Additional Previous Test Scores in the VAM estimates

Quality rating	Increased stability	Minimal increase in stability
4	-	1
3	7	6
1	1	-

Of the 15 studies that examined the stability of teacher effectiveness estimates using the number of previous years' test scores, seven studies suggested that the use of additional test scores adds little to the stability of the estimations. Except for studies that were rated 4 and 1🔒, the rest was rated 3🔒. The eight claimed to have advantages by adding additional prior year test scores in the estimation, but these were rated lower in terms of quality of evidence (seven were rated as 3🔒, and one was given a rating of 1🔒).

The positive studies seem to involve more than three years of data. This may mean that the more years of data used, the more stable the estimates, but because the more years are involved, the greater the loss of data, it is, therefore, difficult to conclude either way. Loss of data is likely to bias the results. More studies rated with 4🔒 are, therefore, needed to confirm the results, but at the moment, there is no very strong evidence that using additional prior test scores increases the stability of teacher effectiveness estimates.

CHAPTER 9

REVIEW OF STUDIES THAT EXAMINE THE STABILITY OF VAMS USING DIFFERENT DATA ANALYSIS METHODS

This chapter reviews studies that examine the stability of teacher effectiveness estimations related to the data analysis methods used in VAMs.

9.1 The Data Analysis Methods Applied

There were 21 studies that evaluate the stability of VAMs in estimating teacher effectiveness that use different methods of data analysis. Nine studies were rated 4🔒, ten had 3🔒, one had 2🔒, and the last one was rated with 1🔒 (see Table 9.1).

Table 9.1 Quality Appraisal of the Studies: The Data Analysis Methods Applied

Author(s) and Year	Design	Smallest Cell	Allocation	Attrition (roughly)	Quality
Castellano (2011)	Longitudinal Comp. Study	25.143 Students	Random	Complete data	4🔒
Goldhaber et al. (2013)	Longitudinal Study	1426Stude nts	Random	3%	4🔒
Guarino et al. (2015a)	Longitudinal Comp. Study	104.441 Students	Random	22%	4🔒
Guarino et al. (2015b)	Longitudinal Comp. Study	2160 Students	Random	68%	4🔒
Guarino et al. (2015c)	Longitudinal Comp. Study	120 Teachers	Random	Complete data	4🔒
Hong (2010)	Longitudinal Comp. Study	1200 Students	Random	Complete data	4🔒
Kurtz (2018)	Longitudinal Comp. Study	1000 Students	Random	Complete data	4🔒
Parsons et al. (2019)	Longitudinal Comp. Study	6000 Teachers	Random	Completed data	4🔒
Shaw (2012)	Mixed Factorial Design	324 Teachers	Random	Complete data	4🔒

Blackford (2016)	Longitudinal Comp. Study	13.087 Students	Non-Random	1%	3🔒
Cunningham (2014)	Longitudinal Comp. Study	1001 Students	Non-Random	20%	3🔒
Dwyer (2016)	Longitudinal Comp. Study	11.215 Students	Non-Random	19%	3🔒
Garai (2017)	Longitudinal Study	1350 Students	Non-Random	Complete data	3🔒
Goldhaber et al. (2014)	Longitudinal Comp. Study	3820 Teachers	Non-Random	54%	3🔒
Sass et al. (2014)	Longitudinal Comp. Study	60.000 Students	Non-Random	18%	3🔒
Schafer et al. (2012)	Longitudinal Comp. Study	5536 Students	Non-Random	1%	3🔒
Schmitz (2007)	Longitudinal Comp. Study	6044 Students	Non-Random	1%	3🔒
Sloat et al. (2018)	Longitudinal Comp. Study	69 Teachers	Non-Random	20%	3🔒
Wei et al. (2012)	Longitudinal Comp. Study	58 Teachers	Non-Random	14%	3🔒
Germuth (2003)	Longitudinal Comp. Study	258 Teachers	Non-Random	66%	2🔒
Newton et al. (2010)	Longitudinal Comp. Study	103 Teachers	Non-Random	Not Reported	1🔒

The first study with the highest-ranking score (4🔒) **(Castellano, 2011)** focused on creating a more accurate alternative model to student growth percentiles (SGPs). Although SGPs are more popular among the growth models, and many states in the USA, such as Michigan, Colorado, Georgia, New Jersey, etc., started to use SGPs in their accountability systems, the existing literature suggested that the model also has a handicap of its performance in small samples and including the number of prior year test scores in the estimates (Culbertson, 2016; Castellano &

Ho, 2013a). The researcher, therefore, proposed an alternative model -the percentile rank of residuals (PRRs)- to be used, especially in small sample sizes. As the study used a longitudinal research design with sufficient sampling size and allowed random allocation with a minimum attrition rate in their estimations (complete data), it was rated 4🔒. The simulated multivariate normal (MVN) data set -drawn from the two state-wide empirical datasets- for students from Grades 3 to 6 was used for comparing estimations. The simulated sample size levels spanned a range from 250 to 10,000 and contained up to three prior years maths and reading test scores. Two separate analyses were conducted to assess the accuracy of SGPs and PRRs' recovery NCGPs (normal conditional growth percentiles) under varying factors of the prior years and the sample size. Under both factors, PRRs consistently better recovered expected growth percentile and NCGPs than SGPs. To compare an average discrepancy between two metrics obtained from analyses, the researcher employed the root mean square difference (RMSD). RMSDs between SGPs and NCGPs were about 2 to 3 times larger than the corresponding RMSDs between PRRs and NCGPs. PRRs were more accurate and stable for small samples like 250 and 1000, but SGPs started to provide fair enough estimates for a sample with at least 5,000. Moreover, to address the robustness of the SGPs and PRRs with regard to scale transformations, five different monotonous scale transformations (i.e., positive/negative skewness, positive/negative kurtosis, and exponential) were applied with one previous year of test score data. In these analyses, SGPs generally have superiority over PRRs, and they substantially provide a higher degree scale invariance than PRRs.

The next study **(Goldhaber et al., 2013)** discussed the value-added models at the high school level. The dataset consisted of a total of 8,002 students (in Grades 9 through 12) linked to 205 teachers from 23 high schools (9 of which are private schools). The data were collected by ACT (the American College Testing Program) as part of a pilot of their QualityCore end-of-course assessment in the Midwest, US, and contained students' test scores in multiple teaching subjects; algebra I and II, biology, chemistry, English, geometry as well as student, teacher, and school characteristics such as student gender and ethnicity information, teachers' college major and GPA, highest degree held, certification, and experience, class size and school the average ACT college entrance score. The study compared teacher effectiveness estimates derived from traditional lagged VAM, which employing pre- and post-test scores in one teaching subject and from the alternative models, using a cross-subject student fixed-effects approaches (student fixed-effects model and student fixed effects with lagged score model). The researchers compared the Value-added models based on the estimated teacher effect size.

It was found that the estimated teacher effects from the traditional lagged model are steadily much higher than for both the student fixed effects models; it was about 1.5- to 2 times higher than the estimated effect sizes derived from the student fixed-effect model, and even much larger than the comprehensive model, student fixed effects with lagged score model. In shortly, the study concluded that the model specification affects both the estimated teacher effect size and the estimates of individual teacher effectiveness.

Another study with the highest quality rate **(Guarino et al., 2015a)** compared SGPs and VAMs' ability to rank the teachers accurately by using simulated and real data from a large anonymous school district in a southern state. Two SGP approaches -SGP-median and SGP-mean- and three Value-added models -dynamic ordinary least-square (DOLS), average residual (AR) and empirical Bayes (EB)- were compared under the following scenarios; a) random grouping and random assignment (RG-RA), b) dynamic grouping coupled with random assignment (DG-RA), c) the dynamic grouping with a positive assignment (students with the lowest prior-year achievement level tend to be assigned to teachers with the lowest effectiveness or vice versa) (DG-PA), and d) the dynamic grouping with a negative assignment (students with the lowest prior-year achievement level tend to be assigned to teachers with the highest effectiveness or vice versa) (DG-NA). The researchers employed the Spearmen rank correlations of the teacher effects estimated in each model with the true teacher effectiveness that was known. Under the scenario of random grouping and random assignment (RG-RA), all models performed fairly similar, and Spearmen rank correlations ranged from .82 to .87. In addition, the researchers checked the misclassified teacher ranking generated by the models, such as the teachers who have a true teacher effect above the 25th percentile, but clarified as in bottom 25%, and the teachers whose true effect below the 25th percentile but rated as above the bottom 25%. While a similar amount of misclassified teacher rate in each model was calculated, the rate was a little high for SGP-median. The percentage of teachers above the bottom 25 per cent in true effect, but misclassified in the bottom 25 per cent was 8 (10% for SGP-median), and the percentage of teachers in the bottom 25 per cent in true effect, but misclassified in the top 75 per cent was around 24 (29% for SGP-median). However, when the assignment of teachers to students is not random, the patterns change dramatically. In the DG-PA scenario, the DOLS estimator maintains a similar Spearmen rank correlation, which was .88, whereas the SGP-Median and SGP-Mean rating correlations were decreased to 0.71 and 0.76, respectively. The study concluded that in situations where students were dynamically grouped based on previous year test scores and were randomly assigned to teachers, the DOLS evaluator maintained a strong relationship with the true teacher influence, while the estimated SGPs performance was weak compared with the DOLS estimator. Using actual data, the researchers also detected that the divergence between the DOLS and SGP estimates was much greater in the non-random grouping scenarios.

The next study **(Guarino et al., 2015b)** examined the relationship between the applied data analysis method and the stability of teacher performance evaluation estimates by comparing empirical Bayes's (EB) estimation with other widely used value-added models under different grouping and assignment scenarios. Simulated data where the true teacher effect is known and real student achievement data were used to compare the ability of EB models to rank teachers properly with other commonly used value-added models, such as mean residual (AR) and dynamic ordinary least-square (DOLS) models. The researchers grouped the students into the classrooms neither randomly, or non-randomly (dynamic grouping (DG) where the students were grouped into classrooms based on their prior academic performance or heterogeneity grouping (HG) where the groupings based on their unobserved heterogeneities), and similarly their teachers were assigned either randomly (RA) or non-randomly (positive assignment (PA) where the worst teachers are assigned to classrooms with the worst students or vice versa, and negative assignment (NA) where the worst teachers are assigned to classrooms with the best students or vice versa). The researchers compared the fixed effect (DOLS) and random effect models (EB and AR) by conducting the Spearmen rank correlation between the true teacher effects and the estimated teacher effects by these estimators. Generally, a very small substantial difference was found between the estimates from fixed and random effect models under random grouping and assignment scenario; the rank correlation was 0.85 for both fixed and random effect models. Again, the researchers also reported the misclassification rate, which indicates the percentage of the teachers whose actual effectiveness is above average but calculated as below average, and the misclassification rate was found as 0.15 roughly. For all non-random grouping and assignment scenarios, DOLS outperformed the other estimators. The researchers reached a conclusion by using simulated data that although EB models generally performed well and similar to other estimators in random grouping and assignment scenario, their performance suffered under all other non-random scenarios. In parallel with the results obtained with the simulated data, the researchers found similar correlations between the models in the real data. The median rank correlation was around 0.99 for between DOLS and AR, and 0.97 for DOLS and EB.

(Guarino et al., 2015c) examined whether widely preferred value-added models provide accurate teacher effectiveness estimates under different grouping and assignment scenarios. For the comparison of six estimators, the researchers employed simulated data. The simulated study involved 10 schools, 120 teachers and 960 pupils. The researcher compared a) dynamic ordinary least-squares (DOLS), b) average residual (AR), c) pools ordinary least-squares (POLS), d) the instrumental variables/Arellano and Bond approach (AB), e) random effect and finally f) fixed effect. The researchers grouped the students into the classrooms randomly or non-randomly (dynamic grouping (DG) where the students were grouped into classrooms based on their prior attainment, or static groupings, one of which is BG where the students were grouped into classrooms based on their baseline test scores, and the other is HG where students were grouped based on their unobserved heterogeneities), and the teachers were assigned to classrooms either randomly (RA) or non-randomly (positive assignment (PA) where the worst teachers are assigned to classrooms with the worst students or vice versa, and negative assignment (NA) where the worst teachers are assigned to classrooms with the best students or vice versa). The researchers estimated individual teacher effects by employing each of the six models and computed the Spearman rank correlation between the estimated individual teacher effects and their known true effectiveness. Under the random grouping and assignment scenario, although DOLS, AR, POLS, and RE estimators provide similar teacher effect estimates with rank correlations of about 0.87, FE had a rank correlation near 0.65 and the correlation for AB being even worse with 0.59. Under the scenarios which the students were grouping non-randomly into classrooms, but their teachers were assigned randomly (DG-RA, BG-RA and HG-RA), the correlations of DOLS, AR, POLS, and RE estimators remained above 0.80; similarly, FE and AB continued to provide worse ranking correlations with around 0.60. Across all non-random grouping and assignment scenarios, DOLS performed better over the other estimators, with a correlation at 0.84 or higher (except HG-NA scenario). The researchers also reported a misclassification percentage measure, which represents the teacher who was misclassified as below average in their estimated effectiveness, even though their true effectiveness were above average. Across all scenarios, DOLS provided the lowest misclassified percentage with below 18 compared to the other estimators (except HG-NA scenario). The study concluded that although none of the estimators accurately estimate true teacher effects across all scenarios, DOLS provided the most accurate teacher effectiveness estimates across all grouping and assignment mechanisms, except the scenario of heterogeneity-based grouping with a negative assignment.

Another study **(Hong, 2010)** evaluated the sensitivity of teacher value-added models by comparing a general VAM with five restricted models. The simulated study contained a total of 1200 students' three consecutive years data in 48 classrooms (each classroom has 25 students) from three different school settings. For creating the heterogeneity among the school settings, three different school settings were generated (80% of the total population in school A were eligible for free and reduced lunch (FRL), this rate was 50% in school B, and 20% in school C). Across three consecutive years, effective and non-effective teachers were assigned into 16 classrooms (Teachers who contribute positively to their students' achievement were considered effective teachers). Within each school setting, four combinations of teacher assignment scenarios into classrooms were applied, which were class NNN, class NNE, class EEN and class EEE (An example for the class EEN, effective teachers were assigned in year 1 and 2, but the non-effective teacher was assigned in year 3). Moreover, for each school setting, the students were randomly assigned into two classes. The researcher compared a multivariate general VAM with two single wave models, which were gain score (GS) and covariate adjustment (CA) models, and three multiple wave models, which were layered (LA), cross-classified (CC) and persistence (PS) models. To explore whether a model provides accurate estimates, the estimated teacher effects were compared with the true teacher effect by employing spearman's rank correlation coefficients. The range of the correlation coefficients was from 0.81 to 0.94. The lowest coefficient (0.81) was found when CA model was employed using school C's data in year 3, and the highest correlation (0.94) was observed when the general VAM in school A in year 2 and 3, in school B in year 3 and in school C in year 1. The researcher concluded that, in general, all models provide satisfactory teacher effectiveness estimates under various scenarios, but the general multivariate model still provides more accurate estimates consistently under all assumptions. Moreover, the mean absolute bias for 16 teachers within each school in each year was estimated. In general, none of the models produced great accuracy; the mean absolute bias ranged from 2.91 to 7.95. Similarly, the general model produced the lowest bias across all years and schools and among the all-other reduced models, PS models produced the second-lowest bias estimates.

Kurz (2018) compared two common models used for teacher effectiveness estimates under different classroom conditions (the researcher called "conditional skewness"), which were the value-added model (based on OLS) and the student growth percentile (SGP) models. To be able to manipulate any allocation bias to occur, a simulated data set was employed in this study, and then the researcher also conducted the same analysis using the observed data from North

Carolina schools (the North Carolina Education Research Data Center (NCERDC) at Duke University) to confirm whether the results are similar. The dataset included 18,821 math teachers who linked to the students with 4th-grade end-of-grade math exams, while the researcher also created 1000 simulated classrooms for simulated analysis. To find out the impact of having a disproportionate number of students in teacher effectiveness estimates, three conditional skewness of classrooms were created; tercile 1 contained an approximately equal proportionate number of students with positive residuals and negative residuals, tercile 2 contained a large proportion of students with positive residuals, and appositely tercile 3 contained a large proportion of students with negative residuals. By using the simulated data, the researcher reported that the models' agreement between standard VAM and SGP within 1 decile was 97.7%, and the total correlation was 0.976. The study also reported that the correlation between the standard models in the first conditional skewness tercile was relatively high at 0.982; however, the correlation measure was the lowest in the third skewness tercile at 0.971. The same analyses were done with employing observed data, and the percentage of the model agreement between standard VAM and SGP within 1 decile was found 84.1%, and the total correlation was 0.936. Similarly, the study also found that while the correlation between the standard models in the first conditional skewness tercile was relatively high at 0.958, the correlation measure dropped to 0.912 in the third skewness tercile. Finally, the researcher also compared the estimated teacher effectiveness with their actual known effect using simulated data. The study reported that the correlation between the actual effects and VAM estimates (0.967) was higher than SGP (0.945). The study concluded that although the standard VAM produced estimates closer to the actual effectiveness, under the classroom condition with a disproportionate number of overachievement students, the VAM provided exaggerated estimates compared to the SGP estimates, or vice versa.

The next study with a 4🔒 **(Parsons et al., 2019)** examined the one-step fixed effect and two-step aggregated residuals models across various sorting scenarios using a simulated data set. The researchers generated the simulated data based on a realistic student-school sorting condition that reflects the real elementary school catchment areas in urban and suburban school districts in Kansas City, Missouri. In short, students were sorted to elementary schools regarding the economic status of their parents. For each school, six teachers were appointed, and there were 100 elementary schools (the total number of teachers were 600). To determine the accuracy of the value-added estimates, the estimates derived from the two models were correlated with the true teacher effectiveness values. Overall, the highest estimates were reported when the continuous income variables are used in both model; the correlation coefficient was 0.700 in comparison to the one-step and 0.706 for two-step value-added estimates. While lower correlations were reported in the estimates when using FRM proxy, the two-step VAM performed better with 0.679 (the correlation was 0.660 for one-step VAM). In addition, the researchers compared their baseline estimates to various simulation scenarios such as various FRM misclassification ratings, teacher sorting (general) and teacher sorting within the school. The overall conclusion of the study was that the two value-added models performed similarly across the various scenarios. Although the differences between the estimates from the two model were generally fractional, two-step VAM performed more accurate estimates under the most reasonable conditions for sorting and quality data, while one-step VAM performed better under extreme conditions.

The last study having the highest quality rate **(Shaw, 2012)**, compared the stability of teacher effectiveness estimates and rankings from univariate and multivariate models under a variety of model combinations. The study examined the consistency of the estimates from the simplest value-added model (called “longitudinally invariant parallel univariate static score” by the researcher) to complex (called “longitudinally non-invariant multivariate latent growth model”) by being carried out with a mixed factorial design. This doctorate thesis utilized the data from Project STAR -teachers and students randomly assigned to three types of classrooms which were a) small classes contain 13 to 17 students, b) regular classes have 22 to 25 students with no aide, and c) regular classes with a paid aide have 22 to 25 students with a full-time teacher aide-, and involved roughly 2,000 students randomly selected from K-2 school setting linked with 327 teachers at the first two time points (kindergarten and grade 1) and 324 teachers at the final time point (Grade 2). In order to compare the consistency of teacher effectiveness estimates and rankings, along with using Spearman’s rho rank-order correlation measure, two additional measures were also employed; lowest quartile rank consistency (Cohen’s Kappa was used) and estimate precision (teachers were grouped into three categories based on their 95% confidence intervals – below expected, expected, and above expected). Although no single model combination produced robust estimates, overall, the study found that multivariate models have the potential to reduce the misclassification of teachers comparing to univariate models. While most of the univariate static and gain score models had a rank-order coefficient over 0.70, these models also produced a remarkable amount of fluctuations across experimental combinations. Although the majority of Kappa measures of multivariate models reached or exceeded a benchmark value of 0.61 (Landis and Koch, 1977), and generally had the most substantial values for scale comparison conditions, the Kappa values in univariate models had a higher range between 0.87 and 0.93 compared to the value range across multivariate models between 0.56 and 0.84. Overall, the study concluded that multivariate models performed well across all experimental conditions compared to the univariate model with having some exceptions.

The next studies rated with 3🔒 **(Blackford, 2016)** compared the teacher effectiveness estimates and rankings across VAMs. This study involved roughly 13,500 fifth grade students linked with 318 maths teachers. The researcher examined the consistency of teacher rankings from four commonly used models, gain score model (GM), covariate-adjusted model (CM), layered model (LM) and equipercentile model (EM), by employing students' third, fourth and fifth-grade end-of-year maths scores and their demographic characteristics across five regions in Arkansas. Regarding the investigation of whether the teachers received similar ranking regardless of the models applied, the study reported strong correlations between all model pairs. While only one correlation was estimated less than 0.80, which was between the equipercentile model and the layered model in 2012 (0.77 – still strong correlation), the strongest agreements were found between the gain score model and the covariate-adjusted model in 2012 (0.98) and 2013 (0.97), and between the gain score model and the equipercentile model in 2014 (0.96). The study also reported the Kappa coefficients (W) between pairs to investigate whether the teachers were classified into similar groups and indicated that intermediate to good agreements were reached across the classification (0.40-0.75). In general, across the models, 65% of the teachers (n= 207) were consistently classified into the same effectiveness category in 2014. Moreover, overall, across all three years, moderate to strong agreements (0.40-0.79) were also found for the teacher ranking correlations within each model. Specifically, the correlation coefficients ranged from 0.49 to 0.59 for LM, 0.49 to 0.55 for EM, 0.51 to 0.61 for GM and 0.55 to 0.63 for CM. Based on the second investigation aspect - teacher effectiveness categories-, each model identified less than 50% of the teachers as the same effectiveness categories across all years, so slightly fair Kappa coefficients were reported. The researcher also ranked all teachers in each effectiveness category for each subpopulation group (based on the teachers working in a district with low, medium, and high poverty and minority students), then compared their rankings in each category across models. The strongest model agreement was found between CM and GM among the subpopulations of poverty and minority groups over the three years (the range of coefficient were 0.829-0.979 for poverty and 0.918-0.979 for minority subpopulations). The weakest model agreement in teacher effectiveness ratings was found between EM and LM among all subpopulation groups across all years. The correlation coefficients were, however, still strong (even very strong in some comparisons) with ranging from 0.68-0.86 for poverty and 0.69-0.83 for minority subpopulations. The study concluded that since each model produced similar results, no one model is superior to the other models regarding their ability to consistently classify teachers. However, the researcher suggested that if each model produces similar results, it might be preferable to use the least expensive model, which requires a minimum amount of data and is easy to understand.

The following two studies rated with 3🔒 were explained in the previous sections. The first study (**Cunningham, 2014)** examined the relationship between modelling preferences and value-added teacher performance estimates by comparing the correlation coefficients between the models. The researcher estimated teacher effectiveness scores derived from five value-added models then compared the rank-ordering of the teachers generated from each model. The five value-added models compared were a) a simple covariate adjustment model (CA1) that use students' prior attainments only, b) a covariate adjustment model (CA2) that used contextual student characteristics along with their prior attainments, c) a gain score model (GAIN) that is underlying the difference between students' current and previous attainments, d) Iowa growth model (IOWA) that is based on the average differences between current years' performance in the vertical scale and the expected performance in the same year estimated by one of the qualities of prior-year test scores, and e) student growth percentiles (SGPs) that use students prior attainments to qualify their current academic performance by utilising quantile regression method. The teacher rank-orderings derived from five VAMs by employing single year data had very strong correlations with each other in a range of 0.908 to 0.993; the highest correlation appeared to arise between CA1 and GAIN (0.993), and oppositely the weakest correlation between IOWA and SGP was reported as 0.908. Similar but weaker results were found for multiple-years analysis with the Spearman's correlations range of 0.834 to 0.972. Although the highest correlation between CA1 and SGP was reported as 0.972, the lowest correlation was obtained again between IOWA and SGP with 0.834. The percentage of movement of teachers between quartiles ranged from 12.3% between CA1 and GAIN, to 34.7% between IOWA and SGP for single-year data analyses, and 14.7% between the CA1 and SGP models to 40.3% between the IOWA and SGP models for multiple-year analyses. Overall, the study concluded that although all models produced similar consistent rank-ordering results, the less consistent rank-ordering results were derived from the IOWA model.

Schafer et al. (2012) also compared six student growth models; quantile regression (QReg), ordinary least square (OLS), growth score difference (DifGr), and three different transition models (value-tables – TUp, TUpDn and TProg). In the student-level analyses, the correlation between paired models ranged from moderate (between TProg and DifGr with a coefficient of 0.51) to very high (between QReg and OLS with a coefficient of 0.95) positive correlation. The lowest correlations were obtained in comparisons with TProg. In general, similar but higher correlations results were obtained among the comparisons in teacher and school levels. The correlation between QReg and OLS raised to 0.98 in teacher-level analyses and 0.99 in the school level analyses. Again, the lowest correlations appeared to arise between TProg and the other growth models. The weakest correlations between TProg and DifGr was reported as 0.45 and 0.56 in teacher- and school-level analyses, respectively. The researcher concluded that the regression-based models, QReg and OLS, produced very similar results, so preferring one model over the other model has very little advantage.

Another doctorate thesis **(Dwyer, 2016)**, ranked 3🔒, examined the degree to which concordance of teacher rankings and classifications derived from value table and the covariate-adjusted regression model, using a longitudinal dataset spanning two school years from 2010 to 2012. The study involved a total of 1,635 maths teachers linked with 60,167 students in Grades 4 through 8. The researcher generated Pearson product-moment correlations for two sets of value-added scores generated from the value-table and the covariate regression model, and strong associations between the two approaches were found ranging from 0.981 to 0.772. Then, the value-added scores were transformed into quintiles, and the agreement/disagreement analyses were conducted regarding the quintile rankings. For fourth grade, out of the 526 teachers, 88 (17%) were assigned to higher quantiles, and 96 (18%) were assigned to lower quantiles in the value table than the covariate-adjusted model. The total disagreement was 35% in the fourth grade. For the second step of the analysis, the two sets of scores were classified into four categories; highly effective, effective, needs improvement and unsatisfactory, and the agreement/disagreement analyses were conducted. Again, for fourth grade, out of the 526 teachers, 15 (3%) were assigned into higher classifications, and 4 (1%) were assigned into lower classification in the value table than the covariate-adjusted model. The total disagreement was 4% in the fourth grade. The study concluded that as the concordance of the two methods ranged from 94% to 99%, the value table model might be preferred as a proxy of the complex statistical model, the covariate-adjusted regression model, in the teacher evaluation classifications.

The next study **(Garai, 2017)** rated with 3🔒 proposed a new multi-stages model as an alternative model to one of the most popular value-added models, the Tennessee Value-Added Assessment System (TVAAS), for estimating teacher effectiveness, especially in small school systems. Although a significant number of educational researchers examined TVAAS in large school systems (Bacher-Hicks et al., 2014; Chetty et al., 2014; Jiang et al., 2015; Rivkin et al., 2005; Sanders and Horn, 1994), the model, like other value-added models, has a handicap of own performance in classroom with a small number of students. The researcher generated a simulated data set involving 5 districts and 6 schools in each district. There were three grades in each school, and different teachers were assigned in each grade (the total number of teachers is 90). The class sizes kept the same, with 15 students in each. As the study used simulated data, the true teacher VA scores were known; therefore, the study examined the accuracy of the scores derived from the traditional TVAAS and the multi-stage TVAAS by comparing them with the true VA scores. First of all, the researcher divided true teacher rankings into deciles, and then the same dividing processes were executed for the rankings generated by the standard and small area method. Later, reported the percentage of agreement/disagreement between the ranking results. Overall, the multi-stage small area method produced more closely ranking results to the true teacher rankings; however, although the alternative model may produce better performance, the model only correctly identified 56% of teachers in the 1st decile and 78% of teachers in the 10th decile. The percentages of teachers identified correctly by the standard TVAAS modelling method in decile 1 were 44% and 67% in decile 10. In parallel with these findings, the researcher concluded that none of the methods performed accurately in regard to the teachers' effectiveness rankings.

Goldhaber et al. (2014) examined the extent to which the teacher effectiveness estimates generated by SGPs agreed with the estimates derived from three VAM specifications. The researchers used longitudinal data from North Carolina, including students standardized test scores in maths and reading in Grades 3 through 5, their background information, and their teachers' credentials and job records spanning thirteen years. After excluding classrooms having less than 10 and more than 29 students (the maximum student per class in the study district) and the students with missing prior-year test scores, the study involved 20,844 maths and reading teachers. Researchers compared the teacher effectiveness estimates derived from median grow percentile (MGP- also known as SGPs), and derived from a) student background VAM (along with employing the students prior math and reading test scores, their background information were included such as gender, ethnicity, educational status of parents, English-language learner, eligibility of free/reduced lunch and disability status), b) classroom characteristics VAM (classroom-level variables such as class size, mean of prior year math and reading performance, percentage of students with FRL and disability, percentage of the students in minority societies, and percentage of students' parents with a bachelor or higher degree in the classroom were added into student background VAM) and c) school fixed effects VAM (school fixed effect was added into student background VAM).

Overall, the study revealed that a one standard deviation increase in teacher effectiveness leads to an increase in student achievement of approximately 0.15-0.25 standard deviation. The effect sizes in math (ranges from 0.22 to 0.25) were found much larger than in reading (ranges 0.15 to 0.20). Thereafter, the Pearson's correlation coefficients between the estimates generated from each model were computed and reported that the correlation coefficients for the teacher effectiveness estimates in math were slightly higher than in reading. Specifically, for each of the model comparisons except the model of school fixed effect VAM, the coefficients were found as over 0.90 for math teachers and over 0.80 for reading teachers. While the highest correlations with 0.99 for both mathematics and reading occurred between VAM specifications where the student- and teacher-level covariates were controlled, the lowest correlations were found between MGP and the school fixed effect VAM, which was 0.61 in math and 0.48 in reading. The researchers also generated the percentile ranking for the teacher based on the estimates from the models in order to examine the relationship between classroom types to which the teachers were assigned and their ranking. Three classroom compositions were created as advantaged, average and disadvantaged classrooms regarding the mean of prior attainment of the students enrolled in the classroom and the percentage of students who are eligible for free/reduced lunch and are in a minority group. Then, the percentile rankings for individual teachers were clustered into the classroom types. The study reported that more effective teachers tended to be assigned to advantaged classrooms. The availability of effective teachers between advantaged and disadvantaged classes differed substantially across models, except in school fixed effect VAM. In other words, while mathematics teachers with an average percentile rank of 61.9 were assigned to advantaged classes in MGP, this rate was 41.4 in disadvantaged classes. As the true teacher effectiveness is unknown, the study could not be stated that a model produces more accurate results than the other models. However, the researcher revealed that each model generated distinctly different teacher effectiveness rankings under the classroom compositions where teachers serve different type of students regarding prior attainment and background characteristics.

Another study **(Sass et al., 2014)** compared the general cumulative model to its various specifications, such as whether accounting prior inputs (student/classroom level), multiple lagged scores, decay in the prior inputs and decay in individual-specific effect. The researchers utilized a longitudinal dataset covering teachers' and students' information and at least three consecutive year test scores of the same students in Florida public schools. A total of 1,951 math teachers' effectiveness was estimated using 196,015 observations. The researchers reported the rank correlations for the teacher effectiveness estimates between pairs, and the highest correlations were obtained between partial persistence models using one and multiple lagged test scores, exceeding 0.90. So, they suggested that including additional prior year test scores does not cause large differences in teacher effectiveness rankings. With considering all other analyses conducted, the researchers reached a conclusion that the estimated teacher effectiveness might be very sensitive to model specifications, so in order to obtain estimates with minimum bias, it was suggested to prefer models with more flexible specifications, such as employing three lagged test scores in addition to three lags of inputs.

Schmitz (2007) investigated whether more sophisticated models can produce substantially different teacher effectiveness estimates than simple models. The study compared seven different value-added models, which were a) simple fixed effect model (model 1), b) unconditional 2-level hierarchical linear model (model 2), c) conditional random intercept 2-level hierarchical model (model 3), d) unconditional 3-level hierarchical linear model (model 4), e) conditional random intercept 3-level hierarchical linear model (model 5), f) unconditional cumulative effect model (model 6), and g) conditional cumulative effect model (model 7). Fourth and fifth-grade maths and reading test scores of a total of 34,099 students, who were linked with 978 math and 945 reading teachers from 1,132 elementary schools, were utilized. In addition to students at least two consecutive year test scores, the data set also contained student-, teacher and school-level predictors. While teacher-level variables were used in the second level of the hierarchical models, school-level variables were clustered into the third level. The study revealed that, except for the adjusted cumulative effect model, all value-added models examined produced very similar teacher effectiveness estimates. The lowest average correlation between the conditional 2-level hierarchical model and the conditional cumulative effect model was estimated, which was 0.905. Based on the correlation analyses between estimates, it was revealed that teachers’ effects were not affected by the absence of the control variables, while the quartile agreement analyses indicated that teachers’ effects were affected by the absence of the control variables. The correlation of the estimates between models 2 and

3 was 0.946 and between models 4 and 5 was 0.945; however, the agreement in the lower quartile was 82% between models 2 and 3 and was 81% between models 4 and 5, and the agreement in upper quartile was 82% between both companions. The research also found that the percentage of variability in the student academic attainments predicted by cumulative impact models attributed to teachers' effectiveness is much greater than other models included in this research.

Sloat et al. (2018) examined the concordance of the teacher effectiveness ratings using six different value-added models in three grades in two subject areas. The study involved 5,496 students in Grades 4, 5, and 6, linked with a total of 221 math and reading teachers. Along with students' math and reading attainment scores spanning two years, the longitudinal data set also contained other student-level variables, including the status of eligible for free/reduced lunch status, home language, English-language learner status, gifted and special education status. The researchers compared student growth percentile (SGP) model to the five models, including value-added linear regression model (VALRM), value-added hierarchical linear model (VAHLM), simple difference (gain) score model, rubric-based performance level (growth) model, and simple criterion (per cent passing) model. The effectiveness ratings derived from the models were transformed into three teachers' effectiveness categories; low, moderate and high, then the categories in which teachers were placed in all models were compared. Among the comparisons to SGP, the highest Spearman's correlations were found between SGP and VALRM; the coefficient values were ranging from 0.82 (for maths and reading in Grade 6) to 0.92 (for reading and maths in Grade 4). Comparisons with the simple criterion model had the lowest correlations among all models; the r_s values ranged from 0.08 (with simple difference (gain) score model in reading in Grade 6) to 0.73 (with rubric-based performance level model in reading in Grade 4). Moreover, although the estimates obtained from other models were consistent with the results of the SGP, the amount of disagreement regarding assigning teachers to one of three effectiveness categories across these models was still substantial. In comparison with SGPs, the Kappa measures were obtained as the highest in Grade 4 mathematics with VALRM (0.79), and the lowest in Grade 4 reading with rubric-based performance level model and in Grade 5 with per cent passing model (0.11). The Kendall tau-c values also ranged from 0.30 in Grade 6 reading (rubric-based performance level model and simple criterion -per cent passing- model) to 0.84 in Grade 4 mathematics (VALRM). On the other hand, it was also revealed that the percentage of disagreement in teachers' effectiveness ratings ranged from 14 to 59 depending on grade, subject area, and method used. After considering all analyses conducted, the researchers reached a conclusion that the classifications of teacher rating substantially varied depending upon the preferred model for evaluating teacher effectiveness.

The last study in the group of the study was ranked with 3🔒 **(Wei et al., 2012)** examined the degree to which the consistency of the teacher effectiveness estimates derived from five value-added models. The researchers utilized a longitudinal dataset from a large urban school district in Texas which contained students' achievement data in maths and English language art (ELA) from Grades 3 to 5 over three years, and their demographic characteristics including gender, race, English language learner status, special education status, and eligibility of free/reduced meal. The study sample consisted of 73 math teachers and 53 ELA teachers in Grade 5. Teacher rank-orderings were generated in each content area on the basis of their effectiveness estimates derived from per cent passing change model (model 1), average score change model (model 2), multiple regression model (model 3), hierarchical linear regression model (model 4) and layered mixed-effects model (model 5). Overall, the correlation of teacher effectiveness rating between pair models ranged from medium to low. For instance, ELA teacher 1 was placed the best ranking (1st) in model 2 and 3 but was assigned a rank of 58 (out of 58) in model 1. The correlation coefficients obtained in the two teaching subjects denoted that the effectiveness rankings from the five value-added models are only moderately associated in the best-case scenario, even negatively associated in some cases. The highest correlation appeared to arise between the average score change model and the multiple regression model (0.670). Negative correlations were also obtained between the per cent passing change model and the hierarchical linear regression model (-0.221) and between the per cent passing change model and the layered mixed-effects model (-0.163). Overall, this study revealed that value-added teacher effectiveness estimates are highly sensitive to model preferences. Moreover, the researchers concluded a suggestion that in addition to value-added models, other measures using in teacher evaluation such as expert/principal observation, portfolio, student survey should be preferred in order to get a complete picture of the impact of the teacher on student learning.

The next doctorate thesis **(Germuth, 2003)**, rated with 2🔒, examined the degree of the consistency of estimates derived from various specifications of HLM and OLS models in identifying effective maths teachers with employing data from middle schools. In this study, four specifications were compared: an OLS model with seven student- and one school-level predictor (model 1), an HLM with the same predictors (model 2), an OLS including more student-, teacher- and school-level predictors (model 3), and an HLM model corresponding to the latest OLS model (model 4). The researcher ranked the teachers based on the estimates obtained from four model specifications, then Kendall’s coefficients of concordance statistics (Kendall’s W) were used to compare the ranking consistency for the 258 maths teachers. No matter which models are compared, Kendall's W showed a high degree of agreement (range 0.974 - 0.999), meaning that all models produced very similar rank-ordering of the teachers. While the highest correlation appeared to arise between model 1 and model 2 (0.999), the weakest correlation was reported between model 2 and model 3 with 0.974. G study covariance was also estimated in order to determine to what extent variance is directly associated with teachers, models, and teacher-model interactions. Findings from G study statistics also supported Kendall' W’s results that although none of the variances in teacher effectiveness rankings was associated with the models themselves, almost all variances obtained were related to the teachers (94.97% - 99.84%). Since the simple model produced similar results with the sophisticated models, the researcher suggested preferring the simplest model (model 1), which is accurate and functional in identifying effective and ineffective teachers as in the other three models in this study.

The last study retrieved in this section, ranked with 1🔒, **(Newton et al., 2010)** examined the stability of effectiveness ratings of teachers in high school across the model specifications, teaching subjects, teaching years. Although it is also a longitudinal study with adequate sampling size, as the attrition rate of the study or the number of cases included in the analysis was not reported directly or indirectly, its ranking was dropped to 1🔒; so, its findings have to be treated with caution. The researchers compared five value-added models, which were an OLS using only prior attainment (model 1), an OLS with prior attainment and student characteristics (model 2), an OLS with prior attainment and school fixed effects (model 3), an OLS with prior attainment, student characteristics and school fixed effects (model 4) and a multilevel mixed-effects model corresponding to model 3 (model 5). This study involved 4,234 students linked with 103 math teachers and 114 English language arts teachers from six high schools in the San Francisco Bay Area, US. Along with students’ math and language art test scores Grades from 9 to 11 in California Standards Tests (CSTs), the dataset also contained other student’ characteristics, including gender, race/ethnicity, on track status (for math), on fast-track status (for math), eligibility for free/reduced lunch status, English language learner status and parent education level. The study found that teachers effectiveness rankings from the four specifications of the OLS model were closely related to each other in both mathematics and English, with over 0.80 in each paired comparison. On the other hand, the multilevel model could not produce noticeable difference rankings comparing to its corresponding model; the correlations between models 4 and 5 were 0.95 and 0.94 in math and ELA, respectively. The effectiveness rankings across teaching courses for teachers who taught more than one teaching courses were also examined in order to reveal the extent to which a teacher's rank in one course is associated with the same teacher's rank in another course. As a result of intra-class correlations, only for mathematics in 2007, positive correlations were found across the models, and for math 2006, ELA 2006 and 2007, negative correlations were reported (except for model 2 in ELA 2007). The highest relationship with 0.72 was found for mathematics teachers in 2007 by model 5; interestingly, the lowest negative correlation with -0.52 was also reported in the same model for ELA teachers in 2006. Last, in order to investigate the consistency of teacher effectiveness ranking among models, the teachers' ranks were converted to deciles, and the percentages of teachers whose rankings varied by one or more, two or more, and three or more decimal in either direction across models, teaching courses, and years were also reported. 59-80% of teachers' rankings fluctuated across models by one or mode decile in either direction, 12-33% of the rankings changed by 2 or more deciles, and 0-14% of the rankings changed by 3 or more deciles. By taking into account all other analysis carried out, the researchers drew a conclusion that teacher effectiveness rankings varied considerably across models, teaching courses and years, but the varieties regarding courses and years were much greater.

In summary, the review suggests no one method of data analysis is better able to consistently predict teacher effectiveness. One study suggested there is an advantage of using OLS (Ordinal Least Square) method over SGPs (student growth percentages), while another showed that that OLS and SGPs might exaggerate teacher effectiveness estimates where a teacher serves a disproportionate number of high- or low-growth students. Different methods of analysis can produce vastly different estimates. The overall findings suggest that teacher performance estimates based on their students' achievement growth substantially varied depending upon the preferred model and would result in vastly different conclusions about the teachers. There is no strong evidence that any single data analysis method is superior to the other methods regarding its power to consistently estimate teachers' effectiveness in various conditions.

SECTION IV
RESULTS OF THE PRIMARY STUDY

This section describes the results of the primary research to assess the consistency of VAMs in estimating the effectiveness of 8th-grade teachers in the province of Samsun in Turkey, across five subjects (maths, Turkish, science, history, and English language). It consists of four chapters (Chapters 10, 11, 12 and 13). Chapter 10 describes the student attainment scores in Grade 8, which is the outcome variable used in the VAM analyses in this study, and the two previous years' grades (Grade 6 and 7), which are used as predictors. It also presents the results of pre-analyses checks on normality, linearity, multicollinearity, and homoscedasticity. The analyses were mainly centred on the consistency of teacher value-added effectiveness estimates regarding model specifications. A series of analyses were conducted to answer each sub-research question, and after explaining which analysis methods were used and data employed for each sub-research question, the results obtained using various model specifications are presented. Chapter 11 analyses the stability of VAMs, which include student, teacher/classroom, and school characteristics as predictors. Chapter 12 describes the results of the analysis that considers the stability of teacher effectiveness estimates over a two-year period of time and by using an additional prior years' test score. Chapter 13 examines the stability of VAMs that use different methods of modelling.

CHAPTER 10
STUDENT OUTCOME VARIABLES USED IN THE ANALYSES AND MODEL ASSUMPTIONS

This chapter starts by presenting descriptive results of students' achievement measures. Model assumptions were then checked to test whether the data meet some of the assumptions for multiple linear regression.

10.1 Descriptive Results of Student Achievement Outcomes

The aim of the primary study is to assess the stability of VAMs in estimating the effectiveness of teachers in Turkey, focusing on 8th-grade mathematics, Turkish, science, history, and English language teachers. Student test scores in Grade 8 in these five subjects are the outcome variables. For this study, prior test scores for each student in the previous two years (Grades 7 and 6) were used for the estimates, along with other contextual variables. A total sample of 1,027 teachers and their 35,435 students who were in Grade 8, were considered for the stability of teacher value-added effectiveness estimates.

The analyses in this study were mainly conducted using student test scores in Grade 8 as the outcome variable. Scores in other previous grade levels were incorporated as predictors in the value-added teacher effectiveness estimates. The overall means and standard deviations of students' available test scores for all teaching subjects are presented in Table 10.1, with the student test scores being depicted in Table 10.2 separately by grades for each teaching subject. It is worth noting here that all student test scores indicate the number of students' correct answers out of 20 in the test. In this study, the longitudinal student data set contains 21,959 students' test results in mathematics, spanning three school years (from 2015 to 2017) and three grades (Grade 6 through to 8), as well as 22,175 records in Turkish, 20,672 records in science, 19,277 records in history, and 19,040 records in English language through the same years and grades. Table 10.1 indicates that the average Turkish attainment of the three grades is the highest among the teaching subjects with 12.85 out of 20.0. The lowest student test scores were calculated in mathematics, with an average of 9.14. The standard deviation of the test scores indicates that students' test scores in English are the most spread out from their average test score (SD= 5.18).

Table 10.2 indicates that the means and standard deviations of the achievement measures are generally consistent across the grades for all teaching subjects. The mean scores demonstrate an overall upward tendency from Grade 6 to Grade 7 for all subjects except English, and from Grade 7 to Grade 8 for all subjects except mathematics and Turkish. The distribution of the standard deviations denoted slight variations across the grade levels.

Table 10.1 Overall Students' Test Scores (number of correct answers in tests out of 20)

Mathematics (N= 21,959)		Turkish (N= 22,175)		Science (N= 20,672)		History (N= 19,277)		English (N=19,040)	
Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
9.14	4.55	12.85	4.36	11.47	4.55	11.48	4.84	10.30	5.18

Table 10.2 Students' Test Scores by Grade

	Grade 6			Grade 7			Grade 8		
	N	Mean	SD	N	Mean	SD	N	Mean	SD
Mathematics	7,186	8.73	4.24	7,230	9.52	4.99	7,543	9.18	4.35
Turkish	7,228	11.45	4.07	7,353	14.17	4.35	7,594	12.90	4.23
Science	6,741	9.88	3.96	6,815	12.14	4.34	7,116	12.32	4.88
History	6,275	10.51	4.73	6,364	10.93	4.44	6,638	12.91	4.99
English	6,221	10.34	5.45	6,275	9.94	5.05	6,544	10.60	5.03

10.2 Model Assumptions

To examine the stability of VAMs in teacher effectiveness estimates, multiple regression analyses were employed. Before such statistical analyses were carried out, it is recommended to check if the data met the basic assumptions that are required in that statistical technique to avoid biased results (Chatterjee and Hadi, 2012; Field, 2013; Osbourne and Waters, 2002). Therefore, although several assumptions are listed in the literature, four common assumptions are tested in this study: normality, linearity, multicollinearity, and homoscedasticity. These all refer to the distribution of scores and the nature of the underlying relationship between the variables. These tests look for the residuals in the scatterplot. Residuals are the differences between the obtained and the predicted dependent variable scores (Pallant, 2001).

10.2.1 Test of Normality Assumption

The first fundamental assumption that needs to be tested is the normality assumption. For the assumption of normality, the residuals(errors) should be normally distributed (Field, 2013;

Williams et al., 2013) around the predicted dependent variable scores (which are the Grade 8 scores in this case). Figure 10.1 shows a normal distribution curve of the residuals for eighth-grade student maths scores. Normality assumptions tested in each teaching subject are given in Appendix F. For the other subjects, the residuals are also normally distributed (see Appendix F). Therefore, analyses of normality confirm that this assumption has been met.

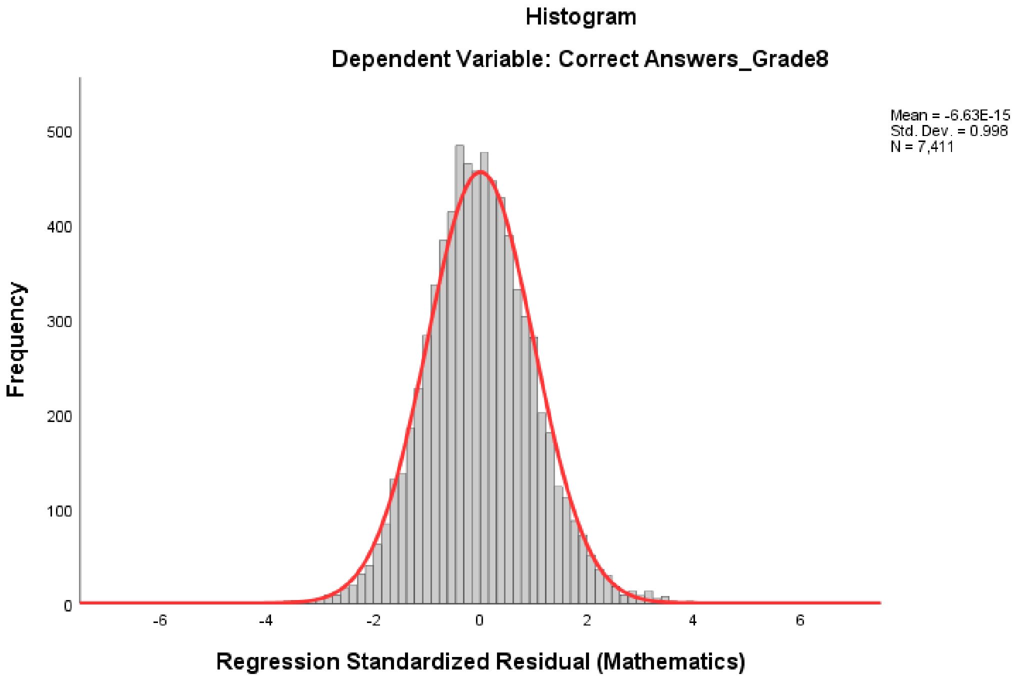


Figure 10.1 Histogram of normality distribution of standardized regression residuals in mathematics

10.2.2 Test of Linearity Assumption

For the test of linearity, the relationship between the residuals and the predicted dependent variable scores should be a straight line. The linearity assumption can be tested by the normal probability plot, also known as the P-P plot for the dependent variable, against the regression coefficients. Figure 10.2 shows the P-P plot for maths. For the other subjects, see Appendix F. The P-P plots for maths and the other subjects indicate that almost all points cluster around the straight line suggesting that assumption of linearity is met for all teaching subjects. The straight line indicates that there are no major deviations from normality.

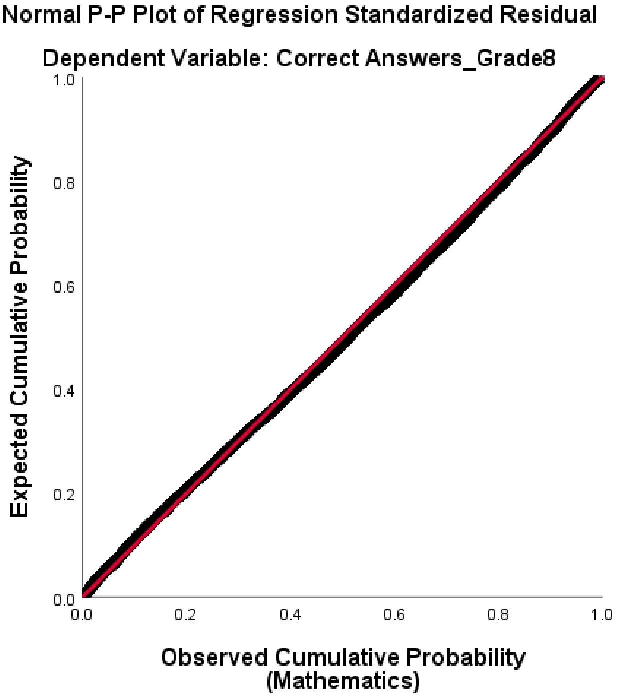


Figure 10.2 P-P plot for testing linearity assumption in mathematics

10.2.3 Test of Multicollinearity

The test of multicollinearity is to check whether the predictors used in the regression models correlate with other predictors. If there is a high correlation between some of the variables, then it presents a problem because it will be difficult to isolate the effects of the predictors since you would not be able to tell if it is predictor A or predictor B that is driving the effect. The higher the multicollinearity, the more difficult it is to interpret the coefficients because it increases the variance of the regression coefficients, making them unstable.

As part of the multiple regression programme, SPSS also performs ‘collinearity diagnostics. The value in the second last column of Table 10.3, labelled *tolerance,* indicates whether the multiple correlations with other variables is high, which suggests multicollinearity. If the value is close to zero, then this indicates that the multiple correlations with other variables are high, suggesting the possibility of multicollinearity. Table 10.3 depicts the results of the bivariate correlation analysis of the independent variables for mathematics. The results for the test of multicollinearity assumptions for the other subjects are also given in Appendix F.

The correlation matrix and the diagnostic statistics of the collinearity show that no predictors used in the equation violated the assumption of no multicollinearity. However, there is an exception in history, where a perfect correlation was found between a teacher’s graduation fields and appointment fields; therefore, the variable of graduation field was not included in the equation for history teachers.

Table 10.3 The Correlation Matrix with the Collinearity Diagnostics for Testing Multicollinearity Assumption in Mathematics

		Correlations																							Collinearity Statistics	
		Prior Attainment (G7)	Prior Attainment (G6)	Students Gender	Language Learner ID	School Category = Regional Boarding	School Category = Vocational	Service Score	Location =Rural	Location =Suburban	School Average_ Grade7	School Average_ Grade6	Teacher's Gender	Class Size	Percentage of female students classroom	Total teaching experience	Experience in current school	Appointment Fiels	Graduation Field	Having Master Degree ?	Master Field= Related	Master Field= Unrelated	Class Average_ Grade7	Class Average_ Grade6	Tolerance	VIF
Pearson Correlation	Prior Attainment (G7)	1.000																							0.447	2.238
	Prior Attainment (G6)	0.683	1.000																						0.444	2.254
	Students Gender	0.094	0.099	1.000																					0.858	1.165
	Language Learner ID	-0.012	-0.015	-0.018	1.000																				0.994	1.006
	School Category= Regional Boarding	-0.054	-0.047	0.002	-0.009	1.000																			0.878	1.139
	School Category= Vocational	-0.045	-0.063	-0.018	-0.001	-0.071	1.000																		0.890	1.124
	Service Score	-0.155	-0.162	0.021	-0.020	0.292	-0.075	1.000																	0.230	4.354
	Location=Rural	-0.180	-0.177	0.039	-0.015	0.164	-0.159	0.622	1.000																0.237	4.226
	Location=Suburban	-0.027	-0.052	-0.008	-0.017	0.097	0.123	0.444	-0.232	1.000															0.327	3.054
	School Average_Grade7	0.432	0.386	-0.014	0.023	-0.127	-0.100	-0.357	-0.415	-0.057	1.000														0.108	9.985
	School Average_Grade6	0.390	0.429	-0.018	0.020	-0.107	-0.146	-0.384	-0.408	-0.132	0.883	1.000													0.109	9.976
	Teacher's Gender	-0.033	-0.049	-0.012	0.016	-0.102	0.017	-0.018	-0.028	-0.030	-0.127	-0.148	1.000												0.887	1.128
	Class Size	0.231	0.225	0.005	0.018	-0.074	-0.010	-0.474	-0.375	-0.271	0.474	0.468	-0.103	1.000											0.584	1.713
	Percentage of female students	0.005	0.014	0.355	0.008	0.006	-0.050	0.058	0.110	-0.022	-0.041	-0.050	-0.033	0.015	1.000										0.833	1.201
	Total teaching experience	0.197	0.195	0.008	0.025	-0.140	0.015	-0.463	-0.438	-0.196	0.441	0.454	-0.073	0.398	0.022	1.000									0.522	1.914
	Experience in current school	0.093	0.091	0.005	0.025	0.006	-0.099	-0.254	-0.218	-0.085	0.251	0.261	-0.175	0.198	0.014	0.486	1.000								0.699	1.430
	Appointment Fiels	-0.044	-0.032	0.019	0.002	0.009	0.017	-0.062	0.021	-0.090	0.002	-0.008	0.048	0.058	0.054	-0.087	0.011	1.000							0.777	1.288
	Graduation Field	-0.008	-0.012	0.003	0.005	0.021	0.040	0.003	0.050	-0.033	-0.028	-0.033	0.036	-0.046	0.007	-0.105	0.024	0.413	1.000						0.807	1.239
	Having Master Degree ?	0.101	0.059	-0.009	0.041	-0.033	-0.065	-0.094	-0.081	-0.031	0.191	0.168	0.116	0.144	-0.025	0.017	-0.068	0.008	0.019	1.000					0.243	4.112
	Master Field= Related	0.121	0.086	-0.022	0.037	-0.025	-0.049	-0.092	-0.061	-0.065	0.195	0.186	0.063	0.230	-0.063	0.035	-0.101	0.006	0.014	0.756	1.000				0.285	3.507
	Master Field= Unrelated	0.030	0.013	0.008	0.036	-0.014	-0.027	-0.051	-0.034	-0.036	0.084	0.070	0.067	-0.034	0.023	0.040	0.034	0.003	0.008	0.415	-0.009	1.000			0.575	1.739
	Class Average_Grade7	0.534	0.465	0.003	0.007	-0.100	-0.078	-0.284	-0.333	-0.043	0.803	0.726	-0.057	0.431	0.009	0.362	0.173	-0.086	-0.016	0.188	0.225	0.053	1.000		0.114	8.756
	Class Average_Grade6	0.463	0.538	0.009	-0.001	-0.083	-0.112	-0.300	-0.324	-0.099	0.713	0.791	-0.087	0.428	0.026	0.362	0.178	-0.087	-0.024	0.117	0.168	0.021	0.865	1.000	0.120	8.325

10.2.4 The Assumption of Homoscedasticity

Homoscedasticity refers to the homogeneity of variance of the residuals in the regression. It simply means having same scatter. The homoscedasticity assumption can be tested by a scatter plot. For meeting the assumption of homoscedasticity, it is expected that the variance of the residuals about the predicted dependent variable scores should be the same for all predicted scores.

Figure 10.3 shows the scatter plot of the residuals obtained from the multiple regression analysis against the value of the predicted outcome for mathematics. Visual inspection of the scatter plot shows that the data values are all scattered or spread out to about the same extent; that is, they exhibit homoscedasticity. The scatter plots for all teaching subjects are also given in Appendix F. They also show that the data points are all scattered about the same extent, i.e., they look rather bunched up.

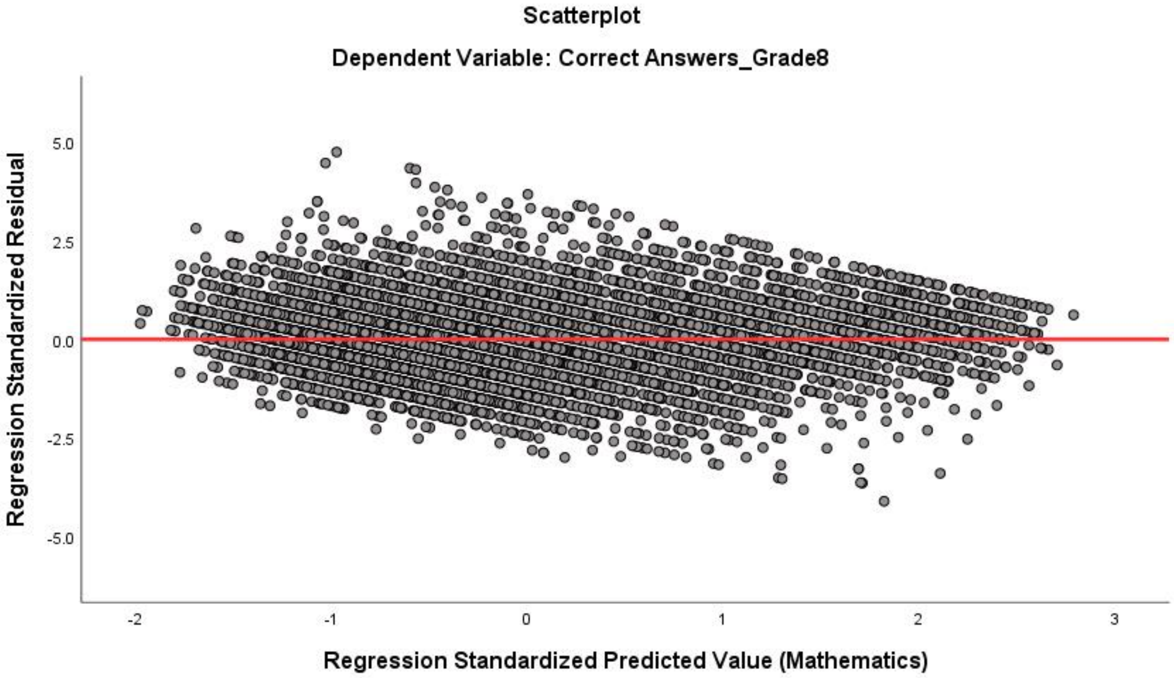


Figure 10.3 Scatterplot of the homogeneity of the variance in the standardized residuals in mathematics

CHAPTER 11

STABILITY OF VAMS IN ESTIMATING TEACHER EFFECTIVENESS USING STUDENT, TEACHER/CLASSROOM AND SCHOOL CHARACTERISTICS

This chapter examines the stability of VAMs in measuring teacher effectiveness that considers student, teacher/classroom, and school characteristics in a regression model.

11.1 Stability of value-added estimates using student characteristics

To estimate the stability of teacher estimates using student characteristics, students' test scores in a range of subjects (mathematics, Turkish, science, history, and English) in Grade 8 were used as the outcome variable, while their 7th-grade test scores (prior attainment, t-1) in the related subject, sex, and language learner identity of the students were employed as predictors.

The records belonging to a total of 35,435 students were examined to explain the value-added estimates for 230 mathematics, 232 Turkish, 204 science, 174 history and 187 English teachers, whose data could be linked to their students. Tables 11.1 and 11.2 summarize the data employed in this analysis.

Table 11.1 Students' Prior Attainment and Their Current Test Scores Used in Mathematics, Turkish, Science, History, and English Teachers' Value-added Estimates

	Mathematics (N= 7,543)		Turkish (N= 7,594)		Science (N= 7,116)		History (N= 6,638)		English (N= 6,544)	
	Mean	Std. deviation	Mean	Std. deviation	Mean	Std. deviation	Mean	Std. deviation	Mean	Std. deviation
Dependent Variable										
Outcome test score in Grade 8	9.18	4.35	12.90	4.23	12.32	4.88	12.91	4.99	10.60	5.04
Independent Variables										
Prior test score in Grade 7 (t-1)	9.52	4.89	14.16	4.28	12.14	4.24	10.93	4.35	9.96	4.95

Table 11.2 Other Contextual Student Characteristics Employed in Estimates in Subgroups

	Mathematics		Turkish		Science		History		English	
	N	Percentage	N	Percentage	N	Percentage	N	Percentage	N	Percentage
Contextual Variables at Student-level										
Sex (Female)	3,643	48.3	3,668	48.3	3,431	48.2	3,203	48.3	3,165	48.4
Language Learner	17	0.2	16	0.2	16	0.2	14	0.2	10	0.2

The data shows that students performed worst for maths both at Grade 8 and Grade 7. Turkish was their strongest subject with a mean of 14.16 (SD= 4.28) at Grade 7 and 12.90 at Grade 8. The sample in each teaching subject is reasonably balanced in terms of sex (roughly 48% are female students), but not for language status. Only 0.2% of the sample consisted of non-native Turkish students in each subgroup. This disproportionate number of students who are non-native Turkish speakers will need to be taken into account in the implication of the findings.

Before conducting the regression analyses, the relationship between the student characteristics and their current test scores was checked by Pearson's correlation for real-number variables and Cohen's effect size for categorical ones (see Tables 11.3 and 11.4). Pearson's r correlation, unsurprisingly, indicated that there is a strong positive relationship between students' current test scores and their prior attainment, with a roughly 0.7 correlation coefficient in each teaching subject. This means that students with higher prior attainment tend to have higher academic performance the following year.

Table 11.3 Correlation between Students' Prior Attainment and Their Current Test Scores

	Mathematics	Turkish	Science	History	English
	Correlation Coefficient *(r)*	*Correlation Coefficient* *(r)*	*Correlation Coefficient* *(r)*	*Correlation Coefficient* *(r)*	*Correlation Coefficient* *(r)*
Prior attainment (t-1)	0.69	0.68	0.69	0.63	0.72

Similarly, Cohen's effect sizes were calculated for categoric student characteristics (sex and language learner status). On average, female students were more successful, regardless of teaching subjects. The attainment differences between boys and girls were more pronounced for Turkish (d= 0.47) and English (d= 0.43), while the difference was less obvious for mathematics (d= 0.18). Students' language learner status is also strongly related to their performance. Native Turkish speakers tend to perform better than the small number of non-native Turkish speakers in all subjects.

Table 11.4 Comparison of Students' Current Attainment (Grade 8) by Sex and Language Learner Status

	Standard deviation	Sex			Language Learner Status		
		Female	Male	Cohen's *d*	Yes	No	Cohen's *d*
Maths	4.35	9.58	8.80	0.18	8.00	9.18	-0.27
Turkish	4.23	13.93	11.94	0.47	9.81	12.91	-0.73
Science	4.88	12.96	11.73	0.25	7.44	12.33	-1.00
History	4.99	13.70	12.18	0.30	7.86	12.92	-1.00
English	5.03	11.72	9.55	0.43	7.90	10.60	-0.54

Having established the relationship between student characteristics and their current test scores, a best-fit regression model (having the largest R-square that can be obtained by using as few student-level variables as possible) was created to find out to what extent teachers' value-added effectiveness estimates can be explained by student characteristics. R-squared represents the proportion of variance in the students' Grade 8 results that can be explained by the independent variables, which are students' prior test scores at Grade 7, their sex and language learner status.

The first stage of analyses was conducted using the data of 8th-grade students for a total of 1,027 teachers, with the results of the regressions displayed in Table 11.5. The summary table provides the R-squared (R^2) values, the total variation in the dependent variables explained by the contextual student-level independent variable(s), and the changes in R^2 values by comparing a new proposed model to the baseline model where the 8th-grade students' test scores in the related teaching subject were regressed on students' prior attainment scores (t-1) alone. The result is displayed as "-" where the R-squared value is not changed. The process of creating a new model stops when a model has already reached the largest R-squared value in the previous step for the relevant teaching course; for instance, Model 3 could not be created created for maths, Turkish, and English language.

The baseline model reveals that prior attainment explains a large proportion of the variance in students' Grade 8 scores. For maths, the R^2 is 0.47, indicating that 47% of the difference in their 8th-grade test score can be accounted for by their prior attainment. Prior attainment explains 52% (R^2= 0.52) of the variability in students' English test scores in Grade 8 and 39% (R^2= 0.39) of students' Grade 8 history results. Including other student characteristics (sex and language learner identity) in the model adds little to explaining any substantial proportion of the differences in teacher effectiveness estimates.

Table 11.5 R-squared Values of the Models Created Employing Student Characteristics

Model	Predictors used	Mathematics		Turkish		Science		History		English	
		R²	Changes in R²	R²	Changes in R²	R²	Changes in R²	R²	Changes in R²	R²	Changes in R²
Baseline Model	Prior attainment	.470		.467		.471		.394		.524	
Full Model	Prior attainment Sex Language Learner ID	.471	.001	.476	.009	.475	.004	.401	.007	.532	.008
Forward method											
Model 1	Prior attainment	.470	-	.467	-	.471	-	.394	-	.524	-
Model 2	Prior attainment Sex	.471	.001	.476	.009	.474	.003	.400	.006	.532	.008
Model 3	Prior attainment Sex Language Learner ID	-		-		.475	.004	.401	.007	-	

Dependent Variable: 8th-grade test score in the related teaching subject

One of the aims of this research is to find out whether this small improvement in the explanation of the variation on the outcome variable is due to the inclusion of both of these two variables or just one of them. Therefore, the final round of regression analysis was conducted by using the forward method of entry. By using the forward method, the regression will automatically exclude variables that make no contribution to the model or are too small to be considered. The forward method suggested a model using the prior attainment and sex variables for mathematics, Turkish, and English, which are the same largest R^2 values that can be achieved with the full model. Therefore, since the language learner identity variable did not contribute to the variance explained in the 8th-grade results in these teaching subjects, or the link was too small to be taken into account, this variable was excluded from the final model created using the student-level variables for these teaching subjects. Again, as stated before, it is worth considering that the disproportionate number of cases for the language learner identity variable (0.2% of students were Turkish language learners) in the data set might have caused this result.

In summary, student characteristics do not account for any noteworthy variation in value-added outcomes once individual prior attainment is accounted for, suggesting that students' prior attainment is the key factor that explains most of the differences in students' current test results. However, it is not possible to make a more general claim about this without having access to a dataset with more background variables.

11.2 Stability of value-added estimates of teacher effectiveness that include school characteristics

To examine the stability of VAMs in estimating teacher effectiveness in models that consider school characteristics in the analysis, students' 8th-grade test scores were again used as the outcome variable, and student characteristics (including their 7th-grade test scores in the five subjects, their sex, and language learner identity [the language learner identity variable is only for science and history]), as well as five school-level variables, were used as predictors (see Table 11.6 and 11.7). The school-level variables employed were school categories, service scores (based on school's infrastructure and facilities), locations, and the school-level average test scores in Grades 7 and 6.

The school-level dataset contained three school categories: general, regional boarding and vocational. The vast majority of the students (around 85%) attended general secondary schools, with only a small minority attended boarding schools (approximately 3% for each subject). In terms of student test score in Grade 8, children in the general schools had the highest academic performance, while those in the boarding schools had the lowest performance in all teaching subjects (see Table 11.8).

All public schools in Turkey are grouped into six service areas in terms of difficulties in assigning and employing teachers and the facilities they have. These service areas are given a score ranging from 1 (highest score) to 6 (lowest score) by the Ministry of Education. The average service scores of the schools are around 2.00, which is the second-highest score for all teaching subjects.

The data used in this analysis was obtained from students in a total of 695 secondary schools located in three locations: urban, suburban, and rural. Over half of the students in each subject attended urban schools, and only approximately 18% of students were from rural schools. In all subjects, children from urban schools performed better in test in Grade 8 on average than those from rural schools and suburban schools.

The last characteristic used is the schools' average attainments. Although the school-level average prior attainment (t-1) was slightly bigger than the school mean scores at the two-prior year (t-2) for mathematics, Turkish, science, and history, this situation was the opposite for the English language.

Table 11.6 School's Service Score and Their Mean Pre-test Scores for Mathematic, Turkish, Science, History, and English Teachers' Value-added Estimates

	Mathematics (N= 7,543)		Turkish (N= 7,594)		Science (N= 7,116)		History (N= 6,638)		English (N= 6,544)	
	Mean	Std. deviation	Mean	Std. deviation	Mean	Std. deviation	Mean	Std. deviation	Mean	Std. deviation
Service Score (1-highest, 6-lowest score)	2.01	1.43	2.01	1.43	1.99	1.42	2.06	1.43	2.08	1.45
Average test score in Grade 7	9.48	2.19	13.99	1.66	12.09	1.68	10.90	1.74	9.92	2.13
Average test score in Grade 6	8.67	1.82	11.39	1.61	9.81	1.62	10.46	1.86	10.23	2.33

Table 11.7 School Categories and Locations, and Their 8th Grade Average Test Scores for Each Subject

		Mathematics		Turkish		Science		History		English	
		N	Percentage	N	Percentage	N	Percentage	N	Percentage	N	Percentage
Categories	General	6,358	84.3	6,488	85.4	6,035	84.8	5,744	86.5	5,744	87.8
	Regional Boarding	264	3.5	183	2.4	183	2.6	183	2.8	167	2.6
	Vocational	921	12.2	923	12.2	898	12.6	711	10.7	633	9.7
Location	Rural	1,347	17.9	1,364	18.0	1,244	17.5	1,252	18.9	1,235	18.9
	Sub-urban	1,497	19.8	1,580	20.8	1,511	21.2	1,387	20.9	1,486	22.7
	Urban	4,699	62.3	4,650	61.2	4,361	61.3	3,999	60.2	3,823	58.4

Table 11.8 Comparisons of Averages for Student Attainment at Grade 8

	Mathematics		Turkish		Science		History		English	
	Mean	Std. deviation	Mean	Std. deviation	Mean	Std. deviation	Mean	Std. deviation	Mean	Std. deviation
School Categories										
General	9.33	4.43	12.94	4.27	12.41	4.90	13.01	4.98	10.71	5.06
Regional	7.77	3.72	11.80	4.49	10.61	4.52	11.44	4.89	8.62	4.49
Vocational	8.55	3.78	12.87	3.90	12.09	4.73	12.46	4.98	10.08	4.70
School Locations										
Rural	7.52	3.45	11.61	4.21	10.93	4.54	11.90	4.91	8.96	4.43
Sub-urban	8.90	4.04	12.77	4.02	12.73	4.64	13.27	4.68	10.60	4.81
Urban	9.74	4.54	13.33	4.23	12.58	4.99	13.10	5.07	11.12	5.18

Before conducting the regression analyses, to examine the relationship between school-level variables and the teachers VAM scores, Pearson's correlation coefficients were calculated for school service scores, school average test scores in Grades 7 and 6 (see in Table 11.9), and Cohen's effect sizes were calculated for each sub-category of the variables of school categories and locations (see in Table 11.10). To calculate the teachers VAMs scores, the individual student residual scores obtained from the final models in the previous section were aggregated at teacher level, and the class averages of these teacher-level residuals were tentatively attributed to teachers' individual value-added effectiveness scores.

Pearson's r coefficient indicated that there is no meaningful relationship between school service score and teacher VAM scores. Interestingly there is a very small but negative relationship between school service and teacher VAM scores for maths, Turkish, and English. It means that higher service score schools had teachers with slightly lower effectiveness scores. Not surprisingly, a medium positive relationship was found for average test scores at school-level for all subjects, except Turkish. However, the school average test score at a two-prior year (t-2, Grade 6) has a slightly better link with teachers' effectiveness scores. For instance, schools with higher average prior test scores tend to have more "effective" math teachers (r= 0.28 for one prior year, and r= 0.39 for two-prior year).

Similarly, Cohen's effect size indicated that, on average, general schools had slightly more "effective" maths and history teachers, while regional boarding schools tended to have less "effective" teachers, especially in English (d= -0.79). The differences in having an effective teacher between school categories were more pronounced in mathematics (d = +0.29 in general, d= -0.44 in regional boarding, and d= -0.22 in vocational schools), whereas in Turkish the differences were less obvious (d = +0.08 in general, d= +0.03 in regional boarding, and d= -0.10 in vocational schools). Similarly, there was a small positive correlation between the schools in urban areas and the effectiveness scores of maths, Turkish and English teachers, while the relationship was more pronounced in rural schools, regardless of teaching subject, but the sign of the relationship was negative. A medium negative effect size (d= -0.5) were calculated for maths and English; the strength of the relationship is slightly less in Turkish and Science and very little in history. The results show that urban schools tended to have more "effective" teachers except in science and history, while schools in rural areas employed less "effective" teachers in all teaching subjects.

Table 11.9 Correlation between School Characteristics and Teacher Value-added Effectiveness Scores

	Mathematics	Turkish	Science	History	English
	Correlation Coefficient (r)	*Correlation Coefficient (r)*	*Correlation Coefficient (r)*	*Correlation Coefficient (r)*	*Correlation Coefficient (r)*
Service score	-0.16	-0.10	0.02	0.03	-0.16
Average test score in Grade 7	0.28	0.01	0.29	0.19	0.21
Average test score in Grade 6	0.39	0.27	0.39	0.30	0.38

Table 11.10 Comparison of Value-added Means for School Characteristics

		Mathematics			Turkish			Science			History			English		
		Yes	No	*d*	Yes	No	*d*	Yes	No	*d*	Yes	No	*d*	Yes	No	*d*
Categories	General	0.06	-0.33	0.29	0.01	-0.09	0.08	0.01	-0.05	0.04	0.06	-0.38	0.24	0.01	-0.06	0.05
	Regional Boarding	-0.57	0.02	-0.44	0.03	-0.00	0.03	-0.39	0.01	-0.28	-0.31	0.01	-0.18	-1.09	0.03	-0.79
	Vocational	-0.26	0.04	-0.22	-0.11	0.01	-0.10	0.02	-0.00	0.02	-0.39	0.05	-0.24	0.21	-0.02	0.17
Locations	Rural	-0.52	0.11	-0.47	-0.26	0.06	-0.26	-0.30	0.06	-0.26	-0.17	0.04	-0.12	-0.63	0.15	-0.55
	Suburban	-0.11	0.03	-0.10	-.012	0.03	-0.12	0.34	-0.09	0.30	0.23	-0.06	0.16	0.11	-0.03	0.10
	Urban	0.18	-0.30	0.36	0.12	-0.19	0.25	-0.03	0.05	-0.06	-0.03	0.04	-0.04	0.16	-0.22	0.27
Std. deviation		1.33			1.23			1.44			1.83			1.41		

To determine to what extent teachers' value-added effectiveness estimates can be explained by school characteristics over and above the student characteristics revealed in the previous section, a best-fit regression model with the largest R-squared value was created by using as few school-level variables as possible. The final model generated using student characteristics was used as a baseline model where the 8th-grade students' test scores were regressed on students' prior attainment scores (t-1), sex, and language learner identity (the language learner identity predictor is only for science and history). The baseline models revealed that the minimum variability in the outcome test score that can be explained using the previous test score, sex, and language learner ID was estimated at 40% (R^2= 0.40) for history, while the maximum variability that can be explained using the previous test score and student sex variables was determined in English with 53% (R^2= 0.53) (see Table 11.11).

To determine the highest R-squared value that can be achieved at school-level, the following school characteristics were included in the baseline model by using the *enter* method: school categories, service scores, locations, and school-level average test scores in Grades 7 and 6. The inclusion of all five school characteristics at one time in the baseline models contributed just under 2 percentage points to the R-squared of each teaching subject. It was determined that mathematics is the course in which the most changes in R-squared with 1.9%. In order to reduce complexity and include only variables that have a predictive power on estimates, it needs to be established whether the improvements in the explained variance in the outcome variables are due to the inclusion of all five school characteristics or only some of them. As before, the same school-level control variables were included in the baseline models using the *forward* method.

The *forward* method suggested a final model with the largest R^2 value using the least variable among the proposed models for each teaching subject. For instance, the *forward* method proposed a final model for Turkish subject using exactly the same variables employed in mathematics: prior attainment, school-level average test scores in Grades 6 and 7, and student sex variables. Since school categories, service scores, and school locations variables did not contribute to the variance explained in the 8th-grade results in maths and Turkish, or the link was too small to be taken into account, these variable were excluded from the final model created for maths and Turkish teachers' value-added effectiveness estimates. These exclusions can also be interpreted as giving no indication that the student's current attainment in maths and Turkish is linked to the school service score, the school location, and the type of school attended once the prior attainment, school-level average test scores in Grades 6 and 7 and student sex have been taken into account.

The *forward* method also proposed a final model for English by including school categories, in addition to the variables identified for the mathematics and Turkish teaching subjects, whereby 55% of the variation in students' 8th-grade English test scores can be explained. Unlike the final model proposed for the mathematics lesson, it was suggested that language learner identity was included in the final models for history and science, while the 7th-grade average school test score was also excluded from the model for history. In addition, the *forward* method also suggested including school location information in the final model for the science subject. By employing the identified predictors into the regression model, 49.1% of the variation in students' current science attainment and 41.6% of students' current test scores for history can be explained.

Overall, the inclusion of school-level variables again makes very little difference to the amount of variation explained in the outcomes once prior attainment is taken into account. And unlike individual student background, this dataset has a reasonable set of school-level indicators.

Table 11.11 R-squared Values of the Models Created Employing School Characteristics

Model	Predictors used	Mathematics		Turkish		Science		History		English	
		R^2	Changes in R^2	R^2	Changes in R^2	R^2	Changes in R^2	R^2	Changes in R^2	R^2	Changes in R^2
Baseline Model	Prior attainment Sex Language Learner ID**	.471		.476		.475		.401		.532	
Full Model	Prior attainment Sex Language Learner ID** School categories Service score School locations 7th grade average school test score 6th grade average school test score	.490	.019	.490	.014	.491	.016	.416	.015	.550	.018
Forward method Final Model	Prior attainment 6th grade average school test score 7th grade average school test score* Student sex Language Learner ID** School locations*** School categories**** Service Score**	.490	.019	.490	.014	.491	.016	.416	.015	.550	.018

Dependent Variable: 8th-grade test score in the related teaching subject

*	Excluded in estimates for science and history
**	Included in estimates for science and history
***	Included in estimates for science
****	Included in estimates for English

11.3 Stability of VAMs in teacher effectiveness estimates that include teacher/classroom characteristics

This section investigates the stability of VAMs in estimating teacher effectiveness that considers teacher/classroom variables over and above the student and school-level variables identified in the previous sections. In other words, in this section, it was tried to find an answer to the question of how much difference do teacher factors and classroom environment make in explaining student outcomes. Seven observable teacher characteristics (sex, number of years of teaching experience, number of years teaching in the current school, teachers' major degree subject, teaching assignment field, and their highest level of qualification and field) and four classroom-level variables (class size, percentage of female students, 7th-grade classroom-level average maths test scores and 6th-grade classroom level average maths test scores) were employed as the teacher/classroom-level predictors in this section. Tables 11.12 and 11.13 summarize the data employed for this analysis.

Table 11.12 Teacher/Classroom Characteristics Used in Mathematics, Turkish, Science, History, and English Teachers' Value-added Estimates

Variables	Mathematics (N= 7,543)		Turkish (N= 7,594)		Science (N= 7,116)		History (N= 6,638)		English (N= 6,544)	
	Mean	Std. deviation	Mean	Std. deviation	Mean	Std. deviation	Mean	Std. deviation	Mean	Std. deviation
Class size	22.93	5.70	22.31	5.62	22.36	5.68	21.99	5.58	21.66	5.30
Percentage of female students	0.48	0.18	0.48	0.18	0.48	0.18	0.48	0.17	0.48	0.16
Total teaching experience	10.13	5.36	11.46	5.79	12.29	7.23	14.88	7.35	9.26	5.46
Experience in the current school	3.33	2.23	3.73	2.38	3.47	1.84	3.85	2.48	3.487	2.70
Classroom average test score in Grade 7	9.48	2.71	13.96	2.02	12.10	2.15	10.90	2.17	9.88	2.72
Classroom average test score in Grade 6	8.66	2.29	11.36	1.93	9.83	2.05	10.43	2.36	10.26	2.90

Table 11.13 Other Contextual Teacher/Classroom Characteristics Employed in Estimates in Subgroups

Variables		Mathematics		Turkish		Science		History		English	
		N	Percentage	N	Percentage	N	Percentage	N	Percentage	N	Percentage
Sex (Female)		4,036	53.5	4,413	58.1	3,742	52.6	2,920	44	4,535	69.3
Graduation field (related to teaching subject)		7,456	98.8	7,563	99.6	70,93	99.7	5,701	85.9	6,338	96.9
Appointment field (related to teaching subject)		7,528	99.8	7,594	100	70,93	99.7	6,570	99	6,544	100
Terminal degree (master/higher degree)		221	2.9	350	4.6	378	5.3	202	3	43	0.7
Field of terminal degree	Related to teaching subject	128	57.9	48	13.7	204	54	44	21.8	0	-
	Not related to teaching subject	39	17.6	107	30.6	19	5	29	14.4	18	41.9
	Unspecified	54	24.4	195	55.7	155	41	129	63.9	25	58.1

More than half of the students were taught by female teachers, except for history. The highest proportion of students taught by female teachers was found in English lesson (69.3%). In each teaching subject, while there was an average of 22 students in each class, 48 per cent of these classes were female students. For all subjects (except for English), students were taught by teachers with over 10 years of teaching experience). On average, teachers had been in their current school for more than 3 years ago. History teachers were the most experienced, with an average of 14.88 years in total and 3.85 years in their current schools.

The bachelor graduation fields of teachers were grouped based on the relation to their teaching field and revealed that for all subjects, almost all students taught by teachers whose teaching subjects are related to their bachelor’s degree. Among other subjects, history is a subject with the highest proportion of teachers who had bachelor's degrees that not related to history. A similar grouping strategy was applied to the variable of teaching appointment subject, and more clustered cases were found. The variable of teachers’ terminal education level indicates whether the teachers have a master’s or higher degree. The highest proportion of students were assigned to science teachers with a master's/higher degree with 5.3%, while the lowest ratio of students had English teachers having a master's/higher degree (0.7%). Lastly, similar to the school-level average score, although the classroom-level average prior attainments (t-1) were higher than the class mean scores at two-year prior (t-2) for mathematics, Turkish, science, and history (slightly higher), this situation is again the opposite for the English teaching subject.

To examine the relationship between teachers’ value-added scores and teacher/classroom characteristics, the individual student residual scores obtained from the final models created using the school and student characteristics were aggregated at teacher level, and the class averages of these teacher-level residuals were tentatively attributed to teachers’ individual value-added effectiveness scores. Then, Pearson’s correlation coefficients were calculated for the continuous teacher/classroom variables, which are class size, percentage of female students, total teaching experience, experience in the current school, and classroom average test scores in Grades 6 and 7.

The results indicate that there is no meaningful relationship between teacher effectiveness scores and the continuous variables in Table 11.14. A very little relationship was found between class size and teachers’ value-added scores, and interestingly, larger classes had teachers with slightly higher effectiveness scores, except in history (r= -0.04). Pupils in crowded classrooms tended to have more effective teachers in value-added modelling terms, even if the difference is very little. Another interesting finding is that classes with a higher female student ratio are taught by less “effective” teachers (except for history). Experience, regardless of whether in total or in their current schools, is negatively associated with their effectiveness scores. In other words, more experienced teachers tended to less “effectiveness” (except for Turkish). Classroom prior attainments are positively rated to teacher effectiveness scores. Interestingly, average classroom attainment at Grade 6 is more closely related to teacher effectiveness estimates than the average score at Grade 7.

Table 11.14 Correlation between Teacher/Classroom Characteristics and Teacher Value-added Effectiveness Scores

	Mathematics	Turkish	Science	History	English
	Correlation Coefficient *(r)*	*Correlation Coefficient* *(r)*	*Correlation Coefficient* *(r)*	*Correlation Coefficient* *(r)*	*Correlation Coefficient* *(r)*
Class size	0.08	0.09	0.08	-0.04	0.08
Percentage of female students	-0.10	-0.10	-0.04	0.01	-0.02
Total teaching experience	-0.03	0.01	-0.12	-0.07	-0.12
Experience in the current school	-0.02	0.01	-0.04	-0.07	-0.14
Classroom average test score in Grade 7	0.06	0.06	0.05	0.01	0.02
Classroom average test score in Grade 6	0.11	0.13	0.06	0.07	0.05

Cohen’s effect sizes were calculated for each sub-category of teacher characteristics. The results of Cohen’s *d* statistics are shown in Table 11.15. On average, female maths and Turkish teachers had slightly worse value-added effectiveness scores than male teachers (*d*= -0.10 and -0.12, respectively). The second set of personal characteristics considered in the study is the graduation field. It was found that maths and history teachers who graduated from a field related to their current teaching subjects tend to have lower effectiveness scores, although the effect size is very small. More interestingly, mathematics and Turkish teachers who were initially appointed as teachers in a field other than their current teaching subjects but later moved to their current teaching area, have remarkably higher value-added effectiveness scores than those originally appointed as mathematics or Turkish teachers (*d*= -0.92, and -0.99, respectively). This result may be due to the disproportionate number in each of the sub-categories; therefore, this result needs to be tested with data containing a balanced sub-categorical distribution. Another interesting finding is that, contrary to what is believed, having a master's degree does not contribute to the effectiveness estimates for mathematics, science and history teachers. Teachers with master’s degrees have, on average, lower effective scores than teachers with just a bachelor's degree. Finally, having a master’s degree in a field related to teaching subject had almost no link to mathematics teachers’ effectiveness estimates (*d*= 0.01), while science and history teachers with a master’s degree had worse effectiveness scores than those who did not (*d*= -0.81 and -0.73).

Table 11.15 Comparison of Value-added Means for Teacher/Classroom Characteristics

	Mathematics			Turkish			Science			History			English		
	Yes	No	*d*	Yes	No	*d*	Yes	No	*d*	Yes	No	*d*	Yes	No	*d*
Sex (Female)	0.05	-0.05	-0.10	-0.05	0.06	-0.12	0.00	-0.00	0.00	0.19	-0.15	0.23	0.08	-0.17	0.26
Graduation field (related to teaching subject)	-0.00	0.08	-0.08		-		0.00	-0.39	0.37	-0.00	0.24	-0.17		-	
Appointment field (related to teaching subject)	-0.00	0.92	-0.92	-0.01	0.94	-0.99	0.00	-0.39	0.37	-0.02	0.13	-0.10	-0.01	0.29	-0.31
Terminal degree (master/higher degree)	-0.25	0.01	-0.25	0.24	-0.01	0.27	-0.39	0.02	-0.39	-0.76	0.02	-0.53	0.61	-0.00	0.63
Field of master's degree															
Related	0.01	0.00	0.01	0.73	-0.01	0.77	-0.84	0.02	-0.81	-1.07	0.01	-0.73		-	
Not Related	-0.97	0.01	-0.96	0.89	-0.01	0.94	0.60	-0.00	0.57	-0.19	0.00	-0.13	-0.78	0.00	-0.80
Unknown	-0.34	0.00	-0.34	-0.23	0.01	-0.25	0.07	-0.00	0.06	-0.78	0.02	-0.54	1.61	-0.01	1.67
Std. deviation		1.01			0.96			1.06			1.48			0.97	

In addition to the correlation analysis, a best-fit regression model was created to see how much teacher/classroom variables contribute to explaining the variation in student attainment for each subject. The results of the regressions are displayed in Table 11.16. The baseline models showed that 41.6 to 55% of the variability in 8th-grade outcome test scores for all subjects could be explained by the identified student and school characteristics in the previous section. Adding all teacher/classroom-level characteristics (sex, class size, percentage of female students, teachers' major degree subject, teaching assignment field, number of years of teaching experience, number of years teaching in the current school, their highest level of qualification and field, and classroom-level average test scores in Grades 7 and 6) to the baseline model at one time increased the prediction by between 0.6 to 1.4 percentage points. This means that all teacher/classroom characteristics contributed an additional 1.4 percentage points to the variance explained for history and only 0.6 percentage points for English.

Another regression analysis was carried out that include only variables that have predictive power on estimates. These variables are displayed in Table 11.17. The simplest final model with six predictors (students' Grade 7 attainment score (t-1), classroom-level average test scores in Grades 6 and 7, sex of student, class size, and percentage of female students) was proposed for maths, and eight predictors were used for the other subjects. For different subjects, different predictors were used because only factors with the strongest predictive powers were used for each subject. Predictors found to be ineffective in previous models were removed. The common predictors employed in all the models are prior attainment, 6th-grade class average test score (t-2), and student sex. Other common predictors used in the final models of at least three teaching subjects were the percentage of female students in the classroom, 7th-grade class average test scores (t-1) and class size.

The analysis also showed that none of the teacher characteristics was found to be considerably related to students' maths attainments, while all classroom level characteristics were included in the eventual model in math. This result shows us that student maths performance is affected more by school and class characteristics than by a teacher's observable characteristics. In other teaching subjects, it was found that some of the teacher characteristics, such as master field, experience had relation to students' achievements (see in Table 11.17).

Table 11.16 R-squared Values of the Models Created Employing Teacher/Classroom Characteristics

Model	Predictors used	Mathematics		Turkish		Science		History		English	
		R²	Changes in R²	R²	Changes in R²	R²	Changes in R²	R²	Changes in R²	R²	Changes in R²
Baseline Model	Prior attainment 6th grade average school test score 7th grade average school test score*** Student sex Language Learner ID* School locations** School categories**** Service Score*	.490		.490		.491		.416		.550	
Full Model	Prior attainment 6th grade average school test score 7th grade average school test score*** Student sex Language Learner ID* School locations** School categories**** Service Score* Teacher sex Class size Percentage of female students Graduation field Appointment field Total teaching experience Experience in the current school Terminal degree Field of terminal degree 7th grade average classroom test score 6th grade average classroom test score	.502	.012	.501	.011	.503	.012	.430	.014	.556	.006

Dependent Variable: 8th grade test score in the related teaching subject

* Included in estimates for science and history
** Included in estimates for science
*** Excluded from estimates in science and history
**** Included in estimates for English

Table 11.17 The Predictors Used in the Final Models for Each Teaching Subject

Mathematics		Turkish		Science		History		English	
Predictors used	R^2	Predictors used	R^2	Predictors used	R^2	Predictors used	R^2	Predictors used	R^2
Prior attainment Ave. class score 6 Ave. class score 7 Student sex Class size % female students	.502	Prior attainment Ave. class score 6 Student sex Class size % female students Ave. school score 7 Ave. school score 6 Master field	.501	Prior attainment Ave. class score 6 Ave. class score 7 Student sex % female students Master field Total teaching exp. Language learner	.503	Prior attainment Ave. class score 6 Student sex Language learner Schl service score Terminal degree Graduation field Exp. current schl	.430	Prior attainment Ave. class score 6 Ave. class score 7 Student sex Class size % female students Exp. current schl Schl categories	.556

Dependent Variable: 8th grade test score in the related teaching subject

The full list of standardised coefficients for the predictors employed in the final model is also shown in Table 11.18. The overall conclusion is that when the other factors are held constant, students’ prior attainment at Grade 7 has the strongest positive relationship with their recent outcomes in each teaching subject, for every one correct answer increase in prior test score, the number of the correct answer in Grade 8 test increases between 0.559 and 0.647. The second-largest relationship was found between the 8th-grade test score and 6th-grade average class test score in the related teaching subject. For each one-point increase in the 6th-grade average classroom test score, the recent attainment would increase, on average, between 0.153 and 0.269. However, for English language and maths, the average classroom Grade 7’s test scores are negatively related to their 8th-grade test scores. This negative relationship means that for each one-point increase in the 7th-grade maths and English average classroom test score, the recent maths and English attainment would decrease on average by 0.12 and 0.09, respectively.

Female pupils appear to outperform boys in recent test regardless of teaching subjects. However, classes with more female students perform less well than classes with more male students for maths, Turkish and English. Another surprising conclusion is that on average, for maths, Turkish, and English, students in large classes tend to do better.

Table 11.18 Standardized Regression Coefficients for the Variables Used in the Final Model

Variables	Mathematics	Turkish	Science	History	English
	Standardized Coefficient β	*Standardized Coefficient β*	*Standardized Coefficient β*	*Standardized Coefficient β*	*Standardized Coefficient β*
Prior attainment (t-1)	0.615	0.630	0.613	0.559	0.647
6th grade average classroom test score	0.269	0.153	0.153	0.180	0.233
7th grade average classroom test score	-0.118	-	0.040	-	-0.094
Student sex (female)	0.039	0.111	0.067	0.080	0.105
Class size	0.026	0.029	-	-	0.021
Percentage of female students	-0.024	-0.032	-0.026	-	-0.024
Language learner	-	-	-0.035	-0.036	-
6th grade average school test score	-	0.054	-	-	-
7th grade average school test score	-	-0.129	-	-	-
Terminal degree (having master/higher degree?)	-	-	-	-0.020	-
Field of terminal degree (Unspecified teaching subject)					
Related to teaching subject	-	0.016	-0.037	-	-
Not related to teaching subject	-	0.027	0.008	-	-
Total teaching experience	-	-	-0.054	-	-
School Service Score	-	-	-	0.045	-
Graduation field (related to teaching subject)	-	-	-	-0.027	-
Experience in the current school	-	-	-	-0.025	-0.032
School categories (General)					
Regional Boarding	-	-	-	-	-0.027
Vocational	-	-	-	-	0.019
R^2	0.502	0.501	0.503	0.430	.556

In summary, there is no notable relationship between teacher/classroom characteristics and teachers’ VAM scores, although a weak relationship was found between teachers’ effectiveness scores and class sizes. The correlation analysis showed that teachers in larger classes tend to have higher VAM scores (except for history). Having a master (or higher) degree in a relevant field to their teaching subject had a negative relationship to science teachers' effectiveness scores, but a positive relationship to Turkish teachers' scores. On the other hand, no relationship was found with the scores of mathematics teachers. The findings also revealed that the eventual models created for each teaching subject were able to explain roughly half of the variation in students’ current attainment, and the models varied in terms of predictors included. Students’ Grade 7 prior attainment (Grade 7), the average class Grade 6 attainment, and student sex were the common contextual predictors employed in each model. The study also showed that students’ current performance were affected more by school and classroom characteristics rather than teacher characteristics, suggesting that teachers make little difference to students’ current performance. Last but not least, the largest positive relationship between student’s prior and current attainments when holding the other characteristics constant was revealed by standardized coefficient analysis.

CHAPTER 12

STABILITY OF VAMS IN TEACHER EFFECTIVENESS ESTIMATES THAT CONSIDER THE NUMBER OF PREVIOUS YEARS' TEST SCORES AND OVER A TWO-YEAR PERIOD

This chapter presents the results exploring the stability of teacher value-added effectiveness estimates over a two-year period and in terms of the number of previous years' test scores used. To determine the consistency in effectiveness estimates of the teachers over two years, the same teacher's current and previous year's effectiveness estimates were compared in the first section, then the consistency of estimates is examined by adding additional prior years' test scores. The results using Grade 7 test scores (one lagged score) and other predictors were compared with those using Grade 6 and Grade 7 (two lagged scores) and the same other predictors.

12.1 Stability of teacher value-added effectiveness estimates over a two-year period of time

This section is focused on the consistency of teacher value-added effectiveness scores over two years across five teaching subjects (maths, Turkish, science, history, and English). In order to compare teacher effectiveness scores in the current year with the effectiveness scores in the previous year, in the dataset used, it was necessary to ensure that the same teachers taught both years and that the school had only one teacher teaching the subject for both years – so that the effectiveness estimates can be attributed to that teacher. Because of this requirement, only 151 teachers whose data could be linked to 2,526 students were used in the estimates. These included 21 mathematics, 32 Turkish, 39 science, 32 history, and 27 English language teachers.

Tables 12.1 and 12.2 show the variables that are used in the analyses to compare teacher effectiveness estimates to see how consistent they are over a two-year period. Students' test scores at Grade 8 and Grade 7, shown in Table 12.1, are the dependent variables used for comparing teacher effectiveness over two years. The independent variables or predictors are students' prior test scores at Grade 6 and Grade 7, the sex of students, the average school attainment at Grades 6 and 7, the school service scores and teacher/classroom variables (these include class size, teachers' experience and qualifications, and the average class attainment at Grades 6 and 7).

Table 12.1 Data Employed for Testing Stability of Teacher Effectiveness Estimates Over Two Years

	Mathematics		Turkish		Science		History		English	
	(N= 332)		(N= 556)		(N= 644)		(N= 542)		(N= 452)	
	Mean	Std. deviation	Mean	Std. deviation	Mean	Std. deviation	Mean	Std. deviation	Mean	Std. deviation
Dependent Variables										
Outcome test score (Grade 8)	7.82	3.38	12.08	4.12	11.86	4.66	12.48	4.66	9.06	4.36
Outcome test score (Grade 7)	8.52	4.52	13.53	4.32	11.41	4.05	10.65	4.16	8.50	4.31
Student-level Independent Variables										
Prior test score (Grade 7)	8.52	4.52	13.53	4.32	11.41	4.05	10.65	4.16	8.50	4.31
Prior test score (Grade 6)	7.29	3.49	10.45	3.62	8.97	3.53	9.81	4.24	8.87	4.69
School-level Independent Variables										
Service Score	-		-		-		2.93	1.35	-	
Average school test score (Grade 7)	-		13.44	1.99	-		-		-	
Average school test score (Grade 6)	-		10.45	1.39	-		-		-	
Teacher/Classroom-level Independent Variables										
Class size	17.35	4.63	18.9	5.28	-		-		18.48	5.53
Percentage of female students	0.50	0.01	0.49	0.10	0.50	0.13	-		0.49	0.10
Total teaching experience	-		-		10.11	4.68	-		-	
Expr. in the current school	-		-		-		5.44	4.08	4.59	2.40
Average class test score (Grade 7)	8.48	1.96	-		11.37	2.17	-		8.47	1.68
Average class test score (Grade 6)	7.24	1.43	10.45	1.39	8.93	1.66	9.88	2.03	8.83	1.84

Table 12.2 Other contextual Characteristics Employed in Stability Estimates in Subgroups

		Mathematics		Turkish		Science		History		English	
		N	Percentage	N	Percentage	N	Percentage	N	Percentage	N	Percentage
Student-level Independent Variables											
	Sex (Female)	165	49.7	270	48.6	321	49.8	271	50	220	48.7
Teacher/Classroom-level Independent Variables											
	Graduation field (related to teaching subject)	-		-		-		493	91.0	-	
Field of terminal degree	Related	-		17	58.6	17	24.6	-		-	
	Not related			12	41.4	19	27.5				
	Unspecified	-		-	-	33	47.8				

Since the teacher VAM scores in two consecutive years are compared, students' individual test scores at Grade 7 and 8 were used as two separate outcome variables for each year estimate. For teachers' current value-added effectiveness scores, students' 8th-grade test scores were used as the outcome variable, while 7th-grade test scores of the same students and other background characteristics shown in Tables 12.1 and 12.2 were included in the model as control variables. Different predictors or independent variables were used for different subjects (see Table 11.17). Only factors that have a relationship with the outcomes scores were included. Those that show no association with the outcomes for each subject were excluded. For example, for maths, students' prior attainment at Grade 6 (or at Grade 7 in previous teacher effectiveness estimates), their sex, class size, percentage of girls in the class, and the average class attainments at Grades 6 and 7 were the predictors used in the final models because these were the factors found to have strong relationships with student outcomes (see Chapter 11).

Correlation analyses were used to establish whether there is any relationship between the individual teacher's latest value-added effectiveness scores for each teaching subject and their previous effectiveness scores. Pearson's r coefficients indicated that there is no meaningful relationship between teachers' current and previous effectiveness scores for all teaching courses (see Table 12.3). In addition, the raw effectiveness scores were grouped into four effectiveness categories by dividing into quartiles: highly effective, effective, partially effective and ineffective. Each teacher was assigned to one of the four possible effectiveness categories based on their current and previous effectiveness scores. Spearman's rho correlations between the effectiveness categories for each teaching subject are also shown in the last column of Table 12.3.

Table 12.3 Correlation Between Teachers Current and Previous Value-added Effectiveness Scores and Categories

	Correlation Coefficient (r) (Score)	*Correlation Coefficient (r_s)* (Category)
Mathematics	-0.028	-0.072
Turkish	-0.133	-0.200
Science	-0.003	0.019
History	-0.133	-0.055
English	-0.018	0.122

Table 12.3 shows that maths teachers' current effectiveness scores are negatively related to their previous effectiveness scores (r= -0.03). A slightly higher negative correlation result was found between effectiveness categories for mathematics teachers (r_s= -0.07). This means that those who scored highly on current effectiveness scores scored low on previous effectiveness scores vice versa. Similarly, those who were categorised as currently highly effective were classified as least effective in their previous ranking. This is similar for all subjects with the exception of science and English language teachers, where teachers who were classified as effective in previous years were also classified as effective in the current year.

In order to closely examine teachers' year to year consistency in teacher effectiveness categories, a transition matrix was created for mathematics teachers in Table 12.4. Year to year consistencies in quartiles for other teaching subjects are also shown in Appendix G.

Table 12.4 A Transition Matrix for Year-to-Year Consistency in Effectiveness Categories for Mathematics Teachers (in Percentages)

		Previous				**Total**
		Highly Effective	Effective	Partially Effective	Ineffective	
Current	Highly Effective	4.8	4.8	9.5	4.8	**5**
	Effective	4.8	4.8	9.5	4.8	**5**
	Partially Effective	4.8	9.5	4.8	4.8	**5**
	Ineffective	14.3	4.8	0.0	9.5	**6**
Total		**6**	**5**	**5**	**5**	**21**

The consistency of mathematics teachers' effectiveness categories over two years involved determining the percentage of teachers that remained and changed their effectiveness categories from one year to the next. Table 12.4 shows that 23.9% of maths teachers were consistently categorised in the same effectiveness degree between the two-time points. Three teachers who were placed in the highly effective category in the previous year were placed in the ineffective category in the current year, while only one teacher was categorised as highly effective in both years.

Table 12.4 also shows that the model used for mathematics teachers also produced inconsistent results when looking at the percentage of teachers who changed at least two categories over two years. The effectiveness categories of 42.9% of mathematics teachers (n= 9) were changed at least two categories up or down from one year to the next. This is also similar for other subjects except for English teachers (Appendix G). For example, the highest 46.9% of Turkish teachers were assigned to at least two upper or lower categories the following year. These results suggest that there is a large inconsistency of effectiveness categories that teachers were placed in, as well as generating very inconsistent effectiveness scores over two years, regardless of teaching subjects.

Looking at the ranking changes on an individual basis, it can be seen more clearly how the rankings of the same teachers have changed year to year. Figure 12.1 illustrates the changes of effectiveness rank ordering from year to year for individual mathematics teachers.

Previous Ranking	Rank-ordering	Current Ranking
Teacher 216	1	Teacher 54
Teacher 87	2	Teacher 102
Teacher 170	3	Teacher 214
Teacher 194	4	Teacher 84
Teacher 191	5	Teacher 21
Teacher 54	6	Teacher 56
Teacher 153	7	Teacher 194
Teacher 26	8	Teacher 215
Teacher 214	9	Teacher 26
Teacher 194	10	Teacher 208
Teacher 76	11	Teacher 216
Teacher 215	12	Teacher 40
Teacher 56	13	Teacher 76
Teacher 102	14	Teacher 103
Teacher 21	15	Teacher 192
Teacher 40	16	Teacher 170
Teacher 32	17	Teacher 153
Teacher 103	18	Teacher 180
Teacher 208	19	Teacher 87
Teacher 180	20	Teacher 191
Teacher 84	21	Teacher 32

Figure 12.1 Effectiveness rank-ordering changes over two years for individual mathematics teachers

Teachers at the top of the effectiveness rank-ordering in the previous year (e.g., Teacher 216) was ranked eleventh in the current year. Similarly, the teacher, who was at the bottom in the rank orderings of the previous year (Teacher 84), was almost at the top in the current. All this reflects the high volatility of teacher effectiveness estimates from year to year.

If teacher effectiveness scores can vary so dramatically from one year to the following year, the models that produce these estimates cannot be relied on for high-stakes personal decisions. This is evidence that such value-added models are highly unreliable.

12.2 Stability of teacher value-added effectiveness estimates when including an additional prior score (t-2)

This section focuses on the consistency of teacher value-added effectiveness scores obtained by adding additional years' prior attainment scores into the eventual models depicted in Table 11.17. Since all teachers could not be directly linked to students' two previous years' test scores (t-2), some teachers had to be excluded as in the previous section. Therefore, the analysis was conducted for 32 mathematics, 32 Turkish, 39 science, 32 revolution history, and 27 English teachers. The variables employed are summarized in Tables 12.1 and 12.2 in the previous section.

Correlation analyses were conducted to determine if there was any contribution of adding additional prior test scores (t-2, Grade 6) to the teachers' value-added effectiveness estimates where the previous year's test score (i.e., Grade 7 attainment) and the other related predictors were already controlled.

The results of Pearson's correlation analyses indicated that teachers' actual effectiveness scores were almost perfectly correlated with their corresponding effectiveness scores for all subjects (see Table 12.5). When a teacher's effectiveness score was high in one estimate, the corresponding teacher's effectiveness score also tended to be high in another estimate, or vice versa.

Table 12.5 Correlation Between Teachers' Actual and the Corresponding Value-added Effectiveness Scores

	Correlation Coefficient (r) (Score)	*Correlation Coefficient (r_S)* (Category)
Mathematics	0.999	1.000
Turkish	0.983	0.941
Science	0.999	1.000

History	0.976	0.893
English	0.997	0.968

In addition to the correlations of the effectiveness scores, four effectiveness categories were also created by grouping the scores into quartiles: highly effective, effective, partially effective, and ineffective. Each teacher was assigned to one of the four possible effectiveness categories based on the actual and the corresponding effectiveness score that was obtained by adding two previous years' test scores (i.e., Grade 6 attainment). The correlations between the effectiveness categories in each teaching subject are also displayed in the last column of Table 12.5. The results also showed that teachers' actual effectiveness categories were also very closely related to their corresponding effectiveness categories generated by using two prior years' test scores combined (i.e., Grade 6 and Grade7 attainments). Using history as an example, the transition matrix (Table 12.6) shows that 25% of teachers classified as highly effective using one prior year test score were also classified highly effective using two prior years' test scores combined. The consistencies in quartiles for other teaching subjects are also shown in Appendix H.

Table 12.6 A Transition Matrix for Consistency of Effectiveness Categories Derived from Using One Prior Year and Two Prior Years' Combined Test Scores for History Teachers (in Percentages)

		By using one prior year				Total
		Highly Effective	Effective	Partially Effective	Ineffective	
By using two prior years combined	Highly Effective	25.0	3.1	0.0	0.0	**9**
	Effective	6.3	12.5	3.1	0.0	**7**
	Partially Effective	0.0	6.3	15.6	3.1	**8**
	Ineffective	0.0	0.0	3.1	21.9	**8**
Total		**10**	**7**	**7**	**8**	**32**

History teachers' movements among the categories were expressed as a percentage of the total number of 32 teachers in the transition matrix. A very strong positive correlation was found between history teachers' effectiveness categories from one estimate to the corresponding estimate (r= 0.893). For other teaching subjects, more consistent results appeared in the differentiation of effectiveness categories obtained under two different models (see Appendix

H). For instance, while all mathematics and science teachers remained in the same effectiveness quartiles from one estimate to the corresponding one, 92.5% of English (n= 25 out of 27) and 81.4% of Turkish teachers (n= 26 out of 32) were assigned to the same effectiveness categories in both estimates. In addition to these results, it was also found that no teacher's effectiveness categories changed up or down by at least two categories from one estimate to the next. This result suggests that the use of additional prior test scores added little to the value-added effectiveness estimates.

CHAPTER 13

STABILITY OF VAMS IN TEACHER EFFECTIVENESS ESTIMATES USING DIFFERENT MODELLING APPROACHES

13.1 Stability of teacher value-added effectiveness estimates using different modelling approaches

This section examines the consistency of teacher value-added effectiveness estimates derived from three common value-added approaches, which are Residual Gain model (RG), Ordinary Least Squared or OLS-based model, and two-level HLM (hierarchical linear model).

Along with using the students' 8th-grade test scores in the relevant teaching subject as the outcome variable in all three approaches, while students' prior attainment was used as a single predictor in the residual gain model, the contextual variables depicted in Table 11.17 for each teaching subject were also employed as predictors in OLS-based and two-level HLM models in the analysis.

The records of a total of 35,435 students were examined to ascertain to what extent the consistent value-added estimates can be achieved for 230 mathematics, 232 Turkish, 204 science, 174 history, and 187 English teachers by using these three common value-added modelling approaches. Tables 13.1 and 13.2 summarize the data employed. As a reminder, contextual variables that showed no relationship with students' achievements were excluded from the model. These variables were indicated with "-".

Table 13.1 Data Employed in the Consistency of Value-added Effectiveness Estimates Derived from Three Modelling Approaches

	Mathematics (N= 7,543)		Turkish (N= 7,594)		Science (N= 7,116)		History (N= 6,638)		English (N= 6,544)	
	Mean	Std. deviation	Mean	Std. deviation	Mean	Std. deviation	Mean	Std. deviation	Mean	Std. deviation
Dependent Variables										
Outcome test score (Grade 8)	9.18	4.35	12.90	4.23	12.32	4.88	12.91	4.99	10.60	5.03
Student-level Independent Variables										
Prior test score (Grade 7)	9.52	4.89	14.16	4.28	12.14	4.24	10.93	4.35	9.88	2.72
School-level Independent Variables										
School Service Score	-		-		-		2.06	1.43	-	
Average school test score (Grade 7)	-		13.99	1.663	-		-		-	
Average school test score (Grade 6)	-		11.39	1.612	-		-		-	
Teacher/Classroom-level Independent Variables										
Class size	22.93	5.70	22.31	5.62	-		-		21.66	5.30
Percentage of female students	0.48	0.18	0.48	0.18	0.48	0.18	-		0.48	0.16
Total teaching experience	-		-		12.29	7.23	-		-	
Expr. in the current school	-		-		-		3.85	2.48	3.49	2.70
Average class test score (Grade 7)	9.48	2.71	-		12.10	2.15	-		9.88	2.72
Average class test score (Grade 6)	8.66	2.29	11.36	1.93	9.83	2.05	10.43	2.36	10.26	2.90

Table 13.2 Other contextual Characteristics Employed in Consistency Estimates in Subgroups

	Variables	Mathematics		Turkish		Science		History		English	
		N	Percentage	N	Percentage	N	Percentage	N	Percentage	N	Percentage
Student-level Independent Variables											
	Sex (Female)	3,643	48.3	3,668	48.3	3,431	48.2	3,203	48.3	3,165	48.4
	Language Learner	-		-		16	0.2	14	0.2	-	
School-level Independent Variables											
School Categories	General	-		-		-		-		5,744	87.8
	Regional Boarding									167	2.6
	Vocational									633	9.7
Teacher/Classroom-level Independent Variables											
	Graduation field (related)	-		-		-		5,701	85.9	-	
	Terminal degree (master/higher)	-		-		-		202	3	-	
Field of terminal degree	Related	-		48	13.7	209	54	-		-	
	Not related			107	30.6	19	5				
	Unspecified			195	55.7	155	41				

Correlation analysis was used to determine how much agreement there was in the individual teachers' value-added effectiveness scores derived from three common statistical approaches. More specifically, the effectiveness scores for each subject derived from the OLS model using the contextual predictors shown in Tables 13.1 and 13.2 were compared with the corresponding effectiveness scores derived from the RG model using student's prior attainment as a unique predictor and from the HLM model where the student-related variables such as prior attainment and sex situated in level-1 were nested within level-2 teacher/classroom (and school, if any) related variables.

The results (as represented by the Pearson's r coefficients) indicated that the effectiveness scores obtained from the OLS model are strongly and positively related to those using RG and HLM models across all subjects (see Table 13.3). But, the results from HLM model have a slightly stronger relationship with OLS model than those from RG.

Table 13.3 Correlation Coefficients Calculated by Comparing with OLS Across Subjects

	N	Raw Score		Ranking		Classification	
		RG	HLM	RG	HLM	RG	HLM
Maths	230	0.83	0.95	0.78	0.98	0.71	0.94
Turkish	232	0.87	0.96	0.83	0.98	0.77	0.94
Science	204	0.86	0.94	0.84	0.98	0.77	0.95
History	174	0.90	0.91	0.88	0.96	0.80	0.91
English	187	0.86	0.95	0.84	0.98	0.77	0.92

In addition to the correlations between effectiveness raw scores, teachers' effectiveness rankings and classifications were also compared across the three statistical approaches. Spearman's rank correlation showed that both RG and HLM models are strongly correlated with the OLS model, although HLM has a slightly stronger relationship than RG. Comparing teacher's effectiveness classification/categories, there was, again, a strong relationship between OLS, RG and HLM, although the correlation between OLS and RG was slightly weaker than that between OLS and HLM.

However, comparing the percentage of teachers that remained in the same effectiveness categories in the three approaches (Table 13.4), it can be seen that HLM model was more closely related to the OLS model in that around 80% of teachers stayed in the same quartile of effectiveness in both models. The OLS and RG models appear to be more divergent. For example, around 50% of maths teachers were classified differently under OLS and RG based estimates.

Table 13.4 shows the percentage of teachers that remained in the same effectiveness categories in HLM and RG models in comparison with the OLS model. In addition to this, transition matrixes were created for each teaching subject to examine the consistencies in the categories more closely (see Appendix I).

Table 13.4 Percentage of Teachers whose Effectiveness Category is Constant by Comparing with OLS Across Subjects

	N	% Stayed in the Same Quartile	
		RG	HLM
Maths	230	49.6	84.4
Turkish	232	55.2	84.5
Science	204	58.4	88.3
History	174	60.4	78.1
English	187	56.7	79.4

Comparing the three approaches in another way, looking at the proportion of teachers whose effectiveness categories changed between the value-added approaches regarding the quartiles (Table 13.5), the results again show that the OLS and HLM approaches are more consistent with each other. There were no teachers whose category changed by two or more quartiles in the corresponding statistical approach (HLM). While only an average of 17% of teachers changed one quartile in classification between OLS and HLM, the proportion of teachers who changed one quartile between OLS and RG was much higher, at an average of 39%.

Table 13.5 Percentage of Teachers whose Effectiveness Category Changed by Comparing with OLS Across Subjects

	N	% Changed one Quartile		% Changed two Quartiles		% Changed three Quartiles	
		RG	HLM	RG	HLM	RG	HLM
Maths	230	44.0	15.7	6.1	0.0	0.3	0.0
Turkish	232	40.8	15.5	3.6	0.0	0.4	0.0
Science	204	36.3	11.8	5.4	0.0	0.0	0.0
History	174	36.8	21.9	2.9	0.0	0.0	0.0
English	187	38.5	20.3	4.8	0.0	0.0	0.0

In comparison with the RG model, the effectiveness categories of one teacher changed three categories up or down (from the top quartile to the bottom, or from the bottom quartile to the top) in RG estimates for maths and Turkish, whereas no teachers changed three categories in the other subjects. For both HLM and RG, most of the misclassifications were only by one category.

Additional analysis was conducted to investigate the degree to which there is an intrinsic concordance of each model between pairs of classrooms, where the same teachers were assigned in the same school year. This was estimated by comparing the percentages of teachers who stayed or changed in the same effectiveness categories in different classes. A total of 510 teachers were identified that taught in multiple classes. As some teachers were assigned more than two classrooms, those classrooms were converted to pairs; for instance, where a teacher was assigned to three classes, three pairs of classrooms were created, such as class A and B, class A and C and class B and C. A total of 939 pairs of classrooms were identified for this comparative analyses.

Table 13.6 shows the percentage of teachers remaining in the same effectiveness categories between paired classrooms across teaching subjects. The transition matrixes that indicate the intrinsic consistencies in the categories for each statistical approach are also shown in Appendix J.

Table 13.6 Percentage of Teachers Remaining in the Same Effectiveness Categories Between Paired Classrooms Across Teaching Subjects

	N	% Stayed in the Same Quartile		
		RG	OLS	HLM
Maths	180	38.3	26.7	12.8
Turkish	172	30.2	27.9	9.9
Science	188	35.1	35.1	19.7
History	216	43.1	38.4	18.5
English	183	35.5	30.6	19.7

Conceptually, the value-added models on teacher evaluation attempt to isolate a particular teacher’s effects (or contributions) on their students’ learning from other factors outside of the teacher’s control. Therefore, the more a model created can achieve this isolation, the more reliable the model is. Where a teacher taught the same subject in different classrooms in the same school year, it is expected that the teacher should have a similar effectiveness score in each classroom in a given year if the model used was able to isolate this teacher's contribution to student achievement from other factors outside the teacher's control. More specifically, if a teacher is classified as "effective" in a class, it is expected that they would be in the same (or similar) category of effectiveness in another class. Based on the understanding that the less mobility there is in the classification of the same teacher between classes, the more trustworthy the model used is. Therefore, by comparing the percentage of teachers who stayed in the same effectiveness categories between paired classrooms across teaching subjects, we could see how reliable the models are.

As seen in Table 13.6, even though the RG model produced somewhat more stable results than other models, none of the models used generated truly consistent results between paired classrooms. As an example, 43.1% of history teachers were assigned to the same effectiveness category in both classes using the RG model, while only 9.9 % of Turkish teachers remained in the same category with HLM. In general, HLM model produced the least consistency in classifying teachers between classes for all subjects.

Another way of looking at consistency between classes is to compare the proportion of teachers whose effectiveness categories changed (Table 13.7). The movements of the total number of 939 teachers between the effectiveness categories (quartiles) were expressed as a percentage in the table. The table shows that HLM model was the least consistent in classifying teachers between paired classes. Around 30% of teachers across all subjects changed three quartiles in effectiveness categories between classes. For instance, 35.6% of maths teachers were classified as "highly effective" in one class and "ineffective" in another class (a change of three quartiles).

Table 13.7 Percentage of Teachers Whose Effectiveness Category Changed Between Paired Classrooms

	N	% Changed one Quartile			% Changed two Quartiles			% Changed three Quartiles		
		RG	OLS	HLM	RG	OLS	HLM	RG	OLS	HLM
Maths	180	34.4	38.9	26.1	17.8	27.8	25.5	9.5	6.6	35.6
Turkish	172	41.2	36.0	33.7	21.0	30.3	24.4	7.6	5.8	32.0
Science	188	39.4	36.2	26.1	17.0	22.3	25.5	8.5	6.4	28.7
History	216	36.9	38.0	29.6	14.4	18.0	20.4	5.6	5.6	31.5
English	183	38.2	35.5	25.7	15.7	24.0	25.1	10.4	9.9	29.5

As some teachers taught multiple classes, the normal ranges of results (i.e., standard deviations) for the same teacher in different classes were calculated for all classroom pairs. After pair scores were created for teachers assigned to multiple classes, the absolute differences between each pair were computed. Then, the mean of these differences and SDs were calculated. The SD of the effectiveness of the same teacher teaching different classes provided the normal range of results that the teacher assigned to these classes could have.

Table 13.8 depicts the means and SDs of the effectiveness results for the teacher teaching different classes in each teaching subject across the value-added approaches. Table 13.9 also shows the proportion of teachers whose effectiveness scores remained within the normal range.

Table 13.8 The Means and SDs of the Effectiveness Results Between Pair Classrooms for the Teacher Teaching Different Classes

	N	RG		OLS		HLM	
		Mean	SD	Mean	SD	Mean	SD
Maths	180	0.38	0.31	0.35	0.26	1.06	0.79
Turkish	172	0.37	0.27	0.34	0.27	0.99	0.80
Science	188	0.39	0.33	0.35	0.30	1.20	1.03
History	216	0.38	0.35	0.34	0.30	1.29	1.11
English	183	0.38	0.33	0.35	0.32	1.16	1.06

Table 13.9 Percentage of Teachers Whose Effectiveness Scores Remained within the Normal Range

	N	RG		OLS		HLM	
		n	%	n	%	n	%
Maths	180	90	50.0	82	45.6	79	43.9
Turkish	172	69	40.1	84	48.8	88	51.2
Science	188	104	55.3	102	54.3	101	53.7
History	216	128	59.3	117	54.2	117	54.2
English	183	102	55.7	104	56.8	107	58.5

Table 13.8 shows that the SD of the effectiveness score for the same maths teacher teaching different classes across statistical approaches ranged between 0.263 to 0.791, suggesting that the teacher effectiveness estimates can vary quite widely depending on the statistical model

used. Table 13.9 shows that 50% of maths teachers' (n= 90) effectiveness scores generated by RG stayed within the acceptable result range, while 45.6 % (n= 82) teachers' effectiveness scores from OLS remained within the normal range, and that for HLM was 43.9 % (n= 79). Similar results were obtained for the other teaching subjects.

In summary, the results show that the effectiveness scores of approximately half of the teachers fluctuated between classes. What this suggests is that none of the models had the potential to produce stable results between classes. However, it is worth mentioning that as the true teacher effectiveness scores are unknown, it was not possible to estimate how close the estimates that statistical approaches can produce are to the true score. Instead, what is done here is to see how similar the results are produced by different statistical approaches. The analysis suggests that there is no advantage in using a more sophisticated statistical approach, such as HLM. For example, the effectiveness raw scores ranking and classification of teacher effectiveness using HLM and OLS models are very similar, with a correlation ranging between 0.907 to 0.984 (see Table 13.3).

SECTION V
CONCLUSION

This is the conclusion section. It is made up of two chapters. Chapter 14 summarises the main findings of the research, bringing together the results from the systematic review and the primary research. The chapter also considers some limitations encountered during the conducting of the research. In Chapter 15, the implications of the findings are addressed for three stakeholders: policymakers and school leaders, researchers, and parents. In addition to this, this chapter discusses some possible suggestions for future research, and the chapter also considers what the findings of this study mean. It questions the purpose of teacher evaluation measures and what we could do instead.

CHAPTER 14
DISCUSSION

The overarching aim of this new study was to examine the stability of teacher value-added effectiveness estimates under different conditions. To achieve this objective, a series of value-added models were developed to see if these estimates of teacher effectiveness are consistent when we consider student, teacher/classroom, and school characteristics, when we include additional students prior years' test results, and if they are consistent over a two-year period and if measured using different statistical approaches. This chapter summarises the findings from both the systematic review and the primary research to answer these questions.

14.1 Summary of the Findings

14.1.1 How stable are teacher effectiveness measured by VAMs that consider student, school, and teacher-classroom characteristics?

The systematic review picked up 50 studies that met the inclusion criteria. Almost all the higher-quality studies identified students' prior attainment as the best predictor of teacher effectiveness. Hu (2015), for example, reported that the nearest prior year's test score alone accounted for an average of 57% and 59% of the variance in students' current academic performance in maths and reading, respectively. Kersting et al. (2013) showed that 68% of the variance in the students' current scores was explained by controlling for only one previous year's test score.

Consistent with the findings of the systematic review, the secondary data analysis revealed that the strongest student-related factor in explaining variation in student's current test scores in each teaching subject is their nearest prior attainment (i.e., maths score at Grade 7) (Aslantas, 2020). The new results showed that approximately half of the variance in the pupils' Grade 8 test score was explained by their Grade 7 results alone (except in history).

The review showed that other student variables (such as sex, English language status, disability status, socio-economic status and school attendance) contributed little to the predictive power of teacher effectiveness. For example, Heistad (1999) found that adding gender to a model that already controlled for student's prior attainment and race increased the explanatory power by only 0.1% to 0.4% depending on the testing year. Tobe (2008) excluded gender in the analyses as it did not make a significant contribution to explaining the variance in their model.

In line with the findings of the systematic review, the primary research also indicates that these other student contextual factors, such as sex, language identity, are not an important factor in explaining differences in the effectiveness of the teachers. These findings suggest that student contextual variables (other than their prior attainment) are not useful in estimating teacher effectiveness.

Findings from the review also provide no robust evidence that teacher/classroom variables (e.g., permanent status, experience, qualification, prior performance, teacher GPA and peer characteristics) and school-level factors are important predictors of teacher effectiveness. Sanders and Horn (1998) reported that racial diversity and the percentage of students receiving free/reduced-price lunches in schools are not linked to cumulative gains in Grades 3–8. Germut (2003) also found that school-level predictors, such as the percentage of students receiving free/reduced-price lunches, crowdedness, racial/ethnic composition, the proportion of students with special educational needs and those who spoke English as a second language (ESL), accounted for very little of the variance in student attainment. Although Ballou et al. (2004) suggested that controlling for the percentage of students eligible for free/reduced-price lunches in class or schools has an impact on TVAAS estimates in some grades and subjects, the authors advised caution in accepting the finding because of a large standard error in the coefficient of the predictor employed.

Regarding teacher-level predictors used in estimates, Nye et al.'s (2004) study found no clear links between teacher characteristics - experience and education - and teacher effectiveness estimates. A few studies have illustrated a link between some specific teacher characteristics and their effectiveness estimates. Kukla-Acevedo's (2009) study identified teachers' undergraduate performance (GPA) as a key predictor, Goel and Barooah (2018) found that teachers' permanent employment status (tenured) was the key predictor, while Tobe (2008) reported that teacher certification by the state is the only important factor. Munoz et al. (2011) suggested that it was teacher experience that mattered. All the other studies in the literature reviewed do not indicate any consistent teacher level factor as important predictors in teacher effectiveness estimates. The strongest studies (rated 4🔒) show that students' previous attainment is the best predictor of teacher effectiveness, and the inclusion of variables at the student and teacher level adds little to the predictive power of teacher performance assessment models.

In accordance with the review findings, the primary research also shows that students' school service score, school location, and the type of school attended accounted for little of the variation in teacher effectiveness after taking into account their prior attainment. Interestingly, teachers' effectiveness scores (except in Turkish) have a slightly negative relationship with both the total experience of teachers in their professional careers and their ongoing experience in their current school. More experienced teachers tend to have lower effectiveness scores on value-added estimates.

Only a modest correlation was found between class size and teacher effectiveness (except for science and history), and, intriguingly, teachers who taught large classes have, on average, higher effectiveness. This finding may contribute to efforts to reconsider the policy of class size reduction to raise student achievement, which involves considerable costs (Rivkin et al., 2005; Hanushek, 2003). Pisa analysis of international performance shows no relationship between class size and student attainment neither within nor across countries (OECD, 2012). Countries like South Korea and China have one of the largest class sizes in the world and yet consistently ranked highest on PISA International tables.

The analysis in the primary research suggested that, consistent with the existing literature, value-added estimates vary from teaching subject to subject depending on the predictors employed. Similar to this current study, Alban (2002) also ran a total of 105 multiple regression analyses in five content areas and found that only students' prior attainment was a significant variable in each analysis for each content area. The significance of other variables used in the equations varied considerably from one content area to another. This inconsistency in estimating teacher effectiveness suggests that such value-added models may not be useful in measuring teacher effectiveness.

14.1.2 How stable are teacher value-added effectiveness estimates over a two-year period of time?

Contrary to the existing literature reviewed (e.g., Aaronson et al., 2007; Bessolo, 2013; McCaffrey et al., 2009; Kane & Cantrell, 2010), which found positive correlations (even too small) for teachers' performance estimates from year to year, this primary research found that there is no meaningful relationship between teachers' current and previous effectiveness scores, regardless of teaching subjects. There was a (very weak) negative relationship between teachers' current and previous scores regardless of teaching subjects; for instance, a teacher's previous effectiveness score tends to be slightly better than his/her current score, on average or vice versa.

Bessolo (2013) examined the movement of mathematics teacher scores across six years and reported that only sixteen to fifty per cent of teachers remained in the same quartile from one year to the next. Newton et al. (2010) supported this finding by examining the teachers' rankings across years. The researchers found seventy and ninety per cent of teaches' effectiveness rankings changed by one or more deciles in either direction.

With evidence from existing literature, this study found that the measures of teacher effectiveness vary substantially across consecutive years. Therefore, it is suggested that value-added measures are unreliable. Presumably, teacher effectiveness, if it means anything, would be a relatively constant characteristic. If a model produces inconsistent results (i.e., teachers are assigned to the upper category this year and to the bottom in the following year or vice versa), the model should not be relied on, especially for high-stakes personnel decisions.

14.1.3 How stable are teacher value-added effectiveness estimates when including an additional prior score (t-2)?

The review of existing literature showed no consistent results with regards to using additional previous year(s) data in teacher effectiveness estimates. Some studies suggested that there are advantages to including additional lagged test scores to improve the stability of value-added estimates (e.g., Goldhaber and Hansen, 2013; Stacy et al., 2018; Potamites et al., 2009; McCaffrey et al., 2009), while others revealed it to be of little benefit (e.g., Ehlert et al., 2014; Johnson et al., 2015; Kersting et al., 2013). Still, others like Heistad (1999) and Goldhaber and Hansen (2010) suggested that using at least three years of data increases the consistency of value-added estimates. However, these were weaker studies because of the potential loss of data with the more years of data used. The stronger studies support the view that additional years' data do not add to the consistency of teacher effectiveness estimates.

The findings from the primary research concur with the stronger studies. The results show that teacher effectiveness scores using one prior year's test scores were closely related to their scores using two prior years' test scores. Teachers who were rated high on effectiveness scores using one prior year test score were also rated high using two prior years test scores combined. When teachers were classified into four categories of effectiveness, the results show that teachers' classification changed little regardless of whether it is one or two prior years' test scores were used. No teacher's effectiveness categories changed by at least two categories from one estimate to the next. These results suggest that there is no advantage in using additional prior test scores in measuring teacher effectiveness.

14.1.4 Do different methods of analyses used in VAMs produce consistent teacher effectiveness estimates?

The existing literature has not reached a large consensus on the relationship of model choice to teacher effectiveness estimates, either. There is some degree of agreement in that teacher performance estimates based on student's achievement growth substantially vary, depending upon the preferred model (Goldhaber et al., 2013; Sass et al., 2014; Sloat et al., 2018; Wei et al., 2012; Newton et al., 2010), while there is no evidence that a single data analysis method is superior to any other method regarding the ability to consistently estimate teachers' effectiveness in a variety of conditions.

Similar to Germuth's (2003) study, which found very high positive correlations between HLM and OLS models, the findings of this primary research also suggested that the OLS model has very strong and positive relationships with both RG and HLM regarding value-added effectiveness estimates for each teaching course. More specifically, very high positive correlations, over 0.90, were estimated between the effectiveness scores generated from OLS and HLM; similarly, the OLS model also produced very similar effectiveness scores with RG. However, the percentage analyses indicated that the relationships were not as strong as those revealed in the correlation analyses. For instance, it was found that only slightly more than half of the maths teachers were assigned to the same effectiveness categories in the RG and the OLS models. Moreover, compared to the OLS model, the effectiveness categories of around forty-five per cent of maths teachers in the RG model changed one category up or down, while six per cent moved two categories and one case moved three effectiveness categories.

Since the true teacher effectiveness scores are unknown, this sub-research question investigated the extent to which similar predictions could be produced across the statistical approaches, rather than investigating how close estimates were to the true scores. Based on the very strong relationship between HLM and OLS models in terms of effectiveness raw scores, ranking, and categories, consistent with the findings of the retrieved studies (Blackford, 2016; Germuth, 2003), this new study clearly suggests that the use of any more sophisticated statistical approach provides very limited advantages for estimating teacher effectiveness. This suggests that the simplest approach should be used, not least because it allows for the widest sceptical readership. On the other hand, somewhat different results emerged between the OLS and RG models (the simplest model used in the analyses), especially in placing teachers in effectiveness categories.

14.2 Limitations

The study has potential limitations that should be considered by future researchers when replicating this study or when the results of this study are used by stakeholders. The limitations are mentioned throughout this thesis, although all of these limitations are discussed collectively in this section.

Missing data is almost always problematic when conducting research utilising longitudinal data, and missing data may affect the findings of this study. The first missing data causing limitations in the study is related to students. Although there was a total of 18,986 students enrolled in grade 8, the total student population in this study is approximately 16,000 for each teaching course (around 16% missing cases). Prior research revealed that the missing data occurs disproportionately for low-achieving students (Gorard & Siddiqui, 2019). Since missing data is not random, it has the potential to cause bias in estimates.

The second major source of missing data arises from the need for teacher-level data in the estimates. The teacher-level data did not exist in the administrative dataset, so the related data was requested from schools' administration offices. The fact that many school directorates did not share the relevant teacher information with the researcher caused attrition of between 54% and 60% in the data obtained. Although the average test scores of the initial student population and the restricted population were close to each other, it is possible that the attrition cases influenced the conclusions that can be drawn from the data. Another missing data issue to be considered is rooted in the analysis method utilised in sub-RQs 2 and 3. Since these sub-research questions required using two lagged test scores, adding the second lagged test scores to the estimates caused the attrition rate to be much higher (more students had missing scores from two years prior). The longitudinal administrative data set did not allow all teachers to directly link with all student samples in the sixth grade (t-2); therefore, in the estimates using the 6th-grade test scores of the students, the selection criteria explained in the relevant section were used. These selection criteria caused the total number of teachers (1,027) to decrease to 151. This large loss in data may cause bias in the results drawn, and this non-random attrition also reduces the strength of the findings.

In addition to the limitation of available student and teacher data, another potential concern is the lack of background data at student-level. The secondary data analysis and existing literature agree that students' prior attainment plays a crucial role in their current attainment. Literature also suggests that students' SES (socio-economic status), which is a measure of their families' poverty, is related to students' attainment (OECD, 2016; Gorard & See, 2009). Moreover, students' academic achievement could be influenced by their family well-being. In this study, instead of directly investigating the relationship between student attainment and predictors, the contribution of the predictors in explaining the variation in the current academic performance of the students was examined. In the literature, there is a general consensus that the inclusion of the SES variable in the estimates makes a very limited contribution to the predictive power of teacher performance assessment models once prior attainment (t-1) is taken into account. However, this consensus could not be tested by this study. The absence of this data limited the contribution of this research to the literature regarding exploring the contribution of SES to the variation explained in students' academic attainment.

CHAPTER 15
IMPLICATIONS AND CONCLUSION

This chapter discusses the implications of the findings for three groups of stakeholders: policymakers and school leaders, researchers, and parents and offers some possible suggestions for future research.

15.1 Implications for Educational Policymakers and School Leaders

A need to reconsider the use of value-added measures in evaluating teachers

Value-added measures are increasingly used by policymakers for school measuring, school improvement, and teacher performance. Schools and teachers can be criticised or penalised, and praised or rewarded based on such measures. Both the systematic review and the primary research have demonstrated that value-added models do not produce consistent enough results for measuring teacher effectiveness. This study revealed that value-added models produced substantially different results when teacher scores were categorised or ranked (which is the common usage of VAM scores in accountability systems). Correlation analysis revealed no clear association between teachers' performance estimates from year to year regardless of teaching content area. Using VAMs, a teacher can be classified as an effective teacher the previous year and ineffective the following year.

Given the lack of stability in such models to accurately classify teachers, performance results achieved through VAMs should not be relied upon, especially for high-risk personnel decisions. It is dangerous, divisive, and demoralising, and there is no robust evidence that such performance measures can improve teacher competency. Therefore, the use of value-added models in evaluating teachers either for promotion or retention should be actively reconsidered.

Interventions should be introduced much earlier in students' school life

Both the literature and secondary data analysis revealed that a prior test score is the best predictor to explain the variation in a student's current test score, which also means that the differential effect of a particular teacher on student outcome is not as great as that of students' prior academic performance. Therefore, in line with UNESCO's 2030 education goal (4.2) which is "all girls and boys have access to quality early childhood development, care and pre-school education so that they are ready for primary education" (UNESCO, 2016), the implication of this finding is that any intervention to improve students' achievement should be introduced much earlier in their school life. However, this does not mean that teachers are not important. Teachers might be key to schools and student learning, even if they are not differentially effective from each other in the local (or any) school system. Therefore, systems that attempt to differentiate “effective” from “ineffective” teachers may not be fair to some teachers, given the method we have available at present. For example, VAMs, the subject of this study, claim to evaluate the effectiveness and success of teachers on the results of the students’ examination. Unfortunately, since this method allows only relative judgements to be made, it cannot be regarded as a precise indicator of the classification of teachers as effective or ineffective. Indeed, this research suggests that any single method used in teacher evaluation cannot accurately measure actual teacher effectiveness and will therefore lead to misclassification of teachers’ performance, so abandoning these methods might be better until a better evaluation method is created.

Re-evaluate the purpose of teacher evaluation

There are others who suggested that perhaps we also should look at other methods of teacher evaluation, such as the use of multiple measures instead of relying on student test scores alone. Admittedly, the results of the current study may have been influenced by many factors, such as study design and data quality. For instance, along with missing data, the dataset does not include a crucial predictor, students’ SES, which might be linked to students’ attainment. These limitations make it necessary to take into account the findings of other studies. Because the educational process is a complex structure, VAMs do not provide information about the strengths and weaknesses of teachers’ classroom practice, while observational assessment fails to distinguish effective teachers from ineffective ones. The use of multi-directional and comprehensive teacher evaluation methods would be helpful in developing teachers. In this way, teachers can realize effective teaching methods that they have applied in their classes, develop these methods and contribute to their professional development. On the other hand, teachers who are considered to be less effective can become more effective by recognizing their shortcomings and obtaining the extra training they need to overcome them. In both cases, the awareness and self-knowledge skills of teachers could contribute to the improvement of the teacher and, of course, the teaching quality. However, this cycle of feedback and improvement has never been demonstrated and must remain just an idea until it is robustly tested.

Similarly, some long-termed studies such as the MET Project (Little et al., 2009) suggested that measurements made with a single tool never provide comprehensive information about teachers’ effectiveness, and the information should be collected from multiple sources, using a combination of classroom observations, student surveys, and value-added. However, while combining several measures can provide more accurate information about teacher's effectiveness, it should be kept in mind that the degree of accuracy depends on the reliability and validity of the underlying components. For instance, as found in this current study, VAMs could not produce consistent enough results for measuring teacher effectiveness, while classroom observations may be biased in measures of teacher effectiveness (Rothstein, 2009, 2010) since teachers are seldom sorted into classes. As Steinberg & Garrett (2016) noted, teacher observation assessment result is strongly and positively related to students' prior attainment, which in turn strongly influenced students' interactions with teachers (see Section 2.2), such observation tools are not necessarily objective ways of assessing teachers. All this suggests that caution has to be taken when implementing teacher evaluation, even it is based on multiple measures.

More importantly, instead of searching for the most accurate way of evaluating teachers, perhaps policymakers and school leaders should reconsider the purpose of such evaluation. Similarly, Robertson-Kraf (2014) stated that teacher evaluation is negatively related to their expectations and does not contribute much to their classroom performance and their decision to remain as a teacher. If the purpose of teacher evaluation is to improve the quality of teachers, which is one of the Sustainable Development Goals of UNESCO 2030 (4.c) (UNESCO, 2016), it might be better to focus on teacher development and training. If VAMs are to be used, they should be as a formative or diagnostic tool providing feedback on how the individual teachers can improve and to identify their needs and what kind of training would make them a better teacher. I believe all teachers want to be effective – no teachers will deliberately not want to help their students do better.

15.2 Implications for Educational Researchers

This new study reveals that the choice of VAM has a substantial link to teacher effectiveness categories derived from their estimates. Although very strong relationships were found between model pairs regarding teacher raw value-added effectiveness scores, on average, 44% of teachers in residual gain model (RG) and 17% in hierarchical linear model (HLM) were classified into different value-added effectiveness categories, compared with ordinary least required model (OLS). This variation suggests that VAM is not the answer to identifying more or less effective teachers.

The high degree of similarity between models in which all contextual variables are included and those with some missing suggests that variables without predictive power should be avoided in order to make models clearer and more understandable. In addition to all these findings, as it is less expensive, more transparent, and more practical, this study would recommend the OLS-based multiple regression model over RG and HLM if a model is needed to be selected from them. Moreover, the internal consistency of each model between pairs of classrooms indicated that the OLS model produced slightly more stable results than the HLM model, while all consistencies between classes were very low.

The study did not provide strong evidence of the superiority of using multi-year student data over a single year in value-added teacher effectiveness estimates. The very high positive correlation between estimates suggests that the use of additional lagged scores gives a very limited advantage to value-added effectiveness estimates.

15.3 Implications for Parents

The findings of this study also have implications for parents with regards to their choice of schools for their children. The findings suggest that the academic performance of students alone are not good predictors of the quality of teachers. Parents should look beyond the academic performance of schools. Teachers who teach to the test may be less effective than teachers who promote a deep conceptual understanding of the curriculum.

Parents apply many different methods when choosing a school/teacher for their children, such as taking into account the opinions of other parents whose children were taught by that teacher, the teachers' experiences, and the exam results of the teacher's previous cohort of students. This study reveals that teacher effectiveness scores based solely on students' exam results can vary from year to year and, as such, are not useful in helping parents to identify effective teachers or effective schools.

Because there are many other factors such as students' readiness, school resources, and family that play key roles in shaping teacher practices that might have an impact on students' test results, parents may need to use other criteria when choosing their children's schools. These factors may include the school culture and ethos, student interests and abilities, and teachers' attitudes, thus creating a more suitable teacher-student match. In other words, the most important criterion to which families need to pay attention in teacher selection should not be "the best" at school, but the one who has the best match with their child.

15.4 Recommendations for Further Research

Researchers should continue their work developing models that are more robust and fair on those evaluated. Since "effectiveness" inherently refers to causality and the design of studies should be suitable to reveal this causality. Regression discontinuity design (RDD) is a promising approach. Only then can the models really provide new information on accountability systems for the literature.

This project was not conducted to determine a more accurate model for teacher value-added effectiveness estimates. Rather, it examined the contribution of contextual predictors to teacher effectiveness estimates and the agreement of teacher value-added effectiveness estimates derived from three value-added approaches. Since this project utilized longitudinal administrative data without knowing about true teacher effects, it is not possible to determine which model generates the more accurate effectiveness score for teachers. A critical area that requires further research is the accuracy of estimates derived from models employed. An appropriate way to investigate this would be to conduct a study based on simulated data where the effects of individual teachers on students' academic attainment are known. By comparing the estimates derived from the tested model with the known teacher effects, it can be estimated how close these models can be in producing value-added estimates.

It is not easy to evaluate teacher performance accurately by any single measure. Therefore, there is a need to develop a more comprehensive teacher evaluation system, taking into account the existing problems with methods and practices used in measuring teacher effectiveness. Despite many doubts about VAMs, education politicians still tend to integrate them into their accountability systems in order to develop their own teacher evaluation system. For this reason, educational researchers should also tend to do more research to increase the reliability and validity of the current evaluation systems or alternative methods. Even though VAMs are included in teacher evaluation systems, supporting VAMs with other alternative evaluation measures such as observation and surveys, as suggested in the previous section, would provide a more comprehensive picture of the teacher performance in the classroom. Therefore, multiple measures to evaluate teachers' effectiveness may be considered in a new, comprehensive accountability system. Kane & Staiger (2012) also suggested that combining multiple measures can provide a wide range of information about teacher performance in the classrooms. Classroom observation and student surveys are the most notable measures among the alternatives examined by previous researchers (Darling-Hammond et al., 2012b; Jia et al., n.d.;

Martinez et al., 2016; Chaplin et al., 2014). Future researchers, therefore, can examine the extent to which statistical estimates based on student test performance are related to observations and student surveys or other indicators of teacher effectiveness such as portfolios. It is expected that there will be a strong correlation among measures used for the same purpose, which is to evaluate teacher performance. For instance, Jacob and Lefgren (2008) analysed the data on the evaluation of 201 teachers teaching Grades 2–6 and found a strong relationship between teachers' evaluations by school principals and the same teachers' VAM scores estimated using students' mathematics and reading scores. The researchers suggested that evaluations based on students' test scores were fairly accurate in predicting which teachers would be in the top 20% the following year, but the accuracy of these estimates was increased by combining the evaluations of the school principals with VAM estimates.

The Effective Teaching Measures (MET) is a notable project on the combination of multiple measurements (test scores, observations, student survey) in a comprehensive teacher assessment that may be a good example for future researchers. However, the use of multiple measures in teacher performance evaluation causes new problems, which also means new research areas for researchers. Using multiple measures raises a new question of how much weight each measure of effectiveness will carry in the comprehensive performance evaluation model to be created. Questions such as whether to give the same weight to each item to create a fairer model or what weights of each of the composite measures in the teacher evaluation model are optimal, are among the burning questions in this field.

Future research could also consider the following:

Test the use of VAMs as a diagnostic rather than an evaluative tool

The potential of VAMs as a diagnostic rather than an evaluative tool to identify areas of improvement and areas of strength to individual teachers to support their development has not been tested. This cycle of feedback and improvement could be robustly tested as an effective teacher development model.

Explore a more comprehensive way to identify variables that potentially influence student performance

VAMs are statistical models based on the principle of predicting the teacher's effectiveness on student attainment by controlling various factors that may affect student scores. The predicted score obtained by controlling student attainment to some degree is subtracted from the actual score. It is also the case that measures of teacher effectiveness may be affected by other factors beyond the control of the teachers, such as peer influence, classroom dynamics, out-of-school tutoring, student's general intelligence, poverty and parental involvement. One of the limitations of VAMs is the inability to control for all the predictors so that the influence of the teacher can be isolated.

For example, in this study, while 47% of the variation could be explained by prior attainment alone (blue part), the contribution of other contextual variables was 3% (orange part). The other 50% (grey) refers to other factors that have an impact on students' attainment but are not included in the equation. Teacher effects are in this grey area. But so are other factors such as the school, family, peers, school effectiveness, and leadership. All have a claim on this 50% along with any random error, plus the inevitable bias caused by missing variables, missing values, measurement error, and the like. The residuals contained a composite effect of all these factors and probably many more. Therefore, it is unknown precisely how much of the unexplained half of the variation in test scores is due to teacher effectiveness, as depicted in Figure 15.1.

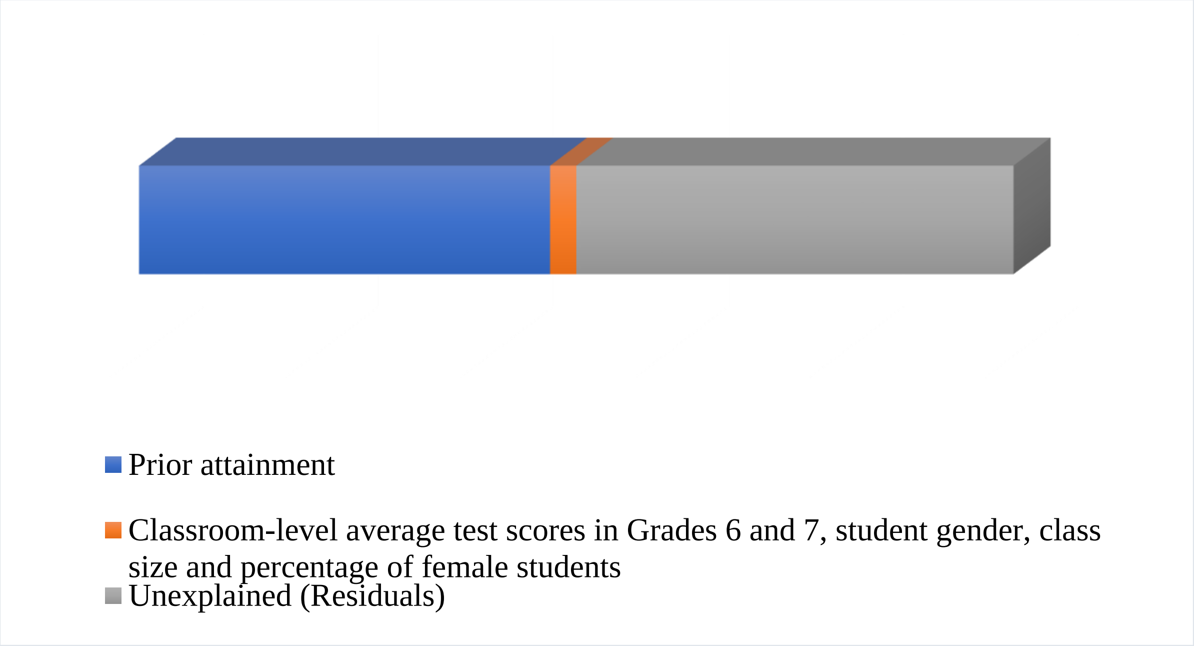


Figure 15.1 Components in explaining the variance in students' current mathematics attainment

For this reason, it cannot be claimed that all of the residuals, which form the difference between the students' actual scores and their estimated scores, are due to teacher effectiveness. In other words, the model does not measure what it was created to measure.

To separate the teacher effects from composite effects, some commentators suggested using the fixed-effect method where dummy variables representing each teacher are added to the model, and the coefficient of each dummy variable denotes the effectiveness of the corresponding teacher. The inclusion of teacher effectiveness as a fixed effect in the model causes the observed teachers to be treated as all teachers in the population of interest (McCaffrey et al., 2003). However, fixed effect estimates are also criticized in that the teacher's value-added effectiveness scores are excessively influenced by extreme student scores, especially in small classes. Therefore, this research study cannot suggest that one model specification is better than another, especially when taking high-stakes decisions; all specifications may produce biased results.

Last, in the systematic review study, some studies may have been missed, and new studies will have emerged. The systematic review in this study is focused on value-added models as a measure of teacher effectiveness. To contribute to understanding existing unknowns about VAMs in the light of new evidence gained from any overlooked or newly published studies, the final recommendation for future researchers is to conduct further systematic review studies.

15.5 Conclusion

A student's academic achievement might be affected by various factors that are student, teacher and school-related, but a growing number of studies have consensus that teacher quality is often considered the most important school-related factor in student achievement (Rice, 2003; Rivkin et al., 2005; Aaronson et al., 2007; Darling-Hammond, 2015). Based on this view, policymakers have tended to develop educational policies that hold schools and teachers accountable for students' achievement. To ensure that qualified teachers are employed in classrooms, they need a performance evaluation system that can determine teacher contributions to student attainment. However, teacher quality (or effectiveness) is not an easy attribute to measure. There are instances where teachers' competencies are measured by school leaders using observation appraisals. This does not always work, as demonstrated in the Bill & Melinda Gates multimillion-dollar initiative where school leaders were reluctant to give teachers a low rating, and few teachers were rated ineffective (Stecher et al., 2018). Some researchers have asked teachers to rate their own teaching efficacy and competence (e.g., de Paor, 2016). Others used teachers' years of experience, teacher test scores, highest degree attained, or National Board Certification as proxies for teacher quality (Darling-Hammond, 2000; Deluca et al., 2016; Feng, 2010; Goldhaber et al., 2004). None have been found to be entirely satisfactory.

Recently, measures of teacher effectiveness have relied on more “objective” measures using student outcomes, such as student performance in high stake tests. Value-added models (VAMs), statistical methods adapted from economics based on student academic achievement growth, are alternative measurements regularly used in teacher accountability. The use of VAMs in measuring teacher performance is one of the most controversial and important issues of educational policy. Although various aspects of VAMs have been criticized by researchers, such as reliability and validity, policymakers’ decisions largely ignore this and VAM estimates are still given credence. Important decisions about teachers such as salary increases, promotion, or termination of employment are made based on such performance evaluations. Schools and teachers are penalised and even shamed based on such measures. They can be damaging, demoralising, and demotivating.

This study aims to provide guidance and advice to policymakers and other stakeholders on the use of VAMs as a teacher performance appraisal. The findings from this study provide no evidence that value-added estimates of teacher performance are useful in measuring teacher effectiveness. They do not produce reliable and consistent results and thus risk misclassifying teachers. They should not be used for making high stake decisions regarding teachers’ promotion, dismissal, or bonuses.

This study reveals that students’ current attainment mostly depend on their performance in early education and therefore suggests that the focus of education and investment should be on the early years. Instead, the issue of measuring the performance of teachers has been one of the leading issues in education policies. However, rather than searching for the most accurate way of evaluating teachers, perhaps policymakers should reconsider the purpose of such evaluation. If the purpose is to differentiate effective teachers from non-effective ones, continuing with such evaluations is unlikely to make teachers effective. If teachers do not know how to improve, appraising them is not going to help. As with students, giving them more tests is not going to help them to improve unless they know what is expected of them and are given the tools to reach that expectation.

If evaluations are to be used, the purpose of teacher evaluation should not be rewarding or punishing them, but rather to be formative (i.e., to help teachers develop) by providing feedback to teachers on how they can improve and identify their needs. It would be more useful to have continuous professional development, regular training of teachers to update them on the latest curriculum requirements, and effective teaching pedagogies if the aim of teacher evaluation is to improve teacher quality. Therefore, more time and money should be spent on training and developing teachers rather than evaluating them.

We need to re-consider the selection and training of teachers. If teachers have gone through proper training, passed teaching training exams, and selected/appointed to teach, they should be qualified to teach. If they are not, it is probably the failure of the selection or training process, not teachers.

BIBLIOGRAPHY

Aaronson, D., Barrow, L. & Sander, W. (2007). Teacher and student achievement in the Chicago public high schools. *Journal of Labor Economics*, 25(1), pp. 95-135.

Achilles, C.M., Bain, H.P., Bellott, F., Boyd-Zaharias, J., Finn, J., Folger, J., Johnston, J. & Word, E. (2008). *Tennessee's Student Teacher Achievement Ratio (STAR) project*. Harvard Dataverse.

Alban, T.R. (2002). *Evaluating school and teacher effectiveness: A comparison of analytic models.* PhD thesis. University of Maryland. Available at: https://search-proquest-com.ezproxy.nottingham.ac.uk/pqdt/docview/305525642/FB74B4E7F9F42F8PQ/1?accountid=8018 (Accessed: 22 May 2019).

American Academy of Special Education Professionals (AASEP) (n.d.). *Progress Monitoring in a RTI.* Available at: http://aasep.org/fileadmin/user_upload/Protected_Directory/BCSE_Course_Files/Course_4/Chapter_5_UNDERSTANDING_RTI.pdf (Accessed:6 August 2020).

American Educational Research Association (AERA) (2015). AERA Statement on Use of Value-Added Models (VAM) for the Evaluation of Educators and Educator Preparation Programs. *Educational Researcher*, 44(8), pp.448–452. doi: 10.3102/0013189X15618385.

American Psychological Association (APA) (n.d.). Reliability. Available at: https://dictionary.apa.org/reliability (Accessed: 21 January 2021).

Amrein-Beardsley, A. & Geiger, T. (2020). *Methodological concerns about the Education Value-Added Assessment System (EVAAS): Validity, reliability, and bias*. Los Angeles, CA: SAGE Publications.

Amrein-Beardsley, A. & Holloway, J. (2019). Value-added models for teacher evaluation and accountability: Commonsense assumptions. *Educational Policy*, 33, pp. 516-542.

Aslantas, I. (2020). Impact of contextual predictors on value-added teacher effectiveness estimates. *Education Sciences*, 10(12), p.390.

Bacher-Hicks, A., Kane, T. & Staiger, D. (2014). *Validating teacher effect estimates using changes in teacher assignments in Los Angeles*. Cambridge, MA: National Bureau of Economic Research.

Bacher-Hicks, A., Chin, M.J., Kane, T.J., & Staiger, D.O. (2019). An experimental evaluation of three teacher quality measures: Value-added, classroom observations, and student surveys. *Economics of Education Review*, 73 https://doi.org/10.1016/j.econedurev.2019.101919

Ball, D.L. & Rowan, B. (2004). Introduction: Measuring instruction. *The Elementary School Journal*, 105(1), pp.3-10.

Ballou, D., Sanders, W.L. & Wright, P. (2004). Controlling for students' background in value-added assessment for teachers. *Journal of Educational and Behavioral Statistics*, 29(1), pp. 37-65.

Beamish, P. & Morey, P. (2013). School choice: What parents choose. *TEACH Journal of Christian Education*, 7(1), p.7.

Berliner, D.C. (2014). Exogenous variables and value-added assessments: A fatal flaw. *Teachers College Record*, 116, pp. 1-31.

Berry, B. (2010). *Teacher Effectiveness: The Conditions that Matter Most and a Look to the Future*. Center for Teaching Quality. Available at: https://files.eric.ed.gov/fulltext/ED509720.pdf (Accessed: 19/05/2020).

Bessolo, J. (2013). *The stability of teacher effects on student math and reading achievement over time: A study of one midsize district*. EdD thesis. University of Kansas. Available at: https://kuscholarworks.ku.edu/bitstream/handle/1808/15129/Bessolo_ku_0099D_12677_DATA_1.pdf?sequence=1 (Accessed: 01 July 2019).

Bettany-Saltikov, J. & McSherry, R. (2016). *How to do a systematic literature review in nursing: A step-by-step guide* (2nd ed.). Maidenhead: McGraw-Hill/Open University Press.

Blackford, K.L. (2016). *Measuring teacher effectiveness: A comparison across VA models utilizing Arkansas data*. PhD thesis. University of Arkansas. Available at: https://scholarworks.uark.edu/cgi/viewcontent.cgi?article=3084&context=etd (Accessed: 01 Jun 2019).

Blackorby, J., Taylor, C., & Wei, X. (2016). *Using growth models to measure child/student outcomes for state systemic improvement plans*. IDEA Data Center. Rockville, MD: Westat.

Blatchford, P., Russell, A., Bassett, P., Brown, P. & Martin, C. (2007). *The role and effects of teaching assistants in English primary schools (Years 4 to 6) 2000–2003. Results from the Class Size and Pupil—Adult Ratios (CSPAR) KS2 Project*. British Educational Research Journal, 33(1), pp.5-26.

Bowerman, B.L. & O'Connell, R.T. (2000). *Linear Statistical Models: An Applied Approach*. Duxbury Press.

Brown, H., Chudowsky, N. & Koenig, J, (Eds). (2010). *Getting value out of value-added: Report of a workshop*. Washington, DC: National Academies Press.

Buddin, R. & Zamarro, G. (2010). *What teacher characteristics affect student achievement? Findings from Los Angeles public schools*. Santa Monica, CA: RAND Corporation.

Buddin, R. (2011). *Measuring teacher and school effectiveness at improving student achievement in Los Angeles elementary schools*. Munich Personal RePEc Archive.

Burgess, S. (2019). *Understanding teacher effectiveness to raise pupil attainment*. IZA World of Labor, 465. doi: 10.15185/izawol.465

Buzick H.M. & Laitusis, C.C. (2010). Using growth for accountability: Measurement challenges for students with disabilities and recommendations for research. *Educational Researcher*, 39(7), pp.537–544.

Camburn, E. & Barnes, C.A. (2004). Assessing the validity of a language arts instruction log through triangulation. *The Elementary School Journal*, 105(1), pp.49-73.

Cantrell, S. & Kane, T. J. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET Project's three-year study*. MET Project, Bill & Melinda Gates Foundation.

Castellano, K.E. & Ho, A.D. (2013a). *A Practitioner's Guide to Growth Models*. Council of Chief State School Officers.

Castellano, K.E. & Ho, A.D. (2013b). Contrasting OLS and Quantile Regression Approaches to Student "Growth" Percentiles. *Journal of Educational and Behavioral Statistics*, 38(2), pp.190–215.

Castellano, K.E. (2011). *Unpacking student growth percentiles: Statistical properties of regression-based approaches with implications for student and school classifications*. PhD thesis. University of Iowa. Available at: https://ir.uiowa.edu/etd/931/ (Accessed: 03 Jun 2019).

Chaplin, D., Gill, B., Thompkins, A. & Miller, H. (2014). *Professional practice, student surveys, and value-added: Multiple measures of teacher effectiveness in the Pittsburgh public schools*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic. Retrieved from http://ies.ed.gov/ncee/ edlabs.

Chatterjee, S. & Hadi, A.S. (2012). *Regression analysis by example*. Somerset: John Wiley & Sons, Incorporated.

Chetty, R., Friedman, J.N. & Rockoff, J.E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, 104(9), pp. 2593-2632.

Clotfelter, C.T., Ladd, H. & Vigdor, J.L. (2006). Teacher-Student Matching and the Assessment of Teacher Effectiveness. *The Journal of Human Resources*, 41(4), pp.778–820.

Coe, R., Aloisi, C., Higgins, S. & Major, L. E. (2014). *What makes great teaching? Review of the underpinning research*. London: Sutton Trust.

Cohen, J. (1994). The earth is round (p<. 05). *American Psychologist*, 49(12), p.997.

Creemers, B. (1994). *The effective classroom*. London: Cassell.

Culbertson, M.J. (2016). *Precision of student growth percentiles with small sample sizes*. RMC Research Corporation. Centennial, CO: Regional Educational Laboratory Central at Marzano Research.

Cunningham, P.L. (2014). *The effects of value-added modeling decisions on estimates of teacher effectiveness*. PhD thesis. The University of Iowa. Available at: https://ir.uiowa.edu/cgi/viewcontent.cgi?article=5486&context=etd (Accessed: 01 Jun 2019).

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy. *Education Policy Analysis Archives*, 8(1), pp. 1-44.

Darling-Hammond, L. (2015). Can value added add value to teacher evaluation?. *Educational Researcher*, 44(2), pp. 132-137.

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E. & Rothstein, J. (2012a). Evaluating teacher evaluation. *Phi Delta Kappan*, 93(6), p. 8–15.

Darling-Hammond, L., Jaquith, A. & Hamilton, M. (2012b). *Creating a comprehensive system for evaluating and supporting effective teaching*. Stanford, CA: Stanford Center for Opportunity Policy in Education.

de Paor, C. (2016). The impact of school-based continuing professional development: Views of teachers and support professionals. *Irish Educational Studies*, 35(3), pp.289-306.

Deluca, C., LaPointe-McEwan, D. & Luhanga, U. (2016). Teacher assessment literacy: A review of international standards and measures. *Educational Assessment, Evaluation and Accountability*, 28(3), pp.251-272.

Denton, C. (2012). Response to Intervention for reading difficulties in the primary grades: Some answers and lingering questions. *Journal of Learning Disabilities*, 45(3), pp. 232-243 doi: 10.1177/0022219412442155

Department for Education (DfE) (2013). *Teachers' Standards*. London: DfE publications.

Department for Education and Employment (DfEE) (2000). *Research into teacher effectiveness: A model of teacher effectiveness.* Research Report No. 216 by Hay McBer. London: DfEE.

Dwyer, T.J. (2016). *A comparison of educational "value-added" methodologies for classifying teacher effectiveness: Value tables vs. covariate regression*. PhD thesis. University of South Florida. Available at: https://scholarcommons.usf.edu/etd/6228/ (Accessed: 21 May 2019).

Education and Science Workers' Union (2016). *It is not possible to accept the performance evaluation imposition of MoNE!*. Available at: https://egitimsen.org.tr/mebin-performans-degerlendirme-dayatmasini-kabul-etmek-mumkun-degildir/ (Accessed: 05 May 2018).

Educators Trade Union (2016). *Principals' evaluation of teachers should be abandoned*. Available at: https://www.ebs.org.tr/manset/3766/mudurlerin-ogretmenleri-degerlendirmesi-uygulamasindan-vazgecilmelidir (Accessed: 05 May 2018).

Ehlert, M., Koedel, C., Parsons, E. & Podgursky, M.J. (2014). The sensitivity of value-added estimates to specification adjustments: Evidence from school- and teacher-level models in Missouri. *Statistics and Public Policy*, 1(1), pp. 19-27.

El Soufi, N. & See, B.H. (2019). Does explicit teaching of critical thinking improve critical thinking skills of English language learners in higher education? A critical review of causal evidence. *Studies in Educational Evaluation*, 60, pp. 140-162.

Ergen, H. & Esiyok, I. (2017). Teacher opinions on school principals' instructional supervision performances. *Journal of Contemporary Administrative Science*, 3(1), pp.1-19.

Everson, K.C. (2017). Value-added modeling and educational accountability: Are we answering the real questions?. *Review of Educational Research*, 87(1), pp. 35-70.

Feng, L. (2010). Hire today, gone tomorrow: New teacher classroom assignments and teacher mobility. *Education Finance and Policy,* 5(3), pp.278-316.

Field, A.P. (2013). *Discovering statistics using IBM SPSS statistics: and sex and drugs and rock 'n' roll (4th ed.)*. London: SAGE.

Figueiredo Filho, D.B., Paranhos, R., Rocha, E.C.D., Batista, M., Silva Jr, J.A.D., Santos, M.L.W.D. & Marino, J.G. (2013). When is statistical significance not significant?. *Brazilian Political Science Review*, 7(1), pp.31-55.

Foegen, A. (2008). Algebra progress monitoring and interventions for students with learning disabilities. *Learning Disability Quarterly*, 31(2), pp. 65–78. doi: 10.2307/20528818.

Foegen, A. & Morrison, C. (2010). Putting algebra progress monitoring into practice: Insights from the field. *Intervention in School and Clinic,* 46(2), pp. 95-103. doi:10.1177/1053451210375302

Fuchs, L.S. & Fuchs, D. (2011). *Using CBM for progress monitoring in reading*. Washington, DC: National Center on Student Progress Monitoring. Available at: https://files.eric.ed.gov/fulltext/ED519252.pdf (Accessed: 11 November 2018).

Gagnon Jr, D.J. (2014). *Understanding the distribution of teacher effectiveness*. PhD thesis. University of New Hampshire. Available at: https://search-proquest-com.ezphost.dur.ac.uk/docview/1617959470?accountid=14533 (Accessed: 02 Jun 2019).

Gallagher, H.A. (2002). *The relationship between measures of teacher quality and student achievement: The case of Vaughn Elementary*. PhD thesis. The University of Wisconsin - Madison. Available at: https://search-proquest-com.ezproxy.nottingham.ac.uk/pqdt/docview/305517553/AA02A32DDCAF4430PQ/1?accountid=8018 (Accessed: 23 May 2019).

Garai, J.M. (2017). *A characterization of a value added model and a new multi-stage model for estimating teacher effects within small school systems*. PhD thesis. The University of Nebraska - Lincoln. Available at: https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1021&context=statisticsdiss (Accessed: 20 May 2019).

Garrett, R. & Steinberg, M.P. (2015). Examining teacher effectiveness using classroom observation scores: Evidence from the randomization of teachers to students. *Educational Evaluation and Policy Analysis*, 37, pp. 224-242.

Germuth, A.A. (2003). *Comparing results from value -added HLM and OLS models to assess teacher effectiveness*. PhD thesis. The University of North Carolina at Chapel Hill. Available at: https://search.proquest.com/docview/305312216/59D8B4387882453BPQ/1?accountid=14533 (Accessed: 01 Jun 2019).

Gibb, N. (2015). *The purpose of education*. Available at: https://www.gov.uk/government/speeches/the-purpose-of-education (Accessed: 15 Jun 2019).

Glass, G. (2004). *Teacher evaluation: Policy brief*. Tempe, Arizona: Education Policy Research Unit (EPRU).

Glazerman, S, Loeb, S., Goldhaber, D., Raudenbush, S. & Whitehurst, G. (2010). *Evaluating teachers: The important role of value-added*. Washington, DC: The Brookings Institution.

Goe, L. (2007). *The link between teacher quality and student outcomes: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.

Goel, D. & Barooah, B. (2018). *Drivers of student performance: Evidence from higher secondary public schools in Delhi*. St. Louis: Federal Reserve Bank of St Louis.

Goldhaber, D. (2007). Everyone's doing it, but what does teacher testing tell us about teacher effectiveness?. *Journal of Human Resources*, 42(4), pp.765-794.

Goldhaber, D. (2010). *When the Stakes Are High, Can We Rely on Value-Added? Exploring the Use of Value-Added Models to Inform Teacher Workforce Decisions*. Center for American Progress. Available at: https://americanprogress.org/wp-content/uploads/issues/2010/12/pdf/vam.pdf (Accessed: 11/02/2020).

Goldhaber, D. (2015). Exploring the potential of value-added performance measures to affect the quality of the teacher workforce. *Educational Researcher*, 44, pp. 87-95.

Goldhaber, D. & Brewer, D. J. (2000). Does teacher certification matter? High school teacher certification status and student achievement. *Educational Evaluation and Policy Analysis*, 22(2), pp.129-145.

Goldhaber, D., Goldschmidt, P. & Tseng, F. (2013). Teacher value-added at the high-school level: Different models, different answers?. *Educational Evaluation and Policy Analysis*, 35(2), pp. 220-236.

Goldhaber, D. & Hansen, M. (2010). *Assessing the potential of using value-added estimates of teacher job performance for making tenure decisions*. Washington, DC: National Center for Analysis of Longitudinal Data in Education Research.

Goldhaber, D. & Hansen, M. (2013). Is it just a bad class? Assessing the long-term stability of estimated teacher performance. *Economica*, 80, pp. 589-612.

Goldhaber, D., Perry, D. & Anthony, E. (2004). The National Board for Professional Teaching Standards (NBPTS) process: Who applies and what factors are associated with NBPTS certification?. *Educational Evaluation and Policy Analysis*, 26(4), pp.259-280.

Goldhaber, D. & Theobald, R. (2012). *Do different value-added models tell us the same things?*. Stanford, CA: Carnegie Knowledge Network.

Goldhaber, D., Walch, J. & Gabele, B. (2014). Does the model matter? Exploring the relationship between different student achievement-based teacher assessments. *Statistics and Public Policy*, 1(1), pp.28-39.

Goldschmidt, P., Roschewski, P., Choi K., Auty, W., Hebbler, S., Blank, R. & Williams, A. (2005). *Policymakers' guide to growth models for school accountability: How do accountability models differ*. A paper commissioned by the CCSSO Accountability Systems and Reporting State Collaborative on Assessment and Student Standards. Washington, DC: CCSSO.

Gorard, S. (2013). What difference do teachers make? A consideration of the wider outcomes of schooling. *Irish Educational Studies*, 32(1), pp. 69-82.

Gorard, S. (2014a). A proposal for judging the trustworthiness of research findings. *Radical Statistics*, 110, pp. 47-59.

Gorard, S. (2014b). Confidence intervals, missing data and imputation: a salutary illustration. *International Journal of Research in Educational Methodology*, 5(3), pp. 693-698.

Gorard, S. (2015). Rethinking 'quantitative' methods and the development of new researchers. *Review of Education*, 3(1), pp. 72-96.

Gorard, S. (2016). Damaging real lives through obstinacy: Re-emphasising why significance testing is wrong. *Sociological Research Online*, 21(1), 2.

Gorard, S. (2018a). *Education Policy: Evidence of equity and effectiveness*. Bristol: Policy Press

Gorard, S. (2018b). Significance testing with incompletely randomised cases cannot possibly work. *International Journal of Science and Research Methodology*, 11(2), pp.42-51.

Gorard, S. (2019). Do we really need confidence intervals in the new statistics?. *International Journal of Social Research Methodology*, 22(3), pp.281-291.

Gorard, S. (2020). Handling missing data in numeric analyses. *International Journal of Social Research Methodology*, 23(6), pp.651-660.

Gorard, S. & See, B.H. (2009). The early impact of SES on participation and attainment in science. *Studies in Science Education*, 45, pp.93-129.

Gorard, S. & Siddiqui, N. (2019). How trajectories of disadvantage help explain school attainment. *SAGE Open*, 9(1), pp.1-14.

Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.B., Poole, C., Goodman, S.N. & Altman, D.G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology,* 31(4), pp.337-350.

Guarino, C.M., Reckase, M., Stacy, B. & Wooldridge, J. (2015a). A comparison of student growth percentile and value-added models of teacher performance. *Statistics and Public Policy*, 2(1), pp.1-11.

Guarino, C.M., Maxfield, M., Reckase, M.D., Thompson, P.N. & Wooldridge, J.M. (2015b). An evaluation of Empirical Bayes's estimation of value-added teacher performance measures. *Journal of Educational and Behavioral Statistics*, 40(2), pp. 190-222.

Guarino, C.M., Reckase, M.D. & Wooldridge, J.M. (2015c). Can value-added measures of teacher performance be trusted?. *Education Finance and Policy*, 10, pp.117-156.

Haertel, E.H. (2013). *Reliability and validity of inferences about teachers based on student test scores.* Princeton, NJ: Educational Testing Service.

Hallgren, K.A. (2012.) Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, 8, pp. 23-34.

Hanushek, E.A. (2003). The failure of input-based schooling policies. *The Economic Journal*, 113(485), pp. F64-F98.

Hawk, P., Coble, C. R. & Swanson, M. (1985). Certification: It does matter. *Journal of Teacher Education*, 36(3), pp.13-15.

Heistad, D.J. (1999). *Stability and correlates of teacher effects in grade two reading achievement*. PhD thesis. University of Minnesota. Available at: https://search.proquest.com/docview/85512549/6DA4C4AB595E471BPQ/1?accountid=14533 (Accessed: 01 Jun 2019).

Hinds, D.P.G. (2018). *Damian Hinds: There are no great schools without great teachers.* Available at: https://www.gov.uk/government/speeches/damian-hinds-there-are-no-great-schools-without-great-teachers (Accessed 11 October 2020).

Hong, Y. (2010). *A comparison among major value -added models: A general model approach*. PhD thesis. The State University of New Jersey. Available at: https://search-proquest-com.ezproxy.nottingham.ac.uk/pqdt/docview/305229053/ACB929A4DC93471EPQ/1 ?accountid=8018 (Accessed: 19 May 2019).

Hu, J. (2015). *Teacher evaluation based on an aspect of classroom practice and on student achievement: A relational analysis between student learning objectives and value-added modeling*. PhD thesis. Boston College. Available at: https://dlib.bc.edu/islandora/object/bc-ir:104148 (Accessed: 01 Jun 2019).

Jacob, B.A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics*, 26(1), 101–136. Available: http://ideas.repec.org/a/ucp/jlabec/v26y2008p101-136.html

January, S.A., Van Norman, E.R., Christ, T.J., Ardoin, S.P., Eckert, T.L. & White, M.J. (2018). Progress Monitoring in reading: Comparison of weekly, bimonthly, and monthly assessments for students at risk for reading difficulties in grades 2–4. *School Psychology Review*, 47(1), pp. 83-94 doi: 10.17105/SPR-2017-0009.V47-1

Jenkins, J.R., Schiller E., Blackorby, J., Thayer, S.K. & Tilly, W.D. (2013). Responsiveness to intervention in reading: Architecture and practices. *Learning Disability Quarterly*, 36(1), pp. 36-46. doi:10.1177/0731948712464963

Jewell, J. (2017). From inspection, supervision, and observation to value-added evaluation: brief history of U.U. teacher performance evaluations. *Drake Law Review*, 65(2), 363-420.

Jia, Y., Cummings, T., Jackson, C., Clifford, M. & Hoch, S. (n.d.). *Analyzing and Improving Multiple Measure Teacher Evaluation Systems*. Strategic Data Project (SDP) Fellowship Capstone Reports.

Jiang, J.Y., Sporte, S.E. & Luppescu, S. (2015). Teacher perspectives on evaluation reform: Chicago's REACH Students. *Educational Researcher*, 44, pp.105-116.

Johnson, M.T., Lipscomb, S. & Gill, B. (2015). Sensitivity of teacher value-added estimates to student and peer control variables. *Journal of Research on Educational Effectiveness*, 8(1), pp. 60-83.

Kane, T.J. & Cantrell, S.M. (2010). *Learning about teaching: Initial findings from the Measures of Effective Teaching Project*. MET Project Research Paper. Seattle, WA: Bill & Melinda Gates Foundation.

Kane, T.J., Kerr, K.A. & Pianta, R.C. (2015). *Designing teacher evaluation systems: New guidance from the measures of effective teaching project*. San Francisco, CA: Jossey-Bass, A Wiley Brand.

Kane, T.J., McCaffrey, D.F., Miller, T. & Staiger, D.O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment*. MET project, Bill & Melinda Gates Foundation.

Kane, T.J. & Staiger, D.O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation*. Cambridge, MA: National Bureau of Economic Research.

Kane, T.J. & Staiger, D.O. (2012). *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains*. MET Project. Seattle, WA: Bill & Melinda Gates Foundation.

Kelly, A., & Downey, C. (2010). Value-added measures for schools in England: looking inside the 'black box'of complex metrics. *Educational Assessment, Evaluation and Accountability*, 22(3), pp.181-198.

Kersting, N.B., Chen, M.-k. & Stigler, J.W. (2013). Value-added teacher estimates as part of teacher evaluations: Exploring the effects of data and model specifications on the stability of teacher value-added scores. *Education Policy Analysis Archives*, 21(7).

King, S.A., Lemons, C.J. & Hill, D.R. (2012). Response to intervention in secondary schools. *NASSP bulletin*, 96(1), pp.5–22.

Ko, J., Sammons, P. & Bakkum, L. (2016). *Effective teaching. Reading, Berkshire: Education Development Trust*. Available at: https://www.educationdevelopmenttrust.com/EducationDevelopmentTrust/files/98/98ad6340-0ef6-4e1d-a541-db6018afce7d.pdf (Accessed: 12 February 2021).

Koedel, C. & Betts, J.R. (2011). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the rothstein critique. *Education Finance and Policy*, 6(1), pp. 18-42.

Koedel, C., Mihaly, K. & Rockoff, J.E. (2015). Value-added modeling: A review. *Economics of Education Review*, 47, 180-195.

Koretz, D. (2008). *Measuring Up: What Educational Testing Really Tells Us*. Cambridge: Harvard University Press.

Korkmaz, M. & Ozdogan, O. (2005). The level of accomplishments of primary school inspectors' guidance duties. *Turkish Journal of Educational Sciences*, 3(4), pp. 431-443.

Kowalsky, J.P.S. (1978). *Evaluating Teacher Performance*. Arlington: Educational Research Service.

Krueger, A.B. (2003). Economic considerations and class size. *The Economic Journal*, 113(485), pp.F34-F63.

Kukla-acevedo, S. (2009). Do teacher characteristics matter? New results on the effects of teacher preparation on student achievement. *Economics of Education Review*, 28, pp. 49-57.

Kupermintz, H. (2002). *Teacher effects as a measure of teacher effectiveness: Construct validity considerations in TVAAS (Tennessee Value-Added Assessment System)*. CSE Technical Report 563. University of California, Los Angeles.

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee Value Added Assessment System. *Educational Evaluation and Policy Analysis*, 25(3), pp. 287–298.

Kurtz, M.D. (2018). Value-added and student growth percentile models: What drives differences in estimated classroom effects?. *Statistics and Public Policy*, 5(1), pp.1-8.

Landis, J.R. & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, pp.159-174.

Lash, A., Makkonen, R., Tran, L. & Huang, M. (2016). *Analysis of the stability of teacherlevel growth scores from the student growth percentile model*. Washington, DC: National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory West.

Legere, E.J. & Conca, L.M. (2010). Response to intervention by a child with a severe reading disability: A case study. *Teaching Exceptional Children*, 43(1), pp. 32–39. https://doi.org/10.1177/004005991004300104.

Leigh, A. (2010). Estimating teacher effectiveness from two-year changes in students' test scores. *Economics of Education Review*, 29, pp.480-488.

Ligon, G.D. (2008). *The optimal reference guide: Comparison of growth and value-add models growth model series – Part II*. ESP Solutions Group.

Linn, R.L. (2001). *The design and evaluation of educational assessment and accountability systems*. CSE Technical Report 539. University of California, Los Angeles.

Little, O.M., Goe, L. & Bell, C.A. (2009). *A Practical Guide to Evaluating Teacher Effectiveness*. Washington, DC: National Comprehensive Center for Teacher Quality.

Loeb, S. & Page, M. (2000). Examining the Link between Teacher Wages and Student Outcomes: The Importance of Alternative Labor Market Opportunities and Non-pecuniary Variation. *Review of Economics and Statistics*, 82(3) pp.393–408.

Lockwood, J.R., McCaffrey, D.F., Hamilton, L.S., Stecher, B., Le, V., & Martinez, J.F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*, 44(1), pp.47–67.

Martínez, J.F., Schweig, J. & Goldschmidt, P. (2016). Approaches for combining multiple measures of teacher performance: Reliability, validity, and implications for evaluation policy. *Educational Evaluation and Policy Analysis*, 38(4), pp.738-756.

McCaffrey, D.F., Lockwood, J.R., Koretz, D.M. & Hamilton, L.S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: The RAND Corporation.

McCaffrey, D.F., Lockwood, J.R., Koretz, D.M., Louis, T.A. & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, 29(1), pp. 67-101.

McCaffrey, D.F., Sass, T.R., Lockwood, J.R. & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy*, 4(4), pp. 572-606.

McCardle, P., Scarborough, H.S., & Catts, H.W. (2001). Predicting, explaining, and preventing children's reading difficulties. *Learning Disabilities Research & Practice*, 16(4), 230–239. https://doi.org/10.1111/0938-8982.00023.

Mchugh, M.L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3), pp. 276-282.

Menard, S.W. (2002). *Applied logistic regression analysis (2nd ed.)*. Thousand Oaks, CA: Sage Publications.

Milanowski, A., Kimball, S., & White, B. (2004). *The relationship between standards-based teacher evaluation scores and student achievement: Replication and extensions at three sites*. Madison: Consortium for Policy Research in Education, University of Wisconsin. Available: http://cpre.wceruw.org/papers/3site_long_TE_SA_AERA04TE.pdf

Ministry of National Education (MoNE) (2015). *Ministry of National Education regulation on teacher appointment and replacement*. Available at: https://www.resmigazete.gov.tr/eskiler/2015/04/20150417-4.htm (Accessed: 10 November 2018).

Ministry of National Education (MoNE) (2017a). *National education statistics: Formal education.* Ankara: Official statistics programme. Available at: http://sgb.meb.gov.tr/www/milli-egitim-istatistikleri-orgun-egitim-20162017/icerik/270 (Accessed: 11 November 2017).

Ministry of National Education (MoNE) (2017b). *Schools and other institutions.* Available at: https://mebbis.meb.gov.tr/KurumListesi.aspx (Accessed: 11 November 2017).

Ministry of National Education (MoNE) (2017c). *Service Areas and Service Scores Chart.* Communiques Journal No. 2713. Available at: http://tebligler.meb.gov.tr/index.php/tuem-sayilar/viewcategory/85-2017 (Accessed: 11 November 2017).

Ministry of National Education (MoNE) (2018). *Teaching Fields, Appointment and Course Teaching Principles*. Communiques Journal No. 2735. Available at: http://ttkb.meb.gov.tr/www/ogretmenlik-alanlari-atama-ve-ders-okutma-esaslari/icerik/201 (Accessed: 20 December 2018).

Moher, D., Liberati, A., Tetzlaff, J, & Altman D.G. (2009). *Preferred reporting items for systematic reviews and meta-analyses: The PRISMA Statement*. BMJ.339: b2535.

Moorman, R.H. & Podsakoff, P.M. (1992). A meta-analytic review and empirical test of the potential confounding effects of social desirability response sets in organizational behaviour research. *Journal of Occupational and Organizational Psychology*, 65(2). pp.131-149.

Morris, T.T., Davies, N.M., Dorling, D., Richmond, R.C. & Smith, G.D. (2018). Testing the validity of value-added measures of educational progress with genetic data. *British Educational Research Journal*, 44, pp.725-747.

Muijs, D. (2006). Measuring teacher effectiveness: Some methodological reflections. *Educational Research and Evaluation*, 12(1), pp.53-74.

Muñoz, M.A. & Chang, F.C. (2007). The elusive relationship between teacher characteristics and student academic growth: A longitudinal multilevel model for change. *Journal of Personnel Evaluation in Education*, 20, pp.147-164.

Muñoz, M.A., Prather, J.R. & Stronge, J.H. (2011). Exploring teacher effectiveness using hierarchical linear models: Student-and classroom-level predictors and cross-year stability in elementary school reading. *Planning and Changing*, 42(3/4), pp.241-273.

Murphy, D. (2012). *Where is the Value in Value-Added Modeling?*. White Paper. s.l.: Pearson Education.

National Board for Professional Teaching Standards (NBPTS) (1987). *National board certification.* Available at: https://www.nbpts.org/national-board-certification/ (Accessed: 09/02/2020).

National Center for Education Statistics (2012). *Growth models: Issues and advice from the states - A guide of the statewide longitudinal data systems grant program*. Available at: https://files.eric.ed.gov/fulltext/ED551302.pdf (Accessed: 20 May 2019).

National Center on Response to Intervention (2010). *What is response to intervention (RTI).* Washington, DC: U.S. Department of Education, Office of Special Education Programs, National Center on Response to Intervention.

National Education Union (NEU) (2020). *Guidance on classroom observation protocol in England.* Available at: https://neu.org.uk/advice/guidance-classroom-observation-protocol-england (Accessed: 10 December 2020).

Newton, X., Darling-Hammond, L., Haertel, E. & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Educational Policy Analysis Archives*, 18(23), pp. 1-27.

Nye, B., Achilles, C., Boyd-Zaharias, J., Fulton, B. & Wallenhorst, M. (1994). Small Is Far Better. *Research in the Schools*, 1, p. 9-20.

Nye, B., Konstantopoulos, S. & Hedges, L.V. (2004). How large are teacher effects?. *Educational Evaluation and Policy Analysis*, 26(3), pp. 237-257.

O'Malley, K.J., Murphy, S., McClarty, K.L., Murphy, D., & McBride, Y. (2011). *Overview of student growth models* (Pearson White Paper). Available at: https://images.pearsonassessments.com/images/tmrs/Student_Growth_WP_083111_FINAL.pdf (Accessed: 29 March 2018).

Opper, I. M. (2019). *Teachers matter: Understanding teachers' impact on student achievement*. Santa Monica, CA: RAND Corporation. https://www.rand.org/pubs/research_reports/RR4312.html.

Organisation for Economic Co-operation and Development (OECD) (2009). *Creating effective teaching and learning environments: First results from TALIS*. Paris: OECD. Available at: https://www.oecd.org/education/school/43023606.pdf (Accessed: 20 April 2018).

Organisation for Economic Co-operation and Development (OECD) (2012). *Does Money Buy Strong Performance in PISA?*, PISA in Focus, No. 13, OECD Publishing, Paris, https://doi.org/10.1787/5k9fhmfzc4xx-en.

Organisation for Economic Co-operation and Development (OECD) (2016). *PISA 2015 results (volume I): Excellence and equity in education.* OECD Publishing. Available at: https://www.oecd.org/publications/pisa-2015-results-volume-i-9789264266490-en.htm (Accessed: 11 February 2019).

Osbourne, J.W. & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research & Evaluation*, 8(2), pp. 1-5.

Ouma, C.A. (2014). *Performance of cart-based value-added model aganist HLM, multiple regressin, and student growth percentile value-added models*. PhD thesis. Florida State University. Available at: http://diginole.lib.fsu.edu/islandora/object/fsu%3A252869 (Accessed: 18 June 2019).

Our Education System (2019). Available at: https://i.pinimg.com/originals/19/59/85/195985660924652b3c007a764e78ce81.jpg (Accessed: 11 February 2019).

Ovenden-Hope, T. & Passy, R. (2020). *Exploring teacher recruitment and retention: Contextual challenges from international perspectives*. Oxon: Routledge.

Ovenden-Hope, T., Blandford, S., Cain, T. & Maxwell, B. (2018). RETAIN early career teacher retention programme: Evaluating the role of research informed continuing professional development for a high quality, sustainable 21st century teaching profession. *Journal of Education for Teaching*, 44(5), pp. 590-607.

Pallant, J. (2001). SPSS survival manual: A step-by-step guid to data analysing using SPSS. Buckingham: Open University Press.

Papay, J. P. (2011). Different tests, different answers. *American Educational Research Journal*, 48(1), pp.163-193.

Parsons, E., Koedel, C. & Tan, L. (2019). Accounting for student disadvantage in value-added models. *Journal of Educational and Behavioral Statistics*, 44(2), pp.144-179.

Paufler, N.A., & Amrein-Beardsley, A. (2014). The random assignment of students into elementary classrooms: Implications for value-added analyses and interpretations. *American Educational Research Journal (AERJ)*, 51(2), 328-362. doi: 10.3102/0002831213508299

Perry, T. (2016a). English value-added measures: Examining the limitations of school performance measurement. *British Educational Research Journal*, 42(6), p. 1056–1080.

Perry, T. (2016b). *The validity, interpretation and use of school value-added measures*. PhD thesis. University of Birmingham. Available at: https://etheses.bham.ac.uk/id/eprint/6773/1/Perry16PhD.pdf (Accessed: 11 April 2018).

Petticrew, M. & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Oxford: Blackwell.

Potamites, L., Booker, K., Chaplin, D. & Isenberg, E. (2009). *Measuring school and teacher effectiveness in the EPIC charter school consortium-Year 2*. Washington, DC: USA: Mathematica Policy Research.

Raudenbush, S.W. & Bryk, A.S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage Publication.

Rice, J.K. (2003). *Teacher quality: Understanding the effectiveness of teacher attributes*. Washington DC: Economic Policy Institute.

Rivkin, S.G., Hanushek, E.A. & Kain, J.F. (2005). Teachers, schools, and academic achievement. *Econometrica*, 73(2), pp. 417-458.

Robertson-Kraft, C. (2014). *Teachers' Motivational Responses to New Teacher Performance Management Systems: An Evaluation of the Pilot of Aldine ISD's inVEST System*. PhD thesis. the University of Pennsylvania. Available at: https://repository.upenn.edu/cgi/viewcontent.cgi?article=3232&context=edissertations (Accessed: 26 May 2020).

RobinB Creative (2018). *Equality, Equity, & Freedom.* Available at: https://artplusmarketing.com/equality-equity-freedom-55a1d675b5d8 (Accessed: 11 February 2019).

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *The American Economic Review*, 94(2), pp.247–252. (Accessed: 11/02/2020).

Rothstein, J. (2007). *Do value-added models add value? Tracking, fixed effects, and causal inference*. Working Papers 1036. Princeton, NJ: Center for Economic Policy Studies, Princeton University.

Rothstein, J. (2009). Student sorting and bias in value added estimation: Selection on observables and unobservables. *Education Finance and Policy*, 4(4), pp. 537-571.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics*, 125(1), pp. 175–214.

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), p. 103–116.

Sammons, P., Kington, A., Lindorff-Vijayendran, A. & Ortega, L. (2014). *Inspiring teaching: What can we learn from exemplary practitioners.* EARLI SIG 18 conference. August 28, Southampton. Available at: http://eprints.worc.ac.uk/5388/1/Inspiring%20Teaching%20What%20we%20can%20learn%20from%20exemplary%20practitioners.pdf (Accessed: 20 December 2020).

Samsun Provincial Directorate of National Education (2014). *Step by Step Achievement project*. Available at: https://samsun.meb.gov.tr/meb_iys_dosyalar/2015_01/15053803_adimadimbaariprojesdzeltlen2.pdf (Accessed: 10 November 2017).

Samsun Provincial Directorate of National Education (2017). *The list of institutions*. Available at: https://mebbis.meb.gov.tr/KurumListesi.aspx (Accessed: 11 November 2017).

Sanders, W.L. & Horn, S.P. (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), pp. 247-256.

Sanders, W.L. & Horn, S.P. (1994). The Tennessee Value-Added Assessment System (TVAAS): Mixed model methodology in educational assessment. *Journal of Personnel Education in Education*, 8, pp. 299-311.

Sanders, W.L. & Rivers, J.C. (1996). *Cumulative and residual effects of teachers on future student academic achievement*. Knoxville: University of Tennessee Value-Added Research and Assessment Center.

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. Millman (Ed.), *Grading teachers, grading schools. Is student achievement a valid evaluation measures?* (pp. 137-162). Thousand Oaks, CA: Corwin Press.

Sass, T.R., Semykina, A. & Harris, D.N. (2014). Value-added models and the measurement of teacher productivity. *Economics of Education Review*, 38, pp.9-23.

Schafer, W.D., Lissitz, R.W., Zhu, X., Zhang, Y., Hou, X. & Li, Y. (2012). Evaluating teachers and schools using student growth models. *The Practical Assessment, Research & Evaluation*, 17(17).

Scheerens, J. (1992). *Effective schooling: Research, theory and practice*. London: Cassell.

Schmitz, D.D. (2007). *An empirical sensitivity analysis of value-added teachers' effect estimates to hierarchical linear model parameterizations*. PhD thesis. University of Northern Colorado. Available at: https://search.proquest.com/openview/59971048402fd49ad9471669c1868871/1?pq-origsite=gscholar&cbl=18750&diss=y (Accessed: 10 June 2018).

Schweig, J. (2019). *Measuring teacher effectiveness: Understanding common, uncommon, and combined methods*. Santa Monica, CA: RAND Corporation.

See, B.H. (2020). Challenges in using research evidence in improving teaching quality. *BERA Research Intelligence,* 144, pp. 20-21.

See, B.H., Gorard, S., Morris, R. & el-Soufi, N. (2020). How to recruit and retain teachers in hard-to-staff areas: A systematic review of the empirical evidence. In: Ovenden-Hope, T. & Passy, R. eds. *Exploring teacher recruitment and retention: Contextual challenges from international perspectives*. Oxon: Routledge, pp. 148–162.

Shaw, L.H. (2012). *Incorporating latent variable outcomes in value-added assessment: An evaluation of univariate and multivariate measurement model structures*. PhD thesis. 74. The University of Nebraska. Available at: http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2013-99190-399&site=ehost-live (Accessed: 20 May 2019).

Siraj-Blatchford, I., Shepherd, D., Melhuish, E., Taggart, B., Sammons, P. & Sylva, K (2011). *Effective pedagogical strategies in English and mathematics in Key Stage 2: A study of Year 5 classroom practice from the EPPSE 3-16 longitudinal study*. DfE Research Report DFE-RR129. London: DfE.

Siraj-Blatchford, I. & Taggart, B. (2014). *Exploring effective pedagogy in primary schools: Evidence from research*. London: Pearson.

Slater, H., Davies, N.M. & Burgess, S. (2012). Do teachers matter? Measuring the variation in teacher effectiveness in England. *Oxford Bulletin of Economics and Statistics*, 74(5), pp.629-645.

Sloat, E., Amrein-Beardsley, A. & Holloway, J. (2018). Different teacher-level effectiveness estimates, different results: Inter-model concordance across six generalized value-added models (VAMs). *Educational Assessment, Evaluation and Accountability*, 30, pp.367-397.

Stacy, B., Guarino, C. & Wooldridge, J. (2018). Does the precision and stability of value-added estimates of teacher performance depend on the types of students they serve?. *Economics of Education Review*, 64, pp. 50-74.

Stecher, B.M., Holtzman, D.J., Garet, M.S., Hamilton, L.S., Engberg, J., Steiner, E.D., Robyn, A., Baird, M.D., Gutierrez, I.A., Peet, E.D., de los Reyes, I.B., Fronberg, K., Weinberger, G., Hunter, G.P., & Chambers, J. (2018). *Improving teaching effectiveness implementation final report: The Intensive Partnerships for Effective Teaching through 2015–2016*. Santa Monica, CA: RAND. Available at: www.rand.org/pubs/research_reports/RR2242.html (Accessed: 18 December 2020).

Steinberg, M.P., & Garrett, R. (2016). Classroom composition and measured teacher performance: What do teacher observation scores really measure?. *Educational Evaluation and Policy Analysis*, 38(2), 293–317.

Stock, J.H. & Watson, M.W. (2012). *Introduction to econometrics* (3rd ed.). Harlow: Pearson Education.

Stronge, J. H. (2006). *Evaluating teaching: A guide to current thinking and best practice* (2nd ed.). California: Corwin Press.

Stronge, J.H. (20018). *Qualities of effective teachers* (3rd ed.), Alexandria, VA.: Association for Supervision and Curriculum Development (ASCD).

Swanson, P.L. (2009). *A quantitative study of school characteristics that impact student achievement on state assessments and those assessments' associations to ACT scores in Tennessee.* EdD thesis. East Tennessee State University. Available at: https://dc.etsu.edu/cgi/viewcontent.cgi?article=3189&context=etd (Accessed: 22 May 2019).

Tichá, R., Espin, C.A. & Wayman, M.M. (2009). Reading progress monitoring for secondary-school students: Reliability, validity, and sensitivity to growth of reading-aloud and maze-selection measures. *Learning Disabilities Research and Practice,* 24(3), pp.132–142.

Tobe, P.F. (2008). *An investigation of the differential impact of teacher characteristics and attitudes on student mathematics achievement using a value-added approach*. EdD thesis. University of Houston. Available at: https://search.proquest.com/docview/304625511?pq-origsite=gscholar&fromopenview=true (Accessed: 02 Jun 2019).

Torgerson, C. (2003). *Systematic reviews*. London: Continuum.

Turkish Education Union (2017). *We decided to take action against freak performance evaluation system*. Available at: https://www.turkegitimsen.org.tr/lib_yayin/468.pdf (Accessed: 05 May 2018).

UNESCO (2016). Education 2030: *Incheon Declaration and Framework for Action for the implementation of Sustainable Development Goal 4: Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all* https://unesdoc.unesco.org/ark:/48223/pf0000245656.

UNESCO Institute for Statistics (UIS) (2016). *The world needs almost 69 million new teachers to reach the 2030 Education Goals* [Fact Sheet] http://uis.unesco.org/sites/default/files/documents/fs39-the-world-needs-almost-69-million-new-teachers-to-reach-the-2030-education-goals-2016-en.pdf.

U.S. Dept. of Education (2002). *No Child Left Behind Act*. Washington, DC: US Department of Education. Available at: https://www.govinfo.gov/content/pkg/PLAW-107publ110/pdf/PLAW-107publ110.pdf (Accessed: 15 November 2018).

U.S. Dept. of Education (2009). *Race to the Top Program: Executive Summary*. Available at: https://www2.ed.gov/programs/racetothetop/executive-summary.pdf (Accessed: 17 November 2018).

Walker, R.J. (2008). Twelve characteristics of an effective teacher: A longitudinal, qualitative, quasi-research study of in-service and pre-service teachers' opinions. *Educational Horizons,* 87(1), pp. 61-68.

Wayne, A.J. & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research*, 73, pp. 89-122.

Wei, H., Hembry, T., Murphy, D.L. & McBride, Y. (2012). *Value-added models in the evaluation of teacher effectiveness: A comparison of models and outcomes*. New York, NY: Pearson.

White, P. & Gorard, S. (2017). Against Inferential Statistics: How and why current statistics teaching gets it wrong. *Statistics Education Research Journal*, 16(1), 55-65.

Williams, M.N., Grajales, C.A. G. & Kurkiewicz, D. (2013). Assumptions of multiple regression: Correcting two misconceptions. *Practical Assessment, Research & Evaluation*, 18(1), pp.1-14.

Worrell, F.C. & Kuterbach, L.D. (2001). The use of student rating of teacher behaviors with academically talented high school students. *Journal of Secondary Gifted Education*, 14(4), pp.236-247.

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*, 11, pp. 57-67.

Yeh, S.S. (2012). The reliability, impact, and cost-effectiveness of value-added teacher assessment methods. *Journal of Education Finance,* 37(4), pp. 374-399.

APPENDICES

Appendix A. Search Strings Used in the Providers, and Search Results

	Providers	Databases	Search Strings	Initial Result Total	Duplicated Cases Found by Software	Duplicated Results found by the researcher	The Final Result Total
1	ProQuest	ProQuest Dissertations & Theses Global: Social Sciences Education Database ERIC International Bibliography of the Social Sciences (IBSS) Social Science Database Applied Social Sciences Index & Abstracts (ASSIA)	ab(((teacher OR educator) AND (effect* OR evaluat* OR quality OR perform* OR appraisal OR assess* OR accountability)) OR ((teacher OR educator) AND performance AND (evaluation OR appraisal OR assessment)) OR "teacher proficiency-rank" OR "teacher judg*" OR "educational effectiveness" OR "teaching effect*" OR "measuring teach*" OR "evaluating teach*") AND ab((academic AND (achievement OR gain*)) OR ((student AND test) AND (score OR performance)) OR achievement OR outcome* OR ((achievement OR outcome*) AND measur*))AND ab(VAM* OR (Value-added AND (model* OR estimat*)) OR ((value AND added) AND (model* OR estimat*)) OR (teacher AND value-added) OR (teacher AND value AND added)) AND ab(stabil* OR concord* OR robust OR sensitiv* OR instabil* OR precis* OR imprecise* OR variat* OR fluctuat* OR persistence OR shrink*)	276	83	48	145

2	EBSCOhost	OpenDissertations British Education Index Business Source Premier Education Abstracts Educational Administration abstracts PsycINFO	(((teacher or educator) and (effect* or evaluat* or quality or perform* or appraisal or assess* or accountability)) or ((teacher or educator) and performance and (evaluation or appraisal or assessment)) or (teacher and proficiency-rank) or (teacher and judg*) or (educational and effect*) or (teaching and effect*) or (measuring and teach*) or (evaluating and teach*)) AND ((academic and (achievement or gain*)) or (student and test and (*score/ or performance)) or achievement or outcome* or ((achievement or outcome*) and measur*)) AND (VAM* or (Value-added and (model* or estimat*)) or (value and added and (model* or estimat*)) or (teacher and value-added) or (teacher and value and added)) AND (stabil* or concord* or robust or sensitiv* or instabil* or precis* or imprecise* or variat* or fluctuat* or persistence or shrink*)	226	126	46	54

3	Web of Science	Web of Science Core Collection Current Contents Connect	# 1 # 2 # 3 # 4 # 5	(TS= (((teacher OR educator) AND (effect* OR evaluat* OR quality OR perform* OR appraisal OR assess* OR accountability)) OR ((teacher OR educator) AND performance AND (evaluation OR appraisal OR assessment)) OR ((teacher AND proficiency-rank) OR (teacher AND judg*) OR (educational AND effect*) OR (teaching AND effect*) OR (measuring AND teach*) OR (evaluating AND teach*)))) AND LANGUAGE: (English) (TS= (((academic AND (achievement OR gain*)) OR ((student AND test) AND (*score OR performance)) OR achievement OR outcome* OR ((achievement OR outcome*) AND measur*)))) AND LANGUAGE: (English) (TS= (VAM* OR (Value-added AND (model* OR estimat*)) OR ((value AND added) AND (model* OR estimat*)) OR (teacher AND value-added) OR (teacher AND value AND added))) AND LANGUAGE: (English) (TS= (stabil* OR concord* OR robust OR sensitiv* OR instabil* OR precis* OR imprecise* OR variat* OR fluctuat* OR persistence OR Shrink*)) AND LANGUAGE:(English) #4 AND #3 AND #2 AND #1	219	131	10	78

4	Elsevier	SCOPUS	TITLE-ABS-KEY ((((teacher OR educator) AND (effect* OR evaluat* OR quality OR perform* OR appraisal OR assess* OR accountability)) OR ((teacher OR educator) AND performance AND (evaluation OR appraisal OR assessment)) OR (teacher AND proficiency-rank) OR (teacher AND judg*) OR (educational AND effect*) OR (teaching AND effect*) OR (measuring AND teach*) OR (evaluating AND teach*)) AND ((academic AND (achievement OR gain*)) OR (student AND test AND (*score/ OR performance)) OR achievement OR outcome* OR ((achievement OR outcome*) AND measur*)) AND (vam* OR (value-added AND (model* OR estimat*)) OR (value AND added AND (model* OR estimat*)) OR (teacher AND value-added) OR (teacher AND value AND added)) AND (stabil* OR concord* OR robust OR sensitiv* OR instabil* OR precis* OR imprecise* OR variat* OR fluctuat* OR persistence OR shrink*))	95	69	6	20

5	SAGE Research Methods	SAGE	for [[[[Abstract teacher] OR [Abstract educator]] AND [[Abstract effect*] OR [Abstract evaluat*] OR [Abstract quality] OR [Abstract perform*] OR [Abstract appraisal] OR [Abstract assess*] OR [Abstract accountability]]] OR [[[Abstract teacher] OR [Abstract educator]] AND [Abstract performance] AND [[Abstract evaluation] OR [Abstract appraisal] OR [Abstract assessment]]] OR [[Abstract teacher] AND [Abstract proficiency-rank]] OR [[Abstract teacher] AND [Abstract judg*]] OR [[Abstract educational] AND [Abstract effect*]] OR [[Abstract teaching] AND [Abstract effect*]] OR [[Abstract measuring] AND [Abstract teach*]] OR [[Abstract evaluating] AND [Abstract teach*]]] AND [[[Abstract academic] AND [[Abstract achievement] OR [Abstract gain*]]] OR [[Abstract student] AND [Abstract test] AND [[Abstract *score] OR [Abstract performance]]] OR [Abstract achievement] OR [Abstract outcome*] OR [[[Abstract achievement] OR [Abstract outcome*]] AND [Abstract measur*]]] AND [[Abstract vam*] OR [[Abstract value-added] AND [[Abstract model*] OR [Abstract estimat*]]] OR [[Abstract value] AND [Abstract added] AND [[Abstract model*] OR [Abstract estimat*]]] OR [[Abstract teacher] AND [Abstract value-added]] OR [[Abstract teacher] AND [Abstract value] AND [Abstract added]]] AND [[Abstract stabil*] OR [Abstract concord*] OR [Abstract robust] OR [Abstract sensitiv*] OR [Abstract instabil*] OR [Abstract precis*] OR [Abstract imprecise*] OR [Abstract variat*] OR [Abstract fluctuat*] OR [Abstract persistence] OR [Abstract shrink*]]	99	18	0	81

6	Taylor & Francis Online	Educational Research Abstracts Online	[[[[All: teacher] OR [All: educator]] AND [[All: effect*] OR [All: evaluat*] OR [All: quality] OR [All: perform*] OR [All: appraisal] OR [All: assess*] OR [All: accountability]]] OR [[[All: teacher] OR [All: educator]] AND [All: performance] AND [[All: evaluation] OR [All: appraisal] OR [All: assessment]]] OR [[All: teacher] AND [All: proficiency-rank]] OR [[All: teacher] AND [All: judg*]] OR [[All: educational] AND [All: effect*]] OR [[All: teaching] AND [All: effect*]] OR [[All: measuring] AND [All: teach*]] OR [[All: evaluating] AND [All: teach*]]] AND [[[All: academic] AND [[All: achievement] OR [All: gain*]]] OR [[All: student] AND [All: test] AND [[All: *score] OR [All: performance]]] OR [All: achievement] OR [All: outcome*] OR [[[All: achievement] OR [All: outcome*]] AND [All: measur*]]] AND [[All: vam*] OR [[All: value-added] AND [[All: model*] OR [All: estimat*]]] OR [[All: value] AND [All: added] AND [[All: model*] OR [All: estimat*]]] OR [[All: teacher] AND [All: value-added]] OR [[All: teacher] AND [All: value] AND [All: added]]] AND [[All: stabil*] OR [All: concord*] OR [All: robust] OR [All: sensitiv*] OR [All: instabil*] OR [All: precis*] OR [All: imprecise*] OR [All: variat*] OR [All: fluctuat*] OR [All: persistence] OR [All: shrink*]] AND [DatabaseType: Educational Research Abstracts Online]	188	41	0	147

7	Google Scholar	teacher effectiveness estimated by value added model stability OR concordance OR robust OR sensitivity OR unstable OR precise OR imprecise OR variation OR fluctuation OR persistence OR Shrinkage	260	22	38	200
8	Personal Contacts		71	2	25	45
9	Hand Search		5			5
TOTAL			1439	492	175	**772**

Appendix B.

Phase I

Title-Abstract Screening Checklist

The 5 checklist questions for Phase I Title-Abstract Screening are used to review the reports as efficiently and systematically as possible. If one of the questions is answered with NO, then the study will be automatically rejected from Phase II screening procedure.

Study #: **Date of review:**

1. Is the study fully available in English?

YES [ ] NO [ ] **(STOP and EXCLUDE)** NOT SURE, YET [ ]

2. Is the study written in education field?

YES [ ] NO [ ] **(STOP and EXCLUDE)** NOT SURE, YET [ ]

3. Does the study take place in K-12 school setting?

YES [ ] NO [ ] **(STOP and EXCLUDE)** NOT SURE, YET [ ]

4. Is this a primary research?

YES [ ] NO [ ] **(STOP and EXCLUDE)** NOT SURE, YET [ ]

5. Is at least one of the populations related to teachers?

YES [ ] NO [ ] **(STOP and EXCLUDE)** NOT SURE, YET [ ]

Decision: ___Exclude ___Include to Phase II screening

Comments:

Appendix C.

Phase II

Full-Text Screening Checklist

The **10** checklist questions for Phase II Full-Text Screening are used to retrieval the previous studies as efficiently and systematically as possible. If one of the questions is answered with NO, then the study will be excluded from this systematic review study.

Study #: **Date of review:**

6. Reported in English YES [ ] NO [ ] **(STOP and EXCLUDE)**
7. Written in education field YES [ ] NO [ ] **(STOP and EXCLUDE)**
8. Take place in K-12 school setting YES [ ] NO [ ] **(STOP and EXCLUDE)**
9. Primary research YES [ ] NO [ ] **(STOP and EXCLUDE)**
10. Focus on teacher evaluation YES [ ] NO [ ] **(STOP and EXCLUDE)**

11. Whether the study focus on stability, operationally defined, of teacher VAM estimates explained clearly in;

a. Research question(s) YES [ ]

b. Aim(s) YES [ ]

c. Findings YES [ ]

d. Implementation/Result YES [ ]

e. OTHERWISE NO [ ] **(STOP and EXCLUDE)**

12. Dependent variable of study;

a. Student test score(s) YES [ ]

b. Students gains make YES [ ]

c. Teacher VAM score NO [ ] **(STOP and EXCLUDE)**

d. OTHERS NO [ ] **(STOP and EXCLUDE)**

13. Focus on stability of estimates based upon the observable characteristics used as a;

a. Individual such as gender, experience etc. YES [ ]

b. Block such as student-, teacher-level NO [ ] **(STOP and EXCLUDE)**

14. Focus on

a. the contribution of predictors to estimates YES [ ]

b. the relationship between outcome and predictors NO [ ] **(STOP and EXCLUDE)**

15. Focus on stability of

a. VAM estimates due to the number of test scores used YES [ ]

b. Teacher effectiveness estimated overtime NO [ ] **(STOP and EXCLUDE)**

Decision: ___Exclude ___Include to Quality Appraisal Phase

Comments:

Appendix D. Data Extraction

Appendix E. Rating of Studies

#No	Study	Design	Scale	Completeness of data	Data quality	Rating
1	Parsons et al. (2019) Accounting for Student Disadvantage in Value-Added Models	Longitudinal comparison study random allocation	Large number of cases per comparison group (Same amount of teacher in each model, n= 600)	No attrition (Simulated data)	Standardised test (the Missouri state-wide exam)	4🔒
		4🔒	4🔒	4🔒	4🔒	
2	Lash et al. (2016) Analysis of the Stability of Teacher-Level Growth Scores from the Student Growth Percentile Model	Longitudinal study non-random allocation	Large number of cases per comparison group (390 Maths, 404 Reading teachers)	Moderate missing data (40% and 46% lost or attrition cases in math and reading, respectively)	Standardised test	3🔒
		3🔒	3🔒	3🔒	3🔒	
3	Goldhaber and Hansen (2010) Assessing the Potential of Using Value-Added Estimates of Teacher Job Performance for Making Tenure Decisions	Longitudinal study random allocation	Large number of cases per comparison group (11,854 teachers in grade 4, and 7,732 teachers in grade 5)	Huge amount of missing data (The number of incompleteness (lost or attrition) of data is 18.977 (97%) - (overall the number of teachers cases used in estimations is at least 603)	Standardised test	1🔒
		3🔒	3🔒	1🔒	1🔒	
4	Guarino et al. (2015) Can Value-Added Measures of Teacher Performance Be Trusted?	Longitudinal comparison study random allocation	Large number of cases per comparison group (Same in all simulation scenarios, n= 120 teachers)	No attrition (Simulated data)	Standardised test	4🔒
		4🔒	4🔒	4🔒	4🔒	
5	Garai (2017) A Characterization of A Value Added Model And New Multi-Stage Model For Estimating Teacher Effects Within Small School Systems	Longitudinal comparison study non-random allocation	Medium number of cases per comparison group (same amount of students records used in each model, n=1,350)	No attrition (Simulated data)	Standardised test	3🔒
		3🔒	3🔒	3🔒	3🔒	

6	Germuth (2003) Comparing Results from Value-Added HLM and OLS Models to Assess Teacher Effectiveness	Longitudinal comparison study non-random allocation	Large number of cases per comparison group (258 teachers in the estimations of each model)	High level of missing data (The percentage of incompleteness (lost or attrition) of data is 66%)	Standardised test	2🔒
		3🔒	3🔒	2🔒	2🔒	
7	Hong (2010) A Comparison among Major Value-Added Models: A General Model Approach	Longitudinal comparison study random allocation	Medium number of cases per comparison group (Same amount of cases in each model, n=1,200 student)	No attrition (Simulated data)	Standardised test (State-wide)	4🔒
		4🔒	4🔒	4🔒	4🔒	
8	Dwyer (2016) A Comparison of Educational "Value-Added" Methodologies for Classifying Teacher Effectiveness: Value Tables vs. Covariate Regression	Longitudinal comparison study non-random allocation	Large number of cases per comparison group (At least 11,215 cases in each group)	Minimal missing data (The percentage of incompleteness (lost or attrition) of data is 19%)	Standardised test (FCAT mathematics scores)	3🔒
		3🔒	3🔒	3🔒	3🔒	
9	Guarino et al. (2015a) A Comparison of Student Growth Percentile and Value-Added Models of Teacher Performance	Longitudinal comparison study random allocation	Large number of cases per comparison group (110,970 students' records in grade 5 and 104,441 records in grade 6)	Some missing data (The percentage of incompleteness (lost or attrition) of data around 22 in fifth grade and 20 in sixth grade)	Standardised test	4🔒
		4🔒	4🔒	4🔒	4🔒	
10	Ballou et al. (2004) Controlling for Student Background in Value-Added Assessment of Teachers	Longitudinal study non-random allocation	Large number of cases per comparison group (120,861 students records in reading, 120,646 in language arts and 120,721 in maths)	Some missing data (The incidence of missing FRL values ranged from 8.5% in 1997 to 14.2% in 1995)	Standardised test (State-wide)	3🔒
		3🔒	3🔒	3🔒	3🔒	

11	Sloat et al. (2018) Different teacher-level effectiveness estimates, different results: inter-model concordance across six generalized value-added models (VAMs)	Longitudinal comparison study non-random allocation 3🔒	Medium number of cases per comparison group (71 teachers in grade 4, 75 in grade 5 and 69 in grade 6) 3🔒	Some missing data (The percentage of incompleteness (lost or attrition) of data is 20) 3🔒	Standardised test (State-wide) 3🔒	3 🔒
12	Kukla-Acevedo (2009) Do teacher characteristics matter? New results on the effects of teacher preparation on student achievement	Longitudinal study non-random allocation 3🔒	Small Number of Cases per comparison group (754 students records used in the estimates of African-American teacher effectiveness, on the other hand 1,522 cases used in European-American teacher effectiveness estimates) 3🔒	Some missing data (The percentage of incompleteness (lost or attrition) of data is 35) 3🔒	Standardised test 3🔒	3🔒
13	Slater et al. (2012) Do teachers matter? Measuring the variation in teacher effectiveness in England	Longitudinal study non-random allocation 3🔒	Large number of cases per comparison group (7,204 records in English, 7,225 records in Maths and 7,095 records in Science) 3🔒	Some missing data (The percentage of incompleteness (lost or attrition) of data is 24) 3🔒	Standardised test 3🔒	3🔒
14	Koedel and Betts (2011) Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique	Longitudinal study non-random allocation 3🔒	Large number of cases per comparison group (with in school model, 595 grade 4 teachers, and 471 grade 5 teachers' records used) 3🔒	Moderate missing data (The percentage of incompleteness (lost or attrition) of data is 49) 3🔒	Standardised test (the Stanford 9 mathematics test) 3🔒	3🔒

15	Goldhaber et al. (2014) Does the Model Matter? Exploring the Relationship Between Different Student Achievement-Based Teacher Assessments	Longitudinal comparison study non-random allocation	Large number of cases per comparison group (7,672 “advantaged” classrooms, 3,820 “average” classrooms, and 8,002 “disadvantaged” classrooms)	Moderate missing data (The percentage of incompleteness (lost or attrition) of data is 54)	Standardised test	3🔒
		3🔒	3🔒	3🔒	3🔒	
16	Stacy et al. (2018) Does the precision and stability of value-added estimates of teacher performance depend on the types of students they serve?	Longitudinal study non-random allocation	Large number of cases per comparison group (the number of teachers in grade 4 is 14,762, and in grade 6 is 5,283)	Some missing data (In fourth grade 22.6% and in sixth grade 12.7% data lost because of various reason (35%))	State criterion-referenced test	3🔒
		3🔒	3🔒	3🔒	3🔒	
17	Goel and Barooah (2018) Drivers of Student Performance: Evidence from Higher Secondary Public Schools in Delhi	Longitudinal study non-random allocation	Large number of cases per comparison group (184, 145 and 144 teachers' records used in the first, second and third models)	Some missing data (The percentage of incompleteness (lost or attrition) of data is around 24)	Standardised test	3🔒
		3🔒	3🔒	3🔒	3🔒	
18	Cunningham (2014) The Effects Of Value-Added Modelling Decisions On Estimates Of Teacher Effectiveness	Longitudinal comparison study non-random allocation	Medium number of cases per comparison group (1,001 students in cohort1, 1,060students in cohort2 and 1,094 students in cohort3)	Some missing data (The percentage of incompleteness (lost or attrition) of data is around 20)	Standardised test	3🔒
		3🔒	3🔒	3🔒	3🔒	
19	Munoz and Chang (2007) The Elusive Relationship Between Teacher Characteristics and Student Academic Growth: A Longitudinal Multilevel Model for Change	Longitudinal study poor sampling non-random allocation	Medium Number of Cases (in unconditional means model 58 teachers, in unconditional growth models 57 teachers, and in conditional growth models 56 teachers’ records used)	Minimal missing data (the minimum number of teachers' records used in estimations is 56. (3% attrition)	Standardised test	2🔒
		2🔒	2🔒	2🔒	2🔒	

20	Schmitz (2007) An Empirical Sensitivity Analysis Of Value-Added Teachers' Effect Estimates To Hierarchical Linear Model Parameterizations	Longitudinal comparison study non-random allocation	Large number of cases per comparison group (Total students are 6332 in maths and 6044 in reading (Cohort 2002, grade 5))	Minimal missing data (The percentage of incompleteness (lost or attrition) of data is below 1)	Standardised test ((SAT9) math and reading score)	3🔒
		3🔒	3🔒	3🔒	3🔒	
21	Leigh (2010) Estimating teacher effectiveness from two-year changes in students' test scores	Longitudinal study non-random allocation	Large number of cases per comparison group (the sample size of cohort 1 is 59,612, of cohort 2 is 60,959, and of cohort 3 is 59,780)	Moderate missing data (The percentage of incompleteness (lost or attrition) of data is around 45)	Standardised test	3🔒
		3🔒	3🔒	3🔒	3🔒	
22	Kane and Staiger (2008) Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation	Longitudinal study random allocation	Large number of cases per comparison group (the sample size of the non-experiment group in experimental schools is 1,785; on the other hand, the sample size of experimental teacher group in experimental schools is 140)	Some missing data (The percentage of incompleteness (lost or attrition) of data is around 36)	Standardised test (the California Achievement Test)	4🔒
		4🔒	4🔒	4🔒	4🔒	
23	Alban (2002) Evaluating School and Teacher Effectiveness: A Comparison Of Analytic Models	Longitudinal comparison study non-random allocation	Large number of cases per comparison group (Sample size for language usage is 5,942, for writing is 5,990, for maths is 5,574, for science is 5,487, and for social studies is 5,902)	Some missing data (The percentage of incompleteness (lost or attrition) of data is around 32)	Standardised test	3🔒
		3🔒	3🔒	3🔒	3🔒	

24	Schafer et al. (2012) Evaluating Teachers and Schools Using Student Growth Models	Longitudinal comparison study non-random allocation	Large number of cases per comparison group (The sample size of cohort 1 is 5689, of cohort 2 is 5536, of cohort 3 is 5567, and of cohort 4 is 5791 in maths in 2018-2019)	Minimal missing data (The percentage of incompleteness (lost or attrition) of data is below 1)	Standardised test (State-wide test)	3🔒
		3🔒	3🔒	3🔒	3🔒	
25	Guarino et al. (2015b) An Evaluation of Empirical Bayes's Estimation of Value-Added Teacher Performance Measures	Longitudinal comparison study random allocation	Large number of cases per comparison group (Simulated data is 2160, real data is 482.031)	High level of missing data (The percentage of incompleteness (lost or attrition) of data is around 68)	Standardised test (State-wide test)	4🔒
		4🔒	4🔒	4🔒	4🔒	
26	Muñoz et al. (2011) Exploring Teacher Effectiveness Using Hierarchical Linear Models: Student- And Classroom-Level Predictors And Cross-Year Stability In Elementary School Reading	Longitudinal study non-random allocation	Large number of cases per comparison group (the number of students 5,837 (Year 1), 5,645 (Year 2), and 5,724 (Year 3) - the numbers of teachers 241 (Year 1), 235 (Year 2), and 236 (Year 3))	Minimal missing data (2,955 student records were removed from the three-year analyses. (17% attrition))	Standardised test (State-wide accountability test)	3🔒
		3🔒	3🔒	3🔒	3🔒	
27	Nye et al. (2004) How Large Are Teacher Effects?	Longitudinal study random allocation	Large number of cases per comparison group (the number of students in kindergarten is 5766, in first class is 6377, in second class is 5968, in third class is 5903)	Minimal missing data (The percentage of incompleteness (lost or attrition) of data is below 5)	Standardised test (the Stanford Achievement Test (SAT))	4🔒
		4🔒	4🔒	4🔒	4🔒	

28	Shaw (2012) Incorporating Latent Variable Outcomes in Value-Added Assessment: An Evaluation of Univariate and Multivariate Measurement Model Structures	mixed factorial design (longitudinal data) random allocation	Large number of cases per comparison group (the number of kindergarten teachers are 327, first grade teachers are 327, and second grade teachers are 324)	No attrition (Complete data)	Standardised test (the Stanford Achievement Test (SAT))	4🔒
		4🔒	4🔒	4🔒	4🔒	
29	McCaffrey et al. (2009) The Intertemporal Variability of Teacher Effect Estimates	Longitudinal study non-random allocation	Large number of cases per comparison group (at least 2070 students in each county)	Some missing data (Although 34204 records obtained, 24232 of them used in estimates (29% attrition))	Standardised test (The Sunshine State Standards Florida Comprehensive Achievement Test (FCAT-SSS))	3🔒
		3🔒	3🔒	3🔒	3🔒	
30	Tobe (2008) An Investigation of The Differential Impact Of Teacher Characteristics And Attitudes On Student Mathematics Achievement Using A Value-Added Approach	Causal comparative research design (longitudinal data) non-random allocation)	Large number of cases per comparison group (6.106 male students, 6.263 female)	Minimal missing data (the number of students declined by 19% for each year of data required prior year test score)	Standardised test (Texas Assessment of Knowledge and Skills mathematics tests)	3🔒
		3🔒	3🔒	3🔒	3🔒	
31	Goldhaber and Hansen (2013) Is it Just a Bad Class? Assessing the Long-term Stability of Estimated Teacher Performance	Longitudinal study non-random allocation	Large number of cases per comparison group (the same amount of observation were used in observed and unobserved	Moderate missing data (The percentage of incompleteness (lost or attrition) of data is 47)	Standardised test	3🔒
		3🔒	3🔒	3🔒	3🔒	

32	Potamites et al. (2009) Measuring School and Teacher Effectiveness in the EPIC Charter School Consortium—Year 2	Longitudinal study non-random allocation	Large number of cases per comparison group (572 teachers in the elementary grade, 233 in the middle school grades, and 103 in the high school grade)	Minimal missing data (Ethnicity, gender, and special education status were missing for less than 1 percent of the final one-year analysis sample. Free or reduced price lunch status was missing for 6 percent and limited English proficiency status was missing for 12 percent)	Standardised test (the National Assessment of Educational Progress (NAEP))	3🔒
		3🔒	3🔒	3🔒	3🔒	
33	Buddin (2011) Measuring teacher and school effectiveness at improving student achievement in Los Angeles elementary schools	Longitudinal study non-random allocation	Large number of cases per comparison group (The number of case used in determination of unobserved heterogeneity in math and ELA teacher effectiveness are same, n=36,484)	Minimal missing data (The percentage of incompleteness (lost or attrition) of data is below 1%).	Standardised test (the California Standards Test (CST))	3🔒
		3🔒	3🔒	3🔒	3🔒	
34	Blackford (2016) Measuring Teacher Effectiveness: A Comparison across VA Models Utilizing Arkansas Data	Longitudinal comparison study non-random allocation	Large number of cases per comparison group (the number of students is 13,200 in 2012, 13,087 in 2013, and 13,485 in 2014)	Minimal missing data (The percentage of incompleteness (lost or attrition) of data is below 1%)	Standardised test	3🔒
		3🔒	3🔒	3🔒	3🔒	
35	Chetty et al. (2014) Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates	Longitudinal comparison study non-random allocation	Large number of cases per comparison group (roughly half of the observations belonged to female students - over 3.5M observation)	Some missing data (The percentage of incompleteness (lost or attrition) of data is 34)	Standardised test (State-wide test)	3🔒
		3🔒	3🔒	3🔒	3🔒	

36	Gallagher (2002) The Relationship between Measures of Teacher Quality and Student Achievement: The Case of Vaughn Elementary	Longitudinal study non-random allocation	Small Number of Cases per comparison group (the number of students in literacy and maths are 584, in language arts is 532)	No attrition (Complete data)	Standardised test (The Stanford-9)	3🔒
		3🔒	3🔒	3🔒	3🔒	
37	Johnson et al. (2015) Sensitivity of Teacher Value-Added Estimates to Student and Peer Control Variables	Longitudinal comparison study non-random allocation	Large number of cases per comparison group (The number of 8th grade teacher in maths is 2778, and in reading is 3344)	No attrition (Complete data)	Standardised test	3🔒
		3🔒	3🔒	3🔒	3🔒	
38	Ehlert et al. (2014) The Sensitivity of Value-Added Estimates to Specification Adjustments: Evidence from School and Teacher-Level Models in Missouri	Longitudinal comparison study non-random allocation	Large number of cases per comparison group (289 and 390 teachers in maths and com art respectively / 20,871 and 21,129 students in maths and com art, respectively)	Minimal missing data (The percentage of incompleteness (lost or attrition) of data is 9)	Standardised test (Missouri Assessment Program (MAP))	3🔒
		3🔒	3🔒	3🔒	3🔒	
39	Heistad (1999) Stability and correlates of teacher effects in grade two reading achievement	Longitudinal study non-random allocation	Large number of cases per comparison group (The number of teacher 1993/94 is 182, in 1994/95 is 197, and in 1995/96 is 206)	Some missing data (The percentage of incompleteness (lost or attrition) of data is 22)	Standardised test (California Achievement Tests - CAT/E and CAT/5)	3🔒
		3🔒	3🔒	3🔒	3🔒	
40	Rothstein (2009) Student sorting and bias in value added estimation: Selection on observables and unobservables	Longitudinal study random allocation	Large number of cases per comparison group (almost all records were used in estimations of predictability of 5th grade reading scores from prior information)	Minimal missing data (The percentage of incompleteness (lost or attrition) of data is 4)	Standardised test	4🔒
		4🔒	4🔒	4🔒	4🔒	

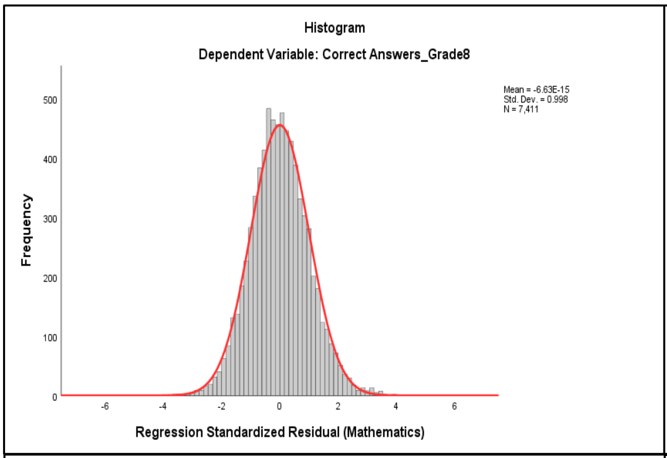
41	Hu (2015) Teacher evaluation based on an aspect of classroom practice and on student achievement: A relational analysis between student learning objectives and value-added modelling	Longitudinal study non-random allocation	Large number of cases per comparison group (1210 and 1239 teachers in maths and reading, respectively)	Minimal missing data (The percentage of incompleteness (lost or attrition) of data is below 5)	Standardised test (the End-of-Grade (EOG))	3🔒
		3🔒	3🔒	3🔒	3🔒	
42	Goldhaber et al. (2013) Teacher Value-Added at the High-School Level: Different Models, Different Answers?	Longitudinal study random allocation	Medium number of cases per comparison group (the range of the students' records in different testing subjects used in the estimations are 1426 to 1840)	Minimal missing data The percentage of incompleteness (lost or attrition) of data is 3)	QualityCore end-of-course assessment	4🔒
		4🔒	4🔒	4🔒	4🔒	
43	Aaronson et al. (2007) Teachers and Student Achievement in the Chicago Public High Schools	Longitudinal study random allocation	Large number of cases per comparison group (the number of male students is 25.299, and female's is 27658)	Moderate missing data (although the initial sample size of teacher is 1132, in the estimations of 589 teachers' effects were calculated (48% attrition))	Standardised test (the Test of Achievement and Proficiency (TAP))	4🔒
		4🔒	4🔒	4🔒	4🔒	
44	Gagnon (2014) Understanding The Distribution Of Teacher Effectiveness	Longitudinal study non-random allocation	Large number of cases per comparison group (14,402 and 15,742 for 5th grade mathematics and ELA, respectively, and 10,657 and 12,610 for 8th grade mathematics, and ELA, respectively)	Minimal missing data (The percentage of incompleteness (lost or attrition) of data is around 3)	Standardised test (the New England Common Assessment Program (NECAP))	3🔒
		3🔒	3🔒	3🔒	3🔒	

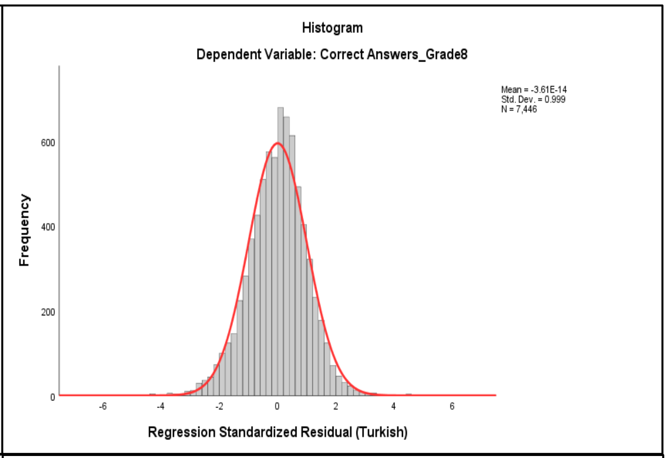
45	Castellano (2011) Unpacking student growth percentiles: statistical properties of regression-based approaches with implications for student and school classifications	Longitudinal comparison study random allocation	Large number of cases per comparison group (The State A dataset contains records for a single cohort of about 25,000 students and the State B dataset has achievement score history for a cohort of about 76,400 students)	No attrition (Complete data)	Standardised test	4🔒
		4🔒	4🔒	4🔒	4🔒	
46	Kurtz (2018) Value-Added and Student Growth Percentile Models: What Drives Differences in Estimated Classroom Effects?	Longitudinal comparison study random allocation	Medium number of cases per comparison group (Same amount of students records used in each model with using simulated data and empirical data (1000 and 18821 respectively))	No attrition (Complete data)	the End-of-Grade (EOG)	4🔒
		4🔒	4🔒	4🔒	4🔒	
47	Newton et al. (2010) Value-Added Modeling of Teacher Effectiveness: An Exploration of Stability across Models and Contexts	Longitudinal comparison study non-random allocation	Large number of cases per comparison group (the number of maths teachers is 103, and English language art is 114)	Not reported	Standardised test	1🔒
		3🔒	3🔒	1🔒	1🔒	
48	Wei et al. (2012) Value-Added Models in the Evaluation of Teacher Effectiveness: A Comparison of Models and Outcomes	Longitudinal comparison study non-random allocation	Medium number of cases per comparison group (73 teachers in math, and 58 in ELA)	Minimal missing data (The percentage of incompleteness (lost or attrition) of data is 14)	Standardised test (the Texas Assessment of Knowledge and Skills (TAKS) tests)	3🔒
		3🔒	3🔒	3🔒	3🔒	

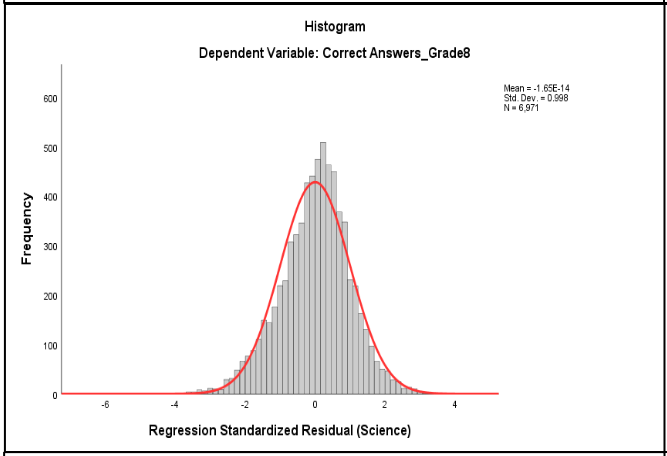
49	Kersting et al. (2013) Value-added Teacher Estimates as Part of Teacher Evaluations: Exploring the Effects of Data and Model Specifications on the Stability of Teacher Value-added Scores	Longitudinal study non-random allocation 3🔒	Large number of cases per comparison group (the number of students in each cohort is at least 38.503) 3🔒	Some missing data (The percentage of incompleteness (lost or attrition) of data is 6 for teacher, 22 for students, and 1 for school) 3🔒	Standardised test 3🔒	3🔒
50	Harris et al. (2014) Value-added models and the measurement of teacher productivity	Longitudinal comparison study non-random allocation 3🔒	Large number of cases per comparison group (196,015 records belong to four cohorts of students) 3🔒	Minimal missing data (The percentage of incompleteness (lost or attrition) of data is 18) 3🔒	Standardised test (State-wide test) 3🔒	3🔒

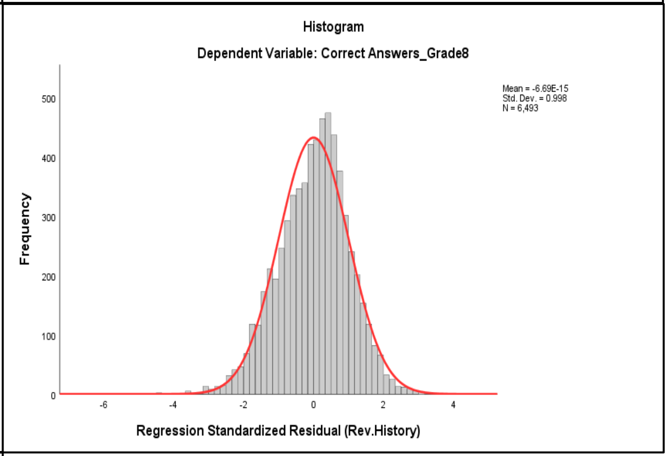
Appendix F. Assumptions Checked

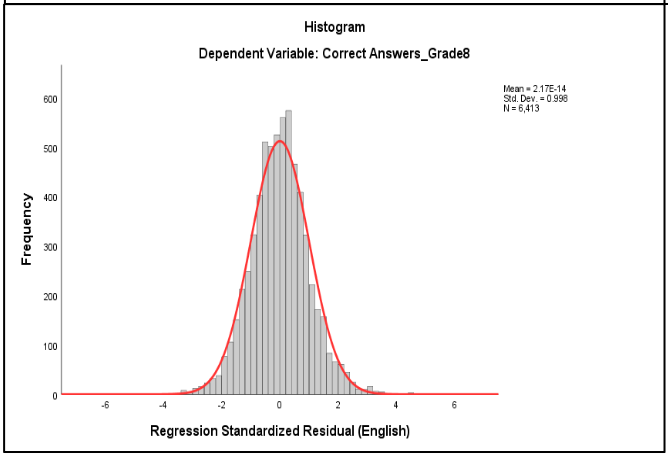
Normality Assumptions



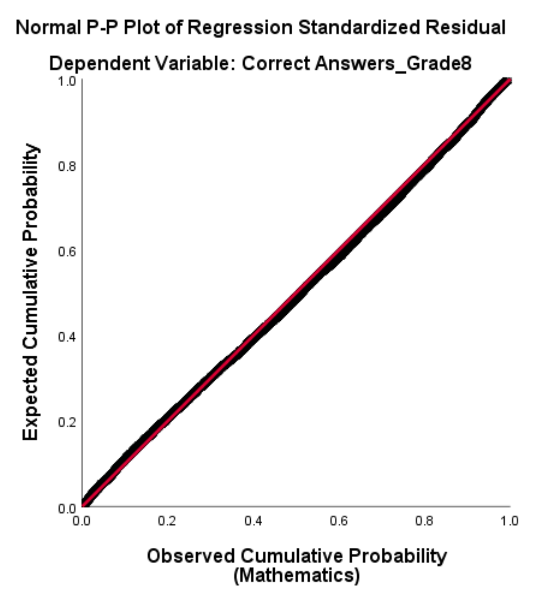


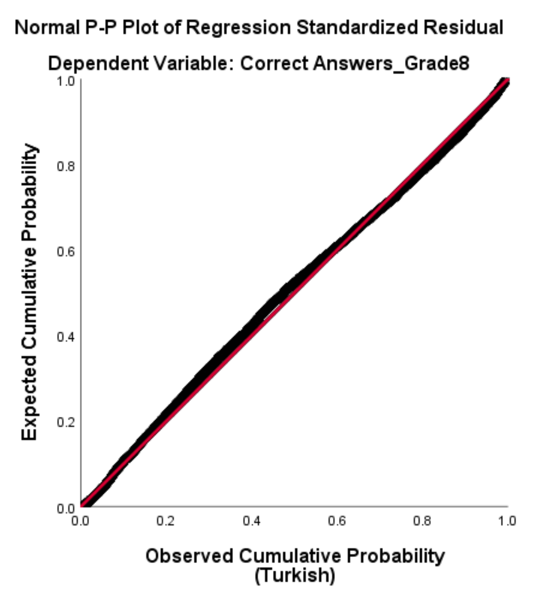


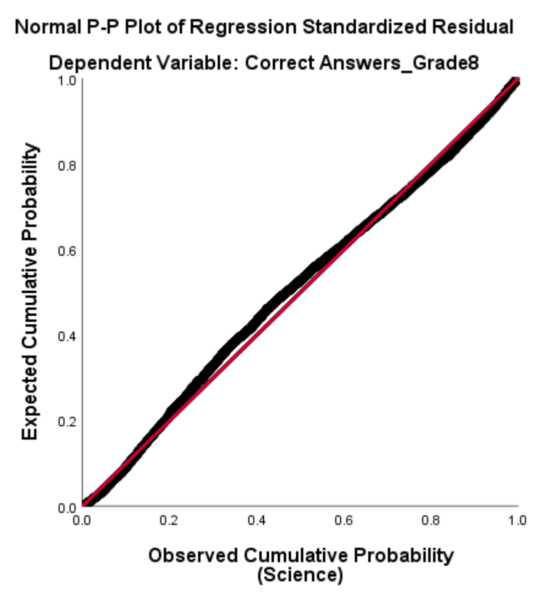


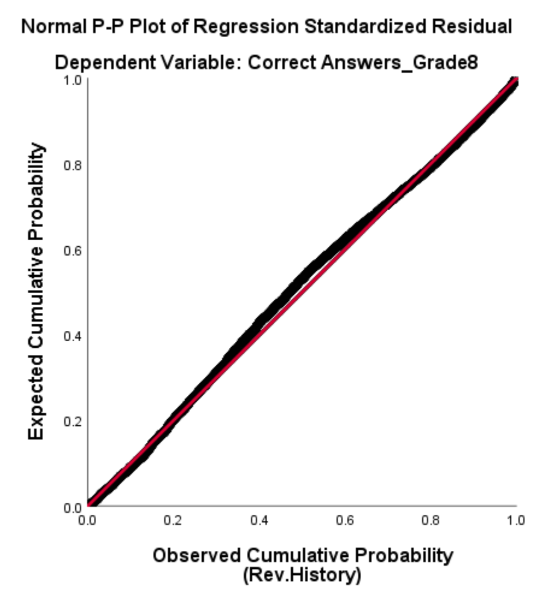


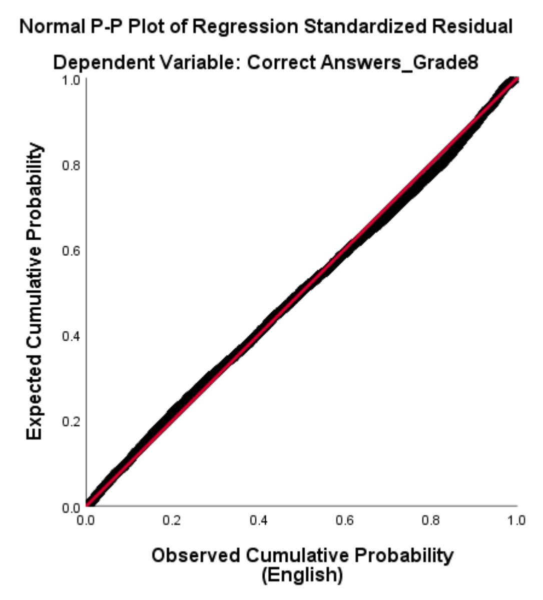
Linearity Assumptions



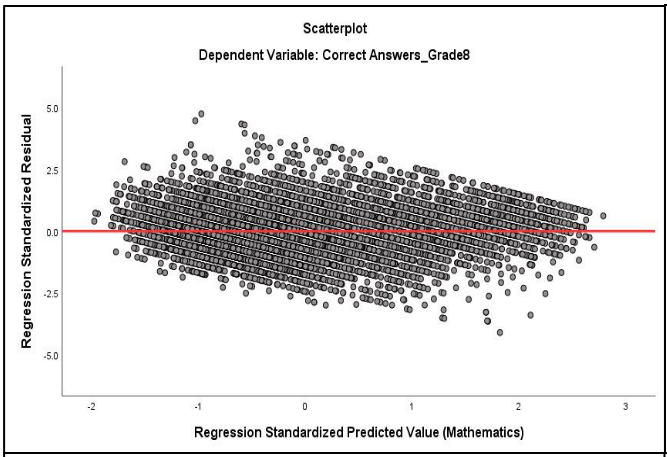


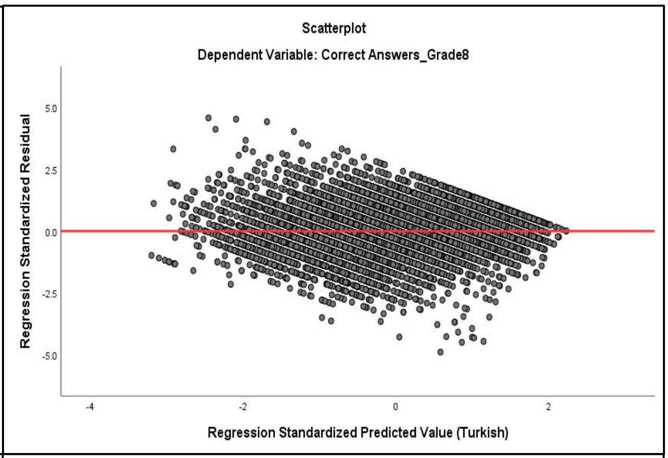


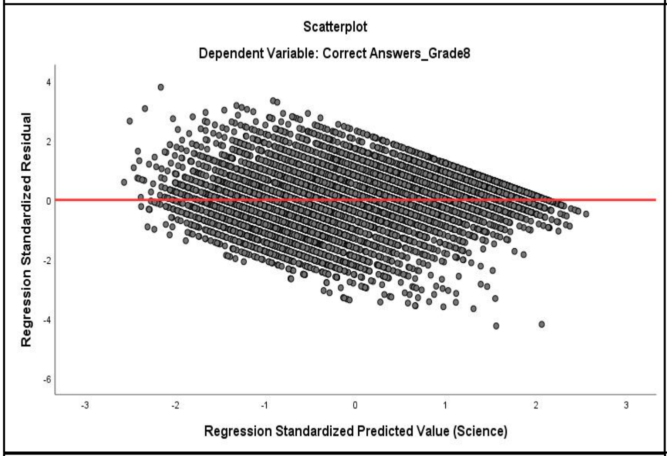


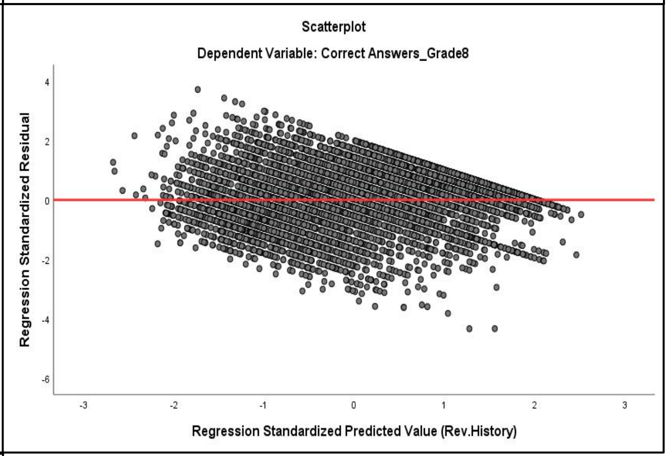


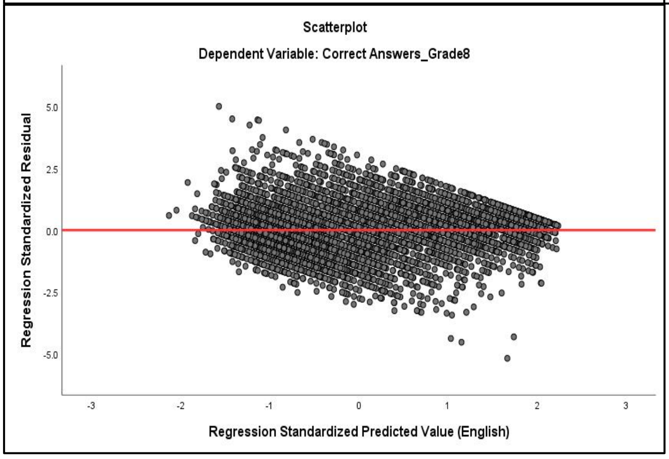
Homoscedasticity











Multicollinearity

		Mathematics																						Collinearity Statistics	
		Prior Attainment (G7)	Prior Attainment (G6)	Students Gender	Language Learner ID	School Category = Regional Boarding	School Category = Vocational	Service Score	Location =Rural	Location =Suburban	School Average_ Grade7	School Average_ Grade6	Teacher's Gender	Class Size	Percentage of female students classroom	Total teaching experience	Experience in current school	Appointment Fiels	Graduation Field	Having Master Degree ?	Master Field= Related	Master Field= Unrelated	Class Average_ Grade7	Class Average_ Grade6	Tolerance	VIF
Pearson Correlation	Prior Attainment (G7)	1.000																							0.447	2.238
	Prior Attainment (G6)	0.683	1.000																						0.444	2.254
	Students Gender	0.094	0.099	1.000																					0.858	1.165
	Language Learner ID	-0.012	-0.015	-0.018	1.000																				0.994	1.006
	School Category= Regional Boarding	-0.054	-0.047	0.002	-0.009	1.000																			0.878	1.139
	School Category= Vocational	-0.045	-0.063	-0.018	-0.001	-0.071	1.000																		0.890	1.124
	Service Score	-0.155	-0.162	0.021	-0.020	0.292	-0.075	1.000																	0.230	4.354
	Location=Rural	-0.180	-0.177	0.039	-0.015	0.164	-0.159	0.622	1.000																0.237	4.226
	Location=Suburban	-0.027	-0.052	-0.008	-0.017	0.097	0.123	0.444	-0.232	1.000															0.327	3.054
	School Average_Grade7	0.432	0.386	-0.014	0.023	-0.127	-0.100	-0.357	-0.415	-0.057	1.000														0.108	9.985
	School Average_Grade6	0.390	0.429	-0.018	0.020	-0.107	-0.146	-0.384	-0.408	-0.132	0.883	1.000													0.109	9.976
	Teacher's Gender	-0.033	-0.049	-0.012	0.016	-0.102	0.017	-0.018	-0.028	-0.030	-0.127	-0.148	1.000												0.887	1.128
	Class Size	0.231	0.225	0.005	0.018	-0.074	-0.010	-0.474	-0.375	-0.271	0.474	0.468	-0.103	1.000											0.584	1.713
	Percentage of female students	0.005	0.014	0.355	0.008	0.006	-0.050	0.058	0.110	-0.022	-0.041	-0.050	-0.033	0.015	1.000										0.833	1.201
	Total teaching experience	0.197	0.195	0.008	0.025	-0.140	0.015	-0.463	-0.438	-0.196	0.441	0.454	-0.073	0.398	0.022	1.000									0.522	1.914
	Experience in current school	0.093	0.091	0.005	0.025	0.006	-0.099	-0.254	-0.218	-0.085	0.251	0.261	-0.175	0.198	0.014	0.486	1.000								0.699	1.430
	Appointment Fiels	-0.044	-0.032	0.019	0.002	0.009	0.017	-0.062	0.021	-0.090	0.002	-0.008	0.048	0.058	0.054	-0.087	0.011	1.000							0.777	1.288
	Graduation Field	-0.008	-0.012	0.003	0.005	0.021	0.040	0.003	0.050	-0.033	-0.028	-0.033	0.036	-0.046	0.007	-0.105	0.024	0.413	1.000						0.807	1.239
	Having Master Degree ?	0.101	0.059	-0.009	0.041	-0.033	-0.065	-0.094	-0.081	-0.031	0.191	0.168	0.116	0.144	-0.025	0.017	-0.068	0.008	0.019	1.000					0.243	4.112
	Master Field= Related	0.121	0.086	-0.022	0.037	-0.025	-0.049	-0.092	-0.061	-0.065	0.195	0.186	0.063	0.230	-0.063	0.035	-0.101	0.006	0.014	0.756	1.000				0.285	3.507
	Master Field= Unrelated	0.030	0.013	0.008	0.036	-0.014	-0.027	-0.051	-0.034	-0.036	0.084	0.070	0.067	-0.034	0.023	0.040	0.034	0.003	0.008	0.415	-0.009	1.000			0.575	1.739
	Class Average_Grade7	0.534	0.465	0.003	0.007	-0.100	-0.078	-0.284	-0.333	-0.043	0.803	0.726	-0.057	0.431	0.009	0.362	0.173	-0.086	-0.016	0.188	0.225	0.053	1.000		0.114	8.756
	Class Average_Grade6	0.463	0.538	0.009	-0.001	-0.083	-0.112	-0.300	-0.324	-0.099	0.713	0.791	-0.087	0.428	0.026	0.362	0.178	-0.087	-0.024	0.117	0.168	0.021	0.865	1.000	0.120	8.325

		Turkish																							Collinearity Statistics	
		Prior Attainment (G7)	Prior Attainment (G6)	Students Gender	Language Learner ID	School Category = Regional Boarding	School Category = Vocational	Service Score	Location =Rural	Location =Suburban	School Average_Grade7	School Average_Grade6	Teacher's Gender	Class Size	Percentage of female students classroom	Total teaching experience	Experience in current school	Appointment Fiels	Graduation Field	Having Master Degree ?	Master Field= Related	Master Field= Unrelated	Class Average_Grade7	Class Average_Grade6	Tolerance	VIF
Pearson Correlation	Prior Attainment (G7)	1.000																							0.496	2.018
	Prior Attainment (G6)	0.653	1.000																						0.492	2.031
	Students Gender	0.209	0.211	1.000																					0.816	1.226
	Language Learner ID	-0.057	-0.037	-0.021	1.000																				0.993	1.007
	School Category= Regional Boarding	-0.065	-0.063	0.013	-0.007	1.000																			0.897	1.115
	School Category= Vocational	0.013	-0.004	-0.017	0.000	-0.058	1.000																		0.896	1.117
	Service Score	-0.162	-0.172	0.020	-0.018	0.245	-0.075	1.000																	0.219	4.564
	Location=Rural	-0.177	-0.177	0.039	-0.014	0.099	-0.159	0.599	1.000																0.242	4.131
	Location=Suburban	-0.001	-0.052	-0.007	-0.016	0.144	0.113	0.468	-0.240	1.000															0.305	3.284
	School Average_Grade7	0.387	0.305	0.011	0.016	-0.165	0.043	-0.372	-0.421	0.012	1.000														0.158	6.346
	School Average_Grade6	0.320	0.392	-0.002	0.020	-0.162	-0.016	-0.450	-0.445	-0.152	0.769	1.000													0.157	6.390
	Teacher's Gender	0.043	0.069	0.014	0.004	0.036	0.015	-0.186	-0.058	-0.168	0.072	0.146	1.000												0.887	1.127
	Class Size	0.173	0.199	0.005	0.018	-0.044	-0.009	-0.478	-0.376	-0.277	0.338	0.450	0.091	1.000											0.600	1.666
	Percentage of female students	0.049	0.046	0.352	0.006	0.037	-0.048	0.058	0.109	-0.019	0.032	-0.006	0.039	0.015	1.000										0.828	1.207
	Total teaching experience	0.151	0.172	-0.015	0.030	-0.159	-0.114	-0.475	-0.341	-0.307	0.307	0.420	0.115	0.456	-0.043	1.000									0.583	1.714
	Experience in current school	0.026	0.052	-0.011	0.009	-0.093	-0.174	-0.291	-0.140	-0.202	-0.004	0.119	0.159	0.189	-0.031	0.372	1.000								0.776	1.289
	Appointment Fiels	There is no case whose appointment fields not related to Turkish																								
	Graduation Field	-0.002	0.008	-0.013	0.003	0.010	0.024	0.045	0.030	0.033	-0.019	0.005	0.013	0.077	-0.035	-0.072	0.009		1.000						0.966	1.035
	Having Master Degree ?	0.023	-0.004	0.001	-0.010	-0.035	-0.082	-0.020	0.097	-0.063	0.048	-0.012	-0.158	-0.024	0.003	-0.032	-0.057		0.014	1.000					0.523	1.913
	Master Field= Related	0.024	0.033	0.013	-0.004	-0.013	-0.030	0.003	0.036	-0.041	0.068	0.065	0.010	0.053	0.036	-0.019	-0.024		0.005	0.363	1.000				0.788	1.270
	Master Field= Unrelated	0.029	0.022	0.005	-0.005	-0.019	-0.044	0.000	-0.021	0.027	0.044	0.039	-0.114	-0.078	0.015	0.039	-0.006		0.008	0.544	-0.010	1.000			0.622	1.609
	Class Average_Grade7	0.463	0.379	0.035	0.001	-0.134	0.038	-0.330	-0.339	-0.021	0.798	0.622	0.067	0.343	0.099	0.274	0.010		-0.015	0.020	0.029	0.006	1.000		0.166	6.017
	Class Average_Grade6	0.376	0.466	0.030	0.011	-0.103	0.001	-0.328	-0.350	-0.093	0.596	0.777	0.149	0.383	0.086	0.311	0.107		0.012	-0.021	0.041	0.039	0.751	1.000	0.182	5.496

		Science																							Collinearity Statistics	
		Prior Attainment (G7)	Prior Attainment (G6)	Students Gender	Language Learner ID	School Category = Regional Boarding	School Category= Vocational	Service Score	Location =Rural	Location= Suburban	School Average_ Grade7	School Average_ Grade6	Teacher's Gender	Class Size	Percentage of female students classroom	Total teaching experience	Experience in current school	Appoint ment Fiels	Graduati on Field	Having Master Degree ?	Master Field= Related	Master Field= Unrelated	Class Average_ Grade7	Class Average_ Grade6	Tolerance	VIF
Pearson Correlation	Prior Attainment (G7)	1.000																							0.537	1.863
	Prior Attainment (G6)	0.587	1.000																						0.522	1.916
	Students Gender	0.112	0.044	1.000																					0.856	1.168
	Language Learner ID	-0.015	-0.015	-0.022	1.000																				0.995	1.005
	School Category= Regional Boarding	-0.066	-0.056	0.014	-0.008	1.000																			0.869	1.151
	School Category= Vocational	-0.029	-0.033	-0.016	0.000	-0.062	1.000																		0.856	1.168
	Service Score	-0.114	-0.100	0.023	-0.018	0.259	-0.067	1.000																	0.224	4.454
	Location=Rural	-0.154	-0.158	0.041	-0.014	0.105	-0.159	0.585	1.000																0.260	3.841
	Location=Suburban	0.010	0.016	-0.007	-0.017	0.146	0.116	0.477	-0.239	1.000															0.323	3.099
	School Average_Grade7	0.378	0.321	0.007	0.028	-0.180	-0.083	-0.308	-0.410	0.021	1.000														0.173	5.767
	School Average_Grade6	0.297	0.410	-0.023	0.021	-0.140	-0.078	-0.256	-0.390	0.023	0.789	1.000													0.169	5.907
	Teacher's Gender	0.034	0.010	0.025	-0.002	-0.171	0.035	-0.056	-0.010	-0.027	0.063	0.000	1.000												0.928	1.078
	Class Size	0.162	0.140	0.006	0.017	-0.046	-0.016	-0.480	-0.369	-0.295	0.379	0.324	-0.007	1.000											0.638	1.568
	Percentage of female students	0.032	-0.015	0.361	0.006	0.038	-0.045	0.064	0.113	-0.020	0.020	-0.065	0.067	0.017	1.000										0.826	1.211
	Total teaching experience	0.129	0.147	-0.005	0.044	-0.139	-0.161	-0.547	-0.337	-0.321	0.325	0.352	-0.054	0.313	-0.013	1.000									0.541	1.847
	Experience in current school	0.032	0.024	0.008	-0.004	-0.108	-0.199	-0.246	-0.075	-0.145	0.072	0.056	0.021	0.079	0.021	0.387	1.000								0.789	1.268
	Appointment Fiels	0.002	-0.025	0.000	0.003	0.009	0.022	0.040	0.026	0.030	0.017	-0.073	0.060	-0.006	0.001	-0.086	0.050	1.000							0.962	1.039
	Graduation Field	0.002	-0.025	0.000	0.003	0.009	0.022	0.040	0.026	0.030	0.017	-0.073	0.060	-0.006	0.001	-0.086	0.050	1.000	1.000							
	Having Master Degree ?	0.012	0.000	0.032	0.002	-0.038	-0.090	-0.073	-0.020	-0.100	-0.009	-0.038	0.098	-0.016	0.090	0.040	0.046	0.013	0.013	1.000					0.401	2.495
	Master Field= Related	0.049	0.049	0.031	0.010	-0.028	-0.065	-0.089	-0.041	-0.089	0.063	0.061	0.069	0.027	0.087	0.089	0.103	0.010	0.010	0.725	1.000				0.428	2.337
	Master Field= Unrelated	0.026	-0.012	0.010	-0.002	-0.008	-0.020	0.074	0.112	-0.027	0.071	-0.026	-0.054	-0.031	0.028	0.016	-0.031	0.003	0.003	0.218	-0.009	1.000			0.832	1.202
	Class Average_Grade7	0.486	0.380	0.024	0.007	-0.141	-0.059	-0.240	-0.316	0.018	0.775	0.610	0.080	0.341	0.066	0.272	0.070	0.003	0.003	0.031	0.103	0.055	1.000		0.202	4.950
	Class Average_Grade6	0.358	0.515	-0.008	0.007	-0.111	-0.064	-0.205	-0.304	0.017	0.620	0.787	0.024	0.283	-0.023	0.289	0.051	-0.052	-0.052	0.007	0.097	-0.021	0.737	1.000	0.194	5.151

		Revolution History																						Collinearity Statistics	
		Prior Attainment (G7)	Prior Attainment (G6)	Students Gender	Language Learner ID	School Category = Regional Boarding	School Category= Vocational	Service Score	Location =Rural	Location= Suburban	School Average_Grade7	School Average_Grade6	Teacher's Gender	Class Size	Percentage of female students classroom	Total teaching experience	Experience in current school	Appointment Fiels	Graduation Field	Having Master Degree ?	Master Field= Related	Master Field= Unrelated	Class Average_Grade7	Class Average_Grade6	Tolerance	VIF
Pearson Correlation	Prior Attainment (G7)	1.000																							0.523	1.911
	Prior Attainment (G6)	0.606	1.000																						0.517	1.935
	Students Gender	0.125	0.094	1.000																					0.857	1.166
	Language Learner ID	-0.019	-0.021	-0.031	1.000																				0.995	1.005
	School Category= Regional Boarding	-0.065	-0.056	0.014	-0.008	1.000																			0.861	1.161
	School Category= Vocational	-0.007	-0.026	-0.007	-0.016	-0.058	1.000																		0.865	1.156
	Service Score	-0.101	-0.104	0.023	-0.018	0.257	-0.031	1.000																	0.198	5.038
	Location=Rural	-0.138	-0.128	0.045	-0.014	0.100	-0.150	0.588	1.000																0.223	4.481
	Location=Suburban	0.021	-0.006	-0.003	-0.016	0.153	0.089	0.496	-0.248	1.000															0.259	3.864
	School Average_Grade7	0.383	0.320	-0.005	0.030	-0.173	-0.025	-0.276	-0.355	0.052	1.000														0.172	5.804
	School Average_Grade6	0.306	0.390	-0.014	0.027	-0.144	-0.075	-0.269	-0.311	-0.039	0.801	1.000													0.183	5.475
	Teacher's Gender	0.062	0.131	-0.001	0.032	-0.073	-0.039	-0.208	-0.091	-0.216	0.129	0.245	1.000												0.853	1.173
	Class Size	0.176	0.150	-0.005	0.015	-0.038	0.017	-0.465	-0.371	-0.258	0.384	0.360	0.150	1.000											0.620	1.613
	Percentage of female students	0.012	-0.003	0.349	-0.011	0.041	-0.020	0.065	0.128	-0.009	-0.015	-0.041	-0.003	-0.014	1.000										0.848	1.180
	Total teaching experience	0.053	0.032	-0.031	0.007	-0.065	0.147	-0.402	-0.318	-0.228	0.135	0.132	0.082	0.225	-0.090	1.000									0.654	1.530
	Experience in current school	0.001	-0.017	-0.007	-0.005	-0.100	-0.181	-0.198	-0.156	-0.077	0.010	-0.017	0.036	0.034	-0.019	0.253	1.000								0.795	1.258
	Appointment Fiels	0.002	0.021	-0.019	0.005	0.017	-0.062	-0.019	-0.066	-0.021	-0.009	0.043	0.090	0.084	-0.053	-0.064	0.063	1.000							0.878	1.139
	Graduation Field	0.030	0.040	0.011	0.009	-0.040	-0.023	0.186	0.093	0.160	0.037	0.013	-0.066	0.002	0.033	-0.385	-0.234	0.251	1.000						0.727	1.376
	Having Master Degree ?	-0.023	-0.020	0.010	0.011	-0.030	-0.061	0.021	-0.020	0.004	-0.009	-0.039	-0.042	-0.037	0.028	0.008	0.108	0.018	0.072	1.000					0.601	1.664
	Master Field= Related	-0.037	-0.054	0.007	-0.004	-0.014	-0.028	0.111	-0.039	0.159	-0.095	-0.138	-0.072	0.000	0.019	-0.008	0.031	0.008	0.033	0.461	1.000				0.685	1.461
	Master Field= Unrelated	-0.036	-0.027	0.000	-0.003	-0.011	-0.023	0.136	0.137	-0.034	-0.088	-0.064	-0.059	-0.079	0.000	0.033	-0.071	0.007	0.027	0.374	-0.005	1.000			0.758	1.319
	Class Average_Grade7	0.475	0.366	0.010	0.014	-0.139	-0.019	-0.218	-0.281	0.039	0.797	0.642	0.135	0.370	0.029	0.107	0.007	0.003	0.063	-0.045	-0.076	-0.071	1.000		0.207	4.829
	Class Average_Grade6	0.348	0.491	0.002	0.022	-0.113	-0.052	-0.216	-0.251	-0.023	0.631	0.782	0.272	0.315	0.005	0.064	-0.016	0.043	0.079	-0.043	-0.108	-0.050	0.733	1.000	0.216	4.629

		English																							Collinearity Statistics	
		Prior Attainment (G7)	Prior Attainment (G6)	Students Gender	Language Learner ID	School Category = Regional Boarding	School Category= Vocational	Service Score	Location =Rural	Location= Suburban	School Average_ Grade7	School Average_ Grade6	Teacher's Gender	Class Size	Percentage of female students classroom	Total teaching experience	Experience in current school	Appoint ment Fiels	Graduati on Field	Having Master Degree ?	Master Field= Related	Master Field= Unrelated	Class Average_ Grade7	Class Average_ Grade6	Tolerance	VIF
Pearson Correlation	Prior Attainment (G7)	1.000																							0.415	2.408
	Prior Attainment (G6)	0.713	1.000																						0.406	2.462
	Students Gender	0.182	0.209	1.000																					0.840	1.191
	Language Learner ID	-0.011	-0.015	-0.030	1.000																				0.995	1.005
	School Category= Regional Boarding	-0.040	-0.040	-0.001	-0.006	1.000																			0.782	1.279
	School Category= Vocational	-0.065	-0.062	-0.017	-0.013	-0.053	1.000																		0.895	1.117
	Service Score	-0.111	-0.109	0.019	-0.010	0.254	-0.036	1.000																	0.242	4.125
	Location=Rural	-0.140	-0.160	0.033	-0.009	0.073	-0.139	0.586	1.000																0.266	3.760
	Location=Suburban	-0.015	-0.004	-0.006	-0.012	0.157	0.150	0.455	-0.261	1.000															0.336	2.980
	School Average_Grade7	0.414	0.378	0.000	0.038	-0.100	-0.162	-0.254	-0.320	-0.042	1.000														0.138	7.268
	School Average_Grade6	0.366	0.430	-0.006	0.034	-0.097	-0.161	-0.269	-0.365	-0.031	0.867	1.000													0.114	8.762
	Teacher's Gender	0.000	0.016	-0.013	-0.008	0.034	-0.042	0.078	0.088	-0.055	0.000	-0.011	1.000												0.876	1.141
	Class Size	0.193	0.185	-0.004	-0.009	-0.014	0.009	-0.449	-0.361	-0.242	0.368	0.383	0.058	1.000											0.639	1.566
	Percentage of female students	0.031	0.026	0.316	-0.011	-0.005	-0.053	0.060	0.103	-0.020	0.000	-0.019	-0.041	-0.012	1.000										0.868	1.151
	Total teaching experience	0.118	0.142	0.008	0.027	-0.116	-0.104	-0.357	-0.248	-0.188	0.288	0.324	-0.233	0.255	0.025	1.000									0.536	1.865
	Experience in current school	0.091	0.109	-0.004	0.002	-0.047	-0.114	-0.188	-0.107	-0.168	0.245	0.246	-0.030	0.094	-0.012	0.572	1.000								0.627	1.595
	Appointment Fiels	All appointment fiels cases are related the teaching subject																								
	Graduation Field	-0.053	-0.051	-0.001	0.007	-0.110	0.059	0.089	0.087	0.045	-0.091	-0.105	0.110	-0.069	-0.002	-0.149	-0.131		1.000						0.833	1.200
	Having Master Degree ?	-0.020	-0.028	0.001	-0.003	0.287	-0.027	0.107	0.048	0.069	-0.061	-0.074	0.054	0.006	0.002	-0.002	0.064		-0.256	1.000					0.427	2.344
	Master Field= Related	There is no case has a master degree in a subject related to English																								
	Master Field= Unrelated	-0.018	-0.018	0.008	-0.002	-0.008	-0.017	0.069	0.109	-0.028	-0.046	-0.053	0.035	-0.036	0.024	-0.029	-0.017		0.009	0.646		1.000			0.499	2.005
	Class Average_Grade7	0.526	0.455	0.020	0.011	-0.076	-0.120	-0.209	-0.262	-0.026	0.763	0.679	0.002	0.371	0.064	0.217	0.161		-0.102	-0.040		-0.035	1.000		0.137	7.305
	Class Average_Grade6	0.459	0.530	0.015	0.013	-0.079	-0.113	-0.206	-0.296	-0.013	0.703	0.805	0.025	0.351	0.046	0.261	0.197		-0.091	-0.064		-0.043	0.854	1.000	0.117	8.527

Appendix G. Transition Matrixes for Year to Year Consistency in Effectiveness Categories (in Percentages)

Mathematics Teachers		Previous				**Total**
		Highly Effective	Effective	Partially Effective	Ineffective	
Current	Highly Effective	4.8	4.8	9.5	4.8	**5**
	Effective	4.8	4.8	9.5	4.8	**5**
	Partially Effective	4.8	9.5	4.8	4.8	**5**
	Ineffective	14.3	4.8	0.0	9.5	**6**
Total		**6**	**5**	**5**	**5**	**21**

Turkish Teachers		Previous				**Total**
		Highly Effective	Effective	Partially Effective	Ineffective	
Current	Highly Effective	3.1	6.3	6.3	9.4	**8**
	Effective	9.4	3.1	9.4	6.3	**9**
	Partially Effective	6.3	3.1	6.3	6.3	**7**
	Ineffective	6.3	12.5	3.1	3.1	**8**
Total		**8**	**8**	**8**	**8**	**32**

Science Teachers		Previous				**Total**
		Highly Effective	Effective	Partially Effective	Ineffective	
Current	Highly Effective	2.6	5.1	7.7	10.3	**10**
	Effective	12.8	10.3	0.0	2.6	**10**
	Partially Effective	5.1	5.1	5.1	10.3	**10**
	Ineffective	2.6	2.6	12.8	5.1	**9**
Total		**9**	**9**	**10**	**11**	**39**

History Teachers		Previous				**Total**
		Highly Effective	Effective	Partially Effective	Ineffective	
Current	Highly Effective	3.1	12.5	3.1	12.5	**10**
	Effective	6.3	0.0	12.5	3.1	**7**
	Partially Effective	12.5	9.4	0.0	0.0	**7**
	Ineffective	0.0	6.3	9.4	9.4	**8**
Total		**7**	**9**	**8**	**8**	**32**

English Teachers		Previous				**Total**
		Highly Effective	Effective	Partially Effective	Ineffective	
Current	Highly Effective	7.4	11.1	3.7	7.4	**8**
	Effective	7.4	7.4	3.7	3.7	**6**
	Partially Effective	7.4	3.7	3.7	7.4	**6**
	Ineffective	7.4	0.0	11.1	7.4	**7**
Total		**8**	**6**	**6**	**7**	**27**

Appendix H. Transition Matrixes for Consistency of Teacher Value-added Effectiveness Categories Derived from Using One Prior Year and Two Prior Years Combined Test Scores (in Percentages)

Mathematics Teachers		By using one prior year				Total
		Highly Effective	Effective	Partially Effective	Ineffective	
By using two prior years combined	Highly Effective	23.8	0.0	0.0	0.0	**5**
	Effective	0.0	23.8	0.0	0.0	**5**
	Partially Effective	0.0	0.0	23.8	0.0	**5**
	Ineffective	0.0	0.0	0.0	28.6	**6**
Total		**5**	**5**	**5**	**6**	**21**

Turkish Teachers		By using one prior year				Total
		Highly Effective	Effective	Partially Effective	Ineffective	
By using two prior years combined	Highly Effective	21.9	3.1	0.0	0.0	**8**
	Effective	3.1	18.8	3.1	0.0	**8**
	Partially Effective	0.0	6.3	18.8	3.1	**9**
	Ineffective	0.0	0.0	0.0	21.9	**7**
Total		**8**	**9**	**7**	**8**	**32**

Science Teachers		By using one prior year				**Total**
		Highly Effective	Effective	Partially Effective	Ineffective	
By using two prior years combined	Highly Effective	25.6	0.0	0.0	0.0	**10**
	Effective	0.0	25.6	0.0	0.0	**10**
	Partially Effective	0.0	0.0	25.6	0.0	**10**
	Ineffective	0.0	0.0	0.0	23.1	**9**
Total		**10**	**10**	**10**	**9**	**39**

History Teachers		By using one prior year				**Total**
		Highly Effective	Effective	Partially Effective	Ineffective	
By using two prior years combined	Highly Effective	25.0	3.1	0.0	0.0	**9**
	Effective	6.3	12.5	3.1	0.0	**7**
	Partially Effective	0.0	6.3	15.6	3.1	**8**
	Ineffective	0.0	0.0	3.1	21.9	**8**
Total		**10**	**7**	**7**	**8**	**32**

English Teachers		By using one prior year				**Total**
		Highly Effective	Effective	Partially Effective	Ineffective	
By using two prior years combined	Highly Effective	29.6	0.0	0.0	0.0	**8**
	Effective	0.0	18.5	3.7	0.0	**6**
	Partially Effective	0.0	3.7	18.5	0.0	**6**
	Ineffective	0.0	0.0	0.0	25.9	**7**
Total		**8**	**6**	**6**	**7**	**27**

Appendix I. Transition Matrixes for Consistency of Teacher Value-added Effectiveness Categories Comparing with OLS Model (in Percentages)

Mathematics Teachers		Residual Gain (RG)				**Total**
		Highly Effective	Effective	Partially Effective	Ineffective	
Ordinary Least Square (OLS)	Highly Effective	15.7	7.8	1.3	0.0	**57**
	Effective	7.4	9.6	7.0	1.3	**58**
	Partially Effective	1.3	5.7	9.6	8.7	**58**
	Ineffective	0.4	2.2	7.4	14.8	**57**
Total		**57**	**58**	**58**	**57**	**230**

Mathematics Teachers		Hierarchical Linear Model (HLM)				**Total**
		Highly Effective	Effective	Partially Effective	Ineffective	
Ordinary Least Square (OLS)	Highly Effective	22.2	2.6	0.0	0.0	**57**
	Effective	2.6	21.3	1.3	0.0	**58**
	Partially Effective	0.0	1.3	20.0	3.9	**58**
	Ineffective	0.0	0.0	3.9	20.9	**57**
Total		**57**	**58**	**58**	**57**	**230**

Turkish Teachers		Residual Gain (RG)				**Total**
		Highly Effective	Effective	Partially Effective	Ineffective	
Ordinary Least Square (OLS)	Highly Effective	16.8	7.3	0.9	0.0	**58**
	Effective	6.9	9.5	7.8	0.9	**58**
	Partially Effective	0.9	7.3	10.8	6.0	**58**
	Ineffective	0.4	0.9	5.6	18.1	**58**
Total		**58**	**58**	**58**	**58**	**232**

Turkish Teachers		Hierarchical Linear Model (HLM)				**Total**
		Highly Effective	Effective	Partially Effective	Ineffective	
Ordinary Least Square (OLS)	Highly Effective	22.4	2.6	0.0	0.0	**58**
	Effective	2.6	21.6	0.9	0.0	**58**
	Partially Effective	0.0	0.9	19.8	4.3	**58**
	Ineffective	0.0	0.0	4.3	20.7	**58**
Total		**58**	**58**	**58**	**58**	**232**

Science Teachers		Residual Gain (RG)				**Total**
		Highly Effective	Effective	Partially Effective	Ineffective	
Ordinary Least Square (OLS)	Highly Effective	17.2	7.8	0.0	0.0	**51**
	Effective	4.9	10.8	7.8	1.5	**51**
	Partially Effective	2.9	5.4	11.8	4.9	**51**
	Ineffective	0.0	1.0	5.4	18.6	**51**
Total		**51**	**51**	**51**	**51**	**204**

Science Teachers		Hierarchical Linear Model (HLM)				**Total**
		Highly Effective	Effective	Partially Effective	Ineffective	
Ordinary Least Square (OLS)	Highly Effective	22.1	2.9	0.0	0.0	**51**
	Effective	2.9	21.1	1.0	0.0	**51**
	Partially Effective	0.0	1.0	22.1	2.0	**51**
	Ineffective	0.0	0.0	2.0	23.0	**51**
Total		**51**	**51**	**51**	**51**	**204**

History Teachers		Residual Gain (RG)				**Total**
		Highly Effective	Effective	Partially Effective	Ineffective	
Ordinary Least Square (OLS)	Highly Effective	16.1	8.6	0.0	0.0	**43**
	Effective	6.3	9.8	8.6	0.6	**44**
	Partially Effective	2.3	6.9	13.2	2.9	**44**
	Ineffective	0.0	0.0	3.4	21.3	**43**
Total		**43**	**44**	**44**	**43**	**174**

History Teachers		Hierarchical Linear Model (HLM)				**Total**
		Highly Effective	Effective	Partially Effective	Ineffective	
Ordinary Least Square (OLS)	Highly Effective	19.0	5.7	0.0	0.0	**43**
	Effective	5.7	17.2	2.3	0.0	**44**
	Partially Effective	0.0	2.3	20.1	2.9	**44**
	Ineffective	0.0	0.0	2.9	21.8	**43**
Total		**43**	**44**	**44**	**43**	**174**

English Teachers		Residual Gain (RG)				**Total**
		Highly Effective	Effective	Partially Effective	Ineffective	
Ordinary Least Square (OLS)	Highly Effective	17.1	6.4	1.1	0.0	**46**
	Effective	5.9	11.8	7.0	0.5	**47**
	Partially Effective	1.6	5.3	10.7	7.5	**47**
	Ineffective	0.0	1.6	6.4	17.1	**47**
Total		**46**	**47**	**47**	**47**	**187**

English Teachers		Hierarchical Linear Model (HLM)				**Total**
		Highly Effective	Effective	Partially Effective	Ineffective	
Ordinary Least Square (OLS)	Highly Effective	20.9	3.7	0.0	0.0	**46**
	Effective	3.7	18.7	2.7	0.0	**47**
	Partially Effective	0.0	2.7	18.7	3.7	**47**
	Ineffective	0.0	0.0	3.7	21.4	**47**
Total		**46**	**47**	**47**	**47**	**187**

Appendix J. Transition Matrixes for Intrinsically Concordance of Each Model Between Pair Classrooms (in Percentages)

Mathematics Teachers Residual Gain (RG)		Classroom A				**Total**
		Highly Effective	Effective	Partially Effective	Ineffective	
Classroom B	Highly Effective	13.9	5.6	3.9	3.9	**49**
	Effective	4.4	7.2	6.1	5.6	**42**
	Partially Effective	3.9	5.6	7.8	4.4	**39**
	Ineffective	5.6	4.4	8.3	9.4	**50**
Total		**50**	**41**	**47**	**42**	**180**

Mathematics Teachers Ordinary Least Square (OLS)		Classroom A				**Total**
		Highly Effective	Effective	Partially Effective	Ineffective	
Classroom B	Highly Effective	10.0	2.8	7.2	2.2	**40**
	Effective	6.1	8.9	7.2	7.8	**54**
	Partially Effective	7.2	6.1	2.8	6.1	**40**
	Ineffective	4.4	5.6	10.6	5.0	**46**
Total		**50**	**42**	**50**	**38**	**180**

Mathematics Teachers Hierarchical Linear Model (HLM)		Classroom A				**Total**
		Highly Effective	Effective	Partially Effective	Ineffective	
Classroom B	Highly Effective	2.8	3.3	6.1	13.9	**47**
	Effective	5.0	3.9	6.1	7.8	**41**
	Partially Effective	5.0	5.0	2.8	1.7	**26**
	Ineffective	21.7	6.7	5.0	3.3	**66**
Total		**62**	**34**	**36**	**48**	**180**

Turkish Teachers Residual Gain (RG)		Classroom A				**Total**
		Highly Effective	Effective	Partially Effective	Ineffective	
Classroom B	Highly Effective	11.6	7.0	7.0	2.9	**49**
	Effective	5.2	7.0	5.2	4.1	**37**
	Partially Effective	6.4	5.8	5.2	8.1	**44**
	Ineffective	4.7	3.5	9.9	6.4	**42**
Total		**48**	**40**	**47**	**37**	**172**

Turkish Teachers Ordinary Least Square (OLS)		Classroom A				**Total**
		Highly Effective	Effective	Partially Effective	Ineffective	
Classroom B	Highly Effective	8.7	5.2	7.0	3.5	**42**
	Effective	5.2	6.4	6.4	8.1	**45**
	Partially Effective	6.4	5.8	7.6	4.1	**41**
	Ineffective	2.3	8.7	9.3	5.2	**44**
Total		**39**	**45**	**52**	**36**	**172**

Turkish Teachers Hierarchical Linear Model (HLM)		Classroom A				**Total**
		Highly Effective	Effective	Partially Effective	Ineffective	
Classroom B	Highly Effective	2.3	7.0	5.8	15.7	**53**
	Effective	3.5	4.7	8.1	5.8	**38**
	Partially Effective	4.7	7.0	2.3	3.5	**30**
	Ineffective	16.3	8.1	4.7	0.6	**51**
Total		**46**	**46**	**36**	**44**	**172**

Science Teachers Residual Gain (RG)		Classroom A				**Total**
		Highly Effective	Effective	Partially Effective	Ineffective	
Classroom B	Highly Effective	9.0	6.9	3.7	3.7	**44**
	Effective	4.8	6.9	6.4	4.8	**43**
	Partially Effective	5.3	5.9	6.9	8.0	**49**
	Ineffective	4.8	3.2	7.4	12.2	**52**
Total		**45**	**43**	**46**	**54**	**188**

Science Teachers Ordinary Least Square (OLS)		Classroom A				**Total**
		Highly Effective	Effective	Partially Effective	Ineffective	
Classroom B	Highly Effective	9.6	4.8	3.7	3.2	**40**
	Effective	6.4	8.5	7.4	7.4	**56**
	Partially Effective	3.2	4.3	6.4	5.9	**37**
	Ineffective	3.2	8.0	7.4	10.6	**55**
Total		**42**	**48**	**47**	**51**	**188**

Science Teachers Hierarchical Linear Model (HLM)		Classroom A				**Total**
		Highly Effective	Effective	Partially Effective	Ineffective	
Classroom B	Highly Effective	4.3	3.2	4.3	14.9	**50**
	Effective	3.2	5.9	6.9	7.4	**44**
	Partially Effective	4.8	6.4	4.3	3.7	**36**
	Ineffective	13.8	9.0	2.7	5.3	**58**
Total		**49**	**46**	**34**	**59**	**188**

History Teachers Residual Gain (RG)		Classroom A				**Total**
		Highly Effective	Effective	Partially Effective	Ineffective	
Classroom B	Highly Effective	12.0	6.9	5.1	2.8	**58**
	Effective	7.4	5.6	6.0	3.7	**49**
	Partially Effective	3.7	5.1	11.6	3.2	**51**
	Ineffective	2.8	1.9	8.3	13.9	**58**
Total		**56**	**42**	**67**	**51**	**216**

History Teachers Ordinary Least Square (OLS)		Classroom A				**Total**
		Highly Effective	Effective	Partially Effective	Ineffective	
Classroom B	Highly Effective	9.7	6.0	7.4	1.9	**54**
	Effective	7.9	6.5	6.9	3.2	**53**
	Partially Effective	3.7	7.9	10.2	3.2	**54**
	Ineffective	3.7	3.7	6.0	12.0	**55**
Total		**54**	**52**	**66**	**44**	**216**

History Teachers Hierarchical Linear Model (HLM)		Classroom A				**Total**
		Highly Effective	Effective	Partially Effective	Ineffective	
Classroom B	Highly Effective	5.1	4.6	7.4	15.7	**71**
	Effective	1.4	4.6	6.9	2.8	**34**
	Partially Effective	5.1	7.4	5.1	3.7	**46**
	Ineffective	15.7	5.1	5.6	3.7	**65**
Total		**59**	**47**	**54**	**56**	**216**

English Teachers Residual Gain (RG)		Classroom A				**Total**
		Highly Effective	Effective	Partially Effective	Ineffective	
Classroom B	Highly Effective	8.7	6.0	3.8	2.7	**39**
	Effective	6.0	8.7	5.5	2.7	**42**
	Partially Effective	7.7	4.9	8.7	7.1	**52**
	Ineffective	7.7	1.6	8.7	9.3	**50**
Total		**55**	**39**	**49**	**40**	**183**

English Teachers Ordinary Least Square (OLS)		Classroom A				**Total**
		Highly Effective	Effective	Partially Effective	Ineffective	
Classroom B	Highly Effective	7.1	5.5	3.8	2.2	**34**
	Effective	6.6	9.8	4.9	6.0	**50**
	Partially Effective	7.7	6.0	6.6	2.2	**41**
	Ineffective	7.7	6.6	10.4	7.1	**58**
Total		**53**	**51**	**47**	**32**	**183**

English Teachers Hierarchical Linear Model (HLM)		Classroom A				**Total**
		Highly Effective	Effective	Partially Effective	Ineffective	
Classroom B	Highly Effective	2.2	2.2	7.7	9.3	**39**
	Effective	4.4	7.7	6.0	4.9	**42**
	Partially Effective	6.6	5.5	4.4	1.6	**33**
	Ineffective	20.2	6.0	6.0	5.5	**69**
Total		**61**	**39**	**44**	**39**	**183**

END OF THE THESIS

(BLANK PAGE)