University of Cyprus

DEPARTMENT OF EDUCATION

ASSESSING TEACHER EFFECTIVENESS: DEVELOPING A MODEL THAT TAKES INTO ACCOUNT THE CLASSROOM CONTEXT EFFECTS IN MEASURING QUALITY OF TEACHING

DOCTOR OF PHILOSOPHY DISSERTATION

ELENA KOKKINOU

2019

DEPARTMENT OF EDUCATION

ASSESSING TEACHER EFFECTIVENESS: DEVELOPING A MODEL THAT TAKES INTO ACCOUNT THE CLASSROOM CONTEXT EFFECTS IN MEASURING QUALITY OF TEACHING

DOCTOR OF PHILOSOPHY DISSERTATION

ELENA KOKKINOU

A Dissertation Submitted to the University of Cyprus in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

April 2019

VALIDATION PAGE

Doctoral Candidate: Elena Kokkinou

Doctoral Thesis Title: Assessing teacher effectiveness: Developing a model that takes into account the classroom context effects in measuring quality of teaching

The present Doctoral Dissertation was submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy at the ***Department of Education*** *and was approved on the 23rd of April 2019 by the members of the* ***Examination Committee.***

Examination Committee:

Research Supervisor: Leonidas Kyriakides, Professor
Department of Education, University of Cyprus

…………..………………….

Committee Members: Charalambos Charalambous, Assistant Professor
Department of Education, University of Cyprus (Chair)

……………………………

Maria Eliophotou, Professor
Department of Education, University of Cyprus

……………………………

George Papakonstantinou, Professor
Faculty of Philosophy, Pedagogy and Psychology, University of Athens

…………………………..

Galini Rekalidou, Professor
Department of Education Sciences in Early Childhood,
Democritus University of Thrace

…………………………..

DECLARATION OF DOCTORAL CANDIDATE

The present doctoral dissertation was submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy of the University of Cyprus. It is a product of original work of my own, unless otherwise mentioned through references, notes, or any other statements.

Elena Kokkinou

………………………..

Abstract

Most countries around the world use measures of quality of teaching, such as classroom observation, as an essential aspect of their teacher evaluation systems. The focus on teaching for evaluating teachers is justified by the findings of effectiveness studies which reveal that what teachers actually do in the classroom is the most important factor at classroom level associated with student outcomes. Over the last decades several theoretical frameworks have been developed to study and analyse teaching and several studies have been carried out to investigate issues related to the validity and reliability of data on quality of teaching. However, what is not yet clear is whether measuring teaching skills depends on the classroom context, and more specifically if teachers exhibit the same generic teaching skills when they teach in different classrooms.

In this context, this study investigates whether we can generate similar scores about secondary school teachers' generic teaching skills through observing them teaching mathematics in different classrooms of the same age group of students, and through asking students from different classrooms to evaluate the classroom behaviour of their teacher. Concerning its theoretical framework, the study made use of the dynamic model of educational effectiveness, which refers to eight generic classroom-level factors that describe teachers' instructional role. Given that previous studies testing the validity of this model were conducted in primary and pre-primary schools, the current study also aims to identify the effect of each classroom-level factor of the dynamic model and its dimensions on achievement gains in mathematics of secondary students. In addition, by collecting data from more than one classroom of the same teachers, this study aims to distinguish between the classroom and teacher effect and investigate which level can explain more variance in student outcomes.

A stage sampling procedure was used. Specifically, 12 lower secondary schools were initially selected. Then, all teachers of the school sample who taught mathematics in

at least two classrooms of students of the same grade (i.e., grade 7 and/or grade 8) were chosen. About 915 students of the class sample (*n*=57) and 26 teachers participated in this study. Student achievement in mathematics was measured at the beginning and at the end of the school year 2014-2015 by using external forms of assessment. The written tests were developed and validated in a pilot study conducted in the previous year (2013-2014). Teacher behaviour in the classroom was measured by classroom observations using low and high-inference observation instruments and through a student questionnaire.

A generalisability study was conducted in order to examine the consistency of teacher behaviour across different classrooms. The results of these analyses revealed that secondary school teachers behave consistently in different classrooms for most classroom-level factors of the dynamic model irrespective of the instrument used to evaluate them. However, the generalisability study showed that the teacher scores regarding the majority of the dimensions of orientation and dealing with misbehaviour are not generalisable at the teacher level but are generalisable at the classroom level. In addition, it was found that the type of instrument used to measure teaching skills can contribute to whether similar judgements will be produced when the same teacher is evaluated across different classrooms, for some dimensions of some factors (e.g., the frequency dimension of orientation).

Multilevel modelling techniques were also used to identify the extent to which each classroom-level factor is associated with achievement gains in mathematics of secondary school students. The empty model of the multilevel analyses revealed that the teacher level explains more variance in student outcomes than the classroom level. Moreover, the results of these analyses revealed that the classroom-level factors of the dynamic model are relevant for promoting secondary school students learning outcomes too, providing support to the generic nature of these factors in different phases of schooling. Curvilinear relations between some dimensions of some factors (e.g., the focus dimension of questioning

techniques) and student achievement were also identified. Implications of the findings for theory and practice, especially for teacher evaluation and professional development, are drawn and suggestions for further research are provided.

Περίληψη

Οι περισσότερες χώρες του κόσμου χρησιμοποιούν μεθόδους μέτρησης της ποιότητας της διδασκαλίας όπως την παρατήρηση στην τάξη, ως απαραίτητο κομμάτι των συστημάτων αξιολόγησής τους. Η έμφαση που δίνεται στη διδασκαλία για την αξιολόγηση των εκπαιδευτικών δικαιολογείται από τα αποτελέσματα ερευνών, που πραγματοποιήθηκαν στο χώρο της εκπαιδευτικής αποτελεσματικότητας, τα οποία έδειξαν ότι το τι πραγματικά κάνουν οι εκπαιδευτικοί στην τάξη είναι ο πιο σημαντικός παράγοντας στο επίπεδο της τάξης που σχετίζεται με τα μαθησιακά αποτελέσματα. Κατά τις τελευταίες δεκαετίες έχουν αναπτυχθεί αρκετά θεωρητικά πλαίσια για τη μελέτη και την ανάλυση της διδασκαλίας και έχουν πραγματοποιηθεί αρκετές έρευνες που διερεύνησαν θέματα που σχετίζονται με την εγκυρότητα και την αξιοπιστία δεδομένων που αφορούν στην ποιότητα της διδασκαλίας. Ωστόσο, δεν κατέστη ακόμη ξεκάθαρο εάν η μέτρηση των διδακτικών δεξιοτήτων εξαρτάται από το συγκείμενο της τάξης και πιο συγκεκριμένα εάν οι εκπαιδευτικοί παρουσιάζουν τις ίδιες γενικευμένες (generic) δεξιότητες διδασκαλίας όταν διδάσκουν σε διαφορετικές τάξεις.

Στο πλαίσιο αυτό, η παρούσα έρευνα διερευνά κατά πόσο μπορούν να παραχθούν παρόμοια σκορ για τις γενικευμένες διδακτικές δεξιότητες των ίδιων εκπαιδευτικών της δευτεροβάθμιας εκπαίδευσης, μέσω της παρακολούθησης της διδασκαλίας Μαθηματικών σε διαφορετικές τάξεις μαθητών (ίδιας ηλικιακής ομάδας) και μέσω των αξιολογήσεων της συμπεριφοράς των εκπαιδευτικών στην τάξη από τους μαθητές των τάξεων αυτών. Όσον αφορά στο θεωρητικό πλαίσιο της έρευνας, χρησιμοποιήθηκε το δυναμικό μοντέλο της εκπαιδευτικής αποτελεσματικότητας, το οποίο αναφέρεται σε οχτώ γενικευμένους παράγοντες του επιπέδου της τάξης που περιγράφουν το διδακτικό ρόλο των εκπαιδευτικών. Δεδομένου ότι οι προηγούμενες έρευνες που εξέτασαν την εγκυρότητα αυτού του μοντέλου πραγματοποιήθηκαν σε Δημοτικά σχολεία και σε Νηπιαγωγεία, επιπρόσθετος στόχος της παρούσας έρευνας είναι η διαπίστωση της επίδρασης του κάθε

παράγοντα που περιλαμβάνεται στο επίπεδο της τάξης, καθώς και των διαστάσεών του, στην πρόοδο των μαθητών της δευτεροβάθμιας εκπαίδευσης. Επιπρόσθετα, μέσω της συλλογής δεδομένων από περισσότερες από μία τάξεις στις οποίες διδάσκουν οι ίδιοι εκπαιδευτικοί, η έρευνα αυτή στοχεύει επίσης στη διάκριση της επίδρασης του επιπέδου της τάξης από την επίδραση του επιπέδου του εκπαιδευτικού, καθώς επίσης και στον εντοπισμό του επιπέδου που μπορεί να ερμηνεύσει το μεγαλύτερο ποσοστό της διασποράς στα μαθησιακά αποτελέσματα των μαθητών.

Το δείγμα της έρευνας επιλέγηκε με τη βοήθεια της μεθόδου της κατά στάδιο δειγματοληψίας. Συγκεκριμένα, επιλέγηκαν αρχικά 12 Γυμνάσια. Στη συνέχεια, επιλέγηκαν όλοι οι εκπαιδευτικοί των σχολείων που συμμετείχαν στην έρευνα, οι οποίοι δίδασκαν μαθηματικά σε τουλάχιστον δύο τάξεις μαθητών Α’ ή/και Β’ Γυμνασίου. Στην έρευνα αυτή συμμετείχαν 26 εκπαιδευτικοί και 915 μαθητές των τάξεων (*n*=57) των εκπαιδευτικών που συμμετείχαν στην έρευνα. Η επίδοση των μαθητών στα Μαθηματικά μετρήθηκε στην αρχή και στο τέλος της σχολικής χρονιάς 2014-2015 με τη χρήση γραπτών δοκιμίων αξιολόγησης. Τα γραπτά δοκίμια αναπτύχθηκαν και εγκυροποιήθηκαν σε πιλοτική έρευνα που είχε διεξαχθεί την προηγούμενη χρονιά (2013-2014). Η συμπεριφορά των εκπαιδευτικών στην τάξη μετρήθηκε μέσω παρατηρήσεων με τη χρήση εργαλείων παρατήρησης χαμηλού (low-inference) και υψηλού συμπερασμού (high-inference), καθώς και μέσω ερωτηματολογίου που δόθηκε στους μαθητές.

Για τη διερεύνηση της συνέπειας της διδακτικής συμπεριφοράς των εκπαιδευτικών σε διαφορετικές τάξεις διεξήχθη μελέτη γενικευσιμότητας (generalisability study). Τα αποτελέσματα των αναλύσεων αυτών κατέδειξαν ότι οι εκπαιδευτικοί της δευτεροβάθμιας εκπαίδευσης δείχνουν συνέπεια στη διδακτική τους συμπεριφορά σε διαφορετικές τάξεις για τους περισσότερους παράγοντες του επιπέδου της τάξης του δυναμικού μοντέλου, ανεξάρτητα από το εργαλείο που χρησιμοποιήθηκε για τη μέτρησή τους. Ωστόσο, η μελέτη γενικευσιμότητας έδειξε ότι για την πλειοψηφία των διαστάσεων των παραγόντων

του προσανατολισμού και της διαχείρισης της απειθαρχίας, τα σκορ των εκπαιδευτικών είναι γενικεύσιμα στο επίπεδο της τάξης και όχι στο επίπεδο του εκπαιδευτικού. Επιπλέον, φάνηκε ότι όταν ο ίδιος ο εκπαιδευτικός αξιολογείται σε διαφορετικές τάξεις, το είδος του εργαλείου που θα χρησιμοποιηθεί για τη μέτρηση των διδακτικών του δεξιοτήτων μπορεί να επηρεάσει το βαθμό στον οποίο μπορούν να προκύψουν παρόμοια συμπεράσματα για τις δεξιότητές του σχετικά με κάποιες διαστάσεις κάποιων παραγόντων (π.χ., τη διάσταση της συχνότητας του προσανατολισμού).

Για τον προσδιορισμό του βαθμού στον οποίο κάθε παράγοντας του επιπέδου της τάξης σχετίζεται με την πρόοδο των μαθητών της δευτεροβάθμιας εκπαίδευσης στα Μαθηματικά, έγινε χρήση των πολυεπίπεδων μοντέλων (multilevel modelling techniques). Το μηδενικό μοντέλο των πολυεπίπεδων αναλύσεων κατέδειξε ότι το επίπεδο του εκπαιδευτικού μπορεί να ερμηνεύσει μεγαλύτερο ποσοστό διασποράς στα μαθησιακά αποτελέσματα από αυτό που μπορεί να ερμηνεύσει το επίπεδο της τάξης. Ακόμη, τα αποτελέσματα των αναλύσεων αυτών υποστηρίζουν τη γενικευμένη (generic) φύση των παραγόντων του επιπέδου της τάξης του δυναμικού μοντέλου, καθώς κατέδειξαν ότι οι παράγοντες αυτοί σχετίζονται και με την προαγωγή των μαθησιακών αποτελεσμάτων των μαθητών της δευτεροβάθμιας εκπαίδευσης πέραν των μαθητών της πρωτοβάθμιας εκπαίδευσης. Τέλος, διαφάνηκε καμπυλόγραμμη σχέση μεταξύ ορισμένων διαστάσεων ορισμένων παραγόντων με τα μαθησιακά αποτελέσματα (π.χ., τη διάσταση της εστίασης των τεχνικών ερωτήσεων). Στο τελευταίο μέρος της παρούσας εργασίας, γίνεται αναφορά στη συνεισφορά των αποτελεσμάτων της έρευνας στη θεωρία και στην πράξη, ειδικά στην αξιολόγηση των εκπαιδευτικών και στην επαγγελματική τους ανάπτυξη. Επιπλέον, παρουσιάζονται εισηγήσεις για περαιτέρω έρευνα.

Acknowledgments

Coming to the end of this journey, I would like to express my sincere gratitude to a number of people without whom this dissertation would not have been completed. First and foremost, I would like to thank my advisor Dr. Leonidas Kyriakides for the continuous support he provided me throughout each stage of my PhD study. His belief in my abilities helped me grow not only as a researcher but also as a person. Without his constant guidance, patience, encouragement and immense knowledge, this thesis would not have been completed or written.

I would also like to thank the rest of the members of my committee: Dr. Charalambos Charalambous, Dr. Maria Eliophotou, Dr. George Papakonstantinou and Dr. Galini Rekalidou, for their time and for providing me with valuable comments and constructive feedback.

My sincere thanks also goes to Florentina Kazantzi who supported me in the process of linguistic editing and who was always willing to answer my questions. I would also like to thank my PhD classmate Ioannis Ioannou who supported me in the process of developing the mathematics achievement tests. I am also grateful to my PhD classmate Andria Dimosthenous, my ‘fellow-traveller’, who made this journey more enjoyable. Her support and friendship during the years of my PhD study have been invaluable.

Last but not least, I would like to express my deepest gratitude to my family and my friends for their endless support, continued patience, understanding and encouragement all these years.

To my parents

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

CHAPTER 1

INTRODUCTION TO THE STUDY

In this chapter the rationale of the study is explained, and an overview of the thesis is given. Specifically, an overview of research in teacher evaluation and educational effectiveness is provided, which helps us identify the importance of searching whether teachers behave similarly when they teach in different classrooms. Additionally, key terms for this research are defined and the research aims of this study are presented. Then, an overview of the way in which this study was conducted is provided and the theoretical contribution of the study and its significance is outlined. Finally, the structure of this thesis is presented in order to facilitate further reading.

Introduction

Findings of effectiveness studies reveal that the classroom level can explain more variance in student learning outcomes than the school level (Muijs et al., 2014). They have also shown that what teachers actually do in the classroom is the most important factor at classroom level associated with student achievement and that teachers vary considerably in their effectiveness (Kyriakides, Christoforou, & Charalambous, 2013; Rivkin, Hanushek, & Kain, 2005; Wright, Horn, & Sanders, 1997). Consequently, teaching practice has become integrated into theoretical models of educational effectiveness which attempt to identify teacher factors associated with student learning outcomes, such as the dynamic model of educational effectiveness (Creemers & Kyriakides, 2008). These findings also seem to be in line with the fact that many countries around the world use measures of quality of teaching as a key component of their teacher evaluation systems (Mihaly & McCaffrey, 2014; OECD, 2013).

In educational effectiveness studies as well as in teacher evaluation systems, classroom observations and/or student questionnaires are usually used as sources of data to measure teacher in-class behaviour (e.g., Doherty & Jacobs, 2013; Kyriakides & Creeemers, 2009; MET Project, 2012). Although several studies have been carried out in regard to the generalisability of observational data (i.e., number of observations per teacher that are needed to make a reliable generalization of a classroom teacher's practice; e.g., Hill, Charalambous, & Kraft, 2012; Praetorius, Pauli, Reusser, Rakoczy, & Klieme, 2014) and the quality of the sources and instruments used to gather data on quality of teaching (e.g., Kyriakides et al., 2014), very little is known about whether measuring teaching skills depends on the classroom context. Some authors (e.g., Pacheco, 2009; Smylie, Miller, & Westbrook, 2008; Spilt, Leflot, Onghena, & Colpin, 2016) argue that classroom context may influence teachers' practices. In other words, individual teachers may respond to different classroom contexts differently and may not show consistency in their teaching behaviour when they teach in different classrooms.

Before proceeding to the importance of examining the classroom context effect on teaching skills, it is necessary here to clarify exactly what is meant by the term "classroom context". It seems that a common definition for classroom context among scholars and researchers does not exist. Particularly, as Turner and Meyer (2000) observe in their literature review, there are nearly as many definitions as studies. Definitions vary extensively depending on the perspective (i.e., educational, sociological, psychological or anthropological) that they have been examined. This study will use the definition provided by Steinberg and Garrett (2016) who saw classroom context as "the settings in which teachers work and the students that they teach" (p. 293). The classroom context can include variables such as the students' socioeconomic backgrounds and student prior achievement (see Chapter 2).

The Importance of Examining the Classroom Context Effect on Teaching Skills for Teacher Evaluation and Educational Effectiveness Research

Teachers are expected to differentiate their instruction according to the specific needs of the students of each class. According to Creemers and Kyriakides (2008), the adaptation to the specific needs of each student or group of students is expected to increase the successful implementation of a teaching factor and eventually maximize its effect on student learning outcomes. However, this study is concerned with the skills of teachers and examines whether teachers are able to demonstrate the same generic teaching skills in teaching different groups of students (e.g., questioning techniques, providing students with application opportunities etc.). This question has important policy implications for establishing valid and fair teacher evaluation systems, especially since observation scores may inform decisions about teachers' hiring, retention, bonus and dismissal (Master, 2014; Mihaly & McCaffrey, 2014). If teachers exhibit the same teaching skills when they teach in different classrooms, then their observation scores will show little variability among different classrooms. However, if teachers are not in a position to demonstrate their skills in specific contexts then observing them teaching in only one classroom may lead to misleading conclusions about their teaching skills with real implications for personnel decisions. This may essentially call into question an evaluation system's ability to effectively and equitably take decisions about teachers' hiring, retention, promotion, improvement, rewards and dismissal.

According to Patrick and Mantzicopoulos (2016), the evaluation process and outcomes are needed to be viewed by teachers as fair and accurate and that these views are backed by strong evidence so as to avoid resentment and dissatisfaction that can corrode workplace morale and cooperation. Recent studies (e.g., Chaplin, Gill, Thompkins, &Miller, 2014; Whitehurst, Chingos, & Lindquist, 2014) have attempted to investigate the extent to which measuring teaching skills by classroom observations and/or student questionnaire is influenced by classroom context variables, such as student achievement or student socioeconomic status. However, most of these studies suffered from some serious methodological weaknesses, as data have been obtained from a single class per teacher per year. Thus, we are unable to determine whether differences in observational ratings are related to student characteristics or to the non-random sorting of teachers to classes of students.

Particular emphasis in the present study is given to students' misbehaviour incidents, where several studies have revealed that teachers consider them as a big problem for their teaching (e.g., OECD, 2009; Public Agenda, 2004). In addition, many studies have shown that teachers tend to attribute the cause of misbehaviour to students or family rather than teaching-related factors (Baron, 1990; Ho, 2004; Koutrouba, 2013; Kyriacou & Martin, 2010). If the teachers are right, then misbehaviour incidents may appear in one classroom and not in another, not as a result of who is the teacher of the classroom, but of who are the students of this classroom. Having this in mind, the factor of student misbehaviour could be seen as a contextual factor which could affect the consistency of teacher behaviour in different classrooms and teacher effects on student achievement. As Marzano, Marzano and Pickering (2003) claim, in situations that students are disorderly and disrespectful and no apparent rules and procedures guide behaviour, teachers struggle to teach and students probably learn much less than they should. Thus, we assume that in those classrooms where many misbehaviour problems may occur, a teacher will not be able to demonstrate her /his other teaching skills unless she/he is good in dealing with misbehaviour. On the other hand, if a teacher is not good in dealing with misbehaviour and she/ he works in a classroom with only a few misbehaviour problems, she/he will not have a problem with demonstrating her/ his other teaching skills.

The question of whether teachers exhibit the same teaching skills when they teach in different classrooms is more relevant when generic factors are used to evaluate them.

Therefore, this research question can help us further test the validity of theoretical frameworks of Educational Effectiveness Research (EER) which refer to generic teaching skills. The generic character of these skills is usually examined by looking at the extent to which these skills are associated with student achievement gains in different subjects and age group of students (e.g., Kyriakides & Creemers, 2009; Seidel & Shavelson, 2007). However, if teacher behaviour in different classrooms varies, then the generic nature of these skills could be questioned. Thus, results about the consistency of teacher behaviour across different classrooms may have implications for differentiated teacher effectiveness (Campbell, Kyriakides, Muijs, & Robinson, 2004). If we found that teacher behaviour in different classrooms is inconsistent, then this gives support to differentiated factors and not to generic factors.

The theoretical framework concerning the selection of teaching skills upon which this study was based, is the dynamic model of educational effectiveness (Creemers & Kyriakides, 2008). The dynamic model is considered as one of the most influential theoretical models of EER (Heck & Moriyama, 2010; Sammons, 2009) and it provides a clear definition of quality of teaching. This model is multilevel in nature and refers at classroom/teacher level to eight factors that are associated with teacher behaviour in the classroom (see Chapter 2). Longitudinal studies testing the validity of the dynamic model demonstrated that the eight classroom-level factors of the model and their dimensions are associated with different types of learning outcomes of students, in different phases of schooling and in different countries (Creemers & Kyriakides, 2015a; Panayiotou et al., 2014). It was therefore argued that these factors can be considered generic. However, these studies took place at primary and pre-primary school level and the extent to which the classroom context affects teacher behaviour could not be examined since teachers who participated taught in a single class only. In this context, this study investigates the extent to which secondary school teachers exhibit the same generic teaching skills (based on the classroom-level factors of the dynamic model) when they teach mathematics in different classrooms.

Investigating teacher behaviour in different classrooms may also provide implications for research comparing teacher and classroom effects in secondary schools. As mentioned before, EER reveals that the classroom level can explain more variance in student learning outcomes than the school level and that a large proportion of this classroom level variance can be explained by teacher behaviour in the classroom. We could also distinguish the classroom and teacher level and compare their effects on student learning outcomes, at the level of secondary education, by collecting data from more than one classroom of the same teacher. Examining the extent to which teachers behave consistently in different classrooms may help us explain why the teacher or the classroom level matters more. If teachers behave similarly (in terms of the classroom-level factors of the dynamic model) in different classrooms, then we expect that the teacher level will explain more variance in student achievement than the classroom level, considering that the classroom-level factors will be associated with the achievement of secondary school students. However, thus far, emphasis is given to the effect that the department has at the level of secondary education (Ko, Hallinger, & Walker, 2015; Ko, Sammons, & Bakkum, 2016; Sammons, Thomas, & Mortimore, 1997) and we are not aware of studies comparing the teacher and classroom level effects.

Research Aims

The main aim of this study is to investigate whether secondary school teachers exhibit the same generic teaching skills when they teach mathematics in different classrooms (of the same age group of students) within a school year. In other words, this study aims to examine whether and to what extent teacher behaviour in the classroom (based on the classroom-level factors of the dynamic model) is affected by the classroom context. It is important to clarify here that teachers are not expected to teach in the same way and provide the same activities in all the classes that they teach. As discussed before in this chapter, the focus of this study is on the teaching skills of teachers (e.g., if they provide students with application opportunities in all the classes that they teach). Given that several methods and instruments for measuring quality of teaching exist, whether the findings are differentiated according to the instrument that is used (i.e., low and high-inference observation instruments and student questionnaire), is also investigated.

In addition, since the studies proving support to the classroom-level factors of the dynamic model took place at primary and pre-primary school level, this study aims to identify the effect of the classroom-level factors on achievement in mathematics of secondary school students too. Furthermore, due to the fact that most previous research studies gathered data on teacher behaviour and student achievement from only one classroom per teacher, it was not possible to distinguish two different levels (i.e., teacher and classroom) and compare the teacher and classroom level effects on student achievement. By collecting data from more than one classroom of the same teacher, this study aims to distinguish between the classroom and teacher effect and explore which level can explain more variance in student achievement of secondary school students.

Taking all the above into account, the main focus of the study is on the investigation of the consistency of teacher behaviour in different classrooms. More precisely, this study aims to answer the research questions that follow:

1. Do secondary school teachers exhibit the same generic teaching skills (based on the classroom-level factors of the dynamic model) when they teach mathematics in different classrooms (of the same age group of students) within a school year? In order to answer this question we investigate the extent to which the observation scores and/or the student questionnaire scores per factor and dimension can be aggregated at the teacher level irrespective of the class that teachers have to teach.

2. Does the type of instrument used to measure teaching skills (i.e., high, low-inference observation instruments and student questionnaire) contribute to whether similar judgments are produced when the same teacher is evaluated across different classrooms?

By collecting data from secondary schools and in particular from more than one classroom of the same teacher, this study also aims to answer the following research questions:

3. To which extent are the classroom-level factors of the dynamic model and their dimensions associated with student achievement in mathematics of secondary school students in Cyprus?
4. Which level explains more variance in student achievement in mathematics of secondary school students, the teacher or the classroom level?

Study Summary

To answer the above research questions, quantitative research methods were used. Stage sampling procedure (Cohen, Manion, & Morrison, 2007) was used to select, at the first stage, 13 Greek Cypriot lower secondary schools and 12 agreed to participate. Then, 26 teachers of the school sample who taught mathematics in at least two classes of the same age group of students (grade 7 or grade 8) and who agreed to participate, as well as 915 students of the class sample (*n*=57), participated in this study. All classes were mixed ability. Secondary schools were selected because there are more possibilities to evaluate the teachers in different classrooms, unlike primary schools.

Data on student achievement were collected at the beginning and at the end of the school year 2014-2015 by using external forms of assessment that are designed to assess knowledge and skills in mathematics. The written tests were developed and validated in a pilot study conducted in the previous year (2013-2014) by taking into consideration the national curriculum of Cyprus for grades 6-8. The extended Logistic Model of Rasch (Andrich, 1988) was used to analyse the data that emerged from each test. Longitudinal research design was chosen for conducting this study as one of its main aims was to investigate the effect of the classroom-level factors included in the dynamic model of educational effectiveness on student achievement of secondary school students. Collecting data from at least two phases may help to draw reliable conclusions regarding relations among factors and outcomes. Permission to collect data was obtained from parents, teachers and schools and all the participants involved were informed that confidentiality would be guaranteed and kept throughout the procedure.

Information was also collected on student gender and ethnicity: students', fathers' and mothers' country of birth and language that students speak at home from a short questionnaire included in the written tests of mathematics. Information regarding students with special educational needs (SEN) of the sample was collected by teachers.

Data about the skills of each teacher were gathered from all his/her classrooms of grade 7 or grade 8 by using external observations and a student questionnaire. Specifically, two observations in each class of the teacher sample (*n*=114) were conducted by a well-trained external observer with the use of one high and two low-inference observation instruments. A questionnaire was administered to the students of all classes of grade 7 and 8 in order to gather data on their teacher's instructional behaviour. The student questionnaire generates data for all the classroom-level factors of the dynamic model. All the observation instruments generate data for all the factors except the assessment. All the instruments have been used and validated in various studies testing the validity of the dynamic model (e.g., Kyriakides & Creemers, 2008; 2009). Some minor amendments were made to adapt the questionnaire to the context of teaching mathematics at secondary school level.

For the purpose of data analysis, a generalisability study (Marcoulides & Kyriakides, 2010) was conducted in order to examine whether similar ratings or judgments were produced when the same teacher was evaluated across different classrooms. Finally, multilevel modelling techniques (Snijders & Bosker, 1999) were used to investigate whether the teacher or the classroom level explains more variance in student achievement, and to search for the effect of the classroom-level factors on student achievement in mathematics, by analyzing the data that emerged from the high and low-inference observation instruments and the student questionnaire.

Contribution to the Theory

This study could provide new insights regarding specific issues such as whether and to what extent the classroom context affects teacher behaviour in the classroom and teacher effects on student achievement. Most of the research that was conducted in recent years mainly examined teacher skills in only one class, ignoring the effect that the classroom context may have upon teacher's instructional behaviour. Therefore, this study may help us further test the validity of theoretical frameworks of EER which refer to generic teaching skills and specifically the generic nature of the classroom-level factors of the dynamic model of educational effectiveness.

Finally, given that previous studies testing the validity of the dynamic model of educational effectiveness were conducted in primary and pre-primary schools, this study is the first attempt to find out whether the classroom-level factors are relevant for promoting secondary school student learning outcomes too.

Significance of the Study

This research is expected to have significant implications for teacher evaluation and EER. The findings may offer guidance to those collecting and interpreting data on teacher in-class behaviour based on classroom observations and/or student questionnaire, either for research or for formative and especially summative purposes of teacher evaluation. Specifically, this study may provide evidence on whether it is necessary to observe teachers teaching different groups of students before we draw conclusions about their teaching skills, or some of them. Moreover, this study could provide evidence on whether the type of instrument used to measure teaching skills (i.e. high, low-inference observation instrument and student questionnaire) could contribute to whether similar judgments will be produced when the same teacher is evaluated across different classrooms. Depending on the results, the fairness and the validity of many current teacher evaluation practices may be questioned. Consequently, this research is not expected to contribute only to theory, but also to policies that can be developed in various countries on issues related to teacher evaluation and evaluation of teaching. Particularly, in Cyprus the results of this study will be significantly relevant as classroom observations are being used by inspectors, since 1976, to evaluate teachers for summative and formative purposes of teacher evaluation (Kyriakides & Campbell, 2003).

In addition, the findings of this study may also be informative to policymakers when design and implement teacher incentive pay programmes that aim to reward teachers for excellent teaching based on performance evaluation system (i.e., performance-based pay plans/ merit-pay plans; e.g., Stedman & McCallion, 2001; Thomas, 1984), and/or reward teachers for acquiring and demonstrating specific knowledge and skills linked to improving student performance (i.e., knowledge- and skills-based compensation systems; e.g., Milanowski, 2002; Odden, Kelley, Heneman, & Milanowski, 2001; and more comprehensive model of teacher pay such as Denver's ProComp; see Koppich, 2008).

These kinds of programmes link teachers' salaries or financial rewards to teacher performance and one of the most commonly used method for evaluating teacher performance, in these programmes, is observation of teaching by administrators or peers (Odden et al., 2001; Stedman & McCallion, 2001).

As discussed earlier, emphasis in this study is given to the effect of misbehaviour incidents on teaching quality. By acknowledging whether and to what extent student misbehaviour may affect the teacher behaviour in classroom, we may be able to make suggestions for the creation of intervention programs that will aim to improve the strategies used by teachers for addressing discipline problems. Training in these strategies may lead to the change of student behaviour and consequently may increase student achievement, a fact that is also supported from the results of several interventions and classroom management programs (Evertson, 1995; Lassen, Steele, & Sailor, 2006; Muscott, Mann, & LeBrun, 2008; Raver et al., 2009; Scott, White, Algozzine, & Algozzine, 2009).

Finally, by identifying which classroom-level factors of the dynamic model may affect student outcomes of secondary school students, implications for policymakers can be drawn. Specifically, we can identify practices that are effective and contribute to the improvement of educational quality in reference to higher average student achievement. These effective practices can be taken into account by policymakers to establish teacher evaluation criteria and form improvement action plans. For instance, a classroom-level factor that is found to have a relatively large impact on student outcomes of secondary school students could constitute a basis upon which evaluation criteria could be established and it might be a priority for improvement in case its functioning is not satisfactory. On the contrary, a factor that is not found to be related to student achievement might not be a priority for creating improvement action plans.

Thesis Structure

The overall structure of the thesis takes the form of five chapters. The first chapter is introductory and presents the research problem addressed and its background, as well as the research questions this study aims to answer. In addition, this chapter highlights the scientific and practical relevance of the study in the field of teacher evaluation and EER. The second chapter aims to provide a literature review of the fundamental concepts and issues related to the study's purpose. Specifically, the basic aspects that need to be considered when developing a comprehensive teacher evaluation system are discussed. Then, a historical overview of TER is provided in order to answer the question of what constitutes effective teaching. Furthermore, the theoretical framework upon which the study is based is presented and described in detail. Moving on, the next section presents a critical review of studies investigating whether measuring teaching skills by classroom observations is influenced by classroom context variables. Moreover, the reasons for choosing misbehaviour as one of the basic contextual factors are discussed. The chapter ends with a summary of the main conclusions drawn from the literature review, together with the research agenda.

The third chapter is concerned with the research methodology used for this study. In Chapter 3, the processes of sampling and data collection are described with particular reference to the data collection instruments and the statistical techniques used. In addition, the main limitations of the study are recognized and discussed in the last section of the third chapter. Continuing, Chapter 4 provides the analysis of the data collected during the study and the research results. The analysis was made so as to provide answers to the research questions of the study, which are presented in the first chapter. Finally, Chapter 5 presents a discussion of the results that occur from the analysis, in accordance to each research question and to the overall aims of this study. Implications of findings for theory, policy and practice are also drawn and suggestions for further research are provided at the end of Chapter 5

CHAPTER 2

LITERATURE REVIEW

Introduction

The purpose of this literature review is to provide the theoretical framework of this study. It also aims to demonstrate links with previous work conducted in the field of teacher evaluation and EER. Through a critical literature review, a framework for the investigation of the research problem and questions stated in Chapter 1 is created. Therefore, this chapter focuses on the provision of the available literature within and across the fields of teacher evaluation and EER, highlighting the need of examining whether teachers exhibit the same generic teaching skills when they teach in different classrooms and how this can have any implications for teacher evaluation policies.

The first section of this chapter synthesizes research and thinking about teacher evaluation. Specifically, the basic aspects that need to be taken into account when developing a comprehensive teacher evaluation system are discussed, giving particular emphasis to the evaluation criteria and to the quality of the sources used to collect relevant data. Then, the definition of effective teaching is examined through a historical overview of TER where the different phases are discussed to demonstrate the growth on the way that effective teaching has been approached through the years. In addition, the rationale of the main models of EER is described. Moving on, the next section presents and describes in detail the theoretical framework used in this study (the dynamic model of educational effectiveness) by giving particular emphasis to the classroom-level factors of this model. Recognizing the need of examining the extent to which the classroom context affects teacher behaviour, the next section presents a critical review of studies investigating whether measuring teaching skills by classroom observations is influenced by classroom context variables. Given that in this study particular emphasis is given to the effect of misbehaviour incidents on teaching quality, in the section that follows, the reasons for choosing misbehaviour as one of the basic contextual factors are discussed. Finally, a summary of the main conclusions drawn from the literature review, together with the research agenda for the present study, are provided in the last section.

Teacher Evaluation

Teacher evaluation is not something new in the educational landscape (Ellett, 1997; Kyriakides & Campbell, 2003*)*. For several reasons, sometimes for professional development, sometimes for accountability and often for both, teacher evaluation has come to be an accepted and expected part to the field of education (Stronge & Tucker, 2003).

According to Stronge (2006), a conceptually sound, well designed and properly implemented teacher evaluation system is a key element of an effective school. However, over the years, the way that teacher evaluation systems have been both designed and implemented in many educational systems has been criticized by a number of writers (e.g., Danielson & McGreal, 2000; Ellett & Garland, 1987; Gitomer & Bell, 2013; Kyriakides, Charalambous, & Demetriou, 2006; Loup, Garland, Ellett, & Rugutt, 1996; The World Bank, 2014). Some problems of teacher evaluation systems that have been reported in the literature are the existence of poor practices and inadequate materials that fail to distinguish good from poor performers and the fact that teacher evaluation very often has been viewed as a superficial function that has lost its meaning rather than as a means for growth and improvement (Peterson, 2000; Stronge, 1997; Stronge & Tucker, 2003). For instance, a survey of approximately 15000 teachers and 1300 administrators that was conducted in the USA (Weisberg, Sexton, Mulhern, & Keeling, 2009) has shown that most teachers receive one of the top two ratings and less than 1% are rated as unsatisfactory. A similar situation to the one described in the USA can be identified in Cyprus where the great majority of teachers are awarded by their inspectors with very high grades (i.e., 35-37 points out of 40) and no teacher has been evaluated as unsatisfactory since 1976 (The World Bank, 2014). This is a cause for considerable concern because poor performance goes unaddressed, excellence goes unrecognized and development is neglected. Another problem of some teacher evaluation systems is that they have not been greatly informed nor influenced by current research into teacher effectiveness and state-of-the-art knowledge bases on teacher evaluation (Danielson & McGreal, 2000; Kyriakides & Campbell, 2003). An exemplary example of this is the case of Cyprus where the existing system is in place since 1976 without any considerable changes (Kyriakides, 2016, The World Bank, 2014).

In recent years, there is renewed interest worldwide on issues related to teacher evaluation and on how countries can develop valid and reliable teacher evaluation systems in order to improve the quality of education (Flores, 2012; Kyriakides & Demetriou, 2007; Liu& Zhao, 2013). This interest comes to a large extent from the realization of the importance of teachers' role on students' learning and to the success of any educational reform effort (Darling-Hammond, 2007; OECD, 2005; Stronge, 2002; 2006). As Stronge and Tucker (2003) point out, "without high quality evaluation systems, we cannot know if we have high quality teachers" (p. 3). Therefore, the establishment of an effective teacher evaluation system is a challenge for researchers but also for policymakers and educational practitioners around the world.

Developing a Comprehensive Teacher Evaluation System

The three basic aspects that need to be taken into account when developing a comprehensive teacher evaluation system are recognized as follows: a) the evaluation purposes, b) the performance criteria and c) the evaluation procedures and sources that will be used for collecting relevant data, analyzing them and interpreting the results (Ellett, Wren, Callender, Loup, & Liu, 1996; Iwanicki, 1990).

Evaluation Purposes

The evaluation purposes state why teachers are evaluated and are the foundation of the teacher evaluation process, as they have a direct effect on the determination of the performance criteria, the selection of evaluation procedures and the interpretation of results (Iwanicki, 1990). Although lists of purposes reported in the literature vary in the content and length (see for example Peterson, 2000; Stronge & Tucker, 2003), they can be divided into two broad categories: those purposes defined as summative (for the purpose of making significant decisions like dismissing incompetent teachers or retaining teachers) and those defined as formative (for the purpose of improving the professional skills of teachers) (Danielson & McGreal, 2000; OECD, 2013). However, the most cited purposes of teacher evaluation are accountability and performance improvement. The accountability purpose (defined as summative in nature) illustrates the need to determine the competence of teachers in order to ensure that services delivered by them are effective and safe. The performance improvement purpose (considered as formative in nature) illustrates the need for professional development of the individual teacher and involves helping teachers to learn about, reflect on, and improve their practice (Stronge, 2006; Stronge & Tucker, 2003). By recognizing individual teachers' strengths and weaknesses, teachers and school leaders can make better informed choices about the professional-development activities that best serve teachers' needs (OECD, 2013).These two broad purposes are not competing, but they are supportive functions of evaluation systems, which are necessary for improvement of educational service delivery (Stronge, 1995). Both accountability and personal growth dimensions are not only desirable to be included in teacher evaluation systems, but they are also essential for evaluation to serve the needs of individual teachers and the school and community at large (Stronge, 1997). Nevertheless, the summative and formative purposes of evaluation are practically impossible to be achieved within a single evaluation system, as the determination of evaluation purposes has an influence on the design of evaluation instruments, their administration and the interpretation of results. Therefore, in order for teacher evaluation systems to serve these two broad purposes, different mechanisms for both of them must be established (Kyriakides & Demetriou, 2007). In addition, in order for teacher evaluation systems to serve the summative and formative evaluation, there must be a rational link between them and this must not allow the summative function of evaluation to dominate the formative function. This link cannot be established if the criteria for both formative and summative evaluations are not based on the same theoretical framework regarding what constitutes an effective teacher (Kyriakides et al., 2006).

Evaluation Criteria

The criteria determine what is expected of teachers in their professional roles (Iwanicki, 1990). It is important for the evaluation criteria to be clear and understandable in order to motivate teachers, otherwise they would not know what is expected from them (Kelly, Ang, Chong, & Hu, 2008). The existence of clear criteria, which are consistently applied by evaluators, is the necessary basis of good practice in teacher evaluation (Santiago & Benavides, 2009). Although the evaluation criteria are a basic aspect that needs to be taken into account when developing a comprehensive teacher evaluation system, there are no universally accepted criteria for measuring teacher effectiveness so far. The criteria used for teacher evaluation differ from country to country (Brandt, Thomas, & Burke, 2008; Eurydice, 2008; OECD, 2013).

As Kyriakides et al. (2006) argue, teacher effectiveness research (TER) and in particular its main theoretical models, could be used as a basis upon which evaluation criteria could be established. Specifically, the seven models (i.e., goal and task model, resource utilization model, working process model, school constituencies satisfaction model, accountability model, absence of problems model and continuous learning model) proposed by Cheng and Tsui (1999) for understanding and ensuring teacher effectiveness could be utilized as sources of developing different evaluation criteria. Each model represents an important perspective that describes and emphasizes some factors which are tightly linked to teachers' performance and contribution in a school. In general, the goal and task model expects teachers to achieve planned goals and assigned tasks in congruence with school goals. The resource utilization model anticipates teachers to use the allocated resources effectively and if needed, to acquire additional resources to perform their job. The working-process model emphasises teachers' contribution to effective teaching and working process. The school constituencies satisfaction model anticipates teachers to satisfy important school constituencies' expectations and demands. The accountability model emphasises teachers' accountability and professional reputation. The absence of problems model expects teachers to identify and avoid possible problems, weakness and dysfunction in teaching and work. Finally, the continuous learning model expects teachers to adapt to the challenges from changing environment (external and internal teaching contexts) and develop themselves through continuous learning (Cheng & Tsui, 1999).

Most countries around the world seem to adopt mainly the working process model but some other countries adopt the goal and task model as well (Doherty & Jacobs, 2015; Kyriakides & Campbell, 2003; OECD, 2013). Thus, several issues related to the use of each of these models for developing criteria for teacher evaluation are discussed below.

Teacher evaluation based on student achievement.

Student learning outcomes are used by some teacher evaluation systems, such as Florida and Washington DC, as sources of evidence for teacher evaluation (Florida Department of Education, n.d.; Lewin, 2010). Particularly, in 2015, 43 states of the USA required measures of student achievement to be included in teacher evaluations (Doherty &

Jacobs, 2015).This kind of evaluation is mainly summative, as teachers receive limited informative feedback regarding the strengths and weaknesses of their teaching (Smith, 2005).

The use of student learning outcomes as an indicator of teacher effectiveness reminds us of the goal and task model. According to this model, a teacher is effective when he/she can achieve the programmed goals and assigned tasks in compliance with school goals. Thus, the extent to which the goals and tasks have been achieved, is often considered as a measure of teacher effectiveness. One of the examples of teacher effectiveness indicators, regarding this model, is student learning outcomes (like the academic achievement in public examinations) (Cheng & Tsui, 1999).

Student achievement is an appealing measure to evaluate teaching performance, as the primary goal of teaching is to improve student learning. Various stakeholder groups in the USA support the idea that learning can be measured adequately through student standardized tests. Supporters of teacher evaluation that is based on student achievement often claim that this kind of evaluation could improve student learning by motivating teachers and by providing information they can use to adapt their teaching. Also, advocates support that student achievement tests could provide accurate information for decision makers to use in placing, supporting and rewarding teachers (Hamilton, 2012). However, the use of student outcomes for teacher evaluation faces considerable statistical challenges and for this reason it has been the source of criticism by several scholars and researchers (e.g., Andrejko, 2004; Braun, 2005; Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012; Goe, 2007; Isoré, 2009; McCaffrey, Lockwood, Louis, Koretz, & Hamilton, 2004; Raundenbush, 2004; Torff & Sessions, 2009). Particularly, there is an emerging consensus in the literature about the fact that student test scores alone (even when value-added modelling is employed) are not adequately reliable and valid indicators of teacher effectiveness to be used in high-stakes personnel decisions (Baker et al., 2010;

Chester & Zelman, 2009). Apart from the concerns about statistical methodology, many scholars suggest that there are also other practical and policy reasons regarding the inappropriateness of the use of student test scores for the evaluation of teachers (Liang, 2013; Stein & Matsumura, 2009). For example, negative effects on teaching and learning may result from evaluating teachers based only on student achievement, such as focus only on the core subjects (e.g., maths instead of arts) and formats which are tested (see Baker et al., 2010; Smith, 2005; Stein & Matsumura, 2009).

Teacher evaluation based on quality of teaching.

Given that there are many caveats against heavy reliance on student test scores to evaluate teachers, there is a broad agreement to evaluate teachers preferably for their teaching practices (Ingvarson, Kleinhenz, & Wilkinson, 2007; Kelly, 2012; Liang, 2013). As Hill and Herlihy (2011) argue, the focus on teaching can create incentives for teachers to improve a factor that they directly control, unlike student outcomes that may be mediated by external factors. Thus, the emphasis on quality of teaching may help policymakers encourage the improvement of teaching. Moreover, effectiveness studies reveal that what teachers actually do in the classroom is the most important factor at teacher level associated with student achievement (Kyriakides et al., 2013; Muijs et al., 2014) rather than factors other than classroom behaviour, such as their beliefs and their background qualifications (Creemers, Kyriakides, & Antoniou, 2013; Palardy & Rumberger, 2008). Therefore, the improvement of quality of teaching may lead to the enhancement of student learning outcomes.

Many countries around the world, including Cyprus, use measures of quality of teaching, such as classroom observations, as an important aspect of their teacher evaluation systems (Doherty & Jacobs, 2013; Kyriakides & Campbell, 2003; Mihaly & McCaffrey, 2014; OECD, 2013). The focus on teaching for setting criteria for teacher evaluation is in line with the main principles of the working process model of teacher effectiveness (Cheng & Tsui, 1999). This model assumes that effective teaching and functional learning processes help teachers to effectively accomplish their assigned tasks, resulting in significant student outcomes. Thus, it assumes that teachers can be considered effective if they can assure the quality of teaching and working processes. It is important to mention that a study conducted in Cyprus by Kyriakides et al. (2006) has shown that primary teachers considered this model as the most appropriate for formative and summative evaluation purposes.

However, when teacher in-class behaviour is used as a criterion for evaluating their effectiveness several issues may emerge. Among these issues are the following: a) the quality of the sources of data and generalisability of observational data; b) what constitutes effective teaching; and c) whether the same teachers exhibit the same generic teaching skills when they teach in different classrooms. Each of these issues will be briefly discussed on the following pages. It should be noted that if teacher evaluation systems fail to give teachers high-quality feedback based on accurate assessments of their teaching, then teaching and learning will not improve. The quality measurement of teaching is also particularly important for school administrators in order not to be left blind when making critical personnel and assignment decisions (Archer, Kerr, & Pianta, 2014).

Quality of the Sources of Data

When quality of teaching is used as a criterion for evaluating teachers, valid and reliable measures of teaching are needed. Several methods for measuring quality of teaching have been reported in the literature, such as classroom observations, (e.g., Pianta & Hamre, 2009), student ratings (e.g., Kyriakides et al., 2014), teacher logs (e.g., Rowan & Correnti, 2009), collecting classroom artefacts such as lesson plans and classroom assignments (e.g., Matsumura, Garnier, Pascal, & Valdés, 2002), teacher self-ratings (e.g., Wilkerson, Manatt, Rogers, & Maughan, 2000; Hiebert et al., 2003; Mayer, 1999) and principal ratings (e.g., Harris, Ingle, & Rutledge, 2014). Each method has its own strengths and limitations. For example, teacher self-ratings are economical and simple to administer, but teachers' responses may be influenced by a social desirability factor (Douglas, 2009). In this study, particular emphasis is given to classroom observations and student ratings. The reasons for choosing these two sources of data and significant issues concerning them are discussed below.

Classroom observation.

Historically, teacher evaluation systems have relied greatly and often solely on direct observation (Mihaly & McCaffrey, 2014; Stronge & Tucker, 2003). The focus on classroom observation for teacher evaluation is justified. Classroom observation is the most direct way to measure quality of teaching (Clare, Valdés, Pascal, & Steinberg, 2001), and specifically, those aspects of teaching that may be directly observed, for instance the interaction between teacher and students and among students (Danielson & McGreal, 2000). Praetorius, McIntyre and Klassen (2017) list three advantages of classroom observation over other measures such as student or teacher ratings. These are: 1) Observers are trained on how to observe and rate the aspects of interest and consequently, should rate them in a more valid way compared to teachers and students. 2) Observers are not involved in teaching at the same time and hence can focus on observing and rating. The teachers themselves may be unaware of their in-class behaviour (Stigler, Gonzales, Kawanaka, Knoll, & Serrano, 1999). 3) Observers usually observe several different teachers and therefore they have a good amount of comparison possibilities. Another advantage is that classroom observations provide teachers with feedback that could help them improve their practice (Pianta & Hamre, 2009). Taylor and Tyler (2011) found that teacher evaluation based on classroom observation can improve the performance of mid-career teachers both during the period of evaluation and in subsequent years. In addition, many resent studies (e.g., Kane & Staiger, 2012; Rockoff & Speroni, 2010; Tyler, Taylor, Kane, & Wooten, 2010) have shown that observation scores are predictive of student learning gains. Moreover, in a meta-analysis conducted by Seidel and Shavelson (2007) it was shown that observational and video analysis measures can produce higher effects on student learning than those obtained through teacher or student questionnaire.

Although, classroom observations can play a significant role in a teacher evaluation system through the provision of information for meaningful feedback, their success depends on quality implementation as they vary greatly in how they are conducted (Goe & Croft, 2009). High- quality observations do not require only good observation instruments, as good tools that are badly implemented will bring little benefit (Kane & Staiger, 2012). Recent evidence suggests that decisions regarding observers and scoring designs (e.g., the number and length of lessons to observe, the number of raters per observation, certification or other rater requirements and observation mode) have considerable consequences on the reliability of teachers' scores (e.g., Casabianca et al., 2013; Hill et al., 2012; Joe, McClellan & Holtzman, 2014; Newton, 2010).

Current practices regarding the number of observations appear limited to two or fewer lessons per teacher/ class per year, each 60 minutes or less (Praetorius et al., 2014; Weisberg et al., 2009). However, according to Kane and Staiger (2012), we cannot have an accurate impression of a teacher's practice from a single observation, even if we had a very precise measure of the quality of teaching of this lesson. This is because according to their findings, a teacher's score varies considerably from lesson to lesson. Similar results about the stability of teachers' scores emerged from the study of Patrick and Mantzicopoulos (2016).

Given that evaluators are limited in regard to the number of observations per teacher/ class that they can afford to carry out, several studies (e.g., Hill et al., 2012;

Newton, 2010; Praetorius et al., 2014) have attempted to find out how many observations per teacher are required to make a reliable generalisation of a teacher's practice in a classroom. Findings varied substantially among studies, but it must be taken into account that each study has used different instruments based on a different theoretical framework and methodology. Specifically, Hill et al. (2012) found that three lessons per teacher rated by two raters were the optimal combination for a reliable estimation of the quality of teaching by using the MQI instrument. Similarly, Whitehurst et al. (2014) recommend, based on their findings, conducting two-to-three annual classroom observations for each teacher, with at least one of those being conducted by a trained observer from outside the teacher’s school. According to their findings, moving from one to two observations increases the reliability of observation scores and their predictive power for value-added scores in the next year as well. On the other hand, Praetorius et al. (2014) found in their study that one lesson per teacher suffices to measure classroom management and personal learning support, while at least nine lessons are needed for cognitive activation by using a rating instrument that measures the aforementioned three dimensions. Therefore, according to these findings, some teaching behaviours are more variable than others. Moreover, Newton (2010) showed that at least six observations and four raters are needed for elementary grades and four observations and four raters would be adequate for secondary grades. Also, she suggested that having more observations per teacher would reduce measurement error more than increasing the number of raters. However, it is not within the scope of this chapter to determine how many observations are needed to produce a reliable measure of a teacher’s behaviour in the classroom.

It is important to note that even if several studies have investigated how many observations per teacher (in a single class) are needed to make a reliable generalisation of a teacher’s behaviour in the classroom, whether the data obtained from a single class can be generalised in all the classes a teacher teaches, seems to be neglected in the research literature. In other words, what is not yet clear is whether teachers exhibit the same teaching skills when they teach in different classrooms. The importance of examining the consistency of teacher behaviour across different classrooms is discussed in detail in a subsequent section of this chapter.

In regard to the use of different observers, a great effort is required in order to ensure that the observers record teachers' behaviour in comparable ways (Stigler et al., 1999). As Gitomer et al. (2014) argue, for a system to be valid, judgments about teaching quality should not be determined by who makes them. Thus, high-quality training and certification of observers are essential to rising inter-rater reliability (Kane & Staiger, 2012).

However, no measure is perfect. Classroom observations suffer from some drawbacks. Some of the disadvantages that have been reported in the literature are that classroom observations are expensive to conduct and time-consuming (Matsamura et al., 2002; Peterson, 2000; Praetorius, Lenske, & Helmke, 2012). Moreover, according to Danielson and McGreal (2000) some important aspects of teaching are not necessarily easily observed in a classroom episode. Additionally, estimates may be affected by a variety of factors such as possible changes in classroom behaviours when an observer is present (Douglas, 2009), rater bias (e.g., leniency/severity bias) (Praetorius et al., 2012) and the number of observations to be conducted (see above).

The importance of using multiple sources of data.

There is some debate in the literature on whether quality of teaching can be measured by using only classroom observation data or whether it is necessary to use other sources of data such as student ratings or teacher self-ratings as well. Specifically, Kane and Staiger (2012) mention that classroom observations by themselves are not highly reliable and they are only moderately associated with student achievement gains.

Moreover, some authors (e.g., Danielson & McGreal, 2000; Stronge, 2006; Stronge & Tucker, 2003) argue that even if classroom observations can be an important data source, teacher evaluation systems should not merely depend on observations if the purpose is to provide a comprehensive picture of teacher performance. In particular, they argue that a teacher evaluation system should use multiple sources of data. Some of the reasons for using multiple sources of data are the limited nature of the sources themselves and the need to address the various responsibilities of teachers (Peterson, 2000). According to Stronge (2006), the integration of multiple data sources in the evaluation system provides a far more realistic picture of actual teacher performance than would be available through a single source of information such as direct observation. The comparisons of various sources of data that are properly employed could increase the internal validity of the evaluation system (Kyriakides & Demetriou, 2007).

A recent study supports the hypothesis that observations alone are not enough. Specifically, in 2013 Bill and Melinda Gates Foundation published a report where the researchers used MET project data to compare variously weighted composites (including student achievement gains based on state assessments, classroom observations and student surveys) and estimate the ability of differently weighted composites to produce consistent results and accurately forecast teachers' impact on different student outcomes. When they tested whether observations alone are enough, the observation-only model performed much worse than any of their multiple measures composites, even with four classroom observations (two by one observer and two by another). However, as noted by Kyriakides and Demetriou (2007), resorting to a plethora of sources to collect relevant data may lead to problems of practicality and validity. Therefore, it is important for researchers to identify those sources of data which are the most appropriate for teacher evaluation.

In a study (see Wilkerson et al., 2000) investigating the relationship of student achievement to teacher performance measures by principals, students and self-ratings by the teachers, it was found that student ratings were the best predictor of student achievement. In addition, student ratings showed the strongest positive relationship to student achievement (in mathematics, language arts and reading) when compared to ratings of principals and teachers. Moreover, the findings have shown that teacher self-ratings were more closely related to student learning compared to ratings of principals. Similarly, a meta-analysis conducted by Cornelius- White (2007) has shown that teacher self-ratings are less predictive of student success than students' ratings and observations.

Teacher self-ratings.

Even the fact that teacher self-ratings are economical, simple to administer (Hiebert et al., 2003) and can provide information useful for planning and teacher improvement, this source of data suffers from some serious drawbacks (Peterson, 2000). One of the limitations of the use of teacher self- ratings to measure classroom processes is memory, as it can be difficult for a teacher to remember aspects of teaching that may happen too quickly to be under the teacher's conscious control. Another problem is comprehension as some questions may not be understood in a consistent way across different teachers (Hiebert et al., 2003; Stigler et al., 1999). In addition, this source of data is subject to error due to judgment and social desirability (Douglas, 2009). As Peterson (2000) claims, the problems that teacher self-ratings face preclude their use in teacher evaluation and especially in summative evaluation.

Student ratings.

In the case of student ratings, this source of data requires minimal training and, just like teacher self-ratings, it is relatively economical in terms of time and personnel (English, Burniske, Meibaum, & Lachlan-Haché, 2016; Peterson, 2000). According to a number of authors, student ratings are defensible sources for evaluating teachers for additional reasons. Students are in a key position to provide information on the quality of teaching and the learning environment in individual classrooms, as they are the direct recipients of the teaching learning process (Goe, Bell & Little, 2008; Kyriakides, 2005). Students can also provide information in regard to the development of motivation in their classroom, opportunities for learning, classroom equity and the degree of rapport and communication developed between teacher and student (Kyriakides et al., 2014). This information can be useful for teacher improvement and can be administered early enough in the year to inform teachers where they need to focus (MET Project, 2012). If the teacher uses this information constructively, then their current students may benefit through an improved teaching and learning environment (Aleamoni, 1999). Another advantage of the use of student ratings is that in contrast to external observers, the experience of students with the behaviour of a certain teacher is often based on a large number of lessons (Den Brok, Brakelmans & Wubbels, 2004; Stronge & Ostrander, 1997). Moreover, students are able to evaluate their teachers and classroom environments as they have closely and recently observed many teachers and have encountered many different situations and contexts (Den Brok et al., 2004; Kyriakides, 2005).

In her review of 154 articles, Aleamoni (1999) identified and discussed 16 of the most common myths regarding student ratings of their instructors from the perspective of the research that has been conducted over a 74-year period. An important finding supported by many studies is that students do not automatically rate teaching skills highly for those teachers who also received a high rating on constructs associated with popularity (e.g., items such as “The instructor seemed to be interested in students as persons”). Moreover, student ratings of their teachers are not highly correlated with grades received from respective teachers. In most instances, the relationships were relatively weak (the median correlation was approximately 0.14, the mean 0.18 and the standard deviation 0.16). In addition, the studies cited by Aleamoni showed that students could make consistent judgments about the teacher and teaching. However, many authors cautioned that the reliability and validity of student ratings depend to some extent on the instrument used; in particular, the content, construction and procedures for the administration of the instrument (Aleamoni, 1999; English et al., 2016; Goe, et. al. 2008; Kyriakides, 2005).

Despite the fact that student ratings are common at the college level, their use for evaluating school teachers was rare (Peterson, 2000). Only recently, an increasing number of states and districts in the USA include student perception surveys as part of their evaluation system (English et al., 2016; MET Project, 2012). However, data from several studies have shown that elementary and secondary school students are capable of providing valid and reliable data on teacher behaviours (e.g., De Jong & Westerhof, 2001; Worrell & Kuterback, 2001). For example, Kyriakides et al. (2014) have shown that younger students (9- and 10- year-olds) from different European countries could provide valid data about the classroom-level factors included in the dynamic model of educational effectiveness. Another study conducted in Cyprus by Kyriakides (2005) revealed that students of year 6 are also capable of providing ratings of teacher behaviour that are reliable and valid, which can help us evaluate the quality of teaching and the interpersonal teacher behaviour. Additionally, this study has found that student ratings of teacher behaviour are highly correlated with value-added measures of student cognitive and affective outcomes. According to Kyriakides, this implies that student ratings compared to value-added measures of student outcomes could be considered as a more practical and valid way to evaluate teachers.

Student ratings have another advantage over other measures such as teacher self-ratings. Specifically, student ratings are considered reliable largely because they often consist of an average of a large number of students who balance each other's biases (Peterson, 2000). Thus, student ratings are only marginally subject to mood swings, personal preferences or other personal factors (Den Brok et al., 2004). Moreover, as

Kyriakides et al. (2014) argue, gathering information from all the students in a class about the behaviour of their teacher, gives to the researchers the opportunity to test the generalisability of the data and identify the extent to which the object of measurement is the teacher. The generalisability of the data may not be easily determined when other sources of data (e.g., classroom observation or teacher self-ratings) are used to measure the quality of teaching. This is attributed to the fact that usually one person rates each teacher.

However, Worrell and Kuterbach (2001) caution that students should not be the sole source of data of evaluation of teaching because students are not aware of the curriculum, classroom management or other areas associated with effective teaching. Nevertheless, many authors (Goe et al., 2008; Kyriakides, 2005; MET Project, 2012) argue that student ratings should be one of the multiple measures of teacher evaluation; "a valuable component of a comprehensive teacher evaluation system" (English et al., 2016, p.11). According to Kyriakides et al. (2014), collecting data from both students and external observers could generate more precise, reliable and valid data on the quality of teaching.

The data from MET Project (see Kane & Staiger, 2012) have shown that combining observation scores with value-added student achievement gains and student feedback improved predictive power and reliability. As Kane and Staiger point out, the combination of these sources capitalizes on their strengths and offsets their weaknesses; each of these sources "shines in its own way" (p. 29). However, in grades and subjects where student achievement gains are not available, they argue that classroom observations should be combined with student ratings.

Examining the Notion of what Constitutes Effective Teaching

Although effectiveness studies show that what teachers actually do in the classroom matters for student learning (Muijs et al. 2014), many scholars (e.g., Cohen& Goldhaber, 2016; Goe et al., 2008; Stodolsky, 1990) have argued that there is no universal agreement about what constitutes effective teaching or quality of teaching (for the purpose of this study the two terms are used interchangeably). The way that effective teaching is defined is important because definitions propose and shape what needs to be measured (Goe et al., 2008).

Most modern definitions of effective teaching focused on teacher behaviour in the classroom. For example, effective teaching is defined by Kyriacou (2009: 7) as “teaching that successfully achieves the learning by pupils intended by the teacher”. According to Kyriacou, the emphasis on the notion of effective teaching is placed on identifying observable behaviour in the classroom that can be associated to observable outcomes. On the other hand, the emphasis on terms like ‘good’ and preferred’ teaching is placed on how an observer feels about the teaching and usually is focused on characteristics of teaching that the observer feels are desirable without necessarily any direct reference to outcomes. However, over the past few decades, thinking about effective teaching has been approached in several different ways.

In the next section, a historical overview of teacher effectiveness research (TER) is presented and the different phases of TER are discussed to demonstrate the growth in the way that effective teaching has been approached through the years. Then, the rationale of the main models of educational effectiveness that have integrated teacher effectiveness factors and school effectiveness findings is described. Finally, the dynamic model of educational effectiveness (which is the theoretical framework where this study is based) is analyzed by giving particular emphasis on the classroom-level factors of this model.

Research on teacher effectiveness: A historical overview.

One of the most influential research traditions about effective teaching is TER (see Creemers et al., 2013, for a critical review of TER). According to Brophy and Good (1986), research on effective teaching was slow to develop due to historical influences on the conceptualization and measurement of teacher effectiveness. Until the 1960s, research on effective teaching was mainly dominated by attempts to identify teachers' personal traits, such as personality characteristics, which may be related to their effectiveness; even though gradually, characteristics more related to education, such as attitude, experience and aptitude/achievement, were also studied (Creemers et al., 2013; Kyriacou, 2009). Specifically, the centre of attention was the teachers themselves and not their behaviour in the classroom (Creemers, 1994). These early studies have been referred to as "presage-product studies" (Creemers & Kyriakides, 2015b). Moreover, these studies (that attempted to relate teacher attributes to educational outcomes) have sometimes been referred to as "black-box" research, as what actually happened in the classroom was completely ignored (Kyriacou, 2009).

Even though this approach produced some consensus on virtues that were considered desirable in teachers, no information was provided regarding the relations between these psychological factors and student performance (Kyriakides, Campbell, & Christoforidou, 2002). In addition, the psychological characteristics of a teacher proved to be poorly related to the teacher's behaviour in the classroom (Borich, 2007). Thus, since the 1960s, researchers have turned to teacher behaviours in the classroom as predictors of student achievement in order to build up a knowledge base on effective teaching (Muijs et al., 2014). The predominant paradigm for research on teaching has been the process-product paradigm (see Brophy & Good, 1986).

The process-product studies were carried out in an effort to identify teacher behaviours (such as teaching skills, techniques or strategies) which predict or cause products (educational results like growth in student knowledge and skills) (Creemers & Kyriakides, 2015b). Specifically, such studies in general use classroom observation to record the frequency of occurrence of various teacher behaviours and aspects of teacher-student interaction (the process variables), and then explore their association with the criteria for effectiveness being used (the product variables) (Kyriacou, 2009). The development of instruments used to measure teacher behaviour was based on theories, paradigms, models or just the researcher's ideas or opinions concerning the relationship between processes in the classroom (especially teacher behaviour) and student outcomes (Creemers, 1994). These studies have led to the identification of a series of behaviours/ indicators (e.g., structuring of lessons, questioning skills and classroom management) that were found to maximize student achievement. Rather than any single teacher behaviour being strongly related to student outcomes, lots of small correlations of different teacher behaviours were found, indicating that effective teaching is not able to do a small number of "big" things right but is rather doing a large number of "little" things well (Reynolds et al., 2014). Many of these findings have been validated experimentally, even though it remains true that experimental findings are weaker and less consistent than correlational findings (Brophy & Good, 1986; Griffin & Barnes, 1986).

A large volume of published reviews has synthesized the findings from the experimental and correlational studies on effective teacher behaviours (e.g., Borich, 2007; Brophy & Good, 1986; Doyle, 1986; Muijs & Reynolds, 2001; Rosenshine, 1983) and has indicated some consensus in TER regarding the importance of certain teacher behaviours for student achievement. In addition, recent meta-analyses, which investigated the impact of generic (e.g., Kyriakides et al., 2013) and domain specific teaching skills (Seidel & Shavelson, 2007) on student outcomes, have been conducted. These meta-analyses took into account effectiveness studies conducted not only in the U.S. but also in Europe (Creemers & Kyriakides, 2015b). The most influential set of recent meta-analyses according to Muijs et al. (2014) were probably those conducted by John Hattie (2009), which synthesized over 800 different meta-analyses relating to the influences on student achievement. Many of the factors identified as having the strongest effect, confirm previous teacher effectiveness findings, like the importance of providing feedback. It is important to add that different approaches of teaching have emerged, such as mastery learning (Block & Burns, 1976) and the active and direct instruction approach (Creemers, 1994; Muijs & Reynolds, 2001), which occurred from an attempt by researchers to combine some factors related to teacher behaviour; as a single factor could not be expected to have large effects on student outcomes (see Creemers et al., 2013).

The process-product paradigm stresses the importance of directly observable teacher behaviour, even though other variables in the general area of teacher variables (like training and experience) have also been considered important. Specifically, over the past three decades, factors other than classroom behaviour have been the focus of considerable research effort. Creemers and Kyriakides (2015b) discuss four of the categories associated to beyond-classroom factors: a) Subject knowledge, b) Knowledge of pedagogy, c) Teacher beliefs and d) Teachers' self-efficacy. However, research on factors other than the teacher behaviours failed to provide empirical support to show that these factors have a direct effect on student outcomes. The studies that reported indirect effects of these factors on student achievement showed that the teacher behaviour in the classroom was the mediating variable and thereby the reported effect sizes of these factors on student outcomes were very small (Creemers et al., 2013). Thus, these findings stress the importance of focusing on teacher behaviour in the classroom for teacher evaluation purposes but also for EER.

The development of integrated models: Moving from teacher effectiveness research to educational effectiveness research.

Teacher effectiveness research has been criticised for a lack of theoretical integration and relatedness to other parts of the education system (Muijs et al., 2014). Although this was true for the earlier studies, over the last decades researchers have attempted to integrate teacher effectiveness factors with findings from school effectiveness research to develop theoretical models (e.g., Creemers, 1994; Scheerens & Creemers, 1989; Scheerens, 1990; Stringfield & Slavin, 1992). These models incorporate a multi-level structure, usually at student, classroom/teacher and school-level; sometimes even extending to context level (e.g., Creemers & Kyriakides, 2008). The variables of these multilevel models of EER are categorized according to an input-process-output framework (Bosker & Scheerens, 1994). The educational processes (teaching and learning) occur at the classroom level and the other levels are supposed to provide the conditions for instruction at the classroom level (Creemers & Reezigt, 1996).

Most theoretical models of EER that emerged during the 1990s relied heavily on the well-known Carroll-model (see Carroll, 1963; 1989). This model was popular because it associated individual student characteristics that are important for learning with characteristics of education that are important for instruction. Furthermore, Carroll considered the factors of time, quantity and quality of instruction as important concepts for learning in schools (Creemers & Kyriakides, 2008). However, as Carroll himself recognized, 25 years after the development of his model, the concept of "high-quality instruction" is rather vague but the model mentions "that learners must be clearly told what they are to learn, that they must be put into adequate contact with learning materials, and that steps in learning must be carefully planned and ordered" (Carroll, 1989, p. 26).

One theoretical model that is based on the Carroll model is the comprehensive model of educational effectiveness (see Creemers, 1994). The only classroom factor in the

Carroll’s model “quality of instruction” has been developed in more detail using the results of TER and put at the core of the comprehensive model of educational effectiveness. However, even if the comprehensive model emphasizes more to the process of teaching than the other integrated models, Kyriakides (2008) argues that the concept of quality of teaching is not defined precisely. According to Kyriakides, the lack of clarity in defining quality of teaching might be ascribed to one of the major weaknesses of EER regarding its assumption that quality is guaranteed whenever an aspect of teaching is able to explain part of the variance of student outcomes. Thus, he claims that researchers in the area of EER must develop a parsimonious model at the classroom level where a clear definition of the quality of teaching will be provided by referring to the most important aspects of effective teaching.

Another weakness of the comprehensive model is the fact that it does not take into consideration the new theories of teaching, since the aspects of quality of teaching taken into account by effectiveness studies conducted to test its validity, mainly referred to the direct teaching approach (Kyriakides, 2008). However, over the past few decades, there has been an increasing interest in the constructivist approach to learning and therefore teaching (Danielson, 1996). Thus, constructivist authors and other supporters of “new learning approach” have developed a set of instructional techniques that are thought to enhance the learning disposition of students like modelling, collaborative teaching and generalization (Creemers, 2006; Muijs & Reynolds, 2011). The constructivist approaches in teaching (Duffy & Cunningham, 1996; Savery & Duffy, 1995; Schoenfeld, 1998) stem from a different view of how learning takes place compared to the more traditional teaching approaches, like the direct instruction approach (from the process-product tradition). Specifically, knowledge and skills are constructed by students themselves during the learning process and are not learned through instruction in which they are delivered by teachers and mastered by students. However, each of these approaches gives emphasis on a single aspect of the teacher's role leading to the provision of a narrowly focused perspective of effective teaching practice (Creemers et al., 2013). Whether or not traditional or constructivist teaching approaches are more effective and whether they benefit all groups of students in the same way, is strongly debated in the literature (Caro, Lenkeit, & Kyriakides, 2016). Nevertheless, the results of recent meta-analyses of teacher effectiveness studies (e.g., Kyriakides et al., 2013; Seidel & Shavelson, 2007) reveal that within each approach there are factors which are related to student achievement. This implies that effective teaching could combine elements of more traditional approaches and elements of constructivist instruction as well.

The dynamic model of educational effectiveness.

A further development of the comprehensive model is the dynamic model of educational effectiveness (see Creemers & Kyriakides, 2008), which is considered as one of the most influential and developed theoretical models of EER (Heck & Moriyama, 2010; Sammons, 2009; Scheerens, 2013). Moreover, it provides a clear definition of quality of teaching through eight generic factors included at the classroom level (these factors are presented in the next section of this chapter). This model, which provided the theoretical basis of the current study, takes into account the main findings of educational effectiveness studies and the strength and weaknesses of previous models of EER (e.g., Creemers, 1994; Scheerens, 1992; Stringfield & Slavin, 1992). In addition, it was developed in order to create stronger links between EER and improvement practice (Creemers & Kyriakides, 2012).

The dynamic model is multilevel in nature and refers to the most important effectiveness factors that operate at four levels: student, classroom (teacher), school and system. However, even if it is multilevel in nature, this model takes as a point of departure the fact that learning has to be explained by the primary processes at the classroom level.

Thus, in this model, teaching and the learning situation are emphasized and the roles of teacher and students are analyzed. It also refers to schools and context-level factors that are expected to have direct and indirect effects on students' outcomes through their effects on the classroom-level factors. Moreover, the dynamic model takes a broad outlook on effectiveness criteria, as the outcomes' measures are not restricted only to the cognitive. Figure 2.1 provides an overview of the dynamic model as proposed by Creemers and Kyriakides (2008) not only by referring to the factors included in each level but also by illustrating the relationships assumed across levels and their relationship with student outcomes.

Another essential element of the dynamic model is that it not only searches for the relationship of factors that operate across levels, but it also assumes that there is a need to closely examine the relationship between the factors that operate at the same level. Specifically, it is based on the assumption that some factors and their dimensions that operate at the same level may be related to each other. Such an approach to modelling educational effectiveness reveals grouping of factors that make teachers or school effective. In this way, specific strategies for improvement could be provided which will be comprehensive in nature and will not focus on the acquisition of an isolated skill.

The above assumption has been explored by Kyriakides, Creemers and Antoniou (2009) in a longitudinal study which was conducted in 50 primary schools in Cyprus. This study revealed that the classroom-level factors and their dimensions of the dynamic model can be grouped into five stages of teacher behaviour which are hierarchically structured regarding the degree of difficulty. Specifically, the teaching skills which were included in the first three stages are mainly related to the direct and active teaching approach (e.g., structuring). In the last two stages, which are more demanding, teaching skills are related to new teaching approaches and differentiation of teaching. Moreover, the findings of this study revealed that transition from one stage of teacher behaviour to the other is not linear and also that the transition to the higher stages (stages 4 -5) is more difficult than the transition between the lower stages (stages 1-3). Additionally, the results of this study showed that the students of teachers who demonstrated skills at the higher stages showed better outcomes, than the students whose their teachers were situated at the lower stages of teacher behaviour. It is important to mention that the study of Kyriakides, Archambault and Janosz (2013), whose was conducted in seven primary schools in Canada, also supports the assumption of the dynamic model; that the classroom-level factors are inter-related. In this study, four stages of teacher behaviour emerged which were similar to the abovementioned study conducted in Cyprus. Moreover, apart from the grouping of factors, this model is based on the assumption that the relation of some factors included in the dynamic model with student achievement may not be linear, but curvilinear (see Creemers & Kyriakides, 2008).

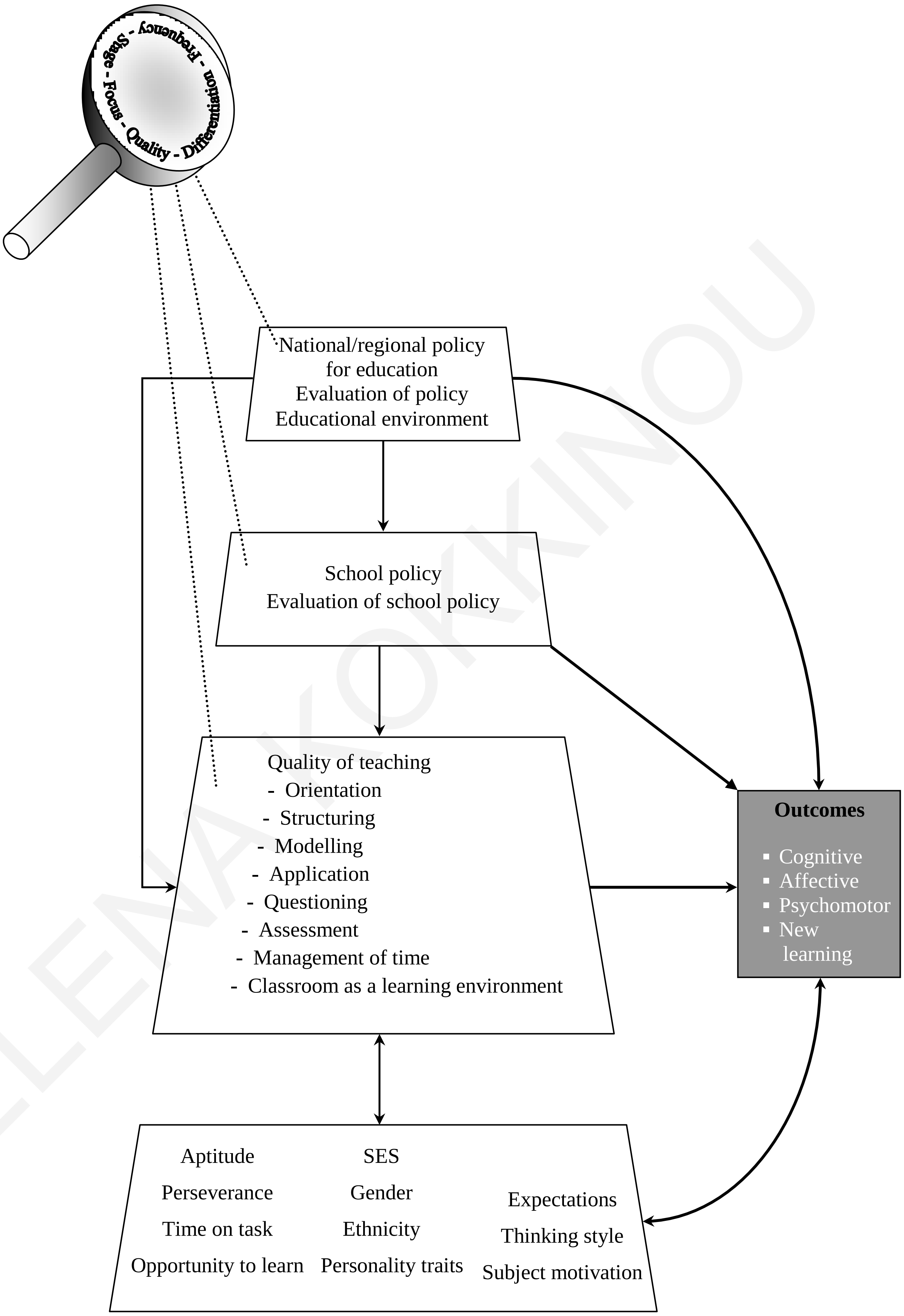


Figure 2.1 The dynamic model of educational effectiveness

The classroom-level factors of the dynamic model of educational effectiveness.

For the purposes of the present study, emphasis is given to the classroom-level factors of the dynamic model. These factors refer to observable instructional behaviour of teachers in the classroom instead of factors that may explain such behaviour. Specifically, based on the main findings of TER (e.g., Brophy & Good, 1986; Darling-Hammond, 2000; Doyle, 1990; Muijs & Reynolds, 2000; Rosenshine & Stevens, 1986; Scheerens & Bosker, 1997) the dynamic model refers to the following eight generic factors that describe the teachers' instructional role and were found to be related to student achievement: a) orientation, b) structuring, c) teaching modelling, d) application, e) questioning, f) assessment, g) management of time and h) classroom as a learning environment. The classroom-level factors do not arise exclusively from one approach, such as the direct teaching approach or the constructivist approach, but they cover, at least to some extent, the main approaches in learning and teaching by adopting an integrated approach in defining quality of teaching. For example, structuring, questioning, application and management of time stem from the major findings of the process-product studies, while modelling and orientation are in line with the constructive theory and its impact on learning (Creemers & Kyriakides, 2015b). A brief description of each classroom-level factor follows, based on the definition of the factors provided by Creemers and Kyriakides (2008). Moreover, an overview of the main elements of these factors translated into teacher behaviours is provided in Table 2.2 (adapted from Creemers et al., 2013).

Orientation: Orientation refers to teacher behaviour in terms of providing the objectives for which a specific task or lesson or series of lessons takes place and/or challenging students to the identification of the reason(s) for which a particular activity occurs in the lesson. Through the orientation process it is anticipated that the tasks/lessons will become meaningful to students, which in turn might foster their active participation in the classroom.

Table 2.1
The main elements of each classroom-level factor included in the dynamic model

Classroom-level factors	**This factor refers to teacher behaviour in terms of:**
Orientation	• providing the objectives for which a specific task or lesson or series of lessons takes place and/or • challenging students to the identification of the reason(s) for which a particular activity occurs in the lesson.
Structuring	• beginning with an overview and/or review of objectives; • outlining the content to be covered and signalling transitions among lesson parts; • calling attention to main ideas and • reviewing main ideas at the end.
Questioning	• offering a mix of product and process questions at appropriate difficulty level; • giving time for students to respond and • dealing with student responses.
Teaching modelling	• encouraging students to use problem-solving strategies and/or • develop their own strategies that can help them solve different types of problems.
Application	• using seatwork or small group tasks in order to provide students with necessary practice and application opportunities and • using application tasks as starting points for the next step of teaching and learning.
The classroom as a learning environment	• contributing to the creation of a learning environment in his/her classroom. This factor takes five elements into consideration: *(a) teacher-student interaction, (b) student-student interaction, (c) students' treatment by the teacher, (d) dealing with classroom disorder and (e) competition between students*.
Management of time	• organising and managing the classroom environment as an efficient learning environment and • maximising student engagement rates.
Assessment	• using appropriate techniques to collect data on student knowledge and skills; • analysing data in order to identify their students' needs; • reporting the assessment results to students and parents and • evaluating their own teaching practices.

Structuring: Structuring is a factor that stems from the process-product studies which had early indications regarding its contribution in maximizing student achievement. Specifically, Rosenshine and Stevens (1986) maintain that student learning is positively influenced when teachers not only actively present materials, but also structure them by: (a) beginning with an overview and/or review of objectives; (b) outlining the content to be covered and signalling transitions between lesson parts; (c) calling attention to main ideas and (d) reviewing main ideas at the end. According to Brophy and Good (1986), overviews and outlines assist the students to develop learning sets to use in assimilating the content as it unfolds. In addition, summary reviews, which are also important, integrate and strengthen the learning of major points. Taken together, the aforementioned structuring elements make memorization of the information easier and also allow for its apprehension as an integrated whole with recognition of the relationships between parts. Furthermore, research has shown that achievement levels tend to be higher when information is presented with a degree of redundancy, especially in the form of repeating and reviewing general views and key concepts. Finally, the structuring factor refers to the ability of teachers to gradually increase the difficulty level of their lessons or series of lessons as well (Creemers & Kyriakides, 2006).

Questioning: The dynamic model defines the questioning factor according to the following five elements; taking into account the results of studies concerned with teacher questioning skills and their association with student outcomes. Firstly, it is supported that effective teachers are expected not only to ask numerous questions and attempt to involve students in class discussion, but also to offer a mix of product questions (i.e., those requiring a single response from students) and process questions (i.e., those expecting students to provide explanations). However, research has shown that effective teachers ask more process questions (Askew & William, 1995; Evertson, Anderson, Anderson, &

Brophy, 1980). Secondly, another element of this factor is the appropriateness of the difficulty level of the question which is largely determined by the developmental level of students. As it is noted by Brophy and Good (1986), most questions (perhaps 75 per cent) should elicit correct answers and most of the other questions should elicit overt, substantive responses (incorrect or incomplete answers) instead of failing to respond at all. Moreover, the optimal question difficulty is expected to vary with context; for instance, basic skills instruction requires a large amount of drill and practice and consequently requires frequent fast-paced review in which most questions are answered rapidly and correctly. On the other hand, when teaching complex cognitive content or trying to get students to generalise, evaluate or apply their learning, effective teachers commonly raise questions that few students can answer correctly or that have no single correct answer at all. Thirdly, the length of pause following questions is taken into account for this factor and it is anticipated to vary according to the level of difficulty of the questions. Fourthly, the clarity of a question and specifically the degree to which students understand what is expected of them to do/find out is another important element of this factor. Finally, the questioning factor refers to the way teachers deal with student responses. Specifically, correct responses should be acknowledged for the purpose of other students' learning. In case of responses that are partially correct or incorrect, then effective teachers acknowledge whatever part may be correct and if they consider there is a good prospect of success, they try to evoke an improved response instead of providing the student with the answer or calling on another student to respond (Rosenshine & Stevens, 1986).

Teaching modelling: During the last two decades increased attention has been given to the teaching and learning activities related to higher order thinking skills and specifically problem-solving because of the emphasis given through policy on the achievement of the new goals of education (Aparicio & Moneo, 2005; Muijs et al., 2014). The teaching modelling factor, which is in line with the new theories of teaching (Creemers, 2006), refers to the ability of teachers to help students use strategies and/or develop their own strategies that can help them solve different types of problems. In this way, it is more likely that the students will develop skills to help them organise their own learning (e.g., self-regulation and active learning). In defining this factor, the dynamic model also addresses the properties of teaching modelling tasks and particularly the role that teachers are anticipated to play in order to help students use a strategy to solve problems, referring to two alternative approaches. Specifically, teachers may either present a problem-solving strategy with clarity or they may invite students to explain how they would approach or solve a particular problem and afterwards use that information for promoting the idea of modelling. The latter approach may encourage the development of the students' own problem-solving strategies.

Application: This factor can be linked to the direct instruction approach and particularly to the process-product studies (Creemers, 1994; Rosenshine, 1983) which emphasise the immediate exercise related to skills and content taught during the lesson. It is supported that effective teachers use seatwork or small group tasks in order to provide students with necessary practice and application opportunities (Borich 1992). In measuring the application factor, it is important to investigate whether students are simply asked to repeat what has already been covered by their teacher or if the application task is more complex than the content covered in the lesson. In addition, the application factor examines whether the application tasks are used as starting points for the next step of teaching and learning. Moreover, this factor refers to teacher behaviour in monitoring, supervising and giving corrective feedback during application tasks.

The classroom as a learning environment: Regarding the factor “classroom as a learning environment”, the dynamic model refers to the teacher’s contribution in creating a learning environment in his/her classroom and it takes five elements into consideration: teacher-student interaction, student-student interaction, students’ treatment by the teacher, competition between students and classroom disorder. The first two of these elements are important aspects of measuring classroom climate, as classroom environment research has shown (e.g., see Cazden 1986; Den Brok et al., 2004; Harjunen 2012). However, the dynamic model concentrates on the types of interactions that exist in a classroom instead of how students perceive their teacher’s interpersonal behaviour. Particularly, this factor is concerned with the immediate impact teacher initiatives have on establishing relevant interactions and it investigates the extent to which teachers are able to establish on-task behaviour through the interactions they promote. The other three elements refer to teachers’ efforts to create a businesslike and supportive environment for learning in the classroom (Walberg 1986). These elements are measured by taking into account the teacher’s ability in establishing rules, persuading students to respect and use the rules and maintaining them in order to create a learning environment in their classroom. The first of these elements refers to more general problems that could occur when students do not believe that they are treated fairly and respected as individual persons by their teacher. The other two elements have to do with specific situations in the classroom (i.e., competition between students and classroom disorder) that might create difficulties in promoting learning. An important feature of this factor is that it examines the impact that the teacher’s behaviour has on solving the problem(s) that occur(s), as measured through students’ behaviour. For instance, a teacher may not use any strategy at all to deal with a classroom misbehaviour incident, may use a strategy that solves the problem only temporarily, or may use a strategy that has a long-lasting effect. Finally, this factor measures the extent to which teachers use different strategies to deal with problems caused by different groups of students. For example, in some cases, when the problem is small, it might be a better strategy not to pay attention, as any reaction from the teacher may promote the continuation of the problem.

Management of time: Effective teachers are anticipated to organise and manage the classroom as an efficient learning environment and in that way to maximize student engagement rates (Creemers & Reezigt 1996). Thus, the main interest of this factor is the extent to which teachers manage to keep students on task and the extent to which they are able to maximize the learning time during the lesson by dealing effectively with any disturbing factors. Therefore, management of time is considered as one of the most important indicators of teacher ability to manage the classroom effectively.

Assessment: Assessment is considered as an integral part of teaching (Stenmark 1991). In particular, formative assessment has been shown to be one of the most important factors associated with effectiveness at all levels, especially at the classroom level (e.g., De Jong, Westerhof, & Kruiter 2004; Shepard 1989). In the dynamic model, the information collected through assessment is expected to be used by the teachers for at least two reasons. The first reason is associated with the identification of their students' needs. The second reason has to do with self-evaluation since information gathered from assessment can be used by the teachers to evaluate their own practice as well. Quality of assessment is measured by looking at the properties of the evaluation instruments used by the teacher, like validity, reliability, practicality and the extent to which the instruments cover the teaching content in a representative way. Quality is also measured by examining the type of feedback the teachers give to their students and the way students use such feedback.

Dimensions of measuring the effectiveness factors.

One of the main weaknesses of the previous models of EER is the fact that they do not explicitly refer to the measurement of each effectiveness factor, implying that the factors represent rather unidimensional constructs (Creemers & Kyriakides, 2006). Contrary to the previous models, the dynamic model defines and measures each effectiveness factor, operating at either level, by using five dimensions (i.e., frequency, focus, stage, quality and differentiation) describing not only quantitative but also qualitative characteristics of the functioning of each factor. According to Creemers and Kyriakides (2015b), the measurement of quantitative and qualitative characteristics of the classroom-level factors can be seen as the development of a promising theory about effective teaching which can guide new research in the area of teaching and teacher professional development. Considering the effectiveness factors as multidimensional constructs, it helps us identify the specific aspects of the functioning of a factor that are related to student outcomes, describe the complex nature of teaching and develop specific strategies for improving educational practice (Kyriakides, 2008; Kyriakides & Creemers, 2009). A short explanation of how each dimension is used to measure each effectiveness factor follows.

Frequency: This dimension is a quantitative means of measuring the functioning of each effectiveness factor. Specifically, the frequency dimension is measured by taking into account the number of tasks/ activities or actions related to an effectiveness factor that take place in an educational setting (e.g., a typical lesson) as well as how long each task takes to complete. These two indicators help us to identify the importance that is attached by the teacher to each effectiveness factor (Creemers et al., 2013). The frequency dimension is perhaps the easiest way to measure the effect of a factor on student outcomes; and most studies in the area of EER have used only this dimension to define effectiveness factors (Creemers & Kyriakides, 2008).

The other four dimensions investigate qualitative characteristics of the functioning of the factors revealing that effectiveness is more complicated than what supposed in previous theoretical models and studies.

Focus: The effectiveness factors in the dynamic model are also measured by taking into consideration the focus of the activities related to each factor. This dimension can be measured by taking into account two different aspects. The first refers to the specificity of the activities which can range from too specific to too general. For example, regarding the specificity of an orientation task, this task may refer to a part of a lesson, to the whole lesson or even to a series of lessons. The second aspect addresses the purpose(s) for which an activity takes place, by looking at whether an activity aims at achieving one or multiple purposes. According to Creemers et al. (2013), research findings have revealed that if all the activities are anticipated to achieve a single purpose, then the chances of success are high, but the effect of the factor may be small, owing to the fact that other purposes are not achieved and/or synergy may not exist since the activities are isolated. On the contrary, if all the activities are anticipated to achieve several purposes, there is a risk that specific purposes will not be addressed in such a way that they can be implemented successfully.

Stage: This dimension refers to the stage at which tasks associated with a factor take place. It is supposed that the effectiveness factors are needed to take place over a long period of time to ensure that they have a continuous direct or indirect effect on student learning. For example, orientation tasks are expected to take place in different parts of a lesson (e.g., introduction, core, ending of the lesson) or series of lessons and not only at a specific part of a lesson (e.g. only in the introduction). Even though measuring the stage dimension gives information about the continuity of the existence of a factor, activities related to this factor may not necessarily be the same.

Quality: This dimension refers to the properties of the specific factor itself, as they are discussed in the literature. The importance of using this dimension arises from the fact that looking only at the quantitative elements of a factor ignores the possibility that the functioning of the factor may vary. In the case of orientation, the measurement of the quality dimension refers to the properties of the orientation task, especially if it is clear for the students and if it has any impact on their learning. For instance, a teacher may present the reasons for doing a task simply because it has to be done and is a part of his/her teaching routine even if it has little effect on student participation. On the other hand, other teachers may encourage students to identify the purposes that can be achieved by implementing a task and as a result increase their students' motivation in relation to a specific task or lesson or series of lessons (Creemers & Kyriakides, 2006).

Differentiation: The dynamic model takes into consideration the findings of research into differential effectiveness (e.g., Campbell et al., 2004). Thus, despite the fact that the dynamic model is expected to be a generic model, it is recognized that the impact of its factors on different groups of students/ teacher/ schools may vary. As a consequence, the dynamic model deals with differentiation as a separate dimension of measuring each factor. This dimension refers to the extent to which activities related to an effectiveness factor are implemented in the same way for all the subjects involved (for all the students regarding the classroom-level). It is expected that the adaptation to the specific needs of each subject or group of subjects will enhance the successful implementation of a factor, therefore, leading to the maximization of its effect on student achievement (Creemers & Kyriakides, 2006). In the case of classroom-level, one way for teachers to differentiate their teaching is to teach according to specific needs of each student or group of students as these are defined from their personal characteristics and background like gender, socio-economic status, ability, thinking style and personality type. Nevertheless, the differentiation dimension does not mean that the students are not expected to achieve the same purposes; contrariwise, adjusting the functioning of each factor to the special needs of each group of students might ensure that all of them will be able to achieve the same purposes (Creemers et al., 2013).

Testing the validity of the dynamic model at classroom level.

The importance of the aforementioned factors and their dimensions is supported by an international study and several national studies, which were conducted in primary and pre-primary schools, and by a recent meta-analysis as well (e.g., Creemers & Kyriakides, 2015a; Kyriakides, Creemers, & Panayiotou, 2018a; Panayiotou et al., 2014). These studies demonstrated that the eight factors included in the dynamic model and their dimensions are associated with different types of learning outcomes of students, in different phases of schooling and in different countries providing support to the generic nature of the classroom-level factors. For instance, Panayiotou et al. (2014) have shown that the classroom-level factors are associated with student achievement gains in mathematics and science in six different European countries. The empirical studies, which have been conducted to test the validity of the dynamic model, have also shown that using all five dimensions to measure the functioning of the classroom-level factors explains a higher percentage of variance in student achievement rather than using a single dimension. It is important to note that in some studies (e.g., Kyriakides & Creemers, 2008, 2009), there are factors which were found to have no statistically significant effect on student achievement when the effect of their frequency dimension was measured, but they were associated with student achievement when other dimensions were taken into account.

Therefore, the findings of these studies reveal that emphasis should be given to all five dimensions of effectiveness factors and not only to the frequency dimension.

Consistency of Teacher Behaviour Across Different Classrooms

As mentioned before, the empirical studies, which have been conducted to test the validity of the dynamic model, provided support to the generic nature of the classroom-level factors in different types of learning outcomes of students, in different phases of schooling and in different countries (Kyriakides et al., 2018a). However, given that these studies took place at primary and pre-primary school level, they were not in a position to investigate the extent to which the classroom context affects teacher behaviour. Most elementary teachers teach only in a single class with the same students over a school year, suggesting that there may be a high degree of similarity in teacher behaviour and in the quality of interactions taking place across the school year. However, if teacher behaviour in different classrooms varies in regard to the classroom-level factors of the dynamic model, then the generic nature of these factors could be questioned. Therefore, there is a need for further research that will be conducted in secondary schools in order to gather data from more than one classroom of the same teacher. Studies conducted at secondary school level may not only provide further support to the dynamic model by investigating the effects of the classroom-level factors in different age groups of students, but may also help to investigate whether there are classroom-level factors the measurement of which is more sensitive to the classroom context and specifically, the student composition of the classroom.

Apart from testing the generic nature of the classroom-level factors of the dynamic model, understanding the extent to which teachers exhibit the same teaching skills when they teach different groups of students is also crucial for policy purposes, given the influence of observational measures in evaluating teachers either for formative and/or for summative reasons. Most teacher evaluation policies around the world seem to presume that teaching is a generic activity and the same teachers would exhibit the same teaching skills in different classrooms. Hence, they would show similar ratings in their teaching skills across different contexts. It can be argued that policymakers may tend to oversimplify the fact that teaching itself is a complex phenomenon to make it more amenable to easier measurement and policy reform (Pacheco, 2009). However, there are doubts between scholars on whether the same teachers could respond to different classroom contexts similarly and show consistency on their teaching behaviour across all contexts. For instance, Whitehurst et al. (2014) discuss the case of a teacher who gets an unfair share of students who are challenging to teach because they are less well prepared academically, aren't fluent in English, or have behavioural problems. As they argue, this teacher is going to have a tougher time performing well, for example on questioning and discussion techniques, than the teacher in the gifted and talented classroom. In addition, Ladson-Bilings (2009) questions whether "being an excellent teacher in a suburban school serving high-income students means that you will also be an excellent teacher in an urban school serving students who are low income, recent immigrants, and/or English language learners" (p.220). Many other authors (e.g., Pacheco, 2009; Smylie et al., 2008) argue that classroom context variables may affect teachers' practices. Nevertheless, none of them had strong evidence to test his/her assumption.

A huge number of different contextual variables has been reported in the literature that may contribute to the variation on teachers' in-class behaviour. Some of these variables are: class size (e.g., Blatchford, Bassett, & Brown, 2011; Smylie et al., 2008), number of adults and students in the classroom and the time of the day (e.g., very beginning or end of the school day), week and year (e.g., Bell et al., 2012; Curby et al., 2011), activity settings-subject (e.g., Curby et al., 2011; Goe, Bell, & Little, 2008) and content domains (Grossman, Cohen, & Brown, 2014), composition/ student characteristics (i.e., academic and language-cultural heterogeneity, the percent of low-income students or students with special needs or relatively older students as compared to the grade-level average etc) (e.g., Grossman et al., 2014; Mihaly & McCaffrey, 2014; Smylie et al., 2008; Steinberg & Garrett, 2016; Whitehurst et al, 2014) and grade (e.g., Goe et al., 2008; Mihaly & McCaffrey, 2014). These contextual variables can be combined in many ways to define a particular classroom context in detail. In addition, as noted by Pacheco (2009), classrooms are not isolated from the larger context of schools and the broader context of community. Both contexts may introduce new variables that can affect the teacher-student interaction. However, the effect of school and community in quality of teaching is outside the scope of the present study.

As Kyriacou (2009) argues, the variety of teaching contexts can create problems for research. An important problem is that each study can take into account only a few aspects of the classroom context at any one time. Another problem is that the influence of one contextual variable on teaching may depend on which other variables are present as well. Therefore, studying classroom context requires a careful design of research. However, even if studying classroom context is complex and difficult, its investigation could help to develop a deeper understanding of teaching in all its complexity. In addition, it may assist the design of an evaluation framework that may be both more responsive to the realities of teaching and more useful in the improvement of teachers' teaching skills (Pacheco, 2009).

This study focuses on the influence of contextual variables related with students on teacher in-class behaviour, as this is measured by classroom observation and/or student questionnaire; since teaching is an interactive process among teachers and students. In other words, this research focuses on whether the same teachers exhibit the same generic teaching skills when they teach different groups of students. When value-added models are used, attempts are made to control contextual variables related with student characteristics that are known to be associated with student test performance (e.g., socioeconomic status and prior achievement levels of students) (McCaffrey et al., 2004). However, this is not the case when classroom observation scores are used in many teacher evaluation systems, as contextual variables are not taken into account (Whitehurst et al., 2014). Moreover, very little is known yet about the influence of contextual variables related with students on the consistency of teacher in-class behaviour in different classrooms.

Several recent studies have attempted to investigate the extent to which measuring teaching skills by classroom observations and/or student questionnaire is influenced by classroom context variables, such as student achievement and student socioeconomic status. These studies found significant correlations between teachers' observation scores and characteristics of the students of the classes they teach. Specifically, by using data from four urban districts of the USA, Whitehurst et al. (2014) found very strong statistical association between the prior achievement level of students and teacher ranking based on observation scores. In other words, they found that teachers with students with higher prior achievement receive observation scores that are higher on overage compared to those received by teachers whose incoming students are at lower achievement levels. Similar results emerged from the study of Lazarev and Newman (2015) who found consistent and pervasive correlations between class-average incoming achievement level and teacher observation scores (from two generic observation tools, FFT and CLASS) by using data from the MET project. However, Stenberg and Garrett (2016), also by using data from the MET project (and specifically from FFT observation tool) found that the incoming achievement of students matters differently for teachers in different classroom settings (ELA teachers compared to math teachers and subject-matter specialists compared to their generalist counterparts). By using data from the same project but focusing on a subject-specific observational tool (PLATO), Grossman et al. (2014) found that the composition of students in the class (i.e., race, income, English language learning status and special education classification) is associated with teacher observation scores. Another recent study (Chaplin et al., 2014) showed that the ratings from both a generic observation tool (RISE) and a student questionnaire (7Cs) are negatively associated with the percentage of low-income and racial/ethnic minority students. Thus, all of the studies presented here support the hypothesis that the classroom context in which teachers work, plays a critical role in determining teachers' performance, based on classroom observation and/or student questionnaire.

However, most of these studies suffered from some serious methodological limitations, as data have been obtained from a single class per teacher per year. Even when data from different years were used in some studies (e.g., Whitehurst et al., 2014), we are unable to determine whether differences in observational ratings are related to student characteristics or to the systematically non-random sorting of teachers to classes of students. According to Braun (2005), in most districts parents often influence to which class and teachers their children are assigned. In addition, data from recent studies showed that schools tend to assign less experienced teachers to classrooms with lower achieving, minority and poor students and the more experienced or effective teachers to higher achieving students (Kalogrides & Loeb, 2013; Kalogrides, Loeb, & Béteille, 2013). As Steinberg and Garrett (2016) found in their study, the non-random process by which teachers are often assigned to classes of students has a significant influence on measured performance based on classroom observation scores.

Therefore, in order to answer such questions, we need to observe teachers teaching in different classes within the same year, since teachers' in-class behaviour may vary over time, especially when teachers participate in professional development programs (see professional development studies e.g., Antoniou & Kyriakides, 2011). In this way we could keep the teacher constant and see whether observation scores change according to the classroom context. Moreover, we need to use a specific theoretical framework, since teachers may be able to demonstrate their abilities in specific factors considered generic and not in others considered as domain specific. Furthermore, given that several methods and instruments for measuring quality of teaching exist, whether the findings are differentiated according to the instrument that is used, should also be investigated. Some instruments may be more sensitive to contextual variables than others. Also, some instruments may give emphasis only on the quantity of behaviours and not on their qualitative characteristics which are also important in describing the complex nature of effective teaching (Kyriakides & Creemers, 2009).

Students' Misbehaviour: A Classroom Context Factor that may Affect the Consistency of Teacher Behaviour in Different Classrooms

In the present study, particular emphasis is given to students' misbehaviour incidents as a classroom contextual factor that may affect teaching quality. Misbehaviour is a term frequently used in the literature, but it is difficult to find a definition which everyone will accept and which will be interpreted and applied consistently (McManus, 1989). The difficulty is due to the fact that behaviour problems are socially disapproved behaviours and the same behaviour can possibly be characterized as problematic by some people and normal by others (Fontana, 1994). According to Kyriacou (2009), although there is a large consensus between teachers concerning some forms of behaviour that constitute misbehaviour (for instance, refusal to do any work or hitting another student), there is a high degree of variation in teachers' judgments for many areas, such as the degree of talking that is allowed. Moreover, there is a possibility a teacher's judgments regarding what constitutes misbehaviour may vary from class to class and from student to student within the same class. Thus, essential to the understanding of behaviour problems is the recognition that any attempt to identify or describe them involves a high degree of subjectivity (Cooper, Smith, & Upton, 1994).

For this study, any kind of student behaviour that “prevents the teacher from teaching and the learner from learning” can be considered misbehaviour or behaviour problem (Montgomery, 1989, p.10). Misbehaviour can take many forms and some examples that are reported in the literature are talking without raising hand, getting out of seat, disrupting others, eating in class, sleeping in class, throwing objects, fighting, use of profanity, vulgar language or obscene gestures, defacing or damaging school property or property of others and so on (Borich, 2007). The reasons for choosing misbehaviour as one of the basic contextual factors which may affect the consistency of teacher behaviour in different groups of students are discussed below.

Several studies reveal that teachers consider students’ misbehaviour as a big problem for their teaching. Specifically, TALIS results reveal that across countries, almost one-third of the teachers on average report that “student interruptions caused the loss of quite a lot of potential teaching time in the classes they teach” (OECD, 2009, p.227). In a study conducted in Cyprus (Kyriakides, 1998), almost 20 years ago, 20 per cent of the teacher sample reveals that they face severe problems with children’s behaviour in classroom. It was also found that teachers spend 25 per cent on average of their teaching time to make remarks to students. Moreover, in a number of studies, a relatively large percentage of teachers seem to believe that they are unprepared to deal with disciplinary problems and they spend more time than they ought in order to address them (Houghton, Wheldall, & Merrett, 1988; Little, 2005; Wheldall & Merrett, 1988). In another study (Public Agenda, 2004) most teachers recognize that their teaching would be more effective, if they didn’t have to spend so much time addressing troublesome behaviours. Therefore, the existence of misbehaviour incidents in a classroom may affect the management of teaching time, reduce students’ time on task and is also possible to affect the quality of teaching.

Causes of students' misbehaviour.

According to Long and Frye (1985, as cited in Marzano et al., 2003) it is a myth to believe that effective teachers can prevent all students' behaviour problems by keeping them interested in learning by using exciting classroom materials and activities; as the potential for misbehaviour exists beyond academics. The literature on the causes of misbehaviour has highlighted the existence of several factors that, in combination, may also provoke students' behaviour problems in classroom. These factors can be classified into two big categories: the internal factors that are related to the students themselves and the external factors that are related to the environmental influences (Charlton & George, 1993).

Internal factors.

Biological factors: The internal factors can include biological factors like the state of the nervous system, hyperactivity, heredity and other genetic considerations (Charlton & George, 1993; Cooper et al., 1994). For instance, as it is mentioned in Poursanidou (2016), 3%-10% of children and adolescents internationally have Attention- Deficit Hyperactivity Disorder (ADHD) and this disorder, which occurs more often in boys than in girls, may be accompanied by learning difficulties but not by reduced mental capacity. ADHD is a brain disorder characterized by an ongoing pattern of inattention and/or hyperactivity-impulsivity that interferes with functioning or development (National Institute of Mental Health, 2016). Children with ADHD cause a constant fuss and at school they are disorganized, they do not pay attention in the class, they do not stay in one place for a long time, they get up often, walk and answer before they listen to the question (Poursanidou, 2016).

Psychological factors: The internal factors also include psychological factors which refer to aspects of the individual's affective and cognitive states (e.g., levels of self-concept, anxiety, motivation and intelligence) (Charlton & George, 1993; Fontana, 1994). For instance, Marciniak (2015) argues that the key issue to understand why misbehaviour occurs in adolescent learners is students' self-esteem. Self-esteem may result partly from teacher approval (particularly for children), from a student's peers (particularly for adolescents) or because of success. A lack of respect from teachers or other students or being asked to do something where they are almost bound to fail could cause students' frustration and feeling upset. This may lead to disruptive behaviour which in this situation seems to be an attractive option. Through misbehaviour students can impress peers, gain the recognition and attention they need and force the teacher to take them seriously (Harmer, 2001). A disruptive student may encourage other students in the classroom to misbehave and thus gradually influence the whole group (Marciniak, 2015).

External factors.

Family: Family is one of the most influential external factors that behaviouristic, psychodynamic and humanistic models of behaviour recognize that affects human behaviour (Charlton & George, 1993). Over time, the occurrence of behaviour problems has been related to family influences, ranging from poor housing conditions, poverty and low social class background through to more sophisticated elements of disrupted parent-child relationships, parental neglect, child-directed physical aggression or sexual abuse, as well as parental discord, divorce and disturbance (Charlton & George, 1993; Cooper et al., 1994; Gustafsson et al., 2014; Kellam, Ling, Merisca, Brown, & Ialongo, 1998; Marciniak, 2015; McManus, 1989; Muijs & Reynolds, 2001; Ntoliopoulou, 2015). According to McManus (1989), students who suffer from parental neglect and societal indifference may find that violence is effective or that teachers are a safe target for the hate they feel for their family. Moreover, domestic experiences can predispose some students to problematic behaviours and strategies like struggles for attention, revenge and seeking refuge from reminders of traumatic experiences in wild behaviour. Some students behave in inappropriate ways at school since they are the norm in a student's family or social sub-group and their life is characterized by acts of anti-social behaviour, violence and aggression (Cooper et al., 1994). Furthermore, sometimes family attitudes to school, learning in general or teachers themselves can dispose students to cause problems (Harmer, 2001). However, as Charlton and David (1993) argue, students from disadvantaged homes can bring their problems with them, but it is very risky to link disadvantaged homes with disturbed children; as some who are extremely disturbed come from good homes and have stable and affectionate parents and parents who have shown almost superhuman patience and tolerance. On the contrary, many well-adjusted and successful students come from extremely unfavourable backgrounds.

Schools and teachers: The effect of schools in general and teachers in particular on their students' behaviour has been highlighted by a number of scholars (e.g., Allen, 2010; Charlton & David, 1993; Cooper et al., 1994; Fontana, 1994; Marciniak, 2015; Muijs & Reynolds, 2001). Specifically, schools and teachers themselves may provoke behaviours they are attempting to eradicate through harsh and punitive discipline methods or when teachers are too authoritarian or lax on discipline (Allen, 2010; Muijs & Reynolds, 2001). As Charlton and George (1993) mention, children need attention and unfortunately, some students find that the only way to secure this attention is to misbehave. Problems become compounded when students learn that such attempts are successful and their subsequent misbehaviour is often reinforced by the teacher, even unintentionally.

Additionally, some other teacher-related variables that are reported in the literature and may cause misbehaviour are: lower-quality teaching; when pairs or groups finish early and are left unattended; when the teacher comes to the class unprepared or being inconsistent when saying that one action is going to be taken; and when the expectations are too low and classroom activities are not challenging for the pupils or when the teacher expects too much and the competitive attitude is promoted by constant testing and imposing high standards (Allen, 2010; Harmer, 2001; Marciniak, 2015). As Muijs and Reynolds (2001) point out, there is a clear relationship between students' achievement and their behaviour in school and low achievement often leads to inappropriate behaviour as students become disappointed with school. As they suggest the provision of a relevant curriculum that allows all students to experience success can limit misbehaviour. Furthermore, some writers have cited that inappropriate curriculum, chosen topic or activity and lessons which are perceived as boring or irrelevant, may be the reason sometimes that students behave badly; as they show their lack of interest in that way (Charlton & George, 1993; Fontana, 1994; Harmer, 2001). According to Kyriacou (2009), reacting to boredom by misbehaving is not restricted only to low-attaining students but occurs throughout the ability range.

Disorganized classroom and school settings, inconsistency between staff in the ways in which they interpret and enforce school rules and weakness in the use of buildings or timetable may also provoke behaviour problems (Allen, 2010; Charlton & David, 1993). Also, previous learning experiences of all kinds can affect students' behaviour. Specifically, even at the level of the "last teacher let me" students are affected by what happened before and their expectations of the learning experience can be influenced by unpleasant memories or by what they were once allowed to get away with (Harmer, 2001).

Other external factors that may affect students' behaviour are noise from outside the classroom and whether the classroom is too hot or too cold, as this may lead to students being too relaxed or too nervy. Teachers, particularly at primary level, notice significant behaviour changes in different weathers too, for instance a high wind tends to make their children go wild (Harmer, 2001).

Classmates/ Peer Group: Many authors claim that the peer group is also an external factor that could affect students' behaviour (Charlton & George, 1993; Fontana, 1994). The effect of peer groups becomes stronger and more pervasive when children grow up. As children attach themselves to a group they must usually accept and behave according to the consensus attitudes of the peer group (Charlton & George, 1993). Moreover, students experience problems with peers during break (which could spill over into the classroom) and in the classroom which often involve the teacher (Long & Frye, 1985 as cited in Marzano et al., 2003).

Teachers and students' views regarding the causes of misbehaviour.

A number of studies have shown that teachers tend to attribute the cause of misbehaviour to students or family rather than teaching-related factors (Baron, 1990; Ho, 2004; Koutrouba, 2013; Kyriacou & Martin, 2010). In particular, teachers seem to believe that students' misbehaviour is likely caused by factors other than teaching and also, misbehaviour incidents may occur in the classroom, in spite of the teacher's successful classroom management (Muijs & Reynolds, 2001). If the teachers are right, then misbehaviour incidents may appear in one classroom and not in another, not as a result of who is the teacher of the classroom, but of who are the students of this classroom. Considering this, the factor of misbehaviour could be seen as a contextual factor which could affect the consistency of teacher behaviour. Specifically, a teacher who may not have the abilities to deal with misbehaviour, may not be able to demonstrate his/her other teaching skills in classrooms that have cases of student misbehaviour. However, this teacher may be able to teach more effectively and demonstrate different teaching skills in classrooms which do not face discipline problems.

On the other hand, Cooper et al.(1994) mention that the researchers who have investigated the perceptions held by disruptive students, have generally found that these students often view their acts of disruption as rational and justifiable responses to poor teaching. Similarly, based on studies interviewing students about when and why they misbehaved, Kyriacou (2009) refers to four situations that students felt they provoked them to misbehave. These are the following: a) teachers being boring, b) teachers who could not teach, c) teachers whose discipline was weak and d) teachers who made unfair comparisons. In these situations, the students often mention that they found that the teacher's behaviour insulted them in some way and that their misbehaviour was in a great extent an attempt to maintain their sense of self-dignity in the circumstances that confronted them. If the students are right and the teachers are responsible for the appearance of in-class misbehaviour incidents, then teachers' instructional behaviour is most likely to be similar across different classrooms. However, there is not much empirical evidence suggesting whether the students or the teachers are responsible for student misbehaviour.

For instance, Stronge, Ward and Grant (2011) found that the disruptive behaviour of students between the classrooms of the top and bottom-quartile teachers was significantly different. Specifically, the top-quartile teachers had fewer teaching disruptions than the bottom quartile teachers. As they have mentioned, it was possible for the teachers from the higher quartile to have students who had less difficulty behaving in schools. Thus, the differences found between the teachers may better be explained by differences in personalities and dispositions of students. But, they seem not to believe that the differences in students are entirely responsible for the differences in teachers. However, given that they have not examined the disruptive behaviour of the students between the classrooms which are taught by the same teachers or the disruptive behaviour of the same students when they are taught by different teachers, there is no clear answer to the question of whether student's misbehaviour may be caused by the students or teachers.

Nevertheless, in the current study attention is given to the teachers' skills in dealing with misbehaviour problems and not who is responsible for student misbehaviour.

It is important to mention the findings of another study (Kyriakides, Creemers & Panayiotou, 2012) which involved six European countries (Cyprus, Belgium, Greece, Germany, Ireland and Slovenia). In this research the classroom level factors of the dynamic model were measured only through a questionnaire (on a likert scale) which was administered to the students. From the findings of this study, two second order factors have been identified which were not related to each other and were found to have a statistically significant effect on student achievement in each subject. Specifically, the factor of teacher ability to deal with student misbehaviour with the factors of management of time and questioning: raising non-appropriate questions was found to belong to a second order factor (referred to quantity of teaching), whereas the other factors of the dynamic model were found to belong to another second order factor (referred to quality of teaching). The findings of this study imply that the teachers who are able to use teaching time effectively, are not necessarily able to maximize the use of teaching time and vice versa. However, this study was conducted in primary schools and was not in a position to investigate whether a teacher who is not able to demonstrate skills which are related to quantity of teaching, will not also be able to exhibit the other teaching skills which are associated with quality of teaching in different classrooms.

In 1986, Brophy published a paper in which he argued that in order to study the quality of teaching, it will be necessary not only to develop more advanced classroom observation instruments that capture qualitative characteristics, but also to hold the quantity of teaching stable. In addition, he assumes that any attempts to make qualitative comparisons will be defeated by confounding with quantitative differences if researchers also include in the sample teachers who lack the classroom management skills to be able to use effectively the teaching time.

Decreasing misbehaviour incidents.

As Hattie (2009) argues, the presence of disruptive pupils can have a negative impact on their own and on all the other pupils' achievement outcomes. However, the solution is not to remove these students from the classroom but for teachers to acquire skills to ensure that no student unnecessarily disrupts their own or the learning of any other students in the class. There have been many studies and meta-analyses that show the existence of effective programs that are aimed at decreasing disruptive behaviours at individual, classroom (e.g., The Classroom Organization and Management Program) and school level (e.g., School-Wide Positive Behaviour Interventions and Supports) (Freiberg & Lapointe, 2011; Hattie, 2009; Muscott et al., 2008). The implementation of such programs can lead to the reduction of students' behaviour problems, the use of effective managerial and instructional practices by the teachers and finally to the improvement of student achievement both at primary and secondary school level (Evertson, 1995; Lassen et al., 2006; Muscott et al., 2008; Raver et al., 2009; Scott et el., 2009).

Main Conclusions from the Literature Review-Research Agenda

The main conclusions emerging from the literature review are presented in the final section of this chapter. Although the evaluation criteria are a basic aspect that need to be considered when developing a comprehensive teacher evaluation system, there are no universally accepted criteria for measuring teacher effectiveness so far. It is argued that the main theoretical models of TER could be used as a basis upon which evaluation criteria could be established. Many educational systems seem to use mainly the teacher in-class behaviour and/ or student learning outcomes as criteria for evaluating teachers' effectiveness. However, the sole use of student outcomes for teacher evaluation faces considerable challenges and for this reason they have been the source of criticism by several scholars and researchers (e.g., Darling-Hammond et al., 2012; Goe, 2007; Raundenbush, 2004; Torff & Sessions, 2009). The importance of focusing on quality of teaching to evaluate teachers is emphasized, especially since effectiveness studies reveal that what teachers actually do in the classroom is the most important factor at teacher-level associated with student achievement. Particular emphasis is given to the use of classroom observations and student ratings to evaluate teachers and the main advantages and disadvantages of both of them are presented. The need to use multiple sources of data to measure quality of teaching is highlighted. Thus, in the present study not only classroom observations but also student questionnaires are used to measure teacher in-class behaviour.

Findings of studies on teacher effectiveness and theoretical and empirical models of EER can provide an answer to the question of what constitutes effective teaching. Examining the notion of what constitutes effective teaching is important, as definitions propose and shape what needs to be measured. The main phases of TER are discussed to demonstrate the growth that this field has met through the years. Then, the rationale of the main models of educational effectiveness that have integrated teacher effectiveness factors and school effectiveness findings is described. It is argued that by moving from Carroll's model to the comprehensive model of educational effectiveness, the concept of quality of teaching has been developed in more detail, using the results of TER, but it is not defined precisely. In addition, an important weakness of the comprehensive model is that only traditional teaching approaches, like the direct teaching approach, have been taken into account and not the new theories of teaching. Findings of meta-analyses of teacher effectiveness studies have shown that within each approach there are factors which are associated with student outcomes. This implies that an integrated approach in defining quality of teaching should be adopted.

The dynamic model, which is a further development of the comprehensive model, takes into account the main findings of EER and the weaknesses of the previous models. This model, which provided the theoretical basis of the current study, provides a clear definition of quality of teaching through eight generic factors included at the classroom level. The main assumptions of the model and the classroom-level factors of the model are presented. Longitudinal studies and meta-analyses provided support to the importance of the classroom-level factors for explaining variation in different types of student learning outcomes, in different phases of schooling and in different countries. It was therefore argued that these factors can be considered generic. However, given that these studies took place at primary and pre-primary school level, they were not in a position to investigate the extent to which the classroom context affects teacher behaviour or to identify the effect of classroom-level factors on achievement gains in mathematics of secondary school students.

The importance of examining whether teachers exhibit the same generic teaching skills when they teach in different classrooms, especially for teacher evaluation, is stressed. The critical review of studies investigating the extent to which measuring teaching skills by classroom observations is influenced by classroom context variables has shown that much uncertainty still exists about the classroom context effects on teacher behaviour. The limitations of these studies are discussed and it is argued that further research is needed to observe teachers teaching in different classrooms in order to determine whether differences in observational ratings are related to student characteristics and not to the non-random sorting of teachers to classes of students. In this study emphasis is given to the effect of misbehaviour incidents (as a contextual factor) on teaching quality. The reasons for choosing misbehaviour as one of the basic contextual factors which may affect the consistency of teacher behaviour in different classrooms are discussed.

Taking all the above into account, the questions that still need further investigation and for which this study aims to provide answers, are whether and to what extent the classroom context affects teacher behaviour in the classroom and teacher effects on student achievement. Moreover, this study aims to investigate further issues, which according to the literature review, need further investigation. Specifically, this research aims to investigate the following questions: (a) Does the type of instrument used to measure teaching skills (i.e., high, low-inference observation instrument and student questionnaire) contribute to whether similar judgments are produced when the same teacher is evaluated across different classrooms? (b) To which extent are the classroom-level factors of the dynamic model and their dimensions associated with student achievement in mathematics of secondary school students? (c) Which level explains more variance in student achievement in mathematics of secondary school students, the teacher or the classroom level? In order to provide answers to these questions a quantitative research was conducted. The research design, the participants, the research instruments and the methods of data analysis are presented in the next chapter.

CHAPTER 3

METHODOLOGY

In this chapter, the methodology used to investigate the research questions is presented. Specifically, the research design used is described and its selection is justified in contrast to other research designs. Then, the research sample, the research variables at student and teacher levels and the research instruments are described in detail. In addition, the statistical techniques employed to analyse the research data, are elaborated. Finally, some limitations of this study are discussed.

Justification of the Research Method Chosen

The selection of a research design depends mainly on the purpose of each study. The experiment is the best design for studies that seek to search for cause-and-effect relations (Slavin, 2010). However, an experimental design was not chosen to provide a basis for the present study as its main aim was not to demonstrate cause-and-effect relations, but to investigate the consistency of teacher behaviour in different classrooms.

In order to examine whether teachers exhibit the same generic teaching skills when they teach mathematics in different classrooms, data about the skills of each teacher were gathered from all of the teacher's classes of the same age group of students (grade 7 or grade 8 but not on both grades). These data were collected by using external observations and a student questionnaire. It was chosen to collect data on teachers' in-class behaviour from more than one classroom within a school year, instead of collecting data of teacher behaviour in different school years. The reason is that teachers' in-class behaviour may vary over time, especially when teachers participate in professional development programs as professional development studies seem to reveal (e.g., Antoniou & Kyriakides, 2011). It is important to note that all the observations were conducted during the same period in all the classes taught by the same teacher in order to keep constant not only the teacher and the subject, but also the time. Moreover, the study was restricted to seventh or eighth grade classrooms to keep constant not only the age of the students, but also the curriculum that teachers were expected to deliver. This is because all classes in Cypriot public lower secondary schools are mixed-ability and all students are taught the same grade-level curriculum (Eurydice, 2004).

As discussed in Chapter 1, this study also aims to examine the effect of the classroom-level factors of the dynamic model on student achievement of secondary students. Therefore, data on student achievement in mathematics were collected at the beginning and at the end of the school year 2014-2015, where this study was conducted. In that way, a longitudinal design (Gustafsson, 2010; Hedeker & Gibbons, 2006) was adopted, since data of the same units (i.e., students) were collected at more than one point in time contrary to the cross-sectional design in which data are collected only once. However, it should be taken into account that in the present study only two measurement periods have been used to collect data on student achievement, since the main focus of the study was on the investigation of the consistency of teacher behaviour in different classrooms within a school year. Therefore, due to this limitation, only the short-term effects of the classroom-level factors of the dynamic model on student achievement of secondary school students could be examined. Longitudinal studies that last for more than one year are needed to investigate the stability of the effects of the classroom-level factors over time and/or the long-term impact of the factors (e.g., Dimosthenous, Kyriakides, & Panayiotou, 2018). Nevertheless, the stability and/or the long-term effects of the factors were beyond the scope of this study.

Research Design

The research design of this study consisted of two main steps described below. It is important to note that the observation instruments and the student questionnaire used in this study to collect data on teachers' in-class behaviour have already been used and validated in previous studies testing the validity of the dynamic model (e.g., Kyriakides & Creemers, 2008, 2009). Therefore, their validity did not need to be examined. Thus, the aim of the first step was the development and validation of the written tests used to measure student achievement in mathematics. Specifically, during the school year 2013-2014 a battery of three written tests (test 6, 7 and 8) were developed to assess knowledge and skills in mathematics which are identified in the national curriculum of Cyprus. A pilot study was carried out between May and June in 2014 in order to examine the construct validity of the tests. The validation study involved the administration of the three tests to 484 students (test 6 to160 sixth grade students, test 7 to 171 seventh grade students and test 8 to 153 eighth grade students). The Extended Logistic Model of Rasch (Andrich, 1988) was used to analyze the data that emerged from each test. Further details on how the tests were developed and validated are presented in the following sections of this chapter.

The second step pertained to the main study where the data collection took place. The data collection was carried out in three phases. In the first phase, which was held at the beginning of the school year 2014-2015, data were collected from the student sample (*n*=915) regarding their achievement in mathematics by using external forms of assessment. As mentioned before, the written tests were developed and validated in the previous year.

The second phase of the data collection process aimed to examine whether teachers exhibit the same generic teaching skills when they teach in different classrooms. Thus, data about the skills of each teacher (*n*=26) were gathered from more than one classroom by using external observations. Two observations in each class (*n*=57) of the teacher sample were conducted (*n*=114) by using one high and two low-inference observation instruments. Specifically, the high-inference observation instrument was used twice in each class of grade 7 or grade 8 of the participating teachers. Each of the two low-inference observation instruments was used only once in each class due to time constraints, as all the observations were conducted by the same well-trained external observer. The classrooms' observations were conducted between November 2014 and March 2015.

The third and final phase of the data collection process aimed to collect data on student final achievement in mathematics in order to examine the student progress over time and determine the effect of the classroom-level factors of the dynamic model on student achievement gains. Thus, data on student achievement were collected at the end of the school year 2014-2015 by using external forms of written assessment. Information about the students' background characteristics (gender: boys or girls and ethnicity: students', fathers' and mothers' country of birth and language that students speak at home) was collected from a short questionnaire included in the written tests of mathematics. Information regarding students with special educational needs (SEN) of the sample was collected from teachers. In addition, a questionnaire was administered to the student sample in order to gather data on their teacher's instructional behaviour. The questionnaire was administered at the end of the school year so that the students would have the time to get used to their teacher's in-class behaviour. The timeframe of the present study is presented in Table 3.1.

Table 3.1

Study timeframe

Study steps	Timeframe	Actions
Step 1: *Pilot study*	February-April 2014	• Development of the three mathematics achievement tests
	May -June 2014	• Validation study of the achievement tests
	June – July 2014	• Analysis of the data emerged from the pilot study • Final version of the tests
Step 2: *Main study*	*Phase 1*: September – October 2014	• Mathematics test administration to the student sample (prior achievement data)
	Phase 2: November 2014 – March 2015	• Two classroom observations in each class of the teacher sample
	Phase 3: April-May 2015	• Mathematics test administration to the student sample (final achievement data) • Student questionnaire administration

Research Sample

Pilot study.

In the validation study of student achievement tests of mathematics (conducted during the school year 2013-2014), a stage sampling procedure (Cohen et al., 2007) was used. At the first stage, six Greek Cypriot primary public schools and five Greek Cypriot lower secondary public schools were selected. A purposive sampling procedure was used to select the schools (easy access) rather than a random sampling. Then, nine sixth grade classes, eight seventh grade classes and eight eighth grade classes were purposively selected. All the students of the class sample (*n*=484: 160 sixth grade students, 153 seventh grade students and 171 eighth grade students) participated in the study. These schools were excluded from the main study.

Main study.

The sample of the main study consisted of teachers who taught mathematics in more than one classes of grade 7 or grade 8, and the students of these classes. It was not chosen to collect data from teachers who taught mathematics in different classes of grade 9 and their ninth grade students in order to develop less written tests to measure student achievement. Regarding the selection of the teacher and the student sample, a stage sampling procedure was used. At the first stage, 13 out of 36 lower secondary public schools of two districts in Cyprus (i.e., Nicosia and Larnaca) were selected and 12 agreed to participate. Particularly, in Larnaca district, 10 out of 11 lower secondary schools were selected and 9 agreed to participate. The reason for not choosing all the lower secondary schools in Larnaca district was to reduce the cost of the research, since one of them was very far from the town of Larnaca where the researcher lived. In Nicosia district, 3 out of 25 schools were selected and all of them agreed to participate. The criterion of choosing these three schools in Nicosia was that they were very close to the town of Larnaca. Therefore, the main criterion of choosing the school sample was to keep the cost as low as possible.

Then, 26 teachers (19 women and 7 men) of the school sample who taught mathematics in at least two classes of the same age group of students (grade 7 or grade 8) participated in the study. All the participating teachers were subject-matter specialists who taught only mathematics. Participating teachers' years of experience ranged from 4 to 23 years, with the mean of the teaching experience estimated at 11.5 (SD= 5.19). It is important to note that for the purpose of this study it was not necessary to have a representative teacher sample of the teacher population of lower secondary schools in Cyprus, as this study does not aim to examine how the secondary school teachers in Cyprus behave in the classroom in general. However, it was examined whether the teacher sample was nationally representative in terms of gender and years of experience, where the

Cyprus Ministry of Education and Culture keeps data. The chi-square test did not reveal any statistically significant difference between the teacher sample and the population in terms of gender (X^2= 0.02, d.f.=1, p=0.88). Moreover, the *t* test did not reveal any statistically significant difference between the teacher sample and the population in terms of their years of experience (t=-1.56, d.f.=25, p=0.13). Therefore, although the teacher sample was not randomly selected and came only from two districts of Cyprus, it has the same characteristics as the national sample in terms of gender and years of experience.

All students from all the classrooms of grade 7 or 8 (*n*=57: 32 classes of grade 7 and 25 classes of grade 8) of the teacher sample, were chosen to participate in this study. The students were required to have parental permission to participate. Almost all the students received permission. Specifically, a total number of 915 students, 426 boys (46.6%) and 489 girls (53.4%) participated in the study. Out of the 915 students, 500 attended grade 7 (54.7%) and 415 attended grade 8 (45.3%). The X^2 revealed that the student sample was representative of the student population of Cyprus in terms of gender (X^2= 1.334, d.f.=1, p=0.25). In addition, the *t* test revealed that the student sample was representative of the student population of Cyprus in terms of the size of the class (*t*=1.205, d.f.=56, p=0.24). It is important to note that regarding the size of the classes of the class sample, all the classes were more or less the same due to the fact that the system in Cyprus is centralised and there are specific regulations about the class size. Moreover, it should be noted that 8.2% of the original sample (75 students) was excluded from the analysis because of the missing post attainment data. Therefore, only 840 out of 915 students were used in the analysis. We assume that these data are missing completely at random due to the fact that the dropout rate in Cyprus is very low (i.e., approximately 0.3% in lower secondary education) (Government of Cyprus, 2018). It should also be noted that all classes in Cypriot public lower secondary schools are organized by age and are mixed-ability. However, students must achieve a minimum level of competence to proceed from one class to another (for more information about the educational system of Cyprus see Ministry of Education and Culture [MOEC], n.d. and Eurydice, 2004).

Data Collection and Research Variables

In order to investigate the research questions set, data concerning teachers' generic teaching skills as well as student achievement in mathematics, were collected. The instruments used were a) low and high-inference observation instruments, b) student questionnaire and c) mathematics achievement tests. The data collection instruments with the corresponding research variables are described below.

Student written tests to measure student achievement.

Since one of the aims of the study was to examine the effect of the classroom-level factors of the dynamic model on student achievement of secondary school students, student tests were needed in order to measure student achievement. Therefore, during the school year 2013-2014, a battery of three written tests (test 6, 7 and 8) were developed to assess knowledge and skills in mathematics of grades 6-8. Mathematics has been chosen because it is a core subject in all countries and relatively culturally free, a fact that gives the possibility for the implementation of some cross-cultural analysis in the future. Specifically, in mathematics, unlike language or history lessons, even if the cultures vary, the content remains similar (Cai & Lester, 2007). In addition, previous studies suggest that schools/ teachers have more impact in the area of mathematics and science than in the area of language (Scheerens, 2016). Moreover, the correction of mathematics tests is relatively more economical in terms of time compared to other subjects like languages. Furthermore, student achievement data in mathematics were also collected in previous studies conducted in primary schools to test the validity of the dynamic model (e.g., Creemers & Kyriakides, 2008; Kyriakides & Creemers, 2008, 2009; Kyriakides et al., 2009).This gives the possibility for a comparison of the results of the present study with the results of the previous studies conducted in primary schools, which were also concerned with the same subject.

The construction of the tests was based on the review of the national curriculum of Cyprus, student books and instruments measuring mathematics skills of Grades 6-8. All tests comprised of tasks classified under two different categories (i.e., (a) Understanding concepts- performing algorithms and computations and (b) solving problems) as well as different content domains. The ratio between tasks related to the two aforementioned categories as well as the ratio of tasks pertained to the different content domains was based on the corresponding ratio of tasks included in the student books of grades 6-8 (In Cyprus all schools use the same mathematics books). For each test, scoring rubrics were developed in order to differentiate between up to three levels of proficiency (i.e., 0-2) in each task, leading to the collection of ordinal data about the extent to which each student had obtained each skill of mathematics. In order to equate the three tests and make test scores comparable enough, common items were used (i.e., approximately 40%- 60% of the number of items of each test) with representative content to be measured (Kolen & Brennan, 2004).

A validation study was carried out at the end of the school year 2014. The pilot study involved the administration of the three tests to 484 students. All the tests were administered and corrected by the researcher in order to increase the reliability of the measurements. The Extended Logistic Model of Rasch (Andrich, 1988; Boone, 2016) was used to analyse the ordinal data emerged from each test, using the computer program Quest (Adams & Khoo, 1996). Three scales which refer to student achievement in mathematics of grade 6, 7 and 8 were created and analysed for reliability, the fit to the model, meaning and validity. Analysis of the data showed that each scale had relatively satisfactory psychometric properties. Specifically, for each scale the indices of cases (i.e., students) and items separation were found to be higher than 0.87 (on a 0 to 1 scale), indicating that the separability of each scale was satisfactory (Wright, 1985). As noted by Wright and Stone (1999), the higher the number, the better the separation that exists and the more precise the measurement. In addition, the mean infit mean squares and the mean outfit mean squares of each scale were found to be near one and the values of the infit t-scores and the outfit t-scores were approximately zero (see more details regarding the Rasch parameter estimates in Appendix A). Thus, each analysis revealed that there was a good fit to the model (Keeves & Alagumalai, 1999) providing support to the validation of the three tests. Five out of 59 items were removed from test 8 based on the review of the “fit” statistics (i.e., MNSQ Item Infit) of the test.

The final version of the tests was used in the main study in order to collect data on student achievement in mathematics. The tests were administered to the student sample at the beginning and at the end of the school year 2014-2015 by the researcher. The initial administrations of the tests (pre-test) aimed to measure the aptitude as an explanatory variable at the student level (phase 1 of the main study). Aptitude has to do with the degree in which a student is able to perform the next learning task (Kyriakides & Creemers, 2008). For the purpose of the present study, aptitude consists of prior knowledge in mathematics at the beginning of the school year (i.e., baseline assessment). Specifically, three tests were used. At the third phase of the main study, post-test for each grade was used, covering in this way the final assessment of students’ mathematics achievement for grade 7 and grade 8. The pre-test for grade 8 was used as a post-test for grade 7. Table 3.3 presents the test administration procedure.

Table 3.2

Procedure of pre- and post-student test administration

Grade	Grade 7	Grade 8
Pre-test	Test 6	Test 7
Post-test	Test 7	Test 8

As in the pilot study, the analysis was made by the Extended Logistic Model of Rasch which revealed that each scale had satisfactory psychometric properties. Particularly, four scales which refer to student achievement in mathematics at each of the two time-points were created (i.e., beginning of grade 7 and grade 8, as well as end of grade 7 and grade 8). Table 3.3 provides a summary of the four scales statistics for the student sample. As can be seen from this table, the mean of students of each scale indicates a quite well matching of the sample ability with the test's difficulty. The standard deviation values for student estimates of each scale indicate that student achievement scores of each grade are quite spread out from the mean. In other words, the standard deviation values for student estimates reveal that the student sample had big variance in terms of their achievement scores. This may be attributed to the fact that all classes were mixed-ability. It should be noted that these values emerged before applying the method of test equating. Reliability was calculated by using the Item Separation Index and the Person (i.e., student) Separation Index. A value of 1 represents high separability in which errors are low as well as item difficulties and students' abilities are well separated along the scale and a value of 0 represents low separability (Wright & Masters, 1981). We can observe that for each scale the indices of student and item separation were higher than 0.83, indicating that the separability of each scale is satisfactory (Wright, 1985). The mean infit mean squares and the mean outfit mean squares of each scale were in all cases very close to the Rasch-model expectations of one and the values of the infit t-scores and the outfit t-scores were approximately zero. Therefore, for each student, two different scores for his/her achievement in mathematics (one for each test: pre-test and post-test) were calculated using the relevant Rasch person estimate in each scale.

Table 3.3

Initial and final Rasch parameter estimates of test scores of the student sample

Parameter Estimates		Initial Data		Final Data	
		Grade 7 (N=461)	Grade 8 (N=379)	Grade 7 (N=461)	Grade 8 (N=379)
Mean	Items	0.00	0.00	0.00	0.00
	Students	-0.64	-0.64	-1.08	-0.99
SD	Items	1.43	1.98	2.03	1.29
	Students	1.24	1.37	1.52	1.60
Reliability	Items	0.99	0.99	0.99	0.99
	Students	0.83	0.93	0.94	0.92
Mean Infit mean square	Items	1.00	0.99	0.99	1.00
	Students	1.00	1.01	1.01	1.00
Mean Outfit mean square	Items	1.06	1.03	0.94	1.03
	Students	1.06	1.03	0.94	1.03
Infit t	Items	0.03	-0.10	-0.02	-0.12
	Students	0.05	-0.05	0.01	0.00
Outfit t	Items	-0.08	0.07	-0.06	0.08
	Students	0.10	0.03	0.08	0.07

Student background characteristics.

Information was collected on students' gender (0=boys, 1=girls), as well as ethnicity: students', fathers' and mothers' country of birth (0=Cyprus, 1=other country) and language that students speak at home (Greek/ Greek and another language/ another language) from a short questionnaire included in the post- tests of mathematics. Cyprus was the country of birth of 87% of the students, 71.9% of the students' mother and 80.8% of the students' father. Regarding the language that students speak at home, 73.1% of the student sample used in the analysis spoke Greek at home, 21.1% spoke Greek and another language and 5.8% of the participants answered that they spoke another language than Greek. Information regarding students with special educational needs (SEN) of the sample

(0=students with no special educational needs, 1=students with special educational needs) was also collected from their teachers. Only 5.4% of the student sample used in the analysis concerned students with special educational need (see more details regarding the student background characteristics of the sample used in the analysis in Appendix B). However, it should be noted that no information was collected for socio-economic status (SES).

Classroom-level factors.

The classroom-level factors of the dynamic model were measured by both an independent observer and students, taking into account the five dimensions of each effectiveness factor. Specifically, one high-inference and two low-inference observation instruments and also a student questionnaire, which have been used in various studies testing the validity of the dynamic model, were used (e.g., Azigwe, Kyriakide, Panayiotou, & Creemers, 2016; Creemers & Kyriakides, 2008; Kyriakides & Creemers, 2008; Kyriakides et al., 2009). It is important to note that support to their validity and reliability has already been provided by the previous studies (see for instance Kyriakides & Creemers, 2008). All these instruments are intended to measure the classroom-level factors of the dynamic model irrespective of grade or subject, making them applicable to all teachers.

In general, low-inference observation instruments require a minimal amount of observer judgment, relying largely on counting the behaviours one wants to study. On the contrary, high-inference observation instruments depend on more observer judgment, as when observers are asked to rate a teacher's behaviour on a rating scale (Muijs, 2006). Judgments that emerge from both kinds of instruments result in numerical scores. The first low-inference observation instrument used in the present study is based on Flanders' system of interaction analysis (Flanders, 1970). Nevertheless, a classification system of teacher behaviour that is based on the way each factor of the dynamic model is measured, was developed. For instance, in order to measure the quality dimension of teacher behaviour in dealing with misbehaviour, which is an element of the classroom as a learning environment factor, the observers are asked to identify any of the types of teacher behaviour in the classroom that follow:

a) The teacher is not using any strategy at all to deal with a disorder problem.

b) The teacher is using a strategy that has a long-lasting effect.

c) The teacher is using a strategy, but the problem is only temporarily solved.

A classification system of student behaviour was also developed. Therefore, the use of this instrument helps us to gather data about the three elements of the factor classroom as a learning environment (i.e., teacher-student and student-student interaction, and teacher behaviour in dealing with misbehaviour) and the management of time factor. The use of the second low-inference observation instrument enables the collection of information regarding the following five factors of the dynamic model: orientation, structuring, teaching modelling, questioning techniques and application. This instrument was designed in a way that enables us to collect more information concerning the quality dimension of these five factors. The high-inference observation instrument covers all the classroom-level factors of the dynamic model except the assessment, and observers are expected to rate the teacher's behaviour on a Likert scale ranging from 1 (minimum point) to 5 (maximum point). The three observation instruments are presented in a book written by Creemers and Kyriakides (2012). As mentioned before, these observation instruments were used in a series of studies conducted in Cyprus and other countries and their construct validity had already been examined by using Structural Equation Modelling (SEM) approaches (see Kyriakides & Creemers, 2008).

In this study the high-inference observation instrument was used twice in each class of grade 7 or grade 8 of the participating teachers. Each of the two low-inference observation instruments was used only once in each class. It is important to clarify that each participating teacher was observed teaching only in all of his/her classes of grade 7 or all of his/her classes of grade 8, but not in his/her classes of both grades (e.g., one seventh grade class and two eighth grade classes).Observations lasted 40-45 min (i.e., one math lesson) and all the observations were conducted during the same period in all the classes taught by the same teacher. The content of the observed lessons in all the classes taught by the same teacher was not the same, since the participating teachers were free to teach whatever content they wanted in each of their class. The classrooms' observations were conducted between November 2014 and March 2015 by the researcher who previously had been trained to the use of the three observation instruments. The observer did not have any kind of relationship with the participating teachers and, thus, did not have to make judgments in the context of being a supervisor or co-worker. It is important to note that the inter-rater reliability of the data emerged from the high-and low-inference observation instruments could not be tested since observations were conducted by only one observer. However, in previous research studies, this observer was part of the team that had used these instruments to collect observation data. In these studies, the inter-rater reliability of the team was tested and was found to be higher than 0.80 (Charalambous, Kyriakides, Tsangaridou, & Kyriakides, 2017; Kyriakides, Christoforidou, Panayiotou, & Creemers, 2017).

A questionnaire was also administered to the students of all classes of grade 7 and 8 of the participating teachers at the end of the school year in order to gather data on their teacher's instructional behaviour in relation to the eight factors and their dimensions of the dynamic model (see Appendix C). Specifically, students were expected to indicate the extent to which their teacher behaved in a certain way in their classroom on a 5-point Likert scale. For instance, an item related to the stage dimension of the structuring asked students to indicate whether the teacher explains how the new lesson is related to previous ones at the beginning of the lesson. A few modifications were made to adapt the questionnaire to the context of teaching mathematics at secondary school level. For example, the question about whether and how often students have tests, was dropped due to the fact that tests in lower secondary education are obligatory, at least one for each trimester (MOEC, n.d.).

Analysis of Data

After the completion of the main study as well as the cleaning and preparation of the data, the research data were analysed by using several statistical techniques. One of the main research questions of this study is whether secondary school teachers exhibit the same generic teaching skills when they teach mathematics in different classrooms within a year and whether the findings are differentiated according to the instrument (low and high-inference observation instruments and student questionnaire) that is used. Thus, a generalisability study (Shavelson, Webb, & Rowley, 1989; Marcoulides & Kyriakides, 2010) was conducted in order to examine the consistency of teacher behaviour (as measured by different instruments) across different classrooms. In addition, multilevel modelling techniques (Snijders & Bosker, 1999) were used in order to answer the other two research questions, to which extent the classroom-level factors of the dynamic model and their dimensions are associated with student achievement in mathematics of secondary school students in Cyprus; and whether the teacher or the classroom level explains more variance in student achievement. The methods used to analyse the research data are described in detail below.

Section A: Investigating the Consistency of Teacher Behaviour by Conducting a Generalisability Analysis

In order to examine whether teachers exhibit the same teaching skills when they teach in different classrooms, a generalisability study was conducted. By conducting a generalisability study, the extent to which observation data and/ or student questionnaire data are generalisable at the level of teacher and/or classroom, can be identified. This is because data have been obtained from different classrooms of the same teacher. According to Shavelson et al. (1989), the Generalisability Theory asks how accurately observed scores allow to generalise about persons' behaviour in a defined universe of situations. If a dimension of a classroom-level factor, as measured by the observation instruments or the student questionnaire, is found to be generalisable at the teacher level, then this will imply that teacher behaviour regarding this dimension of this factor is consistent from classroom to classroom.

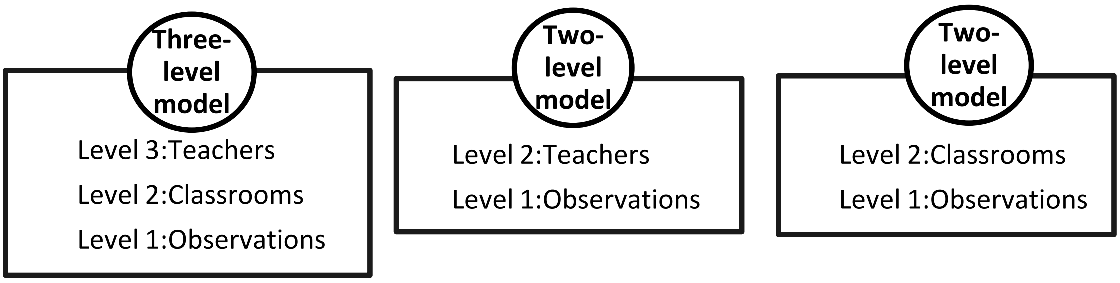
1. Using the analysis of variance (ANOVA) to test the generalisability of the observation data at the level of the teacher.

For the two low-inference observation instruments, which were used only once in each class of grade 7 or grade 8 of the participating teachers, one-way ANOVA was conducted with the use of SPSS software. One-way ANOVA would allow the investigation of search for statistical significant differences, identifying in that way if there is homogeneity in the observation scores obtained from different classrooms which were taught by the same teacher. Thus, ANOVA would help to find out whether the data that emerged from the low-inference observation instruments are generalisable at the level of the teacher.

2. Using multilevel modelling techniques to test the generalisability of data at the level of the teacher and/or classroom.

For the high-inference observation instrument, which was used twice in each class, and the student questionnaire, multilevel analyses of data (Luyten & Sammons, 2010; Snijders& Bosker, 1999) were conducted with the use of MLwiN software (Rasbash, Steele, Browne, & Goldstein, 2012). The only reason for conducting multilevel analysis, for the data emerged from these two instruments was the different nature of these data (e.g., two observation scores per class taught by the same teacher) compared to the data emerged from the low-inference observation instrument (i.e., one observation score per class taught by the same teacher). The multilevel analysis and more specifically the empty model, which only contains random groups and random variation within groups, would allow the investigation of within- group homogeneity in the observation scores and student questionnaire scores. Thus, multilevel analyses of data would help to find out whether the data that emerged from the high-inference observation instrument and the student questionnaire are generalisable at the level of the teacher and/or classroom. Specifically, for each score emerged from the high -inference observation instrument, we ran three alternative empty models: the three-level model; observations within classrooms within teachers, the two level model; observations within teachers and another two level model; observations within classrooms (see Figure 3). Therefore, the repeated observations per teacher on the same factors represent the lowest level.

Figure 3.1. Alternative multilevel models for the analysis of consistency of teacher behaviour in different classrooms



The three-level model was found to fit less well compared to the other two models. This may be attributed to the fact that in most of the cases, teachers who participated taught only in two classes of grade 7 or 8. In regard to the two different two-level models, one might have expected the two-level model (observation within classrooms) to fit better to the data than the other two-level model (observation within teachers) due to the fact that it has larger higher-level sample size (classrooms=57 and teachers=26). Larger samples tend to have smaller standard errors and greater statistical power. However, the two-level model (observations within teacher) was found to fit better to the data than any other model for most factors.

The same approach was followed for all the factors and their dimensions measured by the student questionnaire. In this case, level 1 of each model comprised the students and not the observations. It is important to note that before the creation of factor scores, in the case of the student questionnaire, a generalisability study on the use of students' ratings was conducted (using ANOVA) in order to investigate whether the data collected from all the questionnaire items could be used for measuring the quality of teaching at the level of the classroom. For some dimensions of the classroom-level factors, which are presented in Chapter 4, it was not possible to generate factor scores, as some items of the questionnaire concerned with these dimensions were not found to be generalisable at the level of the classroom, and hence had to be removed. The score for each teacher (in each class) in each of the questionnaire item found to be generalisable, was the mean score emerged from the responses of the students of each class she or he taught.

Section B: Measuring the Impact of the Classroom-Level Factors on Student Achievement by Using Multilevel Analysis

Multilevel modelling techniques (Goldstein, 1999; Snijders, 2011) were used to investigate the short-term effect of the classroom-level factors on student achievement by analyzing the data that emerged from classroom observations and the student questionnaire with the use of MLwiN software. Multilevel analysis was considered appropriate because of the hierarchical structure of the research data (i.e., students within classrooms within teachers). This kind of analysis assists the identification of the contribution of each level to the variance of the student achievement (Luyten & Sammons, 2010).

The first step of this analysis was to run a three-level model (students within classrooms within teachers) without any explanatory variables (empty model) to determine the variance at each level. The empty model contains random groups and random variation within groups. It can be expressed as a model where the dependent variable is the sum of a general mean (β_0), a random effect at the teacher level (V_{0k}), a random effect at the classroom level (U_{0jk}) and a random effect at the individual level (R_{ijk}). This is expressed by the following equation

$$Y_{ijk} = \beta_0 + V_{0k} + U_{0jk} + R_{ijk} \quad \text{(empty model)}$$

Where:

k: is level-3 units (i.e., number of teachers)

j: is level-2 units (i.e., number of classes)

i: is level-1 units (i.e., number of students)

Y_{ijk} = student achievement in mathematics at the end of the school year of student i, who is derived from class j and is taught by teacher k.

The random variables V_{0k}, U_{0jk} and R_{ijk} are assumed to have a mean of 0 (the mean of Y_{ijk} is represented by β_0) and to be mutually independent. The empty model is important as it provides the basic partition in the variability in the data between the three levels (Snijders & Bosker, 1999).

However, the variance at the level of the classroom was not found to be statistically significant. For this reason two different two-level empty models (students within classrooms; and students within teachers) were run. The two-level model, which takes into account the student and teacher level, was found to fit better than any other model (see Table 3.4). The likelihood statistic (X^2) of the two-level model (students within teachers) was smaller than the likelihood statistic of the other two-level model (students within classrooms). Therefore, to test the effect of the classroom-level factors on students' achievement in mathematics, a two-level model with the teacher as the higher and the student as the lower level was used. This is expressed in the equation below:

$$Y_{ijk} = \beta_0 + U_{0j} + R_{ij} \quad \text{(empty model)}$$

Where:

j: is level-2 units (i.e., number of teachers)

i: is level-1 units (i.e., number of students)

Y_{ij} = student achievement in mathematics at the end of the school year of student i, who is taught by teacher j.

Table 3.4

Parameter Estimates and (Standard Errors) for the analysis of mathematics achievements-Empty model

	Three- level model (student-class-teacher)	**Two-level model (student-teacher)**	**Two-level model (student-class)**
Fixed part (Intercept)	-1.088 (0.107)	-1.088 (0.107)	-1.098 (0.083)
Variance components			
Teacher	0.603 (0.083)	0.623 (0.083)	
Class	0.104 (0.084)		0.621 (0.074)
Student	2.457 (0.122)	2.457 (0.122)	2.460 (0.124)
Significance test			
X^2	3176.52	3174. 518	3187.321

Then, students' prior knowledge, SEN, gender and ethnicity: language that students speak at home, were added to the empty model. It should be noted that in order to measure the effect of ethnicity: language that students speak at home on student outcomes, students who spoke only Greek at home were treated as reference group and two dummy variables were entered in Model 1. Model 1 refers to the student background factors that were found to have an effect on student achievement at the end of the school year (after controlling the effect of student achievement at the beginning of the school year). The other variables of ethnicity (students', fathers' and mothers' country of birth) correlated with achievement when they were studied in isolation, but, because of multicollinearity, their effects disappeared when they were studied together with the other variables of ethnicity and were removed from the analysis. Model 1 is expressed by the following equation:

$$Y_{ij} = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \beta_3 X_{3ij} + \beta_4 X_{4ij} + \beta_5 X_{5ij} + U_{0j} + R_{ij} \quad \text{(Model 1)}$$

where X_1= prior achievement and X_2, X_3, X_4, X5 = those student background factors found to be associated with achievement at the end of the school year.

In Model 2, a variable related to the test administration date was added to Model 1. Specifically, given that the post-tests were administered to the students by the researcher, some students might have taken their tests on different dates where a review of all the mathematics lessons of the school year might have preceded in their class compared to other students. Thus, an attempt was made to control the influence of the differences on test administration dates between the classes that might have an effect on the dependent variable by entering the variable post-date (0=27/4-13/5, 1=14/5-19/5 2=20/5-27/5). This variable was found to be associated with achievement at the end of the school year. Model 2 is expressed in the equation below.

$$Y_{ijk} = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \beta_3 X_{3ij} + \beta_4 X_{4ij} + \beta_5 X_{5ij} + \beta_6 X_{6j} + U_{0j} + R_{ij} \quad \text{(Model 2)}$$

Where X_6 = the variable related to the test administration date.

At the next step, different versions of Model 3 were established. In each version of Model 3, factor scores which refer to the classroom-level factors of the dynamic model as emerged from each instrument were added one by one to Model 2 to avoid multicollinearity; since one of the assumptions of the dynamic model is that factors operating at the same level may be related to each other. In addition, this approach was deliberately chosen, as the SEM analyses, which were conducted in a previous study testing the validity of the dynamic model (Creemers & Kyriakides, 2008), have shown that the dimensions of the same factor are interrelated. In this way, the impact of each classroom-level factor and their dimensions was examined separately. The equation below (Model 3a) investigates the impact of the frequency dimension of application (as measured by the low-inference observation instrument) on student achievement at the end of the school year:

$$Y_{ijk} = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \beta_3 X_{3ij} + \beta_4 X_{4ij} + \beta_5 X_{5ij} + \beta_6 X_{6j} + \beta_7 \text{(Application-frequency)}_j + U_{0j} + R_{ij} \quad \text{(Model 3a)}$$

It is important to mention that the explanatory variables, with the exception of grouping variables, were centred as Z-scores with a mean of 0 and a standard deviation of 1. This is a way of centring around the grand mean (Bryk & Raudenbush, 1992) leading to comparable effects. Therefore, each effect shows how much the dependent variable increases or decreases (in case of a negative sign) by each additional deviation on the independent variable (Snijders & Bosker, 2011). Grouping variables were entered as dummies with one of the groups as baseline (e.g., students with no special educational needs=0).

Research Limitations

As described in the previous sections of this Chapter, this study was designed so as to collect data from different classrooms which were taught by the same teacher in order to investigate whether teachers exhibit the same generic teaching skills in different classrooms. In addition, this study was designed to allow the investigation of the effect of the classroom-level factors of the dynamic model on student achievement in mathematics of secondary school students. However, the following limitations should be acknowledged.

One possible limitation of the study is that the data were obtained by a small number of classes of the same teacher, whereas in most of the cases the data were collected from only two classes of the same teacher. This is because it was decided to keep the grade constant and the majority of the participating teachers taught only in two classes of the same age group of students (Grade 7 or Grade 8).

Another possible limitation of the study is the fact that the classroom observations were conducted by the same person, thus there is a potential of rater bias. However, in order to reduce the possibility of bias, a well-trained observer was used, who attended a series of seminars on how to use the three observation instruments. Moreover, not only classroom observations but also a student questionnaire were used to evaluate the generic teaching skills of the participating teachers.

Additionally, the fact that the classroom observations were conducted by the same observer, did not give the opportunity to use the two low-inference observation instruments twice in each class due to time constraints. Only the high-inference observation instrument was used twice in each class. Thus, it was not possible to investigate whether the classroom-level factors of the dynamic model as measured by the low-inference observation instruments are generalisable at the level of the classroom, but only at the level of the teacher. However, previous studies conducted at primary schools, in which more than one observation in each class were conducted with the use of the two low-inference observation instruments (e.g., Creemers & Kyriakides, 2008), have shown that the classroom-level factors of the dynamic model as measured by these two instruments are generalisable at the level of the classroom.

Finally, one more limitation of this study that should be acknowledged is the fact that the student post-tests were administered to the students by the researcher, thus the process of test administration took a long period of time. In order to solve the problem of the possible influence on the results of the differences on test administration dates between the classes, a variable related to the test administration was added to the model measuring the impact of the classroom-level factors on student achievement. This variable was found to be associated with student achievement at the end of the school year. By adding this variable to the model, the problem of long period of test administration is mitigated to some extent, but it should be acknowledged that the problem still remains. However, the fact that the student tests were administered to the students by the same person increases the reliability of the measurements. This chapter has described the research design and methods used in the present study. The next chapter provides the analysis of the research data.

CHAPTER 4

RESEARCH RESULTS

This Chapter is divided into two sections. The first section provides some descriptive information about the classroom-level factors and their dimensions as measured by each instrument used in the study (i.e., the high and low-inference observation instruments and student questionnaire). The results of the generalisability study, which was conducted in order to investigate the consistency of teacher behaviour across different classrooms, are also presented. These are related to the first two research questions of the study. The second section, which is related to the last two questions of the study, presents the results of the multilevel analyses conducted to identify which level (the teacher or the classroom level) explains more variance in student achievement in mathematics of secondary school students. In addition, it presents the results of the multilevel analyses conducted to investigate the short-term effect of the classroom-level factors on student achievement of secondary school students in Cyprus.

Section A: Results Concerning the Consistency of Teacher Behaviour

This section is divided into four smaller parts according to the type of instrument used in this study to collect data on teacher in-class behaviour. Specifically, the first part presents the results of the two low-inference observation instruments. The second part presents the results of the high-inference observation instrument and the third part presents the results of the student questionnaire. Finally, the last part provides a summary of results of all the instruments used in the study regarding the consistency of teacher behaviour across different classrooms.

Low-inference observation instruments.

As mentioned in the previous chapter, for the two low-inference observation instruments one-way ANOVA was conducted to investigate whether the interval data emerged from each of these two instruments are generalisable at the level of the teacher. It was not possible to investigate whether the data that emerged from the low-inference observation instruments are generalisable at the level of the classroom, as each of the two low-inference observation instruments was used only once in each class of the participating teachers. The two low-inference observation instruments are concerned with all the classroom-level factors of the dynamic model except the assessment.

The table that follows (Table 4.1) presents some descriptive information (i.e., mean and standard deviation) of the classroom-level factors and their dimensions as measured by the two low-inference observation instruments and the results of the generalisability analysis. Tables 4.2 and 4.3 provide more information in regard to the dimension of quality of the factors of modelling and questioning techniques, as the quality dimension was measured in more than one way, whereas this dimension was measured in only one way in the other classroom-level factors. The reason for providing some descriptive information about the classroom-level factors and their dimensions in this section is not to compare the performance of the teacher sample in regard to these factors but to identify the variation in the functioning of each factor in order to facilitate the interpretation of the results of the generalisability analysis. Given that the classroom-level factors were not measured by the two low-inference observation instruments in the same way and in the same scale, it is pointless to compare their means.

Table 4.1

Means, standard deviations and the results of the generalisability analysis of factors measured by the two low-inference observation instruments at the teacher level

	Dimensions									
Classroom-level Factors	*Frequency*		*Focus*		*Stage*		*Quality*		*Differentiation*	
	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD
Orientation	0.74^{+}	0.86	0.58	0.68	0.63	0.75	0.65	0.77	0.00	0.00
Structuring	1.89^{++}	0.86	1.98^{++}	0.83	1.37^{++}	0.70	1.85	0.38	0.01	0.07
Application	1.23^{++}	1.24	0.79^{++}	0.73	0.91^{++}	0.83	0.75^{++}	0.64	0.91^{++}	1.24
Modelling	1.19^{++}	1.37	0.74^{++}	0.88	0.74^{++}	0.84	see Table 4.2		0.18^{++}	0.50
Questioning Techniques	24.1^{++}	11.01**	1.91^{++}	0.66	2.96	0.19	see Table 4.3		2.14^{++}	1.61
CLE: ***Dealing with misbehaviour***	1.01	0.97	*N/A*		*N/A*		0.69	0.63	0.09	0.16
CLE: ***Teacher -student interaction***	8.57^{++}	2.31	*N/A*		*N/A*		2.01^{++}	1.18	0.76^{++}	1.38
CLE: ***Student-student interaction***	0.22^{++}	0.61	*N/A*		*N/A*		0.02^{++}	0.12	0.17^{++}	0.53
Management of time*	6.08^{++}	2.34								

+ Statistically Significant at .10 level.

++ Statistically Significant at .05 level.

*Management of time was measured only in terms of the frequency dimension.

**The frequency dimension of questioning techniques was also measured by taking into account the length of pause following questions. This aspect of frequency was also found to be generalisable at the level of the teacher (mean=0.44^{++}, SD=0.70).

N/A= Not applicable

Table 4.2

Means, standard deviations and the results of the generalisability analysis at the teacher level of the quality dimension of the factor of modelling

Classroom -level Factor	Dimensions	Mean	SD
Modelling	**Quality:** *teacher's role (i.e., if the strategy is given by the teacher, students are guided to discover a strategy or students are directed to discover a strategy to solve a problem)*	0.94^{++}	1.00
	Quality: *appropriateness of the model-behaviour strategy (i.e., successful or unsuccessful)*	0.99^{++}	0.99
	Quality: *stage of the lesson during which the model behaviour is observed (i.e., before or after a problematic situation)*	0.91^{++}	0.94

+ Statistically Significant at .10 level.

++ Statistically Significant at .05 level.

Table 4.3

Means, standard deviations and the results of the generalisability analysis at the teacher level of the quality dimension of the factor of questioning techniques

Classroom-level Factor	Dimensions	Mean	SD
	Quality: *type (i.e., product or process questions)*	1.18^{++}	0.15
	Quality: *reaction if no answer from pupils*	3.84^{++}	0.23
	Quality: *feedback-reaction to students' answers (e.g., she/he makes negative comments to incorrect or partially-correct answers)*	3.04^{+}	0.32
	Quality: *feedback-reaction about the answer (e.g., she/he ignores the answer or invites students to comment on the answer)*	2.04	0.07

+ Statistically Significant at .10 level.

++ Statistically Significant at .05 level.

From the three tables above, we can make the following observations. First, the observation data seem to show that the great majority of the teacher sample did not differentiate their teaching in terms of structuring (mean= 0.01, SD= 0.07). Similar results emerged for dealing with misbehaviour, which is an element of the factor concerned with the classroom as a learning environment (mean=0.09, SD=0.19). In addition, no teacher was found to differentiate his/her teaching in terms of orientation. Accordingly, the question of consistency of teacher behaviour across different classrooms regarding the differentiation dimension of the factors of orientation, structuring and dealing with misbehaviour, as measured by the two low-inference observation instruments no longer arises for the teacher sample of this study. In regard to the other classroom-level factors (i.e., application, modelling, questioning techniques and teacher-student and student-student interaction) the teacher sample was found to differentiate their teaching to a greater extent than the factors of orientation, structuring and dealing with misbehavior. For these factors the results of the one-way ANOVA showed that the data emerged from the two low inference observation instruments regarding the differentiation dimension can be generalised at the teacher level, as the between-group variance was higher than the within-group variance ($p<0.05$).

Second, some dimensions of two factors (i.e., the quality dimension of structuring, the stage dimension of questioning techniques and the aspect of feedback-reaction about the answer of the quality dimension of questioning techniques) may not be found generalisable at the teacher level due to the small observed variance that they had. This implies that we did not have enough statistical power to detect significant differences and examine the generalisability of these dimensions at the level of the teacher. Thus, the one-way ANOVA was not in a position to provide answers on whether these data were generalisable or not at the teacher level. Nevertheless, all the other dimensions of the factor of questioning techniques and the frequency, focus and stage dimensions of the factor of structuring were found to be generalisable at the level of the teacher.

Third, it is apparent from the tables above that in two factors (i.e., orientation and dealing with misbehaviour) teacher behaviour was not found to be consistent from classroom to classroom. Specifically, the qualitative characteristics (i.e., focus, stage, quality) of the factor of orientation and the frequency and quality dimensions of the factor of measuring teacher ability to deal with misbehaviour, were not found to be generalisable at the level of the teacher as measured by the two low-inference observation instruments. It is important to note, however, that the frequency dimension of orientation was found to be generalisable at the level of the teacher.

Finally, as it can be seen from the three tables (above), teacher behaviour was found to be consistent from classroom to classroom in regard to all the dimensions of the factors of application, management of time, modelling and some elements of the factor classroom as a learning environment (i.e., teacher-student and student-student interaction) that could be measured.

High-inference observation instrument.

For the high-inference observation instrument, which was used twice in each class, multilevel modelling techniques were used to examine whether the interval data emerged from this instrument are generalisable at the level of the teacher and/or classroom (see Chapter 3). This observation instrument covers all the classroom-level factors of the dynamic model except the assessment, and observers were expected to rate the teacher's behaviour on a 5-point Likert scale. It is important to note that when a factor was found to be generalisable at the level of the teacher, it was also generalisable at the level of the classroom. However, when a factor was found to be generalisable only at the level of the classroom, this means that teacher behaviour regarding this factor was not consistent from classroom to classroom. Table 4.4 presents some descriptive information of the factor scores emerged from the high-inference observation instrument. The results of the generalisability analysis at the level of the teacher are also presented.

Table 4.4

Means, standard deviations and the results of the generalisability analysis of the classroom-factors as measured by the high-inference observation instrument at the level of the teacher

Classroom-level Factors	Mean	SD
Orientation	3.60	0.41
Structuring	3.44^{+}	0.82
Application	3.25^{+}	0.89
Modelling	1.48^{++}	1.63
Questioning Techniques	4.97^{++}	0.13
Management of time	4.41^{++}	0.50
CLE: *Dealing with misbehaviour*	4.23^{++}	0.65
CLE: *Teacher-student interaction*	4.54^{++}	0.52
CLE: *Student-student interaction*	1.30^{++}	0.74

+ Statistically Significant at .10 level.

++ Statistically Significant at .05 level.

From the table above, we can see that all the classroom-level factors measured by the high-inference observation instrument (except the factor of orientation) were found to be generalisable at the level of the teacher. Orientation was not found to be generalisable neither at the level of the teacher nor the classroom. This may be attributed to the fact that orientation had a small standard deviation (SD=0.41). Consequently, we did not have enough statistical power to detect significant differences and examine the generalisability of this factor at the level of the teacher. What is interesting about the data in the table above is that the observed variance of the functioning of the majority of the classroom-level factors was very small. When this study was designed and the size of the teacher sample was selected, it was expected that the standard deviation of each factor measured by the high inference and the student questionnaire would be around one, based on previous studies conducted in primary schools in Cyprus. This is a finding that is discussed in the last chapter.

Student questionnaire.

The student questionnaire measured all the classroom-level factors and their dimensions of the dynamic model and a Likert scale was used to collect data. As mentioned in the previous chapter, a generalisability study on the use of students' ratings was conducted before the creation of factor scores. The ANOVA analysis showed that the data collected from most items of the questionnaire can be generalised at the classroom level. However, some individual items concerned with specific dimensions of the classroom-level factors were not found to be generalisable at the level of the classroom and for this reason they had to be removed. Thus, it was not possible to generate factor scores for specific dimensions of the classroom-level factors (e.g., the frequency dimension of application). These cases are presented in Table 4.5 with the abbreviation N/A (not applicable). This finding is not in line with previous studies conducted in primary schools which have suggested that almost all the student questionnaire items could be used for measuring each classroom-level factor and its dimensions (see Kyriakides & Creemers, 2008; Kyriakides et al., 2009).

When factor scores were generated, the same approach as for the high inference was followed for all the classroom-level factors and their dimensions measured by the student questionnaire in order to examine whether the data emerged from this instrument are generalisable at the level of the teacher and/or classroom. Table 4.5 presents some descriptive information of the factor scores emerged from the student questionnaire. The results of the generalisability analysis at the level of the teacher are also presented.

Table 4.5

Means, standard deviations and the results of the generalisability analysis of the classroom-factors as measured by the student questionnaire at the level of the teacher

	Dimensions									
Classroom-level Factors	*Frequency*		*Focus*		*Stage*		*Quality*		*Differentiation*	
	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD
Orientation	3.57	1.00	*N/A*		*N/A*		4.10	1.00	*N/A*	
Structuring	3.61$^+$	1.23	4.01$^+$	1.07	3.41$^+$	0.91	3.68	0.67	*N/A*	
Application	*N/A*		3.34$^+$	0.88	3.62$^+$	1.26	3.26$^+$	0.65	3.33$^+$	0.79
Modelling	3.60$^+$	1.02	*N/A*		*N/A*		3.51$^+$	0.97	*N/A*	
Questioning Techniques	3.48$^+$	1.17	*N/A*		*N/A*		3.47$^+$	0.70	3.22	0.65
CLE: *Dealing with misbehaviour*	2.46	1.26	3.68	1.32	*N/A*		3.32	0.90	*N/A*	
CLE: *Teacher -student interaction*	3.95$^+$	1.08	*N/A*		*N/A*		3.57$^+$	0.96	*N/A*	
CLE: *Student-student interaction*	2.88	0.61	*N/A*		*N/A*		2.53$^+$	1.41	*N/A*	
Assessment	3.44$^+$	0.86	*N/A*		*N/A*		3.72$^+$	0.48	*N/A*	
Management of time*	3.37$^+$	0.86								

+ Statistically Significant at .05 level.

*Management of time was measured only in terms of the frequency dimension.

N/A= Not applicable

The following observations arise from Table 4.5. First, for four factors (i.e., application, modelling, management of time and assessment) and for the teacher-student interaction, which is an element of the factor concerned with the classroom as a learning environment, all the dimensions were found to be generalisable at the level of the teacher.

Second, for two factors (i.e., structuring and questioning techniques) and for the student-student interaction, which is an element of the factor concerned with the classroom as a learning environment, teacher behaviour was not found to be consistent from classroom to classroom for all their dimensions. Specifically, the quality dimension of structuring, the differentiation dimension of questioning techniques and the frequency dimension of the student-student interaction were found to be generalisable only at the level of the classroom. However, these three cases may not be found to be generalisable at the level of the teacher due to their extremely small observed variance in the functioning of the abovementioned factors (see Table 4.5) and thus, the lack of enough statistical power.

Finally, for two factors (i.e., orientation and dealing with misbehaviour) all the dimensions were found to be generalisable at the level of the classroom. Therefore, teacher behaviour was not found to be consistent from classroom to classroom in regard to the factors of orientation and dealing with misbehaviour as measured by the student questionnaire.

Summary of results.

In order to facilitate the understanding of the results of the generalisability analyses, Table 4.6 was constructed. In this table, one can see the summarized results of all the instruments used in this study in regard to the factors and their dimensions that were found to be generalisable at the level of the teacher.

It is apparent from Table 4.6 that for five factors and the majority of their dimensions (i.e., application, modelling, management of time, structuring except the dimensions of quality and differentiation and questioning techniques except the dimensions of stage and differentiation) as well as some elements of the factor classroom as a learning environment (i.e., the teacher-student interaction and student-student interaction except the dimension of frequency) teacher behaviour was found to be consistent from classroom to classroom irrespective of the instrument used. For two other elements of the factor classroom as a learning environment (i.e., dealing with misbehaviour and the frequency dimension of student-student interaction), as well as for the frequency dimension of orientation and the differentiation dimension of questioning techniques, the findings were differentiated according to the instrument that was used. However, the factor of orientation as measured by the high-inference observation instrument was not found to be generalisable neither at the level of the teacher nor the classroom. As discussed before, the fact that some dimensions of some classroom-level factors were not found to be generalisable at the level of the teacher may be attributed to their extremely small observed variance in the functioning of these factors (see Tables 4.1-4.6). Finally, the factor of assessment was measured only from the student questionnaire and all its dimensions were found to be generalisable at the level of the teacher.

The observation data were those found to be generalisable at the teacher level for most of the factors. Specifically, only the frequency dimension of orientation was found to be generalisable at the teacher level and only in the data gathered with the low-inference observation instrument. Dealing with misbehaviour, which is an element of the factor concerned with the classroom as a learning environment, was found to be generalisable at the teacher level only in the data gathered with the high-inference observation instrument.

Table 4.6

Summarized results of all the instruments used in the study in regard to the classroom-level factors of the dynamic model and their dimensions (i.e., frequency (Freq), focus (Foc), stage (Stag), quality (Qual), and differentiation (Diff) that were found to be generalisable at the level of the teacher)

Classroom-level factors	**Low- inference observation instruments**					**High-inference observation instrument**	**Student questionnaire**				
	Freq	**Foc**	**Stag**	**Qual**	**Diff**		**Freq**	**Foc**	**Stag**	**Qual**	**Diff**
Orientation	+				*N/A*	S→P		*N/A*	*N/A*		*N/A*
Structuring	++	++	++	S→P	S→P	+	++	++	++	S→P	*N/A*
Application	++	++	++	++	++	+	*N/A*	++	++	++	++
Modelling	++	++	++	++	++	++	++	*N/A*	*N/A*	++	*N/A*
Questioning techniques	++	++	S→P	++ *	++	++	++	*N/A*	*N/A*	++	S→P
Management of time	++	*N/A*				++	++	*N/A*			
CLE: ***Dealing with misbehaviour***		*N/A*			S→P	++			*N/A*		*N/A*
CLE: ***Teacher-student interaction***	++	*N/A*		++	++	++	++	*N/A*	*N/A*	++	*N/A*
CLE: ***Student-student interaction***	++	*N/A*		++	++	++	S→P	*N/A*	*N/A*	++	*N/A*
Assessment	*N/A*					*N/A*	++	*N/A*	*N/A*	++	*N/A*

+ Statistically Significant at .10 level.

++Statistically Significant at .05 level.

N/A= Not applicable because it was not measured.

S→P= Observed variance of the functioning of the factor was very small. Consequently, we did not have enough statistical power to detect significant differences and examine the generalisability of the factor at the level of the teacher (see Tables 4.1-4.5).

*The aspect of feedback-reaction about the answer was not found to be generalisable at the teacher level due to the extremely small variance that it had and thus the lack of statistical power. The result regarding the aspect of feedback-reaction to students' answers was statistically significant at .10 level.

The next section presents the results of the multilevel analyses that were conducted to identify the impact of the classroom-level factors and their dimensions of the dynamic model on student achievement in mathematics of secondary school students. The score for each classroom-level factor used in the multilevel analyses in the case of the low and high-inference observation instruments was the mean score of all the observations conducted in all the classes taught by the same teacher or the mean score of all the observations conducted in each class separately based on the results of the generalisability analysis. The factor of orientation as measured by the high-inference observation instrument was not included in the multilevel analyses since it was not found to be generalisable neither at the level of the teacher nor the classroom. As discussed before, it was not possible to investigate whether the classroom-level factors of the dynamic model as measured by the two low-inference observation instruments are generalisable at the level of the classroom, as each of the two low-inference observation instruments was used only once in each class. However, previous studies conducted at primary schools (e.g., Creemers & Kyriakides, 2008) have shown that the classroom-level factors of the dynamic model as measured by the low-inference observation instruments are generalisable at the level of the classroom. Therefore, apart from the mean scores of all the observations conducted in all the classes taught by the same teacher (in the case that a dimension was found to be generalisable at the level of the teacher), the factor scores that emerged from the low-inference observation instruments for each class separately were also used in the multilevel analyses. In the case of the student questionnaire, the score for each classroom-level factor was the mean score of all the students of all classes taught by the same teacher or the mean score of all the students of each class separately according to the results of the generalisability analysis.

Section B: The Impact of the Classroom-Level Factors on Student Achievement in Mathematics: Results of Multilevel Analysis

Multilevel modelling techniques were used to identify the extent to which each classroom-level factor of the dynamic model is associated with student achievement as measured by each instrument (i.e., the high and low-inference observation instruments and student questionnaire) separately. The results of these analyses are presented in Tables 4.7-4.9. These tables present the models that were found to best fit the data. The factors that were not found to have a statistically significant effect on student achievement are not included in these tables.

The first step was to run the empty model to determine the variance at each level. As described in the previous chapter, three different empty models were run (see Table 3.4 in Chapter 3). The two-level model (students within teachers) was found to fit better than any other model. This revealed that the teacher, rather than the classroom level, is more important for explaining variation in student achievement in mathematics of secondary school students. The variance was 20.23% at the teacher level and 79.77% at the student level and was statistically significant in each level (see Tables 4.7-4.9).

Table 4.7

Parameter estimates and (standard errors) for the analysis of student achievement-Classroom-level factors as measured by the two low-inference observation instruments

Factors	**Model 0**	**Model 1**	**Model 2**	**Model 3a**	**Model 3b**	**Model 3c**	**Model 3d**	**Model 3e**	**Model 3f**	**Model 3g**
Fixed Part (Intercept)	-1.09 (0.11)	-1.26 (0.09)	-1.27 (0.09)	-1.26 (0.08)	-1.27 (0.08)	-1.27 (0.08)	-1.27 (0.08)	-1.27 (0.08)	-1.27 (0.09)	-1.26 (0.08)
Student Level										
Prior knowledge in mathematics		0.78 (0.03)	0.78 (0.03)	0.78 (0.03)	0.78 (0.03)	0.78 (0.03)	0.78 (0.03)	0.78 (0.03)	0.78 (0.03)	0.78 (0.03)
Gender		0.54 (0.08)	0.53 (0.08)	0.53 (0.08)	0.53 (0.08)	0.53 (0.08)	0.53 (0.08)	0.53 (0.08)	0.53 (0.08)	0.53 (0.08)
SEN		-0.39 (0.18)	-0.39 (0.18)	-0.39 (0.18)	-0.39 (0.18)	-0.40 (0.18)	-0.40 (0.18)	-0.39 (0.18)	-0.39 (0.18)	-0.38 (0.18)
Ethnicity-Language that students speak at home										
Greek and another language		-0.43 (0.10)	-0.41(0.10)	-0.41(0.10)	-0.41(0.10)	-0.41(0.10)	-0.41(0.10)	-0.41(0.10)	-0.42(0.10)	-0.42(0.10)
Other language		-0.48 (0.17)	-0.46(0.17)	-0.47(0.17)	-0.46(0.17)	-0.46(0.17)	-0.46(0.17)	-0.46(0.17)	-0.45(0.17)	-0.45(0.17)
Teacher Level										
Post-test administration date			0.18 (0.09)	0.16 (0.08)	0.16 (0.08)	0.14 (0.08)*	0.15 (0.08)*	0.18 (0.08)	0.18(0.08)	0.16 (0.08)
Classroom-level factors										
Application (Frequency)				0.14 (0.06)						
Application (Stage)					0.20 (0.09)					
Application (Focus)						0.25 (0.11)				
Application(Quality)							0.27 (0.12)			
Application (Differentiation)								0.10 (0.06)*		
Modelling (Frequency)									0.074 (0.044)*	
Modelling (Stage)										0.24 (0.09)
Variance components										
Teacher	20.23%	11.12%	9.12%	7.05%	6.88%	6.85%	6.73%	7.15%	7.39%	6.85%
Student	79.77%	48.31%	45.21%	45.12%	44.78%	44.82%	44.29%	45.10%	45.02%	44.92%
Explained		40.57%	45.67%	47.83%	48.34%	48.33%	48.98%	47.75%	47.59%	48.23%
Significance test										
X^2	3174.52	2519.63	2515.77	2511.44	2511.02	2510.85	2510.94	2513.71	2514.10	2509.61
Reduction		654.89	3.86	4.33	4.75	4.92	4.83	2.06	1.67	6.16
Degrees of freedom		5	1	1	1	1	1	1	1	1
p-value		0.001	0.001	0.001	0.001	0.001	0.001	0.05	0.08	0.001

Note: For each of the Models 3a up to 3t the reduction is estimated in relation to the deviance of Model 2.

* Statistically Significant at .10 level.

Table 4.7

Parameter estimates and (standard errors) for the analysis of student achievement-Classroom-level factors as measured by the two low-inference observation instruments (continued).

Factors	Model 0	Model 1	Model 2	Model 3h	Model 3i	Model 3j	Model 3k	Model 3l	Model 3m	Model 3n
Fixed Part (Intercept)	-1.09(0.11)	-1.26 (0.09)	-1.27(0.09)	-1.26 (0.08)	-1.27 (0.09)	-1.26 (0.08)	-1.26(0.08)	-1.27 (0.09)	-1.26 (0.08)	-1.26 (0.08)
Student Level										
Prior knowledge in mathematics		0.78 (0.03)	0.78 (0.03)	0.78 (0.03)	0.78 (0.03)	0.78 (0.03)	0.78 (0.03)	0.78 (0.03)	0.78 (0.03)	0.78 (0.03)
Gender		0.54 (0.08)	0.53 (0.08)	0.53 (0.08)	0.53 (0.08)	0.53 (0.08)	0.53 (0.08)	0.53 (0.08)	0.53 (0.08)	0.53 (0.08)
SEN		-0.39 (0.18)	-0.39(0.18)	-0.39 (0.18)	-0.39 (0.18)	-0.39 (0.18)	-0.39(0.18)	-0.39 (0.18)	-0.37 (0.18)	-0.37 (0.18)
Ethnicity-Language that students speak at home										
Greek and another language		-0.43(0.10)	-0.41(0.10)	-0.42(0.10)	-0.42(0.10)	-0.42(0.10)	-0.42(0.10)	-0.42(0.10)	-0.41(0.10)	-0.42(0.10)
Other language		-0.48 (0.17)	-0.46(0.17)	-0.45(0.17)	-0.45(0.17)	-0.45(0.17)	-0.45(0.17)	-0.45(0.17)	-0.46(0.17)	-0.46(0.17)
Teacher Level										
Post-test administration date			0.18 (0.09)	0.16 (0.08)	0.19 (0.08)	0.18 (0.08)	0.18 (0.08)	0.18 (0.08)	0.19 (0.08)	0.19 (0.08)
Classroom-level factors										
Modelling (Focus)				0.20 (0.09)						
Modelling (Quality: teacher's role)					0.12(0.07)*					
Modelling (Quality: appropriateness of the model)						0.16 (0.08)				
Modelling (Quality: stage of the lesson)							0.16(0.09)*			
Modelling (Differentiation)								0.15(0.09)*		
Questioning Techniques (Focus)									-0.22(0.13)*	
Questioning Techniques (Focus-quadratic)										-0.06 (0.03)
Variance components										
Teacher	20.23%	11.12%	9.12%	7.01%	8.12%	7.09%	8.09%	8.04%	8.19%	7.92%
Student	79.77%	48.31%	45.21%	44.13%	44.41%	44.76%	44.91%	44.71%	44.74%	44.81%
Explained		40.57%	45.67%	48.86%	47.47%	48.15%	47.00%	47.25%	47.07%	47.27%
Significance test										
X^2	3174.52	2519.63	2515.77	2511.43	2513.82	2511.74	2512.61	2513.89	2512.82	2512.49
Reduction		654.89	3.86	4.34	1.95	4.03	3.16	1.88	2.95	3.28
Degrees of freedom		5	1	1	1	1	1	1	1	1
p-value		0.001	0.001	0.001	0.05	0.001	0.001	0.06	0.001	0.001

* Statistically Significant at .10 level.

Table 4.7

Parameter estimates and (standard errors) for the analysis of student achievement-Classroom-level factors as measured by the two low-inference observation instruments (continued).

Factors	**Model 0**	**Model 1**	**Model 2**	**Model 3o**	**Model 3p**	**Model 3q**	**Model 3r**	**Model 3s**	**Model 3t**
Fixed Part (Intercept)	-1.09 (0.11)	-1.26 (0.09)	-1.27 (0.09)	-1.26 (0.08)	-1.26 (0.08)	-1.26 (0.08)	-1.27 (0.09)	-1.27 (0.09)	-1.26 (0.08)
Student Level									
Prior knowledge in mathematics		0.78 (0.03)	0.78 (0.03)	0.78 (0.03)	0.78 (0.03)	0.78 (0.03)	0.78 (0.03)	0.78 (0.03)	0.78 (0.03)
Gender		0.54 (0.08)	0.53 (0.08)	0.53 (0.08)	0.53 (0.08)	0.54 (0.08)	0.52 (0.08)	0.52 (0.08)	0.53 (0.08)
SEN		-0.39 (0.18)	-0.39 (0.18)	-0.37 (0.18)	-0.37 (0.18)	-0.38 (0.18)	-0.39 (0.18)	-0.39 (0.18)	-0.38 (0.18)
Ethnicity-Language that students speak at home									
Greek and another language		-0.43(0.10)	-0.41(0.10)	-0.41(0.10)	-0.41(0.10)	-0.41(0.10)	-0.40(0.10)	-0.40(0.10)	-0.42(0.10)
Other language		-0.48(0.17)	-0.46 (0.17)	-0.45(0.17)	-0.45(0.17)	-0.46(0.17)	-0.44(0.17)	-0.44(0.17)	-0.46(0.17)
Teacher Level									
Post-test administration date			0.18 (0.09)	0.16 (0.08)	0.18 (0.08)	0.18 (0.08)	0.16 (0.08)	0.16 (0.08)	0.14 (0.09)*
Classroom-level factors									
Questioning Techniques (Quality: waiting time)				-0.29 (0.12)					
Questioning Techniques (Quality: waiting time-quadratic)					-0.16 (0.07)				
Questioning Techniques (Quality: feedback-reaction to students' answers)						0.63 (0.27)			
Questioning Techniques (Quality: feedback-reaction about the answer)							-1.79(0.68)		
Questioning Techniques (Quality: feedback-reaction about the answer-quadratic)								-0.43 (0.16)	
Management of time									-0.06(0.04)*
Variance components									
Teacher	20.23%	11.12%	9.12%	7.29%	7.10%	7.33%	6.94%	6.77%	7.55%
Student	79.77%	48.31%	45.21%	44.18%	44.09%	44.38%	44.03%	43.93%	44.34%
Explained		40.57%	45.67%	48.53%	48.81%	48.29%	49.03%	49.30%	48.11%
Significance test									
X^2	3174.52	2519.63	2515.77	2510.67	2510.24	2510.75	2508.89	2508.85	2513.80
Reduction		654.89	3.86	5.1	5.53	5.02	6.88	6.92	1.97
Degrees of freedom		5	1	1	1	1	1	1	1
p-value		0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.05

* Statistically Significant at .10 level.

Table 4.8

Parameter estimates and (standard errors) for the analysis of student achievement-Classroom-level factors as measured by the high-inference observation instrument

Factors	**Model 0**	**Model 1**	**Model 2**	**Model 3a**	**Model 3b**	**Model 3c**	**Model 3d**
Fixed Part (Intercept)	-1.09 (0.11)	-1.26 (0.09)	-1.27 (0.09)	-1.26 (0.08)	-1.26 (0.08)	-1.27 (0.09)	-1.26 (0.08)
Student Level							
Prior knowledge in mathematics		0.78 (0.03)	0.78 (0.03)	0.78 (0.03)	0.78 (0.03)	0.78 (0.03)	0.78 (0.03)
Gender		0.54 (0.08)	0.53 (0.08)	0.53 (0.08)	0.53 (0.08)	0.53 (0.08)	0.53 (0.08)
SEN		-0.39(0.18)	-0.39 (0.18)	-0.39 (0.18)	-0.39 (0.18)	-0.36 (0.18)	-0.39 (0.18)
Ethnicity-Language that students speak at home							
Greek and another language		-0.43(0.10)	-0.41(0.10)	-0.42 (0.10)	-0.42 (0.10)	-0.42 (0.10)	-0.41 (0.10)
Other language		-0.48 (0.17)	-0.46(0.17)	-0.45 (0.17)	-0.46 (0.17)	-0.43 (0.17)	-0.47 (0.17)
Teacher Level							
Post-test administration date			0.18 (0.09)	0.17 (0.08)	0.22 (0.08)	0.21 (0.09)	0.21 (0.08)
Classroom-level factors							
Modelling				0.12 (0.06)			
Management of time					0.38 (0.22)*		
CLE: Teacher-student interaction						0.22 (0.14)*	
CLE: Dealing with misbehaviour							0.35 (0.16)
Variance components							
Teacher	20.23%	11.12%	9.12%	6.32%	5.72%	6.12%	6.01%
Student	79.77%	48.31%	45.21%	43.27%	43.37%	43.65%	43.18%
Explained		40.57%	45.67%	50.41%	50.91%	50.23%	50.81%
Significance test							
X^2	3174.52	2519.63	2515.77	2512.42	2512.89	2513.38	2511.19
Reduction		654.89	3.86	3.35	2.88	2.39	4.58
Degrees of freedom		5	1	1	1	1	1
p-value		0.001	0.001	0.001	0.001	0.001	0.001

Note: For each of the Models 3a up to 3d the reduction is estimated in relation to the deviance of Model 2.

* Statistically Significant at .10 level.

Table 4.9

Parameter estimates and (standard errors) for the analysis of student achievement-Classroom-level factors as measured by the student questionnaire

Factors	**Model 0**	**Model 1**	**Model 2**	**Model 3a**	**Model 3b**	**Model 3c**	**Model 3d**	**Model 3e**	**Model 3f**
Fixed Part (Intercept)	-1.09 (0.11)	-1.26 (0.09)	-1.27 (0.09)	-1.27 (0.09)	-1.26(0.09)	-1.28 (0.08)	-1.27 (0.09)	-1.28 (0.09)	-1.27 (0.09)
Student Level									
Prior knowledge in mathematics		0.78 (0.03)	0.78 (0.03)	0.78 (0.03)	0.78 (0.03)	0.78 (0.03)	0.78 (0.03)	0.78 (0.03)	0.78 (0.03)
Gender		0.54 (0.08)	0.53 (0.08)	0.53 (0.08)	0.53 (0.08)	0.53 (0.08)	0.53 (0.08)	0.53 (0.08)	0.53 (0.08)
SEN		-0.39 (0.18)	-0.39 (0.18)	-0.39 (0.18)	-0.39 (0.18)	-0.40 (0.18)	-0.37 (0.18)	-0.35 (0.18)*	-0.39 (0.18)
Ethnicity-Language that students speak at home									
Greek and another language		-0.43(0.10)	-0.41(0.10)	-0.41 (0.10)	-0.41 (0.10)	-0.41 (0.10)	-0.42 (0.10)	-0.40 (0.10)	-0.40 (0.10)
Other language		-0.48(0.17)	-0.46(0.17)	-0.46 (0.17)	-0.46 (0.17)	-0.46 (0.17)	-0.46 (0.17)	-0.42 (0.17)	-0.43 (0.17)
Teacher Level									
Post-test administration date			0.18 (0.09)	0.17 (0.08)	0.18 (0.08)	0.15 (0.08)*	0.15 (0.09)*	0.17 (0.09)	0.16 (0.09)*
Classroom-level factors									
Orientation (Quality)				0.30 (0.17)*					
Application (Quality)					0.62 (0.37)*				
Questioning Techniques (Frequency)						0.38 (0.20)*			
Management of time							0.27 (0.15)*		
CLE: Student-student interaction (Frequency)								0.68 (0.27)	
Assessment(Quality)									0.64 (0.28)
Variance components									
Teacher	20.23%	11.12%	9.12%	7.43%	8.29%	8.17%	8.12%	7.21%	7.10%
Student	79.77%	48.31%	45.21%	44.62%	44.69%	44.54%	44.29%	44.01%	44.57%
Explained		40.57%	45.67%	47.95%	47.02%	47.29%	47.59%	48.78%	48.33%
Significance test									
X^2	3174.52	2519.63	2515.77	2512.73	2514.08	2513.35	2513.02	2509.38	2510.50
Reduction		654.89	3.86	3.04	1.69	2.42	2.75	6.39	5.27
Degrees of freedom		5	1	1	1	1	1	1	1
p-value		0.001	0.001	0.001	0.08	0.001	0.001	0.001	0.001

Note: For each of the Models 3a up to 3l the reduction is estimated in relation to the deviance of Model 2.

* Statistically Significant at .10 level.

Table 4.9

Parameter estimates and (standard errors) for the analysis of student achievement-Classroom-level factors as measured by the student questionnaire (continued).

Factors	**Model 0**	**Model 1**	**Model 2**	**Model 3g**	**Model 3h**	**Model 3i**	**Model 3j**	**Model 3k**	**Model 3l**
Fixed Part (Intercept)	-1.09 (0.11)	-1.26 (0.09)	-1.27 (0.09)	-1.26 (0.09)	-1.26 (0.09)	-1.27 (0.08)	-1.27 (0.09)	-1.27 (0.09)	-1.27 (0.08)
Student Level									
Prior knowledge in mathematics		0.78 (0.03)	0.78 (0.03)	0.78 (0.03)	0.78 (0.03)	0.78 (0.03)	0.78 (0.03)	0.78 (0.03)	0.78 (0.03)
Gender		0.54 (0.08)	0.53 (0.08)	0.53 (0.08)	0.53 (0.08)	0.53 (0.08)	0.53 (0.08)	0.53 (0.08)	0.53 (0.08)
SEN		-0.39 (0.18)	-0.39 (0.18)	-0.39 (0.18)	-0.38 (0.18)	-0.38 (0.18)	-0.38 (0.18)	-0.38 (0.18)	-0.38 (0.18)
Ethnicity-Language that students speak at home									
Greek and another language		-0.43(0.10)	-0.41(0.10)	-0.41 (0.10)	-0.41 (0.10)	-0.41 (0.10)	-0.41 (0.10)	-0.41 (0.10)	-0.41 (0.10)
Other language		-0.48(0.17)	-0.46(0.17)	-0.46 (0.17)	-0.46 (0.17)	-0.46 (0.17)	-0.46 (0.17)	-0.46 (0.17)	-0.46 (0.17)
Teacher Level									
Post-test administration date			0.18 (0.09)	0.16 (0.08)	0.16 (0.08)	0.16 (0.08)	0.18 (0.08)	0.18 (0.08)	0.22 (0.08)
Classroom-level factors									
Structuring (Frequency)				-0.28 (0.16)*		4.20 (3.18)*			
Structuring (Frequency-quadratic)					-0.04 (0.03)*	-0.62 (0.34)*			
Application (Focus)							-0.44 (0.24)*		12.21 (5.89)
Application (Focus-quadratic)								-0.07 (0.04)*	-1.91 (0.89)
Variance components									
Teacher	20.23%	11.12%	9.12%	8.24%	8.19%	8.05%	8.04%	8.01%	7.44%
Student	79.77%	48.31%	45.21%	44.71%	44.56%	44.13%	44.88%	44.18%	44.31%
Explained		40.57%	45.67%	47.05%	47.25%	47.82%	47.08%	47.81%	48.25%
Significance test									
X^2	3174.52	2519.63	2515.77	2513.92	2513.69	2512.02	2513.63	2513.35	2509.34
Reduction		654.89	3.86	1.85	2.08	3.75	2.14	2.42	6.43
Degrees of freedom		5	1	1	1	2	1	1	2
p-value		0.001	0.001	0.06	0.05	0.06	0.001	0.001	0.001

* Statistically Significant at .10 level.

In Model 1 (see Tables 4.7-4.9), the student variables (i.e., prior knowledge, SEN, gender and ethnicity: language that students speak at home) were added to the empty model. Model 1 explains 40.57% of the total variance of student achievement and most of the explained variance is at the student level. The likelihood statistic (χ^2) shows a significant change among the empty model and Model 1 (p <0.001), justifying the selection of Model 1. All the student variables have a statistically significant effect at level .05. As discussed in Chapter 3, the other variables of ethnicity (students', fathers' and mothers' country of birth) were found to be associated with student achievement when they were studied in isolation, but, because of multicollinearity, their effects disappeared when they were studied together with the other student variables and were removed from the analysis. It is important to note that girls were found to have better results compared to boys. This finding is contrary to previous effectiveness studies conducted in Cyprus in primary schools (e.g., Creemers & Kyriakides, 2008; Kyriakides et al., 2009), which have shown that boys have better results in mathematics compared to girls. Moreover, students' prior knowledge in mathematics has the strongest effect in predicting student achievement at the end of the school year. In addition, students with special educational needs have not as good results as students with no special educational needs. Finally, in terms of the variable of ethnicity: language that students speak at home, the group of students who spoke only Greek at home have the best results in mathematics.

In Model 2, a variable related to the test administration date was added to Model 1 (see Chapter 3 for more details regarding this variable). This variable was found to be associated with student achievement at the end of the school year. Model 2 explains 45.67% of the total variance of student achievement and the likelihood statistic (χ^2) shows a significant change between Model 1 and Model 2 (p <0.001).

At the next step of the analyses, for each instrument (i.e., the high, low-inference observation instruments and student questionnaire) different versions of Model 3 were established. In each version of Model 3, factor scores which refer to the classroom-level factors and their dimensions of the dynamic model were added one by one to Model 2. Thus, the fitting of each of these models (i.e., Models 3a-3t for the classroom-level factors measured by the low-inference observation instruments, as shown in Table 4.7, Models 3a-3d for the classroom-level factors measured by the high-inference observation instrument, as shown in Table 4.8 and Models 3a-3l for the classroom-level factors measured by the student questionnaire, as shown in Table 4.9) was tested against Model 2. The classroom-level factors and their dimensions which were not found to have a statistically significant effect on student achievement are not included in Tables 4.7 to 4.9. The likelihood statistic (X^2) shows a significant change between Model 2 and each version of Model 3 presented in Tables 4.7 to 4.9. This implies that the variables measuring the classroom-level factors and their dimensions (included in Tables 4.7 to 4.9) have significant effects on student achievement of secondary school students. Each version of Model 3 explains approximately more than 47% of the total variance.

Low-inference observation instruments.

The following observations arise from Models 3a-3t (as shown in Table 4.7), which refer to the impact of the classroom-level factors and their dimensions measured by the low-inference observation instruments on student achievement. First, for four out of seven classroom-level factors (i.e., application, modelling, questioning techniques and management of time) all their dimensions, apart from the factor of questioning techniques, were found to have a statistically significant effect on student outcomes. Three dimensions (i.e., frequency, stage, differentiation) of the factor of questioning techniques and one aspect of its quality dimension (i.e., type of question) were not found to be associated with student achievement. On the other hand, none of the dimensions of the factors of orientation, structuring and of the three elements of the factor classroom as a learning environment (i.e., teacher-student and student-student interaction and dealing with misbehaviour) were found to have a statistically significant effect on student outcomes. However, the fact that the quality dimension of structuring and the stage dimension of questioning techniques were not found to be associated with student achievement could be attributed to their small standard deviations and as a consequence their lack of enough statistical power to detect the effects (SD=0.38 on a scale from 1 to 2 and SD= 0.19 on a scale from 1 to 3 respectively). In addition, it was not expected to detect an effect for the differentiation dimension of structuring and dealing with misbehaviour, as well as the quality dimension of student-student interaction, since the observation data have shown that the majority of the participating teachers did not exhibit teaching skills concerned with these dimensions of these factors.

Second, some dimensions of the factor concerned with the questioning techniques (i.e., the focus dimension, the aspects of waiting time of the frequency dimension and the aspect of feedback-reaction about the answer of the quality dimension), as well as the frequency dimension of management of time, were found to have a negative effect on student achievement. The fact that the frequency dimension of management of time was found to have a negative effect is justified as it was measured by taking into account how much time was not used for teaching (i.e., the amount of time that students were off-task). However, the fact that the aforementioned dimensions of questioning techniques were found to have a negative effect on student achievement, was an unexpected result. Thus, it was decided to search whether curvilinear relations exist between the abovementioned dimensions of questioning techniques and student achievement. Therefore, separate multilevel analyses were conducted using the quadratic factor scores (i.e., the square values of these dimensions) instead of the non-quadratic scores (see Models 3m and 3n for the dimension of focus, 3o and 3p for the aspect of the waiting time of the quality dimension and 3r and 3s for the aspect of feedback reaction about the answer of the quality dimension in Table 4.7). In all of these cases, a curvilinear relation was identified, which is in line with the assumption of the dynamic model that the relation of some effectiveness factors with student achievement may not be linear (Creemers & Kyriakides, 2008). An alternative model was also examined for each of these dimensions of the questioning techniques where each non-quadratic factor score and its equivalent quadratic factor score were added together in Model 2. However, these models have worse fit to the data compared to the previous models where the non-quadratic factor score was not included in the model and thus, their results are not presented in Table 4.7.

High-inference observation instrument.

The following observations arise from Table 4.8, which refers to the effects of the classroom-level factors measured by the high-inference observation instrument on student achievement. Among the six classroom-level factors measured by the high-inference, two of them (i.e., modelling and management of time) and two out of the three elements of the factor classroom as a learning environment (i.e., dealing with misbehaviour and teacher-student interaction) were found to be associated with student outcomes. Based on the results of the generalisability analysis, the factor of orientation was not included in this multilevel analysis. The factors that were not found to have a statistically significant effect on student achievement are: a) structuring, b) application, c) questioning techniques and d) student-student interaction, which is an element of the factor classroom as a learning environment. However, not identifying statistically significant effects for the questioning techniques and the student-student interaction on student achievement could be attributed to their small standard deviation and thus their lack of statistical power to detect the effects (see Table 4.4).

Student questionnaire.

The following observations arise from Models 3a-3l (as shown in Table 4.9), which refer to the impact of the classroom-level factors and their dimensions measured by the student questionnaire on student achievement. First, for six classroom-level factors out of eight (i.e., orientation, structuring, application, questioning techniques, management of time and assessment) and one of the three elements of the factor classroom as a learning environment (i.e., student-student interaction), at least one of their dimensions was found to be associated with student achievement. Second, the fact that the quality dimension of structuring and questioning techniques and the differentiation dimension of questioning techniques and application were not found to be associated with student achievement, may be attributed to their extremely small standard deviations (see Table 4.5). Specifically, due to their small observed variance in the functioning of these dimensions of these factors, we did not have enough statistical power to detect their effects on student achievement. Third, for the modelling and the other two elements of the factor classroom as a learning environment (i.e., teacher-student interaction and dealing with misbehaviour) none of their dimensions was found to have a statistically significant effect on student outcomes.

Fourth, the frequency dimension of structuring and the focus dimension of application were found to have a negative effect on student achievement. This was a rather unexpected result. Thus, it was decided to search for non-linear relations of the frequency dimension of structuring and the focus dimension of application with student outcomes. Therefore, the same approach was followed, as described above in the case of the dimensions of questioning techniques (as measured by the low-inference observation instrument) that were found to have a negative effect (see Models 3g-3i and 3j-3l in Table 4.9). The fitting of Model 3i was tested against Models 3g and 3h and the fitting of Model 3l was tested against Models 3j and 3k. Based on the likelihood statistic and the reduction observed, one may conclude that Model 3i and Model 3l, which have an extra degree of freedom, have a better fit to the data than both of the previous models where the non-quadratic factor score was not included in the model. Thus, the frequency dimension of structuring and the focus dimension of application were found to be related in a non-linear way to student achievement.

Summary of results concerned with the effects of the classroom-level factors on student achievement.

In order to facilitate the understanding of the results of the multilevel analyses, Table 4.10 was constructed. This table illustrates the identified effects of the classroom-level factors and their dimensions as measured by the three instruments (i.e., high- and low-inference observation instruments and student questionnaire). It is quite revealing in several ways. Looking at the impact that each of the classroom-level factors has on student achievement, we can claim that generally the importance of the eight classroom-level factors is confirmed, since at least one dimension of each of the classroom-level factors was found to be associated with student achievement by using at least one of the three instruments to collect data about the functioning of the factors. However, it is apparent from this table that the only factor that was found to have an impact on student achievement irrespective of the instrument used is the management of time.

In addition, as can be seen from Table 4.10, all the measurement dimensions of application and modelling as measured by the low-inference observation instrument were found to be associated with student achievement. However, there are factors that were found to be associated with student achievement when only one measurement dimension was taken into account (e.g., the student-student interaction as measured by the student questionnaire). In addition, there are factors that were not found to have a statistically significant effect on student outcomes when the impact of their frequency dimension was measured but they had a significant effect on student achievement when other dimensions were taken into account (e.g., the case of orientation and assessment as measured by the student questionnaire). Moreover, a closer inspection of the table shows that the low-inference observation instruments were able to detect effects for more dimensions of the same factors compared to the student questionnaire, but the student questionnaire was able to detect effects in at least one dimension of the majority of the classroom-level factors. In addition, the high-inference observation instrument was the only instrument that was able to detect effects for two elements of the factor classroom as a learning environment (i.e., dealing with misbehaviour, teacher-student interaction).

Furthermore, the lack of impact of the dimensions of the factors that are presented in Table 4.10 with the abbreviation S→P on student achievement could be attributed to the exceptionally small observed variance in the functioning of these factors and their dimensions. Due to their small observed variance (see Tables 4.1-4.5) we did not have enough statistical power to detect their effects on student achievement. In some other cases (e.g., the differentiation dimension of structuring as measured by the low-inference observation instrument) the lack of impact could be attributed to the fact that the participating teachers did not exhibit skills concerned with some dimensions on their teaching (see Table 4.1).

Table 4.10

Overview of the impact of the classroom-level factors of the dynamic model and their dimensions (i.e., frequency (Freq), focus (Foc), stage (Stag), quality (Qual), and differentiation (Diff), as measured by all the instruments used in the study, on student outcomes)

Classroom-level factors	**Low- inference observation instruments**					**High-inference observation instrument**	**Student questionnaire**				
	Freq	**Foc**	**Stag**	**Qual**	**Diff**		**Freq**	**Foc**	**Stag**	**Qual**	**Diff**
Orientation					*N/A*	*N/A*		*N/A*	*N/A*	+	*N/A*
Structuring				S→P	S→P		Curv			S→P	*N/A*
Application	++	++	++	++	+		*N/A*	Curv		+	S→P
Modelling	+	++	++	+*	+	++		*N/A*	*N/A*	S→P	*N/A*
Questioning techniques	Curv	Curv	S→P	**		S→P	+	*N/A*	*N/A*		S→P
Management of time	+	*N/A*				+	+	*N/A*			
CLE: ***Dealing with misbehaviour***		*N/A*			S→P	++			*N/A*		*N/A*
CLE: ***Teacher-student interaction***		*N/A*				+		*N/A*	*N/A*		*N/A*
CLE: ***Student-student interaction***		*N/A*		S→P		S→P	++	*N/A*	*N/A*		*N/A*
Assessment	*N/A*					N/A		*N/A*	*N/A*	++	*N/A*

+ A statistically significant effect (p<0.10) upon student achievement was identified.

++ A statistically significant effect (p<0.05) upon student achievement was identified.

Curv= A curvilinear relationship was identified.

N/A= Not applicable because it was either not generalisable at the level of the teacher and/or classroom or it was not measured.

S→P= Observed variance of the functioning of the factor was extremely small. As a consequence we did not have enough statistical power to detect the effect of the factor on student outcomes (see Tables 4.1-4.5).

* The result regarding the aspect of appropriateness of the model of the quality dimension was statistically significant at .05 level.

** A statistically significant effect (p<0.05) upon student achievement was identified for the aspect of feedback-reaction to students' answers. A curvilinear relationship was identified for the aspect of feedback-reaction about the answer.

Finally, what is surprising is that some dimensions of some factors (i.e., the following dimensions of questioning techniques as measured by the low-inference: focus, the aspect of waiting time of the frequency dimension and the aspect of feedback-reaction about the answer of the quality dimension, as well as the frequency dimension of structuring and the focus dimension of application as measured by the student questionnaire) were found to have a negative effect on student achievement. This finding is not supported by the theory and the results of previous studies testing the validity of the dynamic model. However, when it was investigated whether curvilinear relations exist between the abovementioned dimensions of these factors and student achievement, it was found that they are related in a non-linear way to student outcomes, providing support to the assumption of the dynamic model regarding the existence of curvilinear relations. This implies that searching only for linear relations for some dimensions of some factors could lead to incorrect conclusions.

This Chapter presented the analyses of the research data collected in order to provide answers to the research questions of the study. The first section provided some descriptive information of the classroom-level factors and their dimensions as measured by the three instruments used in the study as well as the results of the generalisability study, which was conducted in order to examine the consistency of teacher behaviour in different classrooms. The second section illustrated the results of the multilevel analysis conducted to identify which level (the teacher or the classroom level) explains more variance in student achievement in mathematics of secondary school students. In addition, it illustrated the results of the multilevel analyses conducted to investigate the impact of the classroom-level factors on student achievement of secondary school students in Cyprus. These results provide support to the assumption regarding the generic nature of the classroom-level factors included in the dynamic model. In the next Chapter, further discussion of the results is presented.

CHAPTER 5

DISCUSSION AND SUGGESTIONS FOR FURTHER RESEARCH

This chapter builds on the findings of this study and draws conclusions in relation to the study's research questions, aiming to provide a better insight to the practical and theoretical contribution of this study. First, the main findings of this study, in relation to the consistency of teacher behaviour in different classrooms and more specifically whether the observation scores and/or student questionnaire scores per factor and dimension are generalisable at the teacher level, are summarised. The main findings of this study regarding the effects of the classroom-level factors of the dynamic model on student achievement are also summarised. Next, the theoretical and methodological implications of the research results are discussed. Moving on, implications of the findings for policy and practice, especially for teacher evaluation and professional development, are drawn. Finally, suggestions for further research are provided.

Introduction

Over the last decade, numerous attempts have been made to investigate issues related to the measurement of the teachers' skills and in particular, the obtainment of valid and reliable data on quality of teaching, such as, the number of observations per teacher that are needed to make a reliable generalization of a classroom teacher's practice, the number of raters per observation, certification or other rater requirements and whether similar inferences emerged from observations made live in the classroom and from video recordings of the same lessons (e.g., Casabianca et al., 2013; Hill et al., 2012; Joe et al., 2014; Newton, 2010; Whitehurst et al., 2014). As Praetorius and Charalambous (2018) argue, obtaining reliable and valid data from any given framework is particularly important not only for the evaluation of instructional quality but also for the investigation of the degree to which instructional quality contributes to student learning.

An important issue concerned with the measurement of quality of teaching, which has received scant attention in the research literature, is whether measuring teaching skills depends on the classroom context. In other words, what is not yet clear is whether teachers exhibit the same teaching skills when they teach different groups of students. It is important to clarify that teachers are expected to differentiate their instruction according to the specific needs of the students of each class. However, this study is concerned with the skills of teachers and examines whether teachers are able to demonstrate the same generic teaching skills, such as providing students with application opportunities (but not the same application tasks), in all the classes that they teach.

Despite the attempts that have been made (e.g., Chaplin et al., 2014; Grossman et al., 2014; Lazarev & Newman, 2015; Whitehurst et al., 2014) to investigate the extent to which measuring teaching skills by classroom observations and/or student questionnaire is influenced by classroom context variables, there is still uncertainty about whether the same teachers behave similarly in different classrooms. As discussed in Chapter 2, this is attributed to the fact that in most of these studies data have been obtained from a single class per teacher per year, and therefore, we are unable to determine whether differences in observational ratings are related to student characteristics or to the non-random sorting of teachers to classes of students. In the present study, this methodological weakness of the previous studies was taken into consideration and data about the skills of each teacher were collected from more than one classroom within a school year in order to examine the extent to which secondary school teachers exhibit the same generic teaching skills when they teach mathematics in different classrooms. More specifically, this study examines the extent to which data emerged from different classrooms about the behaviour of the same teachers are generalisable at the teacher level. It also investigates whether the findings are differentiated according to the instrument that is used by collecting data from high and low-inference observation instruments and a student questionnaire. In this way, the study reported here also aims to add to the discussions in the field of EER and teacher evaluation regarding which source of data is the most appropriate for collecting data on quality of teaching.

Given that the question of whether teachers behave similarly when they teach in different classrooms is more relevant when generic factors are used to evaluate them, this study is based on a specific theoretical framework (i.e., the dynamic model of educational effectiveness) which refers at classroom level (as it is a multilevel model) to eight generic factors that describe the teacher's instructional role. The eight classroom-level factors included in the dynamic model were found to be associated with achievement of primary and pre-primary students (Creemers & Kyriakides, 2015a; Panayiotou et al., 2014). Thus, this model provides the possibility for establishing an evidence-based and theory-driven approach for policy development and specifically, the establishing of teacher evaluation systems that could lead to improvement in student learning outcomes (Creemers & Kyriakides, 2008). However, given that the longitudinal studies testing the validity of the dynamic model took place at primary and pre-primary school level, the study presented in this thesis aims to identify the effect of the eight classroom-level factors on achievement of secondary school students too. Identifying whether the classroom-level factors are relevant for secondary school students too could provide further support to the generic nature of the classroom-level factors included in the dynamic model of educational effectiveness.

By collecting data from more than one classroom of the same teachers, this study also aims to distinguish between the classroom and teacher effect and investigate which level can explain more variance in student achievement of secondary school students. The section that follows summarizes the main findings of this study.

Main Findings of the Study

This study has shown that secondary school teachers behave consistently in different classrooms within a year in regard to the great majority of the measurement dimensions of five factors of the dynamic model (i.e., application, modelling, management of time, structuring and questioning techniques) as well as some elements of the factor classroom as a learning environment (i.e., the teacher-student interaction and student-student interaction) irrespective of the instrument used to evaluate them (see Tables 4.1-4.6). This is because observation data and student questionnaire data regarding the majority of the measurement dimensions of the aforementioned factors were found to be generalisable at the teacher level. Teacher behaviour was also found to be consistent from classroom to classroom in regard to the classroom-level factor of assessment which was measured only by the student questionnaire. This means that by observing a teacher teaching in different classrooms, similar results could emerge regarding the majority of the measurement dimensions of the abovementioned factors.

It should be noted that the dimensions of the aforementioned factors that were not found to be generalisable at the level of the teacher (i.e., the dimensions that teacher behaviour was not found to be consistent across classrooms and they are shown in Table 4.6) had an extremely small standard deviation (i.e., the scores among the participating teachers regarding these dimensions were very similar) in each instrument used. Consequently, we did not have enough statistical power to examine the generalisability of these dimensions of these factors at the level of the teacher. If we had picked a larger sample of teachers, we could have had enough statistical power to examine the consistency of the teacher behaviour in different classrooms regarding the abovementioned dimensions. However, for the differentiation dimension of questioning techniques and the frequency dimension of student-student interaction, the findings were differentiated according to the instrument that was used (i.e., low-inference and student questionnaire). Specifically, they were found to be generalisable at the teacher level when measured by the low-inference observation instrument but they were found to be generalisable at the classroom level when measured by the student questionnaire.

Another important finding was that for two factors (i.e., orientation and dealing with misbehaviour which is an element of the factor classroom as a learning environment), conflicting results emerged. Specifically, the factor concerned with the teacher’s ability to deal with misbehaviour was found to be generalisable at the teacher level only by using the high-inference observation instrument to collect data. The data that emerged from the low-inference observation instrument, which is considered as more precise when measuring teaching skills, as well as the data that emerged from the student questionnaire, showed that teacher behaviour regarding this factor is not consistent from classroom to classroom. Regarding the classroom-level factor of orientation, only the frequency dimension was found to be generalisable at the teacher level and only in the data gathered with the low-inference observation instrument. When it was measured by the student questionnaire, it was found to be generalisable at the level of the classroom, meaning that teachers were not found to behave similarly across classrooms in regard to the frequency dimension of orientation. In regard to the other dimensions of orientation, teacher behaviour was not found to be consistent from classroom to classroom in neither of the instruments used in the study. When the orientation was measured by the high-inference observation instrument, it was not found to be generalisable neither at the level of the teacher nor the classroom, meaning that the teachers were not found to behave similarly across classrooms in regard to this factor not even across lessons in the same class. However, this result may be attributed to the small observed variance that orientation had when measured by the high-inference observation instrument (see Table 4.4). As discussed before, when the observed variance of the functioning of a factor is extremely small, we do not have enough statistical power to examine the generalisability of the factor at the level of the teacher.

Surprisingly, the observation data showed that the great majority of the teacher sample did not differentiate their teaching, especially in terms of structuring and dealing with misbehaviour, which is an element of the factor concerned with the classroom as a learning environment; and also no teacher was found to differentiate his/her teaching in terms of orientation. In other words, activities associated with the majority of the classroom-level factors were found to be implemented in the same way for all the students of each class in most of the cases. In the case of the student questionnaire, a generalisability study on the use of students' ratings was conducted before the creation of factor scores in order to investigate whether there was consensus in student responses in each classroom separately and to identify whether the object of measurement was the teacher. However, some items of the questionnaire concerned with some dimensions of the classroom-level factors (see Table 4.5) were not found to be generalisable at the level of the classroom and hence, had to be removed. Thus, it was not possible to generate factor scores for all the dimensions of the classroom-level factors.

With respect to the question of which level explains more variance in student achievement, the multilevel analysis revealed that the teacher level, rather than the classroom level, explains more variance in student achievement in mathematics of secondary school students. The present study was also designed to investigate the extent to which the classroom-level factors of the dynamic model and their dimensions are associated with student achievement in mathematics of secondary school students in Cyprus. The findings of this study, in general, support the validity of the dynamic model regarding its eight classroom-level factors, since at least one dimension of each of the classroom-level factors was found to be associated with achievement in mathematics of secondary school students in at least one of the three instruments used in the study (see Table 4.10). However, it should be acknowledged that some factors were found to have a statistically significant effect on student achievement when only one measurement dimension was taken into account and not necessarily the frequency dimension.

Another important finding was that different results emerged according to the instrument used in regard to the impact of the classroom-level factors on student achievement. Specifically, the low-inference observation instruments were able to detect effects for more dimensions of the same factors compared to the student questionnaire. However, the student questionnaire was able to detect effects in at least one dimension of most of the classroom-level factors. Nevertheless, in the case of the student questionnaire, as mentioned before, it was not possible to generate factor scores for all the measurement dimensions of the classroom-level factors and search for their effects on student achievement. Regarding the high-inference observation instrument, it was the only instrument that was able to detect effects for two elements of the factor classroom as a learning environment (i.e., dealing with misbehaviour and teacher-student interaction). It is important to note that only the management of time was found to have an impact on student achievement, irrespective of the instrument used.

One unanticipated finding was that the current study found curvilinear relations of some dimensions of some factors with student achievement in mathematics (i.e., the following dimensions of questioning techniques as measured by the low-inference observation instrument: focus, the aspect of waiting time of the frequency dimension and the aspect of feedback-reaction about the answer of the quality dimension, as well as the frequency dimension of structuring and the focus dimension of application as measured by the student questionnaire). It is important to note that before examining for curvilinear relations, these dimensions of these factors were found to have a negative effect on student achievement.

With respect to the student variables (i.e., prior achievement in mathematics, gender, ethnicity and SEN), the results from the multilevel analysis revealed that all of them had statistically significant effects on student achievement. In the following section the theoretical and methodological implications of the study's findings are discussed.

Theoretical and Methodological Implications

Over the last decades, several frameworks and associated instruments have been developed to study and analyze teaching. Each framework seems to capture different aspects of teaching (see Charalambous & Praetorius, 2018; Praetorius & Charalambous, 2018). While a significant work has been done in researching the quality of teaching, much uncertainty still exists about whether measuring teaching skills depends on the classroom context. The current study was based on the dynamic model of educational effectiveness and its findings suggest that the eight classroom-level factors included in this model can be classified into two main categories: a) factors that are expected to occur in every lesson and every class taught by the same teacher (i.e., teacher behaviour is consistent from classroom to classroom) and b) factors that are not expected to occur in every lesson and/or every class taught by the same teacher (i.e., teacher behaviour in different classrooms may be inconsistent).

According to the findings, the first category includes six factors (i.e., structuring, application, modelling, questioning techniques, management of time and assessment) and some elements of the factor classroom as a leaning environment (the teacher-student interaction and student-student interaction). These factors have two things in common: a) the majority of their measurement dimensions were found to be generalisable at the teacher level irrespective of the instrument used to evaluate them (except the frequency dimension of student-student interaction and the differentiation dimension of questioning techniques) and b) the observed variance of the functioning of the dimensions that were not found to be generalisable at the level of the teacher was extremely small and therefore, we did not have enough statistical power to detect for significant differences and examine the generalisability of these dimensions at the level of the teacher (see Table 4.6). Therefore, collecting data from only one classroom per teacher suffices to measure teaching skills concerning the aforementioned factors and the majority of their dimensions, irrespective of the instrument used (i.e., high and low-inference observation instruments and student questionnaire).

The second category consists of the classroom-level factor of orientation and the teacher's ability to deal with misbehaviour, which is an element of the factor classroom as a learning environment. The results of the present study indicate that special attention is needed before drawing conclusions for these two factors, since the majority of their measurement dimensions were not found to be generalisable at the level of the teacher in the majority of the instruments used in this study. In other words teachers were not found to behave similarly across classrooms in regard to these two factors in most of the instruments used. Therefore, the results of this study suggest that data from more than one classroom per teacher are needed to measure teaching skills concerning the abovementioned factors. Moreover, these findings imply that previous studies which have collected data on these two factors from only one classroom per teacher might have drawn incorrect conclusions about teachers' skills concerning these factors. The same could have happened with teacher evaluation results regarding these factors, an issue that is discussed in the next section of this chapter. This inconsistency in teacher behaviour across different classrooms regarding the factors of orientation and dealing with student misbehaviour may be attributed to some limitations in measuring the teachers' skills concerning these two factors, which are discussed below.

Regarding the factor of dealing with misbehaviour, the inconsistency in teacher behaviour across different classrooms may be attributed to the fact that in one or more of the classes taught by the same teacher, no student misbehaviour incident might have taken place, contrary to his/her other class/ classes. As it is acknowledged by Kyriakides et al.

(2018a), when no misbehaviour incident takes place, the observer can say nothing about the ability of a teacher to deal with student misbehaviour, since no data about this factor can be generated. In order to evaluate the skills of a teacher in dealing with misbehaviour, student misbehaviour incidents must take place in his/her classroom. Thus, data from more than one classroom of the same teacher are needed to measure the factor of dealing with student misbehaviour before drawing conclusions for this factor.

As discussed in Chapter 2, a large body of literature on student misbehaviour pays particular attention to the causes of misbehaviour incidents by investigating teachers' and students' perceptions. Several studies have shown that students tend to attribute the causes of misbehaviour to teachers and teaching (e.g., Cooper et al., 1994; Kyriacou, 2009), whereas teachers tend to attribute the causes to students or family rather than teaching-related factors (e.g., Baron, 1990; Ho, 2004; Koutrouba, 2013; Kyriacou & Martin, 2010). In the present study, attention is given to the teachers' skills in dealing with student misbehaviour and to whether the existence of student misbehaviour problems may affect the consistency of teacher behaviour in different classrooms regarding the other teaching skills and not who is responsible for student misbehaviour. Specifically, in Chapter 1, it was assumed that in those classrooms where many misbehaviour problems occurred, a teacher would not be able to demonstrate his/her other teaching skills unless she/he was good in dealing with student misbehaviour. Additionally, it was assumed that if a teacher was not good in dealing with misbehaviour and he/she worked in a classroom with only a few misbehaviour problems, he/she would not have a problem with demonstrating his/her other teaching skills.

Surprisingly, although teacher behaviour was not found to be consistent across classrooms in regard to the factor of dealing with misbehaviour in this study, teacher behaviour in most of the other classroom-level factors was found to be consistent from classroom to classroom irrespective of the instrument used. This finding might be explained by the fact that previous studies (see Kyriakides et al., 2018a) have shown that the factor of dealing with misbehaviour is situated at stage 1 of effective teaching. Specifically, previous studies supported the assumption that the classroom-level factors of the dynamic model and their dimensions are inter-related, and revealed that they can be classified into stages of effective teaching, structured in a developmental order (see Chapter 2). The fact that the factor of dealing with misbehaviour was found to be situated at stage 1 in these studies implies that a teacher who does not have the skills to deal with student misbehaviour effectively, is also not effective in terms of the other classroom-level factors that were found to be situated in the other stages of effective teaching. Thus, this might be a possible explanation in regard to the fact that in the present study, teacher behaviour in most of the classroom-level factors was found to be consistent in different classrooms, even if teacher behaviour regarding the factor of dealing with misbehaviour was found to be inconsistent. As discussed before, the inconsistency of teacher behaviour in different classrooms regarding the factor of dealing with misbehaviour may be due to the fact that in one of the classes taught by the same teacher, who might not have the abilities to deal with student misbehaviour effectively, no student misbehaviour incidents had taken place, contrary to his/her other class/ classes.

According to the results, orientation is the other classroom-level factor that needs special attention. A possible explanation for this might be that orientation may not occur in every lesson (which may not necessarily be problematic since student learning may still take place) and one could therefore argue that the two observations that were conducted in each class may not be enough to measure this factor. This is because orientation might have occurred in one classroom but not in another, as we might have not visited this classroom on the same day and the orientation might have preceded. Another possible explanation regarding the fact that orientation was not found to be consistent across classrooms, especially as measured by the student questionnaire, might be that it is related to the students' motivation (Kyriakides et al., 2018a). It is important to note here that students are not assigned randomly in the classrooms. Thus, teachers in some classrooms may provide more and different types of orientation according to the motivation of the students of each class. Therefore, the student-level factor of subject motivation, which is included in the dynamic model, may affect teacher behaviour in regard to the classroom-level factor of orientation. In the dynamic model of educational effectiveness, relations among factors operating at the student level and factors operating at the classroom level are expected to exist and further research into differential effectiveness could help us identify these relations (Creemers & Kyriakides, 2008). However, the present study is not in a position to answer why teacher behaviour was not found to be consistent from classroom to classroom in regard to the factor of orientation. This could be examined with further research that will investigate the relationship of the classroom-level factor of orientation with the student-level factor of subject motivation.

The findings of the current study, also suggest that the type of instrument used to measure teaching skills (i.e., high, low-inference observation instruments and student questionnaire) can contribute to whether the data are generalisable at the level of the teacher and/or classroom for some dimensions of some factors (i.e., the frequency dimension of orientation, the differentiation dimension of questioning techniques and two elements of the factor classroom as a learning environment: dealing with misbehaviour and the frequency dimension of student-student interaction). The observation data (data collected from both the high and the low-inference observation instruments) were those found to be generalisable at the level of the teacher for most of the classroom-level factors of the dynamic model compared to the data measured by the student questionnaire. In some cases (i.e., the frequency dimension of student-student interaction and the differentiation dimension of questioning techniques) the difference among the results of factors measured by the student questionnaire and the low-inference observation instruments may be attributed to the small observed variance of the data collected by the student questionnaire. Thus, one could argue that the low-inference observation instruments are able to detect differences more easily among teachers regarding the frequency dimension of student-student interaction and the differentiation dimension of questioning techniques than the student questionnaire. In some other cases (i.e., orientation and dealing with misbehaviour) the difference among the findings may be attributed to the fact that the instruments used in this study were designed to collect data concerned with different aspects of the classroom-level factors of the dynamic model (Kyriakides et al., 2018a).

With respect to the student questionnaire, as discussed in Chapter 4, it was not possible to generate factor scores for all the dimensions of the classroom-level factors as some items were not found to be generalisable at the level of the classroom. A possible explanation for the disagreement among students regarding some items might be that some of the items (e.g., I need some time to think before I answer one of my teacher's questions) have to do with the ability of the student (i.e., high or low achieving student) who responds to these items. The results of the generalisability study concerning students' ratings are not in line with the results of previous studies conducted in primary schools which showed that almost all the student questionnaire items could be used for measuring the quality of teaching of each teacher (see Kyriakides & Creemers, 2008; Kyriakides et al., 2009). Therefore, the results of this study suggest that students should not be the sole source of data for measuring the quality of teaching and specifically the classroom-level factors of the dynamic model. The use of multiple sources of data and in particular the complementary use of classroom observations could provide the opportunity to collect more precise data on all the dimensions of the classroom-level factors, gaining a more comprehensive picture of teacher performance.

Regarding the level that explains more variance in student achievement, it was found that the teacher level explains more variance in student achievement than the classroom level. This finding implies that teachers are equally effective in promoting learning outcomes of students (of the same age) in different classrooms. One could attribute this finding to the fact that teachers were found to perform similarly for most of the classroom-level factors included in the dynamic model when being observed teaching in different classrooms and/or when their students in different classrooms were asked to evaluate their teacher's behaviour. However, this finding needs to be interpreted with caution as this study, as far as we know, is the only study which compared the teacher and classroom level effects on student achievement and took place in a single country where the effects of the levels were examined in only one cognitive subject (i.e., mathematics). So far, as mentioned in Chapter 1, more emphasis is given to the effect that the department has at the level of secondary education (Ko et al., 2015; Ko et al., 2016; Sammons et al., 1997).

Thus far, previous studies testing the validity of the dynamic model were conducted in primary and pre-primary schools (Kyriakides et al., 2018a). The results of this study indicate that the classroom-level factors included in the dynamic model are relevant for secondary school students too. This implies that the classroom-level factors of the dynamic model could be considered as generic as they were found to be associated with student achievement in different phases of schooling. The fact that some factors were found to be associated with student achievement when only one measurement dimension was taken into account is in line with the results of a previous study conducted in primary schools (i.e., Creemers & Kyriakides, 2008) which showed that there are possibilities for creating a more parsimonious model by indicating factors that should be measured across all the measurement dimensions and factors for which it is not necessary to measure across the five dimensions. Nevertheless, it should be noted that the lack of impact of some dimensions of some factors on student achievement in this study (see Table 4.10) could be attributed to the following two reasons: a) the lack of enough statistical power to detect the effect, since the standard deviation of some dimensions of some factors was exceptionally low, such as the quality dimension of structuring and b) the participating teachers did not exhibit skills concerned with some dimensions on their teaching (e.g., the differentiation dimension of structuring, see Table 4.1) and this might be considered a context specific result. In regard to the first reason, a larger sample of teachers could perhaps help to have enough statistical power to detect the effect of these dimensions.

Most effectiveness studies in the past used predominantly the frequency dimension to measure the effectiveness factors (i.e., how frequently an activity associated with a factor took place), taking, in that way, into account only the quantitative characteristics of a factor (Kyriakides et al., 2018a). One interesting finding of the present study is that there are factors that were not found to have a statistically significant effect on student achievement when the impact of their frequency dimension was measured; but they had a statistically significant effect on student outcomes when other dimensions were taken into account. Thus, this study confirms that emphasis should be given not only to the frequency dimension but also to other dimensions of effectiveness factors (Creemers & Kyriakides, 2008; Kyriakides & Creemers, 2009). In addition, since this study is the first attempt to test the validity of the dynamic model at secondary school level, it suggests that the use of different dimensions to measure the functioning of the classroom-level factors could be used to describe more precisely the effective teaching for secondary education too.

As in the case of the consistency of teacher behaviour across different classrooms, different results emerged regarding the effects of the classroom-level factors on student achievement according to the type of instrument used to measure teaching skills (i.e., high, low-inference observation instruments and student questionnaire). Thus, these findings revealed the importance of using all the instruments together to measure the quality of teaching. Moreover, these findings imply that previous studies that have used only one instrument to measure the classroom-level factors might have drawn incorrect conclusions about the impact of a factor on student achievement.

Furthermore, as mentioned in the literature review, there is some debate in the literature in regard to whether or not traditional or constructive teaching approaches are more effective and whether they benefit all groups of students (Caro et al., 2016). The results of this study further support the idea of using an integrated approach in defining quality of teaching, since the multilevel analyses revealed that factors belonging to different teaching approaches (e.g., the factor of modelling which is associated with the constructivist approach and the factor of application which is associated with a more traditional approach) are related to student achievement of secondary school students. These findings are in line with those of recent meta-analyses (e.g., Kyriakides et al., 2013; Seidel & Shavelson, 2007) which showed that within each approach there are factors which are associated with student achievement.

Additionally, this study seems to reveal that some differences exist between different phases of schooling by comparing its results with the results of previous studies conducted in primary schools (i.e., Creemers & Kyriakides, 2008; Kyriakides & Creemers, 2008, 2009; Kyriakides et al., 2009). However, the generalisability of the findings of this study should be tested as it is the first study testing the validity of the dynamic model in secondary schools. As mentioned in the literature review, the dynamic model assumed that the relationship of some effectiveness factors with student achievement may not be linear but curvilinear (Creemers & Kyriakides, 2006). However, curvilinear relations were not found among classroom-level factors and student outcomes in mathematics in the previous national studies conducted in primary and pre-primary schools (see Creemers & Kyriakides, 2008; Kyriakides & Creemers, 2009). Only two such relations were identified (i.e., the frequency dimension of the factor of questioning techniques and the frequency dimension of the factor of assessment) and only in relation to the teaching of the Greek language. As Creemers and Kyriakides (2008) argue, the difficulty of demonstrating curvilinear relations in their study may be attributed to the difficulty of establishing enough variation in the functioning of the classroom-level factors. Surprisingly, the current study found curvilinear relations for some dimensions of some factors with student achievement in mathematics (i.e., the following dimensions of questioning techniques as measured by the low-inference: focus, the aspect of waiting time of the frequency dimension and the aspect of feedback-reaction about the answer of the quality dimension, as well as the frequency dimension of structuring and the focus dimension of application as measured by the student questionnaire), providing support to the assumption of the dynamic model regarding the existence of curvilinear relations. This finding may be attributed to the fact that the variance of these dimensions of these classroom-level factors in secondary schools was much higher than in the primary and pre-primary schools. In other words, many more differences may exist among the behaviour of secondary school teachers who teach mathematics regarding the abovementioned factors compared to the primary school teachers. A more detailed discussion of the findings concerning the existence of curvilinear relations follows.

Regarding the focus dimension of questioning techniques, this was measured by looking at the relation between each question and the tasks that took place during a lesson (i.e., if the question was related to a specific task only, the whole lesson or the unit/ a number of lessons). The results of this study suggest that asking too many questions related to the unit/number of lessons may reach an optimal point. A possible explanation for this might be that the extensive provision of questions related to previous lessons and previous knowledge may deprive students of the time to learn new lesson content. Therefore, suggestions can be made for teachers for trying to maintain a balance between the different types of questions (i.e., more specific questions related to a specific part of the lesson and the more general questions related to the whole lesson or a number of lessons) that they use in their teaching. It is important to note that this finding is not in line with the findings of previous studies conducted in primary schools, where no curvilinear relations were found among the focus dimension of questioning techniques and student achievement in mathematics. A possible reason for this might be that secondary school teachers may use a much larger number of questions in their teaching than primary school teachers and this may be due to the fact that teaching in secondary school classrooms may be more teacher-centred (Toh, Ho, Chew, & Riley, 2003) than in primary school classrooms.

Similar results with the focus dimension of questioning techniques emerged for the focus dimension of application, as it was found to reach an optimal point where the provision of application tasks related to the unit/ a number of lessons has no further effect on student achievement. Therefore, providing too many application tasks related to previous lessons may not allow sufficient time to provide application tasks to students that will help them practice in what they have learnt during that lesson and eventually, progress in terms of acquiring sufficient amount of knew knowledge. Thus, having a balance among the application tasks related to a specific part of the lesson and tasks related to the whole lesson or a number of lessons could be suggested.

The aspect of waiting time of the frequency dimension of questioning techniques was also found to be related in a not linear way to student achievement. This aspect has to do with the time that elapses after a question is posed by the teacher and before the teacher asks a student to answer this question (Creemers & Kyriakides, 2012). It is important to note that this aspect of questioning techniques was not able to be measured in primary schools as the length of pause following questions was too small in order to be captured. The current study not only managed to capture this aspect of questioning techniques but it has also shown that the length of that pause could reach an optimal point. This suggests that although teachers are expected to wait before they ask a student to answer a question in order to give the students time to think (Brophy & Good, 1986), too much waiting time may negatively affect the teaching time and eventually, the student achievement. A possible explanation for the difference between the teaching behaviour of secondary teachers and the behaviour of primary teachers regarding the length of pause following questions might be that the mathematics curriculum of secondary education is more cognitively demanding than the curriculum of primary education. Thus, the secondary teachers may implement more cognitively demanding instructional tasks (see Smith & Stein, 1998) and ask more process questions than the primary teachers. As Brophy and Good argue, the length of pause following questions is expected to vary according to the difficulty level of the questions and especially their complexity or cognitive level.

The other dimension of questioning techniques that was found to have a curvilinear relation with student learning outcomes is the aspect of feedback-reaction about the answer of the quality dimension. This aspect was measured by looking at whether a teacher's reaction about an answer that a student has given belonged to one of the three following categories: 1) teacher ignores the answer, 2) teacher indicates that the answer is correct or partly correct or incorrect, 3) students are invited to give comments on the answer. This study has shown that this aspect may reach an optimal point where it can have a negative effect on student outcomes. A possible explanation for this might be that even if teachers are expected to promote interactions among students (Creemers & Kyriakides, 2008), giving too much time to students to comment on the answers of other students may act at the expense of teaching new learning content (based on the fact that teaching time is, at a certain degree, restricted).

The last factor that was found to have a curvilinear relation with achievement is the factor of structuring and specifically its frequency dimension. Rosenshine and Stevens (1986) note that student achievement is maximised not only when teachers actively present materials but also when they structure their lessons. Structuring activities aim to help students develop links between the different parts of lessons, instead of dealing with them in an isolated way (Kyriakides et al., 2018a). However, the current study has found that structuring may reach an optimal point where it can no longer be beneficial for student achievement of secondary school students. It can therefore be assumed that spending too much time on structuring activities may act at the expense of students' time on task; which in turn may lead to less chance for the learning aims to be achieved. As Creemers and Kyriakides (2008) mention, the time on task has to do with the time during which students are really involved in learning. This variable is included in the student level of the dynamic model. It is argued that this variable along with the student-level variable of opportunity to learn, influence learning directly. Therefore, less time on task may lead to less impact that a specific lesson can have on student learning. As discussed in Chapter 4, before examining for curvilinear relations, the dimensions of the factors mentioned before may be found to have a negative effect on student achievement. This implies that searching only for linear relations for some dimensions of some factors may lead to misleading conclusions regarding their relationship with student achievement.

Finally, regarding the student variables used (i.e., prior achievement in mathematics, ethnicity and SEN), the results of this study are in line with previous research findings which stress that the student-level variables have a significant effect on student learning outcomes (e.g., Kyriakides & Creemers, 2008, Creemers & Kyriakides, 2008). With respect to gender, in contrast to earlier findings conducted in primary schools (e.g., Kyriakides et al., 2009), this study has shown that girls have better results in mathematics compared to boys. However, in general, it seems that inconsistent results emerged from studies conducted in Cypriot primary schools regarding the effect of gender on student achievement. For instance, Demosthenous et al. (2018) have shown that gender does not have a statistically significant effect on student achievement in mathematics of primary school students, whereas other studies (e.g., Creemers & Kyriakides, 2008) have shown that gender has a statistically significant effect on student achievement and boys have better results compared to girls. Nevertheless, what seems to be more important is not which group of students (i.e., girls or boys) has the best results in mathematics, but what primary and secondary schools and teachers do to promote not only quality, but also equity in education. As Kyriakides, Creemers and Charalambous (2018b) argue, the dimension of equity demands that the expected learning outcomes of students should depend only on their own efforts and capacity, and not on considerations over which they have no influence (e.g., gender and family socio-economic level). The effectiveness status of teachers or schools in terms of equity can be measured by looking at the extent to which differences in student achievement between groups of students with different characteristics (e.g., gender) are reduced.

Concluding, the findings of the present study, as these are discussed above, provide new insights on the measurement of the classroom-level factors of the dynamic model, as well as the impact of these factors on student achievement of secondary school students and their relation with student learning outcomes (i.e., linear and curvilinear). Despite the implications of findings for the further development of the dynamic model at the classroom level, the results of this study also have important policy implications especially for teacher evaluation and professional development. These implications are discussed in the section below.

Implications for Policy and Practice

In the past few years there has been an increasing interest in many countries in issues related to teacher evaluation in order to improve the quality of education (Flores, 2012; Kyriakides & Demetriou, 2007; Liu& Zhao, 2013). In Cyprus, particularly, where this study was conducted, the Ministry of Education and Culture has initiated a new structured dialogue with the relevant stakeholders, during the last year, regarding the modernization of the teacher evaluation system, since both policy-makers and stakeholders seem to accept that the existing evaluation system does not contribute to educational improvement (MOEC, 2018). The existing teacher evaluation system is in place since 1976 without any important changes (Kyriakides & Campbell, 2003; Kyriakides, 2016). In the section below, implications of the methodology used in this study and its findings are drawn for the development of new teacher evaluation systems.

As discussed in the literature review, the evaluation purposes, the performance criteria and the evaluation procedures and sources that will be used for collecting relevant data are the three basic aspects that must be taken into account when developing a comprehensive teacher evaluation system (Ellett, et al., 1996; Iwanicki, 1990). With respect to the process of generating criteria, apart from the teachers' job description and/or the professional code, it is supported in the literature that the results of TER and its main theoretical models could also be used as a foundation upon which evaluation criteria could be established (Kyriakides et al., 2006; Kyriakides & Demetriou, 2007). Many countries around the world, including Cyprus (Kyriakides & Campbell, 2003), seem to adopt mainly the working process model of Cheng and Tsui (1999) as they give emphasis on the quality of teaching for teacher evaluation (see Chapter 2). The first step that a country, which adopts the working process model, should take is to define the teacher factors that will be used as a basis for the development of criteria. The findings of this study suggest that the eight classroom-level factors included in the dynamic model could help policy-makers generate criteria for both formative and summative evaluation regarding the generic teaching skills that help teachers to be effective, since at least one dimension of each of the classroom-level factors was found to be related to achievement in mathematics of secondary school students. The importance of the classroom-level factors is also supported by both longitudinal studies conducted in primary and pre-primary schools and meta-analyses (see Creemers & Kyriakides, 2015a).

Once a country adopts the eight classroom-level factors of the dynamic model or some of them and/or other factors, the next step should be to develop the instruments that will be used for collecting relevant data and also to test the psychometric properties of the instruments that will be developed (Creemers & Kyriakides, 2008). As discussed in the previous section of this chapter, the findings of this study provide support to the idea of using multiple sources of data for documenting teacher performance (Kyriakides et al., 2014; Kane & Stanger, 2012). Then, a similar research to the research described in this thesis should be conducted in order to examine whether the classroom context affects the measuring of teaching skills concerning the selected factors or some of them. In other words, it should be investigated whether the same teachers exhibit the same teaching skills regarding the selected factors when they teach in different classrooms and more specifically whether their scores regarding the selected factors can be aggregated at the teacher level irrespective of the class they have to teach. Through this research, the selected factors may be classified into categories as those emerged in the current study (i.e., a) factors that are expected to occur in every lesson and every class taught by the same teacher and b) factors that are not expected to occur in every lesson and/or every class taught by the same teacher). According to the results, data from only one classroom per teacher will be needed to measure the factors that will be found in the first category and data from more than one classroom will be needed before drawing conclusions for the factors that will be found in the second category. If a country is reluctant to gather data from more than one classroom of the same teacher for teacher evaluation due to practical and financial issues, then a study must be conducted to examine the added value and efficiency of this policy.

The findings of the current study have shown that special attention will be needed before drawing conclusions for teacher evaluation for orientation and for the teacher’s ability to deal with misbehaviour. Even if these two factors were found to be relevant for promoting secondary school students learning outcomes, teacher behaviour regarding the majority of the dimensions of orientation and dealing with misbehaviour was not found to be consistent from classroom to classroom. This implies that data from more than one classroom per teacher are needed before drawing conclusions for the skills of teachers regarding these factors for formative and especially summative purposes of teacher evaluation. If a teacher evaluation system cannot afford to obtain data from more than one classroom of the same teachers within a school year (e.g., because of a small number of inspectors), then orientation and dealing with misbehaviour cannot be used as criteria for evaluating teachers, especially for summative purposes. However, by choosing to not take into account the factors of orientation and dealing with misbehaviour for summative evaluation, there is a risk of sending the wrong message that these two factors are not important factors; and this may eventually lead teachers to not pay attention to them in their teaching. The findings of this study also imply that the existing teacher evaluation systems that obtain data from only one classroom of the same teachers regarding teaching skills that are related to the factors of orientation and dealing with misbehaviour, may not generate representative scores for the skills of teachers in these two factors. To avoid this problem there are two options for teacher evaluation systems: a) either to take these two factors into account for teacher evaluation purposes but collect data from different classrooms of the same teachers before drawing conclusions for the skills of teachers; or b) to not take them into account for summative reasons but ensure that teachers will be supported to improve their skills concerned with these two factors in teacher professional development programs. Consequently, the insights gained from this study could be taken into consideration by policy-makers when they develop new teacher evaluation systems but also when they revise the existing teacher evaluation systems.

The fact that teacher behaviour was not found to be consistent from classroom to classroom in regard to the factor of dealing with misbehaviour may indicate the need for teachers to participate in relevant intervention programs aimed at decreasing students' behaviour problems and improving the strategies used by teachers for addressing discipline problems (e.g., Freiberg & Lapointe, 2011; Hattie, 2009; Muscott et al., 2008). In addition, this finding may draw attention to the need for support, from the part of the schools to the teachers through direct assistance and appropriate training in techniques of addressing discipline problems. Furthermore, an adaptation of the universities' curricula may be suggested in order to help future teachers acquire useful skills for the prevention and the addressing of discipline problems. As previous studies revealed, a relatively large percentage of teachers seem to believe that they are unprepared to deal with disciplinary problems and they spend more time than they ought in order to address them (Houghton et al., 1988; Little, 2005; Wheldall & Merrett, 1988).

With respect to the dimension of differentiation in general, the results of this study showed that the majority of the participating teachers did not differentiate their instruction in terms of most of the classroom-level factors, since the mean of the dimension of differentiation of the majority of the factors was relatively low. This is a cause of considerable concern due to the fact that all classes in Cyprus, where this study was conducted, are mixed-ability and it was expected that teachers would differentiate their teaching (i.e., adapt their teaching to the specific learning needs of each student or groups of students). Thus, this finding raises questions regarding the way and the criteria that inspectors in Cyprus used to evaluate teachers, since all teachers in Cyprus are awarded by their inspectors with grades above 32 points out of 40 and the great majority of teachers are awarded with very high grades (i.e.,35-37, see Kyriakides & Campbell, 2003; The World Bank, 2014). The fact that the majority of secondary school teachers do not differentiate their instruction may be one of the reasons that Cyprus has poor results in mathematics in international studies like PISA (OECD, 2018).

As discussed in Chapter 2, previous studies examined the assumption that the classroom-level factors of the dynamic model and their dimensions are inter-related, and showed that they can be classified into stages of effective teaching (see Kyriakides et al., 2018a). The findings of the present study regarding the dimension of differentiation may be explained by the fact that in the abovementioned previous studies the differentiation dimension of the classroom-level factors was found to be situated at the last two stages of effective teaching, which are the most demanding. For instance, the differentiation dimension of structuring was found to be situated at stage four and the differentiation dimension of orientation was found to be situated at stage five. Previous studies (e.g., Kyriakides et al., 2009) revealed that teachers who use more advanced types of teaching behaviour (i.e., teaching skills included in the last two stages, as the differentiation dimension of orientation) have better student learning outcomes. This may therefore indicate the need for policy-makers to develop relevant training courses that will help teachers to acquire skills concerned with the differentiation dimension of the classroom-level factors.

Apart from the dimension of differentiation, the results of the current study could be used to inform policy-makers about other effective practices at classroom level that could contribute to the improvement of student achievement of secondary school students in order to implement professional development courses that will promote these effective practices. More specifically, the findings of this study could be used by policy-makers to identify priorities for improvement of factors that were found to be associated with student learning outcomes and their functioning was not satisfactory. For instance, the factor of modelling and specifically its stage dimension could constitute a priority for improvement as it was found to be associated with student achievement and the observation data (see Table 4.1) revealed that its functioning was not satisfactory as its mean was quite small (i.e., mean= 0.74, SD= 0.84, on a scale ranging from 0 to 3). Therefore, the results of this study support that the classroom-level factors of the dynamic model and their dimensions may help policymakers not only to generate teacher evaluation criteria, but also to form improvement action plans by indicating ways of improving educational practice other than just increasing the presence of the classroom-level factors in the classroom. It is important to note that the practical use of the dynamic model for improvement purposes has already been proven in previous studies (e.g., Antoniou, 2013; Antoniou & Kyriakides, 2011, 2013; Antoniou, Kyriakides & Creemers, 2011; Kyriakides et al., 2017).

Research Limitations and Suggestions for Further Research

A number of frameworks and associated instruments with different foci have been developed over the past two decades to study and analyze teaching (see Charalambous & Praetorius, 2018; Praetorius & Charalambous, 2018). This study was based on the dynamic model of educational effectiveness and has shown that teacher behaviour concerning two factors included in this model (i.e., orientation and teacher's ability to deal with misbehaviour) may not be consistent from classroom to classroom, and also that the type of instrument used to measure teaching skills could contribute to whether similar judgements will be produced when the same teacher is evaluated across different classrooms. However, further research is required to explore whether teachers exhibit the same teaching skills when they teach in different classrooms and more specifically whether their scores can be aggregated at the teacher level irrespective of the class they have to teach by using instruments based on a different theoretical framework. In addition, given that this study was based on a relatively small sample of teachers who teach mathematics to seventh or eighth grade students in Cyprus, further research is needed to test the generalisability of the findings by collecting data on teacher behaviour in different subjects, in other grades and in different educational contexts.

Even if this study has shown that there may be inconsistency in teacher behaviour concerning the factors of orientation and dealing with misbehaviour when the same teachers teach in different classrooms, this study was not in a position to examine why and under what conditions there is inconsistency in teacher behaviour regarding these two factors. Thus, further research is needed to investigate the conditions producing these differences in teacher behaviour across different classrooms by using multiple sources of data and not only the classroom observations and the student questionnaire. For instance, in order to examine why and under what conditions there is inconsistency in teacher behaviour regarding the factor of dealing with misbehaviour, apart from the classroom observations, school records about students' behaviour could also be used in order to examine how severe the problem of student misbehaviour is in each class and how often misbehaviour incidents occur in each class, since misbehaviour incidents may not appear in every lesson.

As mentioned in a previous section of this chapter, a possible explanation regarding the fact that orientation was not found to be consistent across classrooms might be that this factor is related to the students' subject motivation. Therefore, further research is needed to examine whether the variation in teacher behaviour across different classrooms that is observed regarding the factor of orientation, could be explained by the fact that differences in students' subject motivation exist among classrooms that are taught by the same teacher. Specifically, the subject motivation of the students of each class of the same teacher could be measured. Then, a generalisability study could be conducted in order to examine whether the data that will emerge are generalised at the classroom level. After that, it should be examined whether students' subject motivation of each class could explain the variation in teacher behaviour across different classrooms regarding the factor of orientation.

Furthermore, it is important to note that the data used in the generalisability analyses come from a research study where the participation of the teachers was voluntary and the observer did not have any relationship with the teachers in the study. Given that observation scores may inform decisions about teachers’ hiring, retention, bonus and dismissal (Master, 2014; Mihaly & McCaffrey, 2014), it is imperative for countries to engage in this kind of research in order to investigate whether the criteria used to measure teaching skills in their teacher evaluation systems are more or less sensitive to the classroom context by using real-life teacher evaluation data.

The present study was also designed to allow the investigation of the effect of the classroom-level factors of the dynamic model on student achievement in mathematics of secondary school students. However, due to the fact that the main focus of the study was the investigation of the consistency of teacher in-class behaviour in different classrooms within a year, only the short-term effects of the classroom-level factors on student achievement could be examined. Further research is needed to investigate the long-term impact of the factors on learning outcomes of secondary school students. It is important to note that a recent study which was conducted in primary schools in Cyprus has shown that the long-term effect of the classroom-level factors on student achievement was stronger than the short-term effect (see Dimosthenous et al., 2018).

Moreover, the current study examined the effect of the classroom-level factors of the dynamic model only on students’ cognitive outcomes in mathematics. More studies need to be carried out in order to investigate the impact of the classroom-level factors in other cognitive (e.g., language and science) and non-cognitive (e.g., affective and psychomotor) outcomes (Knuver & Brandsma, 1993; Stankov, Morony & Lee, 2014), as well as in meta-cognition (Boström & Lassen, 2006; Kuyper, Van der Werf & Lubbers, 2000; Kyriakides, Anthimou, & Charalambous, 2016). For instance, previous studies have shown that teachers usually have more impact in the area of mathematics than in the area of language (Scheerens, 2016). Therefore, if the subject of language was chosen for this study instead of mathematics, different results may have emerged regarding the effect of the classroom-level factors on student achievement of secondary school students. Different results may have also emerged regarding the level which explains more variance in student achievement (i.e., teacher level or classroom level).

Finally, given that no information was collected about the socio-economic status of the students, this study was not able to address issues associated with equity (see Kyriakides et al., 2018b). More studies are needed to examine whether the classroom-level factors of the dynamic model have differential effects on learning outcomes of secondary school students coming from different socio-economic backgrounds.

References

Adams, R. J., & Khoo, S. T. (1996). *Quest: The interactive test analysis system*. Camberwell, Vic : Australian Council for Educational Research.

Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education*, *13*(2), 153-166.

Allen, K. P. (2010). Classroom management, bullying, and teacher practices. *Professional Educator*, *34*(1), 1-15.

Andrejko, L. (2004). Value-added assessment: A view from a practitioner. *Journal of Educational and Behavioral Statistics*, *29*(1), 7-9.

Andrich, D. (1988). A general form of Rasch's extended logistic model for partial credit scoring. *Applied Measurement in Education, 1*(4), 363–378.

Antoniou, P. (2013). A longitudinal study investigating relations between stages of effective teaching, teaching experience and teacher professional development approaches. *Journal of Classroom Interaction, 48*(2), 25-40.

Antoniou, P., & Kyriakides, L. (2011). The impact of a dynamic approach to professional development on teacher instruction and student learning: Results from an experimental study. *School Effectiveness and School Improvement*, *22*(3), 291-311.

Antoniou, P., Kyriakides, L., & Creemers, B. P. M. (2011). Investigating the effectiveness of a dynamic integrated approach to teacher professional development. *Center for Educational Policy Studies Journal, 1*(1), 13-42.

Aparicio, J. J., & Moneo, M. R. (2005). Constructivism, the so-called semantic learning theories, and situated cognition versus the psychological learning theories. *Spanish Journal of Psychology, 8*(2), 180-198.

Archer, J., Kerr, K., & Pianta, R. (2014). Why measure effective teaching. In T. Kane, A. Kerri & R. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the measures of effective teaching project* (pp. 1-5). San Francisco, CA: John Wiley.

Askew, M., & William, D. (1995). *Recent research in mathematics education* 5–16. London, England: HMSO.

Azigwe, J. B., Kyriakides, L., Panayiotou, A., & Creemers, B. P. M. (2016). The impact of effective teaching characteristics in promoting student achievement in Ghana. *International Journal of Educational Development*, *51*, 51-61.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., … Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers* (EPI Briefing Paper 278). Washington, DC: Economic Policy Institute.

Baron, M. A. (1990). Who's to blame for misbehavior in our school?. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, *63*(7), 333-334.

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment*, *17*(2-3), 62-87.

Bill & Melinda Gates Foundation. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project's three-year study*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from ERIC database. ERIC Document ED540958.

Blatchford, P., Bassett, P., & Brown, P. (2011). Examining the effect of class size on classroom engagement and teacher–pupil interaction: Differences in relation to pupil prior attainment and primary vs. secondary schools. *Learning and Instruction*, *21*(6), 715-730.

Block, J. H., & Burns, R. B. (1976). Mastery learning. *Review of Research in Education*, *4*(1), 3-49.

Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how?. *CBE-Life Sciences Education*, *15*(4), 1-7. https://doi.org/10.1187/cbe.16-04-0148

Borich, G. D. (1992) *Effective teaching methods* (2nd ed.). New York, NY: Macmillan.

Borich, G. D. (2007). *Effective teaching methods: Research-based practice*. Upper Saddle River, N.J: Pearson Merrill/Prentice Hall.

Bosker, R. J., & Scheerens, J. (1994). Alternative models of school effectiveness put to the test. *International Journal of Educational Research*, *21*(2), 159-180.

Boström, L., & Lassen, L. M. (2006). Unraveling learning, learning styles, learning strategies and meta-cognition. *Education + Training, 48*(2/3), 178-189.

Brandt, C., Thomas, J., & Burke, M. (2008). *State policies on teacher evaluation practices in the Midwest Region* (REL Technical Brief, REL 2008- No. 004). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Midwest. Retrieved from httm://ies.ed.gov/ncee/edlabs.

Braun, H. I. (2005). *Using student progress to evaluate teachers: A primer on value-added models.* Princeton, NJ: Educational Testing Service.

Brophy, J. (1986). Teaching and learning mathematics: Where research should be going. *Journal for Research in Mathematics Education, 17(*5), 323–346.

Brophy, J., & Good, T. L. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 328–375). New York, NY: Macmillan.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models*. Newbury Park, CA: Sage.

Cai, J., & Lester, F. (2007). Contributions from cross-national comparative studies to the internationalization of mathematics education: Studies of Chinese and U.S. classrooms. In B. Atweh, A. C. Barton, M. C. Borda, N. Gough, C. Keitel-Kreidt, C. Vistro-Yu & R, Vithal (Eds.), *Internationalisation and globalisation in mathematics and science education* (pp. 269-283). Dordrecth, The Netherlands: Springer.

Campbell, R. J., Kyriakides, L., Muijs, R., D., & Robinson, W. (2004). *Assessing teacher effectiveness: Developing a differentiated model.* London, England: RoutledgeFalmer.

Caro, D. H., Lenkeit, J., & Kyriakides, L. (2016). Teaching strategies and differential effectiveness across learning contexts: Evidence from PISA 2012. *Studies in Educational Evaluation*, *49*, 30-41.

Carroll, J. B. (1963). A model of school learning. *Teachers College Record*, *64*(8), 723-733.

Carroll, J. B. (1989). The Carroll model: A 25-year retrospective and prospective view. *Educational Researcher*, *18*(1), 26-31.

Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., Hamre, B. K., & Pianta, R. C. (2013). Effect of observation mode on measures of secondary mathematics teaching. *Educational and Psychological Measurement*, *73*(5), 757-783.

Cazden, C. B. (1986). Classroom discourse. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 432–463). New York, NY: Macmillan.

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* (REL 2014–024). 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.

Charalambous, C. Y., & Praetorius, A. K. (2018). Studying mathematics instruction through different lenses: Setting the ground for understanding instructional quality more comprehensively. *ZDM-Mathematics Education*, *50*(3), 355-366.

Charalambous, C. Y., Kyriakides, E., Tsangaridou, N., & Kyriakides, L. (2017). Exploring the reliability of generic and content-specific instructional aspects in physical education lessons. *School Effectiveness and School Improvement*, *28*(4), 555-577.

Charlton, T., & David, K. (1993). Ensuring schools are fit for the future. In T. Charlton & D. Kenneth (Eds.), *Managing misbehavior in schools* (pp. 3-16). London, England: Routledge.

Charlton, T., & George, J. (1993). The development of behavior problems. In T. Charlton & D. Kenneth (Eds.), *Managing misbehavior in schools* (pp. 17-52). London, England: Routledge.

Cheng, Y. C., & Tsui, K. T. (1999). Multimodels of teacher effectiveness: Implications for research. *The Journal of Educational Research, 92*(3), 141-150.

Chester, M. D., & Zelman S. T. (2009). Approximations of teacher quality and effectiveness: View from the state education agency. In D. H. Gitomer (Ed.), *Measurement issues and assessment for teaching quality* (pp.131-149). Thousand Oaks, CA: Sage.

Clare, L., Valdés, R., Pascal, J., & Steinberg, J. R. (2001). Teachers' assignments as indicators of instructional quality in elementary schools (CSE technical report No.545). Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing, University of California. Retrieved from ERIC database. ERIC Document ED457169.

Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher, 45*(6), 378-387.

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education* (6th ed.). London, England: Routledge.

Cooper, P., Smith, C., & Upton, G. (1994). *Emotional and behavioural difficulties: Theory to practice.* London, England: Routledge.

Cornelius-White, J. (2007). Learner-centered teacher-student relationships are effective: A meta-analysis. *Review of Educational Research*, *77*(1), 113-143.

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

Creemers, B. P. M. (2006). Combining different ways of learning and teaching in a dynamic model of educational effectiveness. *Journal of Basic Education, 15*(2), 1-38.

Creemers, B. P. M., & Kyriakides, L. (2008). *The dynamics of educational effectiveness: A contribution to policy, practice and theory in contemporary schools*. London, England: Routledge.

Creemers, B. P. M., & Kyriakides, L. (2012). *Improving quality in education: Dynamic approaches to school improvement.* London, England: Routledge.

Creemers, B. P. M., & Reezigt, G. J. (1996). School level conditions affecting the effectiveness of instruction. *School Effectiveness and School Improvement*, *7*(3), 197-228.

Creemers, B. P. M., Kyriakides, L., & Antoniou, P. (2013). *Teacher professional development for improving quality in teaching.* Dordrecht, the Netherlands: Springer.

Creemers, B. P., & Kyriakides, L. (2006). Critical analysis of the current approaches to modelling educational effectiveness: The importance of establishing a dynamic model. *School Effectiveness and School Improvement*, *17*(3), 347-366.

Creemers, B., & Kyriakides, L. (2015a). Developing, testing, and using theoretical models for promoting quality in education. *School Effectiveness and School Improvement, 26* (1), 102-119. doi: 10.1080/09243453.2013.869233

Creemers, B., & Kyriakides, L. (2015b). Process-product research: A cornerstone in educational effectiveness research. *Journal of Classroom Interaction, 50(2),*107-119.

Curby, T. W., Stuhlman, M., Grimm, K., Mashburn, A., Chomat-Mooney, L., Downer, J., ...Pianta, R. C. (2011). Within-day variability in the quality of classroom interactions during third and fifth grade: Implications for children’s experiences and conducting classroom observations. *The Elementary School Journal*, *112*(1), 16-37.

Danielson, C. (1996). *Enhancing professional practice: A framework for teaching*. Alexandria, VA: Association for Supervision and Curriculum Development.

Danielson, C., & McGreal, T. L. (2000). *Teacher evaluation to enhance professional practice*. Alexandria, VA: ASCD.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives, 8*(1), 1-44. http://dx.doi.org/10.14507/epaa.v8n1.2000.

Darling-Hammond, L. (2007). Recognizing and enhancing teacher effectiveness: A policymaker's guide. In L. Darling-Hammond & C. D. Prince (Eds.), *Strengthening teacher quality in high-need schools-policy and practice* (pp. 1-26). Washington, DC: The Council of Chief State School Officers.

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

De Jong, R., & Westerhof, K. J. (2001). The quality of student ratings of teacher behaviour. *Learning Environments Research*, *4*(1), 51-85.

De Jong, R., Westerhof, K. J., & Kruiter, J. H. (2004). Empirical evidence of a comprehensive model of school effectiveness: A multilevel study in mathematics in the 1st year of junior general education in the Netherlands. *School Effectiveness and School Improvement, 15*(1), 3 – 31.

Den Brok, P., Brekelmans, M., & Wubbels, T. (2004). Interpersonal teacher behaviour and student outcomes. *School Effectiveness and School Improvement*, *15*(3-4), 407-442.

Dimosthenous, A., Kyriakides, L., & Panayiotou, A. (2018). *Short- and long- term effects of the home learning environment and teachers on student achievement in mathematics: A longitudinal study*. Manuscript submitted for publication.

Doherty, K. M., & Jacobs, S. (2013). *State of the States 2013: Connect the dots: Using evaluations of teacher effectiveness to inform policy and practice.* Retrieved from https://www.nctq.org/dmsView/State_of_the_States_2013_Using_Teacher_Evaluations_NCTQ_Report

Doherty, K. M., & Jacobs, S. (2015). *State of the States 2015: Evaluating teaching, leading and learning*. Washington, DC: National Council on Teacher Quality. Retrieved from ERIC database. ERIC Document ED581451.

Douglas, K. (2009). Sharpening our focus in measuring classroom instruction. *Educational Researcher*, *38*(7), 518-521.

Doyle, W. (1986). Classroom organization and management. In M.C. Wittrock (Ed.), *Handbook of Research on Teaching* (3rd ed., pp. 392–431). New York, NY: Macmillan.

Doyle, W. (1990). Classroom knowledge as a foundation for teaching. *Teachers College Record*, *91*(3), 347-360.

Duffy, T. M., & Cunningham, D. J. (1996). Constructivism: Implications for the design and delivery of instruction. In D. J. Jonassen (Ed.), *Handbook of research for educational communication and technology* (pp. 170–198). New York, NY: McMillan.

Ellett, C. D. (1997). Classroom-based assessments of teaching and learning. In J. H. Stonge (Ed.), *Evaluating teaching: A guide to current thinking and best practice* (pp.107-128). Thousand Oaks, CA: Corwin Press.

Ellett, C. D., & Garland, J. S. (1987). Teacher evaluation practices in our largest school districts: Are they measuring up to 'state-of-the-art'systems?. *Journal of Personnel Evaluation in Education*, *1*(1), 69-92.

Ellett, C. D., Wren, C. T., Callendar, K. E., Loup, K. S., & Liu, X. (1996). Looking backwards with the personnel evaluation standards: An analysis of the development and implementation of a statewide teacher assessment system. *Studies in Educational Evaluation*, *22*(1), 79-113.

English, D., Burniske, J., Meibaum, D., & Lachlan-Haché, L. (2016). *Using student surveys as a measure of teaching effectiveness*. Retrieved from https://education.ohio.gov/getattachment/Topics/Teaching/EducatorEvaluation-

System/Ohio-s-Teacher-Evaluation-System/AlternativeComponents/Student-Surveys-in-Teacher-Evaluation-FINAL-for-Ohio-060816.pdf.aspx

Eurydice. (2004). *Structures of educational, vocational training and adult education systems in Europe: Cyprus.* Retrieved from http://www.refernet.org.cy/images/media/assetfile/Eurydice.Educ&Train&Adult.Educ.03&04.EN.pdf

Eurydice. (2008). *Levels of autonomy and responsibilities of teachers in Europe.* Brussels, Belgium: Eurydice.

Evertson, C. M. (1995).*Classroom organization and management program: Revalidation submission to the program effectiveness panel.* Nashville: TN: U.S. Department of Education. Retrieved from ERIC database. ERIC Document ED403247.

Evertson, C. M., Anderson, C. W., Anderson, L. M., & Brophy, J. E. (1980). Relationships between classroom behaviors and student outcomes in junior high mathematics and English classes. *American Educational Research Journal, 17*(1), 43-60.

Flanders, N. A. (1970). *Analyzing teaching behavior*. Reading, MA: Addison-Wesley.

Flores, M. A. (2012). The implementation of a new policy on teacher appraisal in Portugal: How do teachers experience it at school?. *Educational Assessment Evaluation and Accountability, 24*(4), 351–368.

Florida Department of Education. (n.d.). *Educator quality update on teacher evaluation.* Retrieved from http://www.fldoe.org/profdev/pdf/OverviewFloridasTeacherEvaluationSystem.pdf

Fontana, D. (1994). *Managing classroom behavior* (2nd ed.). Leicester, England: The British Psychological Society.

Freiberg, H. J., & Lapointe, J. M. (2011). Research-based problems for preventing and solving discipline problems. In C. M. Evertson & C. S. Weinstein (Eds.), *Handbook of* *classroom management: Research, practice, and contemporary issues* (pp. 735-786). New York, NY: Routledge.

Gitomer, D. H., & Bell, C. A. (2013). Evaluating teachers and teaching. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology* (Vol. 3, pp. 415–444). Washington, DC: American Psychological Association.

Gitomer, D., Bell, C., Qi, Y., McCaffrey, D., Hamre, B. K., & Pianta, R. C. (2014). The instructional challenge in improving teaching quality: Lessons from a classroom observation protocol. *Teachers College Record*, *116*(6), 1-32.

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

Goe, L., & Croft, A. (2009). *Methods of evaluating teacher effectiveness.* Washington, DC: Comprehensive Center for Teacher Quality. Retrieved from ERIC database. ERIC Document ED543666.

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis.* Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved from ERIC database. ERIC Document ED521228.

Goldstein, H. (1999). *Multilevel statistical models.* Retrieved from http://www.bristol.ac.uk/media-library/sites/cmm/migrated/documents/multbook1995.pdf

Government of Cyprus. (2018). *Europe 2020: Cyprus national reform programme 2018.* Retrieved from https://ec.europa.eu/info/sites/info/files/2018-european-semester-national-reform-programme-cyprus-en.pdf

Griffin, G. A., & Barnes, S. (1986). Using research findings to change school and classroom practices: Results of an experimental study. *American Educational Research Journal*, *23*(4), 572-586.

Grossman, P., Cohen, J., & Brown, L. (2014).Understanding instructional quality in English language arts: Variations in PLATO scores by context and context. In T. Kane, A. Kerri & R. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the measures of effective teaching project* (pp. 303-331). San Francisco, CA: John Wiley.

Gustafsson, H. C., Barnett, M. A., Towe-Goodman, N. R., Mills-Koonce, W. R., Cox, M. J., & Family Life Project Key Investigators. (2014). Family violence and children's behavior problems: Independent contributions of intimate partner and child-directed physical aggression. *Journal of Family Violence*, *29*(7), 773-781.

Gustafsson, J. E. (2010). Longitudinal designs. In. B. P. M. Creemers, L. Kyriakides & P. Sammons (Eds). *Methodological advances in educational effectiveness research* (pp. 77-101). London, England: Routledge.

Hamilton, L. S. (2012). Measuring teaching quality using student achievement tests. In S. Kelly (Ed.), *Assessing teacher quality: Understanding teacher effects on instruction and achievement* (pp.49-75). New York, NY: Teachers College Press.

Harjunen, E. (2012). Patterns of control over the teaching–studying–learning process and classrooms as complex dynamic environments: A theoretical framework. *European Journal of Teacher Education*, *35*(2), 139-161.

Harmer, J. (2001). *The practice of English language teaching* (3rd ed.). Essex, England: Longman.

Harris, D. N., Ingle, W. K., & Rutledge, S. A. (2014). How teacher evaluation methods matter for accountability: A comparative analysis of teacher effectiveness ratings by principals and teacher value-added measures. *American Educational Research Journal*, *51*(1), 73-112.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* London, England: Routledge.

Heck, R. H., & Moriyama, K. (2010). Examining relationships among elementary schools' contexts, leadership, instructional practices, and added-year outcomes: A regression discontinuity approach. *School Effectiveness and School Improvement, 21*(4), 377–408.

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

Hiebert, J., Gallimore, R., Garnier, H., Givvin, K. B., Hollingsworth, H., Jacobs, J., … Stigler, J. (2003). *Teaching mathematics in seven countries: Results from the TIMSS 1999 video study*. Washington, DC: National Center for Education Statistics.

Hill, H. C., Charalambous, C. Y., & Kraft, M. (2012). When rater reliability is not enough: Teacher observation systems and a case for the G-study. *Educational Researcher, 41* (2), 56-64.

Hill, H.C., & Herlihy, C. (2011). Prioritizing teaching quality in a new system of teacher evaluation. *Education Outlook*, *9,* 1-6.

Ho, I. T. (2004). A comparison of Australian and Chinese teachers' attributions for student problem behaviors. *Educational Psychology*, *24*(3), 375-391.

Houghton, S., Wheldall, K., & Merrett , F. (1988). Classroom behaviour problems which secondary school teachers say they find most troublesome. *British Educational Research Journal*, *14*(3), 297-312.

Ingvarson, L., Kleinhenz, E., & Wilkinson, J. (2007). *Research on performance pay for teachers.* Retrieved from http://research.acer.edu.au/workforce/1

Isoré, M. (2009). Teacher evaluation: Current practices in OECD countries and a literature review. *OECD Education Working Papers,* (23). http://dx.doi.org/10.1787/223283631428

Iwanicki, E. F. (1990). Teacher evaluation for school improvement. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook for teacher evaluation: Assessing* *elementary and secondary school teachers* (pp. 158-171). Newbury Park, CA: Sage Publication.

Joe, J. N., McClellan, C. A., & Holtzman, S. L. (2014). Scoring design decisions: Reliability and the length and focus of classroom observations. In T. Kane, A. Kerri & R. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the measures of effective teaching project* (pp. 415-443). San Francisco, CA: John Wiley.

Kalogrides, D., & Loeb, S. (2013). Different teachers, different peers: The magnitude of student sorting within schools. *Educational Researcher*, *42*(6), 304-316.

Kalogrides, D., Loeb, S., & Béteille, T. (2013). Systematic sorting: Teacher characteristics and class assignments. *Sociology of Education*, *86*(2), 103-123.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains.* Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from ERIC database. ERIC Document ED540960.

Keeves, J. P., & Alagumalai, S. (1999). New approaches to measurement. In G. N. Masters & J. P. Keeves (Eds.), *Advances in measurement in educational research and assessment* (pp. 23-42). Oxford, England: Pergamon.

Kellam, S. G., Ling, X., Merisca, R., Brown, C. H., & Ialongo, N. (1998). The effect of the level of aggression in the first grade classroom on the course and malleability of aggressive behavior into middle school. *Development and Psychopathology*, *10*(2), 165-185.

Kelly, K. O., Ang, S. Y. A., Chong, W. L., & Hu, W. S (2008). Teacher appraisal and its outcomes in Singapore primary schools. *Journal of Educational Administration*, *46*(1), 39-54.

Kelly, S. (2012). Understanding teacher effects: Market versus process models of educational improvement. In S. Kelly (Ed.), *Assessing teacher quality: Understanding* *teacher effects on instruction and achievement* (pp.7-32). New York, NY: Teachers College Press.

Knuver, A. W., & Brandsma, H. P. (1993). Cognitive and affective outcomes in school effectiveness research. *School Effectiveness and School Improvement, 4*(3), 189-204.

Ko, J., Hallinger, P., & Walker, A. (2015). Exploring whole school versus subject department improvement in Hong Kong secondary schools. *School Effectiveness and School Improvement, 26* (2), 215-239. doi: 10.1080/09243453.2014.882848

Ko, J., Sammons, P., & Bakkum, L. (2016). *Effective teaching*. Reading, Berkshire: Education Development Trust.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.).New York, NY: Springer.

Koppich, J. (2008). *Towards a more comprehensive model of teacher pay*. Retrieved from https://my.vanderbilt.edu/performanceincentives/ncpi-publications/design-and-implementation-of-incentive-pay-systems/toward-a-more-comprehensive-model-of-teacher-pay/

Koutrouba, K. (2013). Student misbehaviour in secondary education: Greek teachers' views and attitudes. *Educational Review, 65*(1), 1-19.

Kuyper, H., Van der Werf, M. P. C., & Lubbers, M. J. (2000). Motivation, meta-cognition and self-regulation as predictors of long term educational attainment. *Educational Research and Evaluation, 6*(3), 181-205.

Kyriacou, C. (2009). *Effective teaching in schools: Theory and practice* (3rd ed.). Cheltenham, England: Nelson Thornes.

Kyriacou, C., & Martín, J. L. O. (2010). Beginning secondary school teachers' perceptions of pupil misbehaviour in Spain. *Teacher Development: An International Journal of Teachers' Professional Development, 14*(4), 415-426.

Kyriakides, L. (1998). Οι αντιλήψεις των δασκάλων και των γονιών για την πειθαρχία στο δημοτικό σχολείο [Parents' and teachers' perceptions of school discipline]. *Παιδαγωγική Επιθεώρηση, 27*, 203-222.

Kyriakides, L. (2005). Drawing from teacher effectivess research and research into teacher interpersonal behaviour to establish a teacher evaluation system: A study on the use of student ratings to evaluate teacher behaviour. *The Journal of Classroom Interaction*, *40*(2), 44-66.

Kyriakides, L. (2008). Testing the validity of the comprehensive model of educational effectiveness: A step towards the development of a dynamic model of effectiveness. *School Effectiveness and School Improvement*, *19*(4), 429-446.

Kyriakides, L. (2016). *An independent view from an academic expert* [PowerPoint slides]. Retrieved from https://ec.europa.eu/cyprus/sites/cyprus/files/ec_2016_education_and_training_monitor_l.kyriakides_2.pptx

Kyriakides, L., & Campbell, R. J. (2003). Teacher evaluation in Cyprus: Some conceptual and methodological issues arising from teacher and school effectiveness research. *Journal of Personnel Evaluation in Education*, *17*(1), 21-40.

Kyriakides, L., & Creemers, B. P. M. (2008). Using a multidimensional approach to measure the impact of classroom-level factors upon student achievement: A study testing the validity of the dynamic model. *School Effectiveness and School Improvement, 19*(2), 183–205.

Kyriakides, L., & Creemers, B. P. M. (2009). The effects of teacher factors on different outcomes: Two studies testing the validity of the dynamic model. *Effective Education, 1*(1), 61–86.

Kyriakides, L., & Demetriou, D. (2007). Introducing a teacher evaluation system based on teacher effectiveness research: An investigation of stakeholders' perceptions. *Journal of Personnel Evaluation in Education*, *20*(1-2), 43-64.

Kyriakides, L., Anthimou, M., & Charalambous, C. Y. (2016, April). *Searching for the impact of teacher behavior on promoting students' cognitive and metacognitive skills.* Paper presented at the American Educational Research Association (AERA) Conference 2016, Washington, DC.

Kyriakides, L., Archambault, I., & Janosz, M. (2013). Searching for stages of effective teaching: A study testing the validity of the dynamic model in Canada. *Journal of Classroom Interaction*, *48*(2), 11-24.

Kyriakides, L., Campbell, R. J., & Christofidou, E. (2002). Generating criteria for measuring teacher effectiveness through a self-evaluation approach: A complementary way of measuring teacher effectiveness. *School Effectiveness and School Improvement*, *13*(3), 291-325.

Kyriakides, L., Christoforidou, M., Panayiotou, A., & Creemers, B. P. M. (2017). The impact of a three-year teacher professional development course on quality of teaching: Strengths and limitations of the dynamic approach. *European Journal of Teacher Education*, *40*(4), 465-486.

Kyriakides, L., Christoforou, C., & Charalambous, C. Y. (2013). What matters for student learning outcomes: A meta-analysis of studies exploring factors of effective teaching. *Teaching and Teacher Education*, *36*(0), 143-152. http://dx.doi.org/10.1016/j.tate.2013.07.010

Kyriakides, L., Creemers, B. P. M., & Antoniou, P. (2009). Teacher behaviour and student outcomes: Suggestions for research on teacher training and professional development. *Teaching and Teacher Education, 25*(1), 12–23.

Kyriakides, L., Creemers, B. P. M., Panayiotou, A., Vanlaar, G., Pfeifer, M., Gašper, C., & McMahon, L. (2014). Using student ratings to measure quality of teaching in six European countries. *European Journal of Teacher Education, 37*(2), 125-143.

Kyriakides, L., Creemers, B. P., & Panayiotou, A. (2018a). Using educational effectiveness research to promote quality of teaching: The contribution of the dynamic model. *ZDM-Mathematics Education, 50*(3), 381-393.

Kyriakides, L., Creemers, B., & Charalambous, E. (2018b). *Equity and quality dimensions in educational effectiveness*. Dordrecht, the Netherlands: Springer.

Kyriakides, L., Creemers, B.P.M., & Panayiotou, A. (2012). *Report of the data analysis of the student questionnaire used to measure teacher factors: Across and within country results* (ESF project: Establishing a knowledge base for quality in education: testing a dynamic theory for education 08-ECRP-012). Nicosia, Cyprus: University of Cyprus.

Kyriakides, L., Demetriou, D., & Charalambous, C. (2006). Generating criteria for evaluating teachers through teacher effectiveness research. *Educational Research, 48*(1), 1-20.

Landson-Billings, G. (2009). Opportunity to teach: Teacher quality in context. In D. H. Gitomer (Ed.), *Measurement issues and assessment for teaching quality* (pp.206-222). Thousand Oaks, CA: Sage.

Lassen, S. R., Steele, M. M., & Sailor, W. (2006). The relationship of school-wide positive behavior support to academic achievement in an urban middle school. *Psychology in the Schools*, *43*(6), 701–712.

Lazarev, V., & Newman, D. (2015). *How teacher evaluation is affected by class characteristics: Are observations biased?.* Palo Alto, CA: Empirical Education Inc. Retrieved from ERIC database. ERIC Document ED558567.

Lewin, T. (2010, July 23). School chief dismisses 241 teachers in Washington. *The New York Times.* Retrieved from https://www.nytimes.com/2010/07/24/education/24teachers.html

Liang, G. (2013). Teacher evaluation and value-added: Do different models give us the same answer. *Journal of Postdoctoral Research, 1*(5), 42-43.

Little, E. (2005).Secondary school teachers' perceptions of students' problem behaviours. *Educational Psychology, 25,* 369–77.

Liu, S., & Zhao, D. (2013). Teacher evaluation in China: Latest trends and future directions. *Educational Assessment, Evaluation and Accountability, 25*(3), 231–250.

Loup, K. S., Garland, J. S., Ellett, C. D., & Rugutt, J. K. (1996). Ten years later: Findings from a replication of a study of teacher evaluation practices in our 100 largest school districts. *Journal of Personnel Evaluation in Education*, *10*(3), 203-226.

Luyten, H. & Sammons, P. (2010). Multilevel Modelling, In B. P. M. Creemers, L. Kyriakides & P. Sammons (Eds), *Methodological advances in educational effectiveness research* (pp. 246-276). London, England: Routledge.

Marciniak, A. (2015). Effective ways of dealing with discipline problems when teaching adolescent learners. *World Scientific News*, (1), 53-72.

Marcoulides, G. A., & Kyriakides, L. (2010). Using generalizability theory. In B. P. M. Creemers, L. Kyriakides & P. Sammons (Eds.), *Methodological advances in educational effectiveness research* (pp. 219- 245). London, England: Routledge.

Marzano, R., Marzano, J. S., & Pickering, D. J. (2003). *Classroom management that works: Research-based strategies for every teacher*. Alexandria, VA : ASCD.

Master, B. (2014). Staffing for success: Linking teacher evaluation and school personnel management in practice. *Educational Evaluation and Policy Analysis*, *36*(2), 207-227.

Matsumura, L. C., Garnier, H., Pascal, J., & Valdés, R. (2002). Measuring instructional quality in accountability systems: Classroom assignments and student achievement. *Educational Assessment*, *8*(3), 207-229.

Mayer, D. P. (1999). Measuring instructional practice: Can policymakers trust survey data?. *Educational Evaluation and Policy Analysis*, *21*(1), 29-45.

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

McManus, M. (1989). *Troublesome behaviour in the classroom: A teachers' survival guide.* London, England: Routledge.

MET Project. (2012). *Asking students about teaching: Student perception surveys and their implementation.* Retrieved from http://k12education.gatesfoundation.org/resource/asking-students-about-teaching-student-perception-surveys-and-their-implementation/

Mihaly, K., & McCaffrey, D. F. (2014). Grade level variation in observational measures of teacher effectiveness. In T. J. Kane, K. A. Kerr & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the measures of effective teaching project* (pp. 9–49). San Francisco, CA: Jossey-Bass.

Milanowski, A. (2002). *The varieties of knowledge and skill-based pay design: A comparison of seven new pay systems for K-12 teachers* (CPRE Research Reports RR-050). Retrieved from https://repository.upenn.edu/cpre_researchreports/29

Ministry of Education and Culture. (2018, May 29). *Έναρξη διαλόγου για εκσυγχρονισμό του συστήματος αξιολόγησης των εκπαιδευτικών και του εκπαιδευτικού έργου* [Initiation of a dialogue for the modernization of the teacher evaluation system and educational work]. Retrieved from http://enimerosi.moec.gov.cy/archeia/1/ypp7598a

Ministry of Education and Culture. (n.d.). *A guide to education in Cyprus*. Retrieved from http://www.moec.gov.cy/odigos-ekpaidefsis/documents/english.pdf

Montgomery, D. (1989). *Managing behaviour problems*. London, England: Hodder & Stoughton.

Muijs, D. (2006). Measuring teacher effectiveness: Some methodological reflections. *Educational Research and Evaluation, 12*(1), 53-74.

Muijs, D., & Reynolds, D. (2000). School effectiveness and teacher effectiveness in mathematics: Some preliminary findings from the evaluation of the mathematics enhancement programme (primary). *School Effectiveness and School Improvement*, *11*(3), 273-303.

Muijs, D., & Reynolds, D. (2001). *Effective teaching: Evidence and practice.* London, England: Sage.

Muijs, D., & Reynolds, D. (2011). *Effective teaching: Evidence and practice*. (3rd ed.). London, England: Sage.

Muijs, R. D., Kyriakides, L., van der Werf, G., Creemers, B.P.M., Timperley, H., & Earl, L. (2014). State of the art-teacher effectiveness and professional learning. *School Effectiveness and School Improvement, 25*(2), 231-256.

Muscott, H. S., Mann, E. L., & LeBrun, M. R. (2008). Positive behavioral interventions and supports in New Hampshire: Effects of large-scale implementation of schoolwide positive behavior support and student discipline and academic achievement. *Journal of Positive Behavior Interventions, 10*(3), 190–205.

National Institute of Mental Health. (2016). *Attention deficit hyperactivity disorder*. Retrieved from https://www.nimh.nih.gov/health/topics/attention-deficit-hyperactivity-disorder-adhd/index.shtml

Newton, X. A. (2010). Developing indicators of classroom practice to evaluate the impact of district mathematics reform initiative: A generalizability analysis. *Studies in Educational Evaluation*, *36*(1-2), 1-13.

Ntoliopoulou, E. (2015). Οι επιπτώσεις της φτώχειας γενικότερα και στα παιδιά ειδικότερα και πιθανοί τρόποι παρεμβάσεων για την πρόληψη και τη μείωσή της [The impacts of poverty in general and on children in particular and possible ways of interventions for its prevention and reduction]. *Έρευνα στην Εκπαίδευση*, *3*, 97-123. http://dx.doi.org/10.12681/hjre.8849

Odden, A., Kelley, C., Heneman, H., & Milanowski, A. (2001*). Enhancing teacher quality through knowledge- and skills-based pay* (CPRE Policy briefs RB-34). Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1007.7214&rep=rep1&type=pdf

OECD. (2005). *Teachers matter: Attracting, developing and retaining effective teachers.* Retrieved from https://www.oecd.org/education/school/34990905.pdf

OECD. (2009). *Creating effective teaching and learning environments: First results from TALIS*. Retrieved from https://www.oecd.org/education/school/43023606.pdf

OECD. (2013). *Teachers for the 21st century: Using evaluation to improve teaching.* Retrieved from http://www.oecd.org/site/eduistp13/TS2013%20Background%20Report.pdf

OECD. (2018). *PISA 2015: Results in focus.* Retrieved from https://www.oecd.org/pisa/pisa-2015-results-in-focus.pdf

Pacheco, A. (2009). Mapping the terrain of teacher quality. In D. H. Gitomer (Ed.), *Measurement issues and assessment for teaching quality* (pp.160-178). Thousand Oaks, CA: Sage.

Palardy, G. J., & Rumberger, R. W. (2008). Teacher effectiveness in first grade: The importance of background qualifications, attitudes, and instructional practices for student learning. *Educational Evaluation and Policy Analysis*, *30*(2), 111-140.

Panayiotou, A., Kyriakides, L., Creemers, B. P. M., McMahon, L., Vanlaar, G., Pfeifer, M., Rekalidou, G., & Bren, M. (2014). Teacher behavior and student outcomes: Results of a European study. *Educational Assessment, Evaluation and Accountability*, *26*(1), 73-93.

Patrick, H., & Mantzicopoulos, P. (2016). Is effective teaching stable?. *The Journal of Experimental Education*, *84*(1), 23-47.

Peterson, K.D. (2000). *Teacher Evaluation: A comprehensive guide to new directions and practice.* Thousand Oaks, CA: Corwin Press.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38*(2), 109-119.

Poursanidou, E. (2016). Προβλήματα συμπεριφοράς στην τάξη και παρέμβαση του δασκάλου [Classroom behavior problems and the teacher's intervention]. *Έρευνα στην Εκπαίδευση*, *5*(1), 62-75. http://dx.doi.org/10.12681/hjre.9380

Praetorius, A. K., & Charalambous, C. Y. (2018). Classroom observation frameworks for studying instructional quality: Looking back and looking forward. *ZDM-Mathematics Education*, *50*(3), 535-366.

Praetorius, A. K., Lenske, G., & Helmke, A. (2012). Observer ratings of instructional quality: Do they fulfill what they promise?. *Learning and Instruction, 22*(6), 387-400.

Praetorius, A. K., McIntyre, N. A., & Klassen, R. M. (2017). Reactivity effects in video-based classroom research: An investigation using teacher and student questionnaires as well as teacher eye-tracking. *Zeitschrift für Erziehungswissenschaft*, *20*(1), 49-74.

Praetorius, A. K., Pauli, C., Reusser, K., Rakoczy, K. & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction, 31,* 2-12.

Public Agenda. (2004). *Teaching interrupted: Do discipline policies in today's public schools foster the common good?.* Retrieved from http://www.publicagenda.org/files/teaching_interrupted.pdf

Rasbash, J., Steele, F., Browne, W. J., & Goldstein, H. (2012). *A user's guide to MLwiN, v2.26*. Retrieved from http://www.bris.ac.uk/media-library/sites/cmm/migrated/documents/manual-web.pdf

Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice?. *Journal of Educational and Behavioral Statistics, 29*(1), 121–129.

Raver, C. C., Jones, S. M., Li-Grining, C., Zhai, F., Metzger, M. W., & Solomon, B. (2009). Targeting children's behavior problems in preschool classrooms: A cluster-randomized controlled trial. *Journal of Consulting and Clinical Psychology*, *77*(2), 302.

Reynolds, D., Sammons, P., De Fraine, B., Van Damme, J., Townsend, T., Teddlie, C., & Stringfield, S. (2014). Educational effectiveness research (EER): A state-of-the-art review. *School Effectiveness and School Improvement*, *25*(2), 197-230.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417-458.

Rockoff, J. E., & Speroni, C. (2010). Subjective and objective evaluations of teacher effectiveness. *American Economic Review*, *100*(2), 261-66.

Rosenshine, B. (1983). Teaching functions in instructional programs. *The Elementary School Journal*, *83*(4), 335-351.

Rosenshine, B., & Stevens, R. (1986). Teaching functions. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 376–391). New York, NY: Macmillan.

Rowan, B., & Correnti, R. (2009). Studying reading instruction with teacher logs: Lessons from the study of instructional improvement. *Educational Researcher*, *38*(2), 120-131.

Sammons, P. (2009). The dynamics of educational effectiveness: A contribution to policy, practice and theory in contemporary schools. *School Effectiveness and School Improvement*, *20*(1), 123–129.

Sammons, P., Thomas, S., & Mortimore, P. (1997). *Forging links: Effective schools and effective departments.* London, England: Paul Chapman.

Santiago, P., & Benavides, F. (2009). *Teacher evaluation: A conceptual framework and examples of country practices.* Retrieved from http://www.oecd.org/edu/school/44568106.pdf

Savery, J. R., & Duffy, T. M. (1995). Problem based learning: An instructional model and its constructivist framework. *Educational Technology*, *35*(5), 31-38.

Scheerens, J. (1990). School effectiveness research and the development of process indicators of school functioning. *School Effectiveness and School Improvement*, *1*(1), 61-80.

Scheerens, J. (1992). *Effective schooling: research, theory and practice*. London, England: Cassell.

Scheerens, J. (2013). The use of theory in school effectiveness research revisited. *School Effectiveness and School Improvement*, *24*(1), 1-38.

Scheerens, J. (2016). *Educational effectiveness and ineffectiveness: A critical review of the knowledge base*. Dordrecht, The Netherlands: Springer.

Scheerens, J., & Bosker, R.J. (1997). *The foundations of educational effectiveness.* Oxford, England: Pergamon.

Scheerens, J., & Creemers, B. P. (1989). Conceptualizing school effectiveness. *International Journal of Educational Research*, *13*(7), 691-706.

Schoenfeld, A.H. (1998). Toward a theory of teaching in context. *Issues in Education, 4*(1), 1–94.

Scott, J. S., White, R., Algozzine, B. & Algozzine, K. (2009). Effects of positive unified behavior support on instruction. *The International Journal on School Disaffection*, *6*(2), 41-48.

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77*(4), 454-499.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, *44*(6), 922-932.

Shepard, L. A. (1989). Why we need better assessments. *Educational Leadership, 46*(7), 4-9.

Slavin, R. E. (2010). Experimental studies in education. In. B. P. M. Creemers, L. Kyriakides & P. Sammons (Eds). *Methodological advances in educational effectiveness research* (pp. 102-114). London, England: Routledge.

Smith, K. (2005). New methods and perspectives in teacher evaluation. In D. Beijaard, P. Meijer, G. Morine-Dershimer & H. Tillema (Eds.), *Teacher professional development in changing conditions* (pp. 95 – 114). Dordrecht, the Netherlands: Springer.

Smith, M. S., & Stein, M. K. (1998). Selecting and creating mathematical tasks: From research to practice. *Mathematics Teaching in the Middle School, 3*(5), 344-350.

Smylie, M. A., Miller. C. L., & Westbrook, K. P. (2008). The work of teachers. In T. L. Good (Ed.), *21st century education: A reference handbook (Vol.1)* (pp.3-11). Thousand Oaks, CA: Sage.

Snijders, T. (2011). Multilevel analysis. In M. Lovric (Ed.), *International encyclopedia of statistical science* (pp. 879-882). Berlin, Germany: Springer.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* London, England: Sage.

Spilt, J. L., Leflot, G., Onghena, P., & Colpin, H. (2016). Use of praise and reprimands as critical ingredients of teacher behavior management: Effects on children's development in the context of a teacher-mediated classroom intervention. *Prevention Science, 17*(6), 732-742.

Stankov, L., Morony, S., & Lee, Y. P. (2014). Confidence: The best non-cognitive predictor of academic achievement?. *Educational Psychology, 34*(1), 9-28.

Stedman, J. B. & McCallion, G. (2001). *Performance-based pay for teachers* (CRS Report No. RL30217). Retrieved from http://digitalcommons.ilr.cornell.edu/key_workplace/45/

Stein, M. K., & Matsumura, L. C. (2009). Measuring instruction for teacher learning. In D. H. Gitomer (Ed.), *Measurement issues and assessment for teaching quality* (pp.179-205). Thousand Oaks, CA: Sage.

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.

Stenmark, J. K. (1991). *Mathematics assessment: Myths, models, good questions, and practical suggestions*. Reston, VA: NCTM.

Stigler, J. W., Gonzales, P., Kwanaka, T., Knoll, S., & Serrano, A. (1999). *The TIMSS videotape classroom study: Methods and findings from an exploratory research project on eighth-grade mathematics instruction in Germany, Japan, and the United States* (Report No. NCES 99-074). Washington, DC: U.S. Department of Education, National Center for Education Statistics. Retrieved from ERIC database. ERIC Document ED431621.

Stodolsky, S. (1990). Classroom observation. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook for teacher evaluation: Assessing elementary and secondary school teachers* (pp. 175-190). Newbury Park, CA: Sage Publication.

Stringfield, S. C., & Slavin, R. E. (1992). A hierarchical longitudinal model for elementary school effects. In B. P. M. Creemers & G. J. Reezigt (Eds), *Evaluation of educational effectiveness* (pp.35-69). Groningen, the Netherlands: ICO.

Stronge, J. H. (1995). Balancing individual and institutional goals in educational personnel evaluation: A conceptual framework. *Studies in Educational Evaluation, 21,* 131-151.

Stronge, J. H. (1997). Improving schools through teacher evaluation. In J. H. Stonge (Ed.), *Evaluating teaching: A guide to current thinking and best practice* (pp.1-23). Thousand Oaks, CA: Corwin Press.

Stronge, J. H. (2002). *Qualities of effective teachers.* Alexandria, VA: Association for Supervision and Curriculum Development.

Stronge, J. H. (2006). Teacher evaluation and school improvement: Improving the educational landscape. In J. H. Stronge (Ed.), *Evaluating teaching: A guide to current thinking and best practice* (2nd ed., pp. 1-23). Thousand Oaks, CA: Corwin Press.

Stronge, J. H., & Ostrander, L. P. (1997). Client surveys in teacher evaluation. In J. H. Stonge (Ed.), *Evaluating teaching: A guide to current thinking and best practice* (pp.129-161). Thousand Oaks, CA: Corwin Press.

Stronge, J. H., & Tucker, P. D. (2003). *Handbook on Teacher evaluation: Assessing and improving performance*. Larchmont, NY: Eye On Education.

Stronge, J. H., Ward, T. J., & Grant, L. W. (2011). What makes good teachers good? A cross-case analysis of the connection between teacher effectiveness and student achievement. *Journal of Teacher Education, 62*(4), 339-355.

Taylor, E. S., & Tyler, J. H. (2011). *The effect of evaluation on performance: Evidence from longitudinal student achievement data of mid-career teachers* (Report No. w16877). Cambridge, MA: National Bureau of Economic Research. Retrieved from ERIC database. ERIC Document ED517378.

The World Bank. (2014, May 22). *Teacher policies in the Republic of Cyprus*. Retrieved from http://www.paideia-news.com/content/files/58097732.pdf

Thomas, E. I. (1984). *Merit pay for teachers. ERIC Clearinghouse on Educational Management: ERIC Digest, Number Ten.* Washington, DC: National Institute of Education. Retrieved from ERIC database. ERIC Document ED259453.

Toh, K. A., Ho, B. T., Chew, C. M., & Riley, J. P. (2003). Teaching, teacher knowledge and constructivism. *Educational Research for Policy and Practice, 2*(3), 195-204.

Torff, B., & Sessions, D. (2009). Principals' perceptions of the causes of teacher ineffectiveness in different secondary subjects. *Teacher Education Quarterly*, *36*(3), 127-148.

Turner, J. C., & Meyer, D. K. (2000). Studying and understanding the instructional contexts of classrooms: Using our past to forge our future. *Educational Psychologist*, *35*(2), 69-85.

Tyler, J. H., Taylor, E. S., Kane, T. J., & Wooten, A. L. (2010). Using student performance data to identify effective classroom practices. *American Economic Review*, *100*(2), 256-260.

Walberg H.J. (1986). Syntheses of research on teaching. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 214–229). New York, NY: Macmillan.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness* (2nd ed.). Brooklyn, NY: The New Teachers Project. Retrieved from ERIC database. ERIC Document ED515656.

Wheldall, K., & Merrett, F. (1988). Which classroom behaviours do primary school teachers say they find most troublesome?. *Educational Review, 40* (1), 13-27.

Whitehurst, G. J., Chingos, M. M., & Lindquist, K. M. (2014). *Evaluating teachers with classroom observations: Lessons learned in four districts*. Washington, DC: Brown Center on Education Policy at Brookings.

Wilkerson, D. J., Manatt, R. P., Rogers, M.A. & Maughan, R. (2000). Validation of student, principal, and self-ratings in 360^0 feedback for teacher evaluation. *Journal of Personnel Evaluation in Education*, *14*(2), 179-192.

Worrell, F. C., & Kuterbach, L. D. (2001). The use of student ratings of teacher behaviors with academically talented high school students. *Journal of Secondary Gifted Education*, *12*(4), 236-247.

Wright, B. D. (1985). Additivity in psychological measurement. *Measurement and Personality Assessment*, 101-112.

Wright, B. D., & Masters, G. N. (1981). *The measurement of knowledge and attitude*. Chicago, IL: Statistical Laboratory, Department of Education, University of Chicago.

Wright, B. D., & Stone, M. (1999). *Measurement essentials* (2nd ed.). Wilmington, DE: Wide Range.

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*(1), 57–67.

APPENDIX A: Rasch Parameter Estimates of the Three Tests Scores of the Student Sample-Pilot Study

Parameter Estimates		Test 6 (Grade 6) (N=160)	Test 7 (Grade 7) (N=171)	Test 8 (Grade 8) (N=153)
Mean	Item	0.00	0.00	0.00
	Students	0.16	0.31	-0.32
SD	Item	1.21	1.61	1.24
	Students	1.37	1.30	1.32
Reliability	Item	0.97	0.96	0.93
	Students	0.87	0.91	0.90
Mean Infit mean square	Item	0.99	0.99	1.01
	Students	1.00	1.01	1.08
Mean Outfit mean square	Item	1.08	1.04	1.05
	Students	1.09	1.09	1.07
Infit t	Item	-0.10	-0.11	-0.09
	Students	0.04	0.02	0.08
Outfit t	Item	0.04	0.04	0.01
	Students	0.03	0.07	0.06

APPENDIX B: Number and Percentage of Different Groups of Students per Gender and per Grade

Number and percentage of different groups of students per gender

Variable	Boys		Girls		Total	
	n	%	*n*	%	*n*	%
Special educational needs						
Students with no special educational needs	363	43.2%	432	51.4%	795	94.6%
Students with special educational needs	25	3.0%	20	2.4%	45	5.4%
Ethnicity: language that students speak at home						
Greek	284	34.3%	322	38.8%	606	73.1%
Greek and other language	75	9.0%	100	12.1%	175	21.1%
Other language	22	2.7%	26	3.1%	48	5.8%
Ethnicity: students' country of birth						
Cyprus	336	40.5%	386	46.5%	722	87.0%
Other country	46	5.5%	62	7.5%	108	13.0%
Ethnicity: mothers' country of birth						
Cyprus	279	33.7%	317	38.2%	596	71.9%
Other country	103	12.4%	130	15.7%	233	28.1%
Ethnicity: fathers' country of birth						
Cyprus	318	38.5%	350	42.3%	668	80.8%
Other country	62	7.5%	97	11.7%	159	19.2%
Grade						
Grade 7	224	26.7%	237	28.2%	461	54,9%
Grade 8	164	19.5%	215	25.6%	379	45.1%

Number and percentage of different groups of students per grade

Variable	**Grade 7**		**Grade 8**		**Total**	
	n	*%*	*n*	*%*	*n*	*%*
Special educational needs						
Students with no special educational needs	433	51.5%	362	43.1%	795	94.6%
Students with special educational needs	28	3.3%	17	2.0%	45	5.4%
Ethnicity: language that students speak at home						
Greek	341	41.1%	265	32.0%	606	73.1%
Greek and other language	88	10.6%	87	10.5%	175	21.1%
Other language	28	3.4%	20	2.4%	48	5.8%
Ethnicity: students' country of birth						
Cyprus	397	47.8%	325	39.2%	722	87.0%
Other country	61	7.3%	47	5.7%	108	13.0%
Ethnicity: mothers' country of birth						
Cyprus	322	38.8%	274	33.1%	596	71.9%
Other country	136	16.4%	97	11.7%	233	28.1%
Ethnicity: fathers' country of birth						
Cyprus	366	44.3%	302	36.5%	668	80.8%
Other country	90	10.9%	69	8.3%	159	19.2%

APPENDIX C: Student Questionnaire

Αγαπητέ μαθητή/ αγαπητή μαθήτρια,

Διεξάγουμε μια έρευνα και θα θέλαμε να μάθουμε την άποψή σου για τη διδασκαλία **του μαθήματος των Μαθηματικών. Μη γράψεις πουθενά το όνομά σου.** Σε παρακαλούμε να απαντήσεις σε **όλες** τις ερωτήσεις.

ΜΕΡΟΣ Α

Αφού διαβάσεις προσεκτικά την κάθε πρόταση, βάλε σε κύκλο τον αριθμό:

1 : αν η κατάσταση που περιγράφεται δε συμβαίνει **ποτέ** στην τάξη σας

2 : αν η κατάσταση που περιγράφεται συμβαίνει **σπάνια** στην τάξη σας

3 : αν η κατάσταση που περιγράφεται συμβαίνει **μερικές** φορές στην τάξη σας

4 : αν η κατάσταση που περιγράφεται συμβαίνει **συχνά** στην τάξη σας

5 : αν η κατάσταση που περιγράφεται συμβαίνει **σχεδόν πάντα** στην τάξη σας

		Ποτέ	**Σπάνια**	**Μερικές φορές**	**Συχνά**	**Σχεδόν πάντα**
1.	Όταν εκτελώ μια δραστηριότητα, γνωρίζω τι προσπαθώ να πετύχω.	1	2	3	4	5
2.	Ο/Η καθηγητής/τρια βρίσκει τρόπο να μας εξηγήσει πώς συνδέονται τα καινούρια πράγματα που μαθαίνουμε με αυτά που ήδη γνωρίζουμε.	1	2	3	4	5
3.	Στην αρχή του μαθήματος των Μαθηματικών, ο/η καθηγητής/τρια συνδέει το μάθημα με προηγούμενα μαθήματα.	1	2	3	4	5
4.	Ο/Η καθηγητής/τρια των Μαθηματικών μας βοηθά να καταλάβουμε πώς οι δραστηριότητες που κάνουμε σε ένα μάθημα συνδέονται μεταξύ τους.	1	2	3	4	5
5.	Υπάρχουν στιγμές που δεν καταλαβαίνω ποια σχέση έχει μια εργασία που κάνω με την προηγούμενη εργασία που έκανα.	1	2	3	4	5

		Ποτέ	Σπάνια	Μερικές φορές	Συχνά	Σχεδόν πάντα
6.	Ζητώ περισσότερες πληροφορίες από τον/την καθηγητή/τρια των Μαθηματικών, όταν κάνουμε διαγώνισμα, γιατί δεν καταλαβαίνω όλες τις οδηγίες.	1	2	3	4	5
7.	Λίγες μέρες πριν, ο/η καθηγητής/τρια των Μαθηματικών μας δίνει ασκήσεις παρόμοιες με αυτές που θα μπουν στο διαγώνισμα.	1	2	3	4	5
8.	Όταν οι γονείς μου επισκέπτονται τον/την καθηγητή/τρια μου, τους λέει πόσο καλός/καλή είμαι, σε σχέση με τους άλλους συμμαθητές μου.	1	2	3	4	5
9.	Όταν ελέγχουμε την κατ' οίκον εργασία μας, ο/η καθηγητής/τρια μας εντοπίζει (βρίσκει) τα σημεία που δυσκολευόμαστε και μας βοηθά να ξεπεράσουμε τις δυσκολίες μας.	1	2	3	4	5
10.	Οι ασκήσεις που υπάρχουν στα διαγωνίσματα που μας δίνει ο/η καθηγητής/τρια μας είναι πιο δύσκολες από τις ασκήσεις που λύνουμε στην τάξη.	1	2	3	4	5
11.	Γνωρίζω κάθε φορά σε πιο μέρος του μαθήματος (αρχή, μέση και τέλος) βρισκόμαστε.	1	2	3	4	5
12.	Ξεκινάμε το μάθημα των Μαθηματικών με πιο απλές δραστηριότητες και όσο προχωράμε γίνονται πιο δύσκολες.	1	2	3	4	5
13.	Κατά τη διάρκεια του μαθήματος αφιερώνουμε, συνήθως, αρκετό χρόνο για τις δραστηριότητες του καινούριου μαθήματος.	1	2	3	4	5
14.	Για να λύσουμε ασκήσεις που μας βάζει ο/η καθηγητής/τρια μας πρέπει να θυμηθούμε πράγματα που διδαχθήκαμε σε προηγούμενα μαθήματα.	1	2	3	4	5
15.	Ο/Η καθηγητής/τρια μας βάζει ασκήσεις στην αρχή του μαθήματος για να ελέγξει αν έχουμε μάθει το προηγούμενο μάθημα.	1	2	3	4	5
16.	Για κάθε νέο πράγμα που ο/η καθηγητής/τρια μας διδάσκει, μας δίνει ασκήσεις που έχουν σχέση με αυτό το πράγμα που μας είπε.	1	2	3	4	5

		Ποτέ	Σπάνια	Μερικές φορές	Συχνά	Σχεδόν πάντα
17.	Στο τέλος του μαθήματος των Μαθηματικών, λύνουμε ασκήσεις στην τάξη που αφορούν το μάθημα της ημέρας που κάναμε.	1	2	3	4	5
18.	Με τις ασκήσεις που μας δίνει ο/η καθηγητής/τρια να κάνουμε στην τάξη επαναλαμβάνουμε αυτό που έχουμε προηγουμένως διδαχθεί.	1	2	3	4	5
19.	Όταν ασχολούμαι με μια δραστηριότητα και δυσκολεύομαι, ο/η καθηγητής/τρια έρχεται αμέσως να με βοηθήσει.	1	2	3	4	5
20.	Βρίσκω πολύ εύκολες τις δραστηριότητες που μου ζητά ο/η καθηγητής/τρια των Μαθηματικών να κάνω.	1	2	3	4	5
21.	Ο/Η καθηγητής/τρια μας δίνει την ευκαιρία σε όλους τους μαθητές να συμμετέχουν στο μάθημα.	1	2	3	4	5
22.	Ο/Η καθηγητής/τρια όταν κάνει το μάθημα των Μαθηματικών, αφήνει να συμμετέχουν περισσότερο κάποιοι μαθητές.	1	2	3	4	5
23.	Κατά τη διάρκεια του μαθήματος των Μαθηματικών, ο/η καθηγητής/τρια μας παροτρύνει να συνεργαζόμαστε με τους συμμαθητές μας.	1	2	3	4	5
24.	Στην τάξη μου συνεργάζονται μεταξύ τους μόνο κάποιοι μαθητές, ενώ κάποιοι άλλοι όχι.	1	2	3	4	5
25.	Ο/Η καθηγητής/τρια των Μαθηματικών μας κάνει να νιώθουμε άνετα στην τάξη για να ζητήσουμε τη βοήθεια ή τη συμβουλή του/της.	1	2	3	4	5
26.	Κατά τη διάρκεια του μαθήματος, ο/η καθηγητής/τρια μας ενθαρρύνει να κάνουμε ερωτήσεις για ό,τι δεν καταλαβαίνουμε.	1	2	3	4	5
27.	Ο/Η καθηγητής/τρια συγχαίρει τους μαθητές, όταν προσπαθούν να κάνουν μια δραστηριότητα (π.χ. μας λεει «μπράβο»).	1	2	3	4	5
28.	Όταν κάποιος μαθητής δώσει μια λανθασμένη απάντηση, ο/η καθηγητής/τρια μας τον βοηθά να καταλάβει το λάθος του και να βρει τη σωστή απάντηση.	1	2	3	4	5

		Ποτέ	Σπάνια	Μερικές φορές	Συχνά	Σχεδόν πάντα
29.	Οι περισσότερες ερωτήσεις που υποβάλλει ο/η καθηγητής/τρια των Μαθηματικών μας ζητούν να δώσουμε μια απάντηση και όχι να εξηγήσουμε τον τρόπο που βρήκαμε αυτή την απάντηση.	1	2	3	4	5
30.	Ο/Η καθηγητής/τρια μας είναι δίκαιος με όλους τους μαθητές.	1	2	3	4	5
31.	Στο μάθημα των Μαθηματικών προσπαθούμε να ξεπεράσουμε ο κάθε μαθητής τον άλλο.	1	2	3	4	5
32.	Όταν εργαζόμαστε σε ομάδες στο μάθημα των Μαθηματικών, ο/η καθηγητής/τρια μας ενθαρρύνει να συναγωνιζόμαστε η μια ομάδα την άλλη.	1	2	3	4	5
33.	Στην τάξη μου, κάποιοι μαθητές κρύβουν τις ασκήσεις και τις απαντήσεις τους για να τις ξέρουν μόνο αυτοί.	1	2	3	4	5
34.	Στο μάθημα των Μαθηματικών ο/η καθηγητής/τρια βαθμολογεί τη συνεργασία μας.	1	2	3	4	5
35.	Κατά τη διάρκεια του μαθήματος των Μαθηματικών υπάρχουν παιδιά που κοροϊδεύουν άλλους συμμαθητές τους.	1	2	3	4	5
36.	Γνωρίζω πως εάν παραβιάσω κάποιο από τους κανονισμούς της τάξης μου θα τιμωρηθώ.	1	2	3	4	5
37.	Στην τάξη μας το μάθημα διακόπτεται από διάφορες αταξίες που κάνουν κάποιοι μαθητές.	1	2	3	4	5
38.	Όταν κάποιος μαθητής κάνει λάθος ορισμένα παιδιά βρίσκουν την ευκαιρία να τον κοροϊδέψουν.	1	2	3	4	5
39.	Ο/Η καθηγητής/τρια καταφέρνει να σταματήσει τις αταξίες που γίνονται στην τάξη.	1	2	3	4	5
40.	Υπάρχουν φορές που δεν έχουμε τα κατάλληλα υλικά για να γίνει το μάθημα των Μαθηματικών.	1	2	3	4	5
41.	Κατά τη διάρκεια του μαθήματος των Μαθηματικών αφιερώνουμε, συνήθως, λίγο χρόνο στην αρχή, για την εισαγωγή του μαθήματος.	1	2	3	4	5

		Ποτέ	Σπάνια	Μερικές φορές	Συχνά	Σχεδόν πάντα
42.	Υπάρχουν φορές που το κουδούνι κτυπά για διάλειμμα ή για να σχολάσουμε και το μάθημα των Μαθηματικών δεν έχει τελειώσει.	1	2	3	4	5
43.	Όταν τελειώσω μια εργασία πιο νωρίς από τους συμμαθητές μου, ο/η καθηγητής/τρια μου αναθέτει αμέσως κάτι άλλο.	1	2	3	4	5
44.	Όταν ο/η καθηγητής/τρια κάνει κάποια παρατήρηση σε κάποιους, αυτοί μπορεί σε λίγο να ξανακάνουν αταξία.	1	2	3	4	5
45.	Κατά τη διάρκεια του μαθήματος των Μαθηματικών αφιερώνουμε, συνήθως, χρόνο στο τέλος για την ανακεφαλαίωση.	1	2	3	4	5
46.	Χρειάζεται να σκεφτώ αρκετά πριν να απαντήσω κάποια ερώτηση που κάνει ο/η καθηγητής/τρια μας.	1	2	3	4	5
47.	Υπάρχουν στιγμές κατά τη διάρκεια του μαθήματος των Μαθηματικών που δεν έχω κάτι συγκεκριμένο να κάνω.	1	2	3	4	5
48.	Ο/Η καθηγητής/τρια μου δίνει την ευκαιρία να συμμετέχω στο μάθημα.	1	2	3	4	5
49.	Ο/Η καθηγητής/τρια των Μαθηματικών μας κάνει ερωτήσεις, στις οποίες πρέπει να πούμε τη γνώμη μας για ένα θέμα.	1	2	3	4	5
50.	Στην αρχή του μαθήματος, ο/η καθηγητής/τρια μας ρωτά ερωτήσεις, για να θυμηθούμε αυτά που μελετήσαμε στο προηγούμενο μάθημα.	1	2	3	4	5
51.	Όταν ο/η καθηγητής/τρια μας κάνει ερωτήσεις, χρησιμοποιεί εκφράσεις που είναι δύσκολες και δεν τις καταλαβαίνω.	1	2	3	4	5
52.	Αν δεν καταλαβαίνουμε μια ερώτηση, ο/η καθηγητής/τρια μας τη λέει με άλλο τρόπο ώστε να την κατανοήσουμε.	1	2	3	4	5
53.	Όταν ο/η καθηγητής/τρια μας ρωτά μια ερώτηση, μας δίνει **αρκετό** χρόνο για να σκεφτούμε.	1	2	3	4	5
54.	Όταν ένας μαθητής απαντήσει λάθος σε μια ερώτηση, ο/η καθηγητής/τρια μας βάζει άλλο μαθητή να απαντήσει την ερώτηση.	1	2	3	4	5

		Ποτέ	Σπάνια	Μερικές φορές	Συχνά	Σχεδόν πάντα
55.	Όταν δώσω μια λανθασμένη απάντηση, ο/η καθηγητής/τρια με βοηθά να καταλάβω το λάθος μου και να βρω τη σωστή απάντηση.	1	2	3	4	5
56.	Ο/Η καθηγητής/τρια μας επαινεί το ίδιο όλους τους μαθητές, όταν απαντούν μια ερώτηση σωστά.	1	2	3	4	5
57.	Ο χρόνος που δίνει ο/η καθηγητής/τρια μου για να απαντήσουμε μια ερώτηση είναι πολύ λίγος και μόνο οι καλοί μαθητές προλαβαίνουν να σκεφτούν, για να βρουν την απάντηση.	1	2	3	4	5
58.	Όταν αντιμετωπίζουμε κάποιο εμπόδιο ή δυσκολευόμαστε να λύσουμε τις ασκήσεις ή τα προβλήματα που έχουμε στο μάθημα των Μαθηματικών, ο/η καθηγητής/τρια μας βοηθά δείχνοντάς μας εύκολους τρόπους ή «κόλπα» για να λύσουμε αυτές τις ασκήσεις και προβλήματα.	1	2	3	4	5
59.	Ο/η καθηγητής/τρια μας αφήνει να σκεφτόμαστε και μας βοηθά με τον τρόπο του να ανακαλύψουμε εύκολους τρόπους ή «κόλπα» για να λύσουμε τις ασκήσεις ή τα προβλήματα που έχουμε στα Μαθηματικά.	1	2	3	4	5
60.	Στα Μαθηματικά, οι τρόποι ή τα «κόλπα» που μας μαθαίνει ο/η καθηγητής/τρια μπορούν να χρησιμοποιηθούν και σε άλλα μαθήματα της ενότητας.	1	2	3	4	5
61.	Ο/Η καθηγητής/τρια μας ενθαρρύνει να βρίσκουμε τρόπους ή «κόλπα», για να λύσουμε τις ασκήσεις και τα προβλήματα που μας δίνει.	1	2	3	4	5
62.	Όταν ο/η καθηγητής/τρια των Μαθηματικών μου μιλά στους γονείς μου για την πρόοδό μου είμαι και εγώ παρών.	1	2	3	4	5
63.	Όταν κάνουμε διαγώνισμα, στα Μαθηματικά, τελειώνω στο χρόνο που μας δίνεται.	1	2	3	4	5

ΜΕΡΟΣ Β

Στο μέρος αυτό περιλαμβάνονται κάποιες δηλώσεις. Για κάθε δήλωση **κύκλωσε** την απάντηση που αντιπροσωπεύει το τι γίνεται στην τάξη σου στο μάθημα των Μαθηματικών.

Ο/Η καθηγητής/τρια των Μαθηματικών μας επιστρέφει διορθωμένα τα διαγωνίσματα που κάνουμε

Α. το πολύ σε μια εβδομάδα

Β. το πολύ σε δύο εβδομάδες

Γ. το πολύ σε τρεις εβδομάδες

Δ. σε ένα μήνα

Ε. δεν μας τα επιστρέφει ποτέ.

Ο/Η καθηγητής/τρια μας εξηγά τι αναμένει να μάθουμε από το μάθημα των Μαθηματικών που θα μας διδάξει. Αυτό γίνεται:

Α. σε κάθε μάθημα

Β. στα περισσότερα μαθήματα

Γ. κάποιες μόνο φορές

Δ. πολύ σπάνια

Ε. σε κανένα μάθημα.

Ο/Η καθηγητής/τρια μας ζητά να σκεφτούμε τι μας βοήθησε να μάθουμε το μάθημα των Μαθηματικών που κάναμε. Αυτό γίνεται

Α. σε κάθε μάθημα

Β. στα περισσότερα μαθήματα

Γ. κάποιες μόνο φορές

Δ. πολύ σπάνια

Ε. σε κανένα μάθημα.

ΜΕΡΟΣ Γ

Πιο κάτω υπάρχουν κάποιες δηλώσεις. Σημείωσε ✓ στο κουτάκι του ΝΑΙ, στις δηλώσεις εκείνες που γράφουν το τι συμβαίνει στην τάξη σου στο μάθημα των Μαθηματικών, και ✓ στο κουτάκι του ΟΧΙ, στις δηλώσεις που δεν περιγράφουν αυτό που συμβαίνει στην τάξη σου, στο συγκεκριμένο μάθημα.

Όταν ο/η καθηγητής/τρια επιστρέφει τα διαγωνίσματα:		ΝΑΙ	ΟΧΙ
Α.	συζητά μαζί μου και μου εξηγεί τα λάθη μου.		
Β.	συζητούμε τα λάθη που έκαναν οι περισσότεροι μαθητές της τάξης.		
Γ.	λέει ποιοι μαθητές πήραν τους ψηλότερους ή χαμηλότερους βαθμούς.		
Δ.	πάνω στο διαγώνισμα έχει σχόλια για το πόσο καλά τα πήγα σε σύγκριση με το προηγούμενό μου διαγώνισμα.		
Ε.	συζητάμε και αποφασίζουμε τι πρέπει να κάνω για να γίνω καλύτερος/καλύτερη.		
ΣΤ	μας ζητά να κάνουμε στο σπίτι ασκήσεις παρόμοιες με αυτές που κάναμε λάθος στο διαγώνισμα.		
Ζ.	μας δίνει παρόμοιες ασκήσεις με όσες κάναμε λάθος για περισσότερη εξάσκηση.		
Η.	δεν τα σχολιάζουμε καθόλου.		

Παρακάτω μπορείς να γράψεις τις παρατηρήσεις σου για το ερωτηματολόγιο ή για τη διδασκαλία των Μαθηματικών στην τάξη σου.

....

....

....

....

....

....

....

....

....

Ευχαριστούμε πολύ για τη συνεργασία σας