UNIVERSITY OF CYPRUS

DEPARTMENT OF EDUCATION

TEACHER PROFESSIONAL DEVELOPMENT IN CLASSROOM ASSESSMENT:

Using the Dynamic Model of Educational Effectiveness to Improve Assessment Practice

MARGARITA CHRISTOFORIDOU

PhD THESIS

2013

TEACHER PROFESSIONAL DEVELOPMENT IN CLASSROOM ASSESSMENT: USING THE DYNAMIC MODEL OF EDUCATIONAL EFFECTIVENESS TO IMPROVE ASSESSMENT PRACTICE

Margarita Christoforidou

Submitted to the Department of Education,
in part completion of the requirements
for the degree of Doctor of Philosophy
Department of Education
University of Cyprus
May 2013

UNIVERSITY OF CYPRUS
DEPARTMENT OF EDUCATION

Dissertation Acceptance

This is to certify that the dissertation prepared

By Margarita Christoforidou

Entitled Teacher Professional Development in Classroom Assessment: Using the Dynamic Model of Educational Effectiveness to Improve Assessment Practice

Complies with the University regulations and meets the standards of the University for originality and quality

For the degree of Doctor of Philosophy

The dissertation was successfully presented to the examining committee on Monday, 20th of May, 2013.

Major Professor: Leonidas Kyriakides, Associate Professor
Department of Education, University of Cyprus

Advising Committee: Maria Eliophotou Menon, Associate Professor
Department of Education, University of Cyprus

Mary Koutselini, Professor
Department of Education, University of Cyprus

…….. ……..

Leonidas Kyriakides Maria Eliophotou Menon

……..

Mary Koutselini

Examining Committee
Maria Eliophotou Menon (Chair)
Associate Professor, Department of Education, University of Cyprus
Leonidas Kyriakides
Associate Professor, Department of Education, University of Cyprus
Mary Koutselini
Professor, Department of Education, University of Cyprus
Daniel Muijs
Professor, Southampton Education School, University of Southampton
Ioannis Katsillis
Professor, Department of Education, Univeristy of Patras

Abstract

Current teaching practices emphasize the integration of teaching and assessment, and recognize assessment as a key effectiveness factor. Educational researchers estimate that teachers spend a great percentage of their teaching time in assessment-related activities. In contrast to this, the research literature reveals that teachers are inadequately prepared to design, perform and use in-classroom assessment. In addition, whereas the literature highlights the role of teacher professional development in any attempt to change teachers' classroom practices, so far there has been inadequate solid empirical evidence to describe the change in teachers' actual assessment practice resulting from the received professional development. Taking the above into consideration, this study examines teachers' skills in assessment and how these can be developed though professional development. During the first phase of the study, a framework of teacher assessment skills is proposed and an instrument to measure teachers' skills in assessment is developed. This instrument is used to examine whether developmental stages can be identified, when investigating teachers' skills in assessment. The results of the first phase of the study provided support to the validity of the proposed framework as well as to the construct validity of the instrument developed. In addition, four stages of teacher assessment behavior are identified. The second phase of the study moves a step forward and compares the impact of the Dynamic Integrated Approach and the Competency-Based Approach to professional development on teacher assessment skills and student outcomes. It was found out that teachers, who use more advanced types of assessment behaviour, were more effective than those who demonstrate the relatively easy types and also that the Dynamic Integrated Approach had greater impact on both teacher assessment skills and student outcomes. Implications of findings in relation to teacher professional development in assessment are further drawn.

Περίληψη

Η αξιολόγηση του μαθητή αναγνωρίζεται ως αναπόσπαστο μέρος της διδασκαλίας και σημαντικός παράγοντας της εκπαιδευτικής αποτελεσματικότητας. Η διεθνής βιβλιογραφία επισημαίνει ότι οι εκπαιδευτικοί αφιερώνουν σημαντικό ποσοστό του διδακτικού τους χρόνου σε δραστηριότητες που αφορούν στην αξιολόγηση του μαθητή. Την ίδια στιγμή, επισημαίνεται ότι παρά τις διάφορες προσπάθειες για επαγγελματική επιμόρφωση των εκπαιδευτικών σε θέματα αξιολόγησης, μεγάλη μερίδα των εκπαιδευτικών δεν είναι ακόμη σε θέση να αξιολογήσει αποτελεσματικά τους μαθητές. Με βάση τα πιο πάνω, η παρούσα έρευνα εξέτασε τις δεξιότητες των εκπαιδευτικών στην αξιολόγηση και πώς αυτές μπορούν να αναπτυχθούν μέσα από προγράμματα επαγγελματικής επιμόρφωσης. Αρχικά, προτάθηκε ένα θεωρητικό πλαίσιο βάσει του οποίου αναπτύχθηκε ένα εργαλείο για μέτρηση των δεξιοτήτων των εκπαιδευτικών στην αξιολόγηση. Το εργαλείο αυτό χρησιμοποιήθηκε για να διερευνηθεί η ύπαρξη διαφορετικών επιπέδων στις δεξιότητες των εκπαιδευτικών στην αξιολόγηση. Τα αποτελέσματα της πρώτης φάσης της έρευνας εγκυροποίησαν το θεωρητικό πλαίσιο και το εργαλείο μέτρησης που προτάθηκαν. Επιπλέον, αναγνωρίστηκαν τέσσερα επίπεδα δεξιοτήτων αξιολόγησης. Η δεύτερη φάση της έρευνας σύγκρινε τη Δυναμική Προσέγγιση (Dynamic Integrated Approach) επαγγελματικής επιμόρφωσης με αυτή της Προσέγγισης στη Βάση Μεμονωμένων Δεξιοτήτων (Competency-Based Approach) σε σχέση με την επίδρασή τους στις δεξιότητες αξιολόγησης των εκπαιδευτικών και στα μαθησιακά αποτελέσματα. Τα αποτελέσματα έδειξαν ότι οι εκπαιδευτικοί που χρησιμοποιούσαν πιο ανεπτυγμένες μορφές συμπεριφοράς στην αξιολόγηση ήταν πιο αποτελεσματικοί και ότι η Δυναμική Ενδιάμεση Προσέγγιση επαγγελματικής επιμόρφωσης είχε μεγαλύτερη επίδραση τόσο στις δεξιότητες των εκπαιδευτικών στην αξιολόγηση όσο και στα μαθησιακά αποτελέσματα.

To my father…

"The righteous man walks in his integrity; his children are blessed after him."

(Proverbs 20:7)

Acknowledgments

As this journey has finally come to an end, I would like to express my gratitude to all those who have made this possible.

First and foremost, I want to thank my advisor Dr. Leonidas Kyriakides. Throughout my studies he provided me with support, encouragement and constructive advice. The confidence he showed in me helped me to surpass challenges and to stretch my abilities. It appears that the joy and enthusiasm he has for his research is contagious! I could not have wished for a better advisor.

I would also like to thank the members of my committee: Dr. Maria Eliophotou Menon, Dr. Mary Koutselini, Dr. Daniel Muijs and Dr. Ioannis Katsillis, for their time, valuable comments and constructive feedback.

I am most indebted to my family (Mom, Dad, Andreas, Lexas, Zoe and Porgos) for their constant support and unconditional love. Mom and Dad, you have always inspired me to do my best and have supported me in every decision I have made. I am proud to be a member of the Christoforides family!

I would also like to thank my PhD "therapy group". Andri, Stavroulla and Yiannakis our meetings have been a great support system all these years.

I am also grateful to my best friend throughout the years. Koumera, thank you for putting up with me and for always knowing who I am and reminding me of that when I forget. Your support and friendship over these years has been invaluable.

Last but not least, my deepest thanks and love go to my partner in life, Minas. Thank you for your unshakable faith in me and for supporting me even when you didn't understand what I was doing! You mean the world to me…

TABLE OF CONTENTS

LIST OF TABLES…… ix

LIST OF FIGURES…… ix

CHAPTER 1: INTRODUCTION TO THE STUDY……..1

Introduction……1

Research Aims and Questions…….....3

Study Summary...…….....….4

Contribution to the Theory……...…...6

Significance of the Study: Implications for Policy and Practice……8

Thesis Structure…….10

CHAPTER 2: LITERATURE REVIEW..……...……12

Educational Assessment…….13

Defining the Purpose of Assessment…….14

Summative Assessment…….14

Formative assessment……....16

Defining Formative Assessment…….……...16

Basic Principles of Formative Assessment…….……...20

The Phases of Assessment…….……....23

Planning and Construction of Assessment Tools…….…….25

Administration of Assessment Instruments……...….26

Recording and Analysing Data……..……27

Reporting Results to Students and Parents……..……...28

Educational Assessment: Issues of Effectiveness…….…..……29

Teachers' Skills in Assessment…….…….30

Teacher Training and Professional Development……..……35

Main Approaches to Teacher Professional Development……..……35

The Holistic/Reflective Approach to Teacher Professional Development…….36

Competency-Based Approach (CBA)……...…..39

*Developmental Stages and Professional Develop*ment…….……43

The Dynamic Integrated Approach (DIA) to Teacher Professional Development……...47

The Main Steps of the DIA ……..…….…..…...48

Step 1: Identify needs and priorities for improvement through empirical investigation ……...……...…48

Step 2: Provide guidelines for improvement: The role of the A&RTeam……...……49

Step 3: Establish formative evaluation mechanism……..……..49

Step 4: Establish summative evaluation mechanism……..……49

A framework for Investigating the Factor of Assessment…….…….....52

*a) Main Phases of the Assessment Proc*ess…….…….52

b) Assessment Techniques…….…….……..53

c) Measurement Dimensions…….…….……...54

Research Agenda…….…….…….56

CHAPTER 3: METHODOLOGY…….…….…..…..59

Research Design and Justification of the Methods Chosen……...…….…….59

Phase 1……..…...…….…….59

Phase 2……..……..…..…….62

Research Instruments……...……..……..…..64

Teacher Questionnaire…….……65

Teacher Interviews……..…….…….…..67

Student Written Tests……..……..72

Research Sample…….…….……74

Phase 1…….…...74

Phase 2…….……...74

The Intervention…….…….……76

Step 1: Initial Evaluation of teachers' assessment skills and student outcomes and allocation of teachers into treatment groups……..…….…....76

Step 2: Training sessions……..…78

Session 1 (Experimental Group A+B)……..….....78

Experimental Group A: DIA approach……..…...79

Session 2……....79

Focus Area 1- Working group A (Stage 1)……..…80

Focus Area 2- Working group B (Stage 2)……..…..…81

Focus Area 3- Working group C (Stage 3)……..…..…81

Focus Area 4- Working group D (Stage 4)……..…..…82

Sessions 3-6……..…..83

Experimental Group B: CBA approach……..…...83

Session 2……..…..…85

Sessions 3-6……..…...85

Session 7 (Experimental Group A+B)……..…..…86

Step 3: Final evaluation of teachers' assessment skills and student outcomes……..……..86

Analysis of Data……...……...…….87

The Rasch and Saltus models……...……87

Qualitative Analysis……..……...91

Multilevel Analysis……..…….92

Research Limitations……..…..……..93

CHAPTER 4: RESEARCH RESULTS……..…….…...…...96

Searching for Stages of Teacher Skills in Assessment……..…..…..96

*Using the Rasch Model to Specify the Hierarchy of Item Diffic*ulty…….…..……96

Using Cluster Analysis to Specify Levels of Difficulty……...…..…...100

Using the Saltus model to Specify the Developmental Structure of Assessment Skills…….....….100

*Type 1: Using written tests to measure basic skills in mathematics for summative reason*s…….……...104

Type 2: Using different techniques of assessment to measure basic skills in mathematics ……...……..…104

Type 3: Using assessment techniques to measure more complex educational objectives for formative reasons …….……...…...105

Type 4: Differentiation in Assessment: Applying assessment in and for different occasions and students …….…….……...105

The Impact of the Interventions on Teacher Assessment Skills and Student Achievement……..….106

Impact on Teacher Assessment Skills…….…….……...106

Impact on Student Achievement…….…….……..109

CHAPTER 5: DISCUSSION AND SUGGESTIONS FOR FURTHER RESEARCH…….……116

Measurement of Teacher Assessment Skills……..……116

Teacher Professional Development in Assessment……..……...123

Implications for Policy and Practice in the context of Cyprus……..……..130

Suggestions for Further Research……..……...121

REFERENCES…..……...…….135

APPENDIX A……..……..……....….165

APPENDIX B……...…..……..……...173

APPENDIX C……...…..……..……...174

APPENDIX D…….…….……...178

APPENDIX E…….…….……...……..179

LIST OF TABLES

Table 1. Study timeframe……..60

Table 2. Pre and Post- student test administration……73

Table 3. Teachers' allocation into groups…….77

Table 4. Statistics relating the questionnaire measuring assessment skills, based on the Rasch analysis of the whole sample and of Rasch analysis of each gender group separately…….98

Table 5. Rasch and Saltus parameter estimates for the 87 items of the teacher questionnaire grouped into four levels of assessment skills……...101

Table 6. Means and standard deviations of teacher scores measuring assessment skills of the control and the experimental group before and after the intervention……107

Table 7. Parameter Estimates and (Standard Errors) for the analysis of student achievement in mathematics(Students within classes, within schools)…….……..112

Table 8. Parameter Estimates and (Standard Errors) emerged from separately analyzing achievement of students taught by teachers situated at the same level……...113

Table 9. Effect of employing each approach expressed as Cohen's d per group of students taught by teachers situated at the same stage and for the whole sample…….....114

LIST OF FIGURES

Figure 1. The assessment cycle illustrating the phases of assessment ……..……53

Figure 2. A framework for measuring teacher assessment skills ……..56

Figure 3. True experimental pretest-post-test control group design……..64

Figure 4. Rasch Scale of teacher's skills in assessment……....99

CHAPTER 1

INTRODUCTION TO THE STUDY

This chapter presents an overall view of the study. The research purpose of the study is stated and specific research aims and questions are set. The importance of the study is justified based on its theoretical and practical relevance. Finally, a summarized version of the study's outline is included in order to facilitate further reading.

Introduction

Classroom assessment research appears to be of high priority in the field of education. Formative assessment, in particular, has been prevalent in the educational discourse over the past decades, shifting the attention towards assessment practices that aid the learning and teaching process. A study conducted in Cyprus in the mid-90s, measuring teacher perceptions towards assessment, found that Cypriot teachers are in favor of formative assessment (Kyriakides, 1997). Based on this finding as well as on international claims concerning the effectiveness of formative assessment (Black & Wiliam, 1998; Hattie & Temperley, 2007; Wiliam, Lee, Harrison, & Black, 2004), further studies were conducted in Cyprus in order to examine how the use of assessment affects teachers' effectiveness (Kyriakides, 2005; Kyriakides & Creemers, 2008a). The first study demonstrated that primary school teachers, who conduct assessment for formative reasons, are more effective in terms of promoting student learning outcomes (both cognitive and affective outcomes were taken into account) than those who conduct assessment for summative reasons (Kyriakides, 2005). Multilevel analysis of the data which emerged from this study also revealed, that schools promoting formative assessment in their policy are more effective than schools with no policy on assessment. In this way, formative assessment at classroom level and school policy on assessment were seen as factors associated with student achievement gains. Moreover, two

other studies demonstrated that not only the extent to which teacher assessment takes place for formative reasons but also qualitative aspects of the functioning of teacher assessment such as the focus and quality of assessment tasks, are associated with student learning outcomes (see Kyriakides & Creemers, 2008a; Creemers & Kyriakides, 2009). These studies provided empirical support to the validity of the dynamic model of educational effectiveness, which refers to five dimensions that can be used to measure the functioning of each factor (including teacher assessment). The series of effectiveness studies conducted in Cyprus, during the last 15 years, provided empirical support to the impact that formative assessment can have on student learning outcomes. However, despite the positive attitudes of Cypriot teachers towards formative assessment, only a limited number of teachers actually implement it in their everyday practice (Creemers, Kyriakides & Antoniou, 2013). This finding is in line with international research suggesting that classroom everyday assessment practice still appears to be outcome-oriented (Earl & Katz, 2000; Herman, Osmundson, Ayala, Schneider, & Timms, 2006; Lock & Munby, 2000).

Taking the above into consideration, this study shifts its attention towards teacher assessment skills, bringing together research findings from the areas of Educational Effectiveness Research and Teacher Professional Development Research (Creemers, Kyriakides & Antoniou, 2013). In particular, this research aims to investigate in more detail the development of teachers' skills in assessment. Considering recent findings concerning the validity of the developmental stages of teaching skills (Dall'Alba & Sandberg, 2006), this study aims to investigate whether teachers could be classified in developmental stages, in relation to their assessment skills, and whether this classification could act as a guide for offering needs-oriented professional development in classroom assessment. Therefore, the first phase of the study aims to investigate whether developmental stages can be identified when examining teachers' assessment skills. The second phase of the study aims to implement and

compare two approaches to professional development. The first approach is competency - based and it aims to offer training in assessment skills to all participating teachers despite their stage of development. This approach is in line with the assessment literacy movement, recognized in the literature (Popham, 2004; 2009; Schafer, 1991; Stiggins, 1991; 1999). On the other hand, the second approach is in line with the Dynamic Integrated Approach (Creemers, Kyriakides & Antoniou, 2013) recently proposed and it aims to offer differentiated training to teachers of each group by taking into account their skills and their current stage. In an overall, the purpose of this research is to provide information related to the effectiveness of the two approaches as this is measured through students' learning outcomes. In this context the following questions arise.

Research Aims and Questions

The first question arising is whether stage classification can be identified when examining teachers' assessment skills and whether these stages represent teachers' overall assessment practice or each phase of the assessment process. Given that this classification is empirically justified, the next question deserving attention is how professional development can aid stage progression more effectively. In order to answer this question, an experimental study will be conducted to examine the impact of two different professional development approaches on teachers' assessment skills and students' achievement. Both approaches will address teacher skills in classroom assessment, however, with fundamental differences to their content and structure. The first approach is competency based and it will offer all teachers the same training on assessment skills in order for them to acquire the necessary competencies in assessment. The second approach, the dynamic integrated approach, adopts a theory driven evidenced based approach to professional development, advocating the provision of professional development adjusted to teachers' developmental stage (Creemers & Kyriakides, 2010). More precisely, this study aims to answer the following research questions:

1) Can teachers be classified in distinctive developmental stages based on their assessment skills? And if so,

   1.1) How can these stages be defined? More specifically, are these stages created based on teachers' assessment skills across the four aspects of the assessment process and/or across the five measurement dimensions of the dynamic model?

   1.2) Do these stages describe the overall assessment practice across the four aspects of the assessment process or are there differentiations between each aspect of assessment?

   1.3) To which extent can teachers' stages in assessment be associated with student achievement?

2) Are teachers, who use more advanced types of behaviour, more effective than those who demonstrate the relatively easy types?

3) Which of the two professional development approaches has greater impact on: a) the improvement of teachers' assessment skills and b) the learning outcomes of their students?

Study Summary

To answer the research questions set, a two phase study was conducted. The first phase of the study examined the identification of developmental stages of teacher assessment skills. In order to examine the factor of classroom assessment in more detail, a framework based on the assessment process as described in the literature was developed. First, the necessary skills across all phases of the assessment process were identified, in order to create a comprehensive view of what teachers should be able to do in relation to classroom assessment. In addition, traditional as well as alternative assessment techniques were taken into consideration, since the literature supports the use of a combination of assessment techniques to assess student

learning. Finally, a measurement framework, developed within the field of Educational Effectiveness Research (EER), was adopted. Based on the framework proposed, a teacher questionnaire measuring assessment skills was developed. The questionnaire consists of 87 items, designed to measure teachers' assessment skills across the three aspects of the framework and it was administered to a representative sample of 178 primary school teachers. Moreover, semi-structured interviews were conducted in order to match responses and further ensure the internal validity of the results. Matching teachers' responses from the interviews with the questionnaire data provided support for the internal validity of the study. Using the Rasch and Saltus models to analyze the data, it was found that assessment skills can be grouped into four types of assessment behavior, which are discerned in a distinctive way and move gradually from skills, associated with everyday assessment routines, to more advanced skills, concerned with differentiation in assessment.

Based on the results of the first phase of the study, a decision was taken to investigate the extent to which different approaches to professional development can be used for improving teachers' skills in assessment as well as student outcomes. The second phase of the study was based on an experimental design and it examines the impact of two professional development approaches upon teacher assessment skills and student achievement. In particular, the second phase of our study aims to compare the impact of a teacher professional development program in mathematics assessment based on the Dynamic Integrated Approach (DIA) with the impact of a program using the Competency Based Approach. Teachers, who participated in the first phase of the study (n=178), were invited to attend a teacher professional development program. The program was to be completed through seven three-hour meetings, from November 2010 to May 2011. All meetings were scheduled in non-working time and participation was on a volunteer basis. Out of the 178 teachers, 76 teachers agreed to use their free time to attend this course. Teachers, who agreed to participate in the

teacher professional development program and were found to be at a certain developmental stage, were randomly allocated and evenly divided into two groups. As a result of the random assignment, two groups of teacher of comparable assessment ability were created. In addition, data on student achievement were collected by using external written forms of assessment, designed to assess knowledge and skills in mathematics. The intervention took place from November 2010 to May 2011. During this time, teachers participated in a series of seven training sessions that aimed at improving teachers' assessment skills, using the professional development approach employed. The first group employed the Dynamic Integrated Approach and the second the Competency Based Approach. Teachers, who did not attend any INSET course (n=102), were treated as members of the control group. Teacher questionnaires as well as student tests in Mathematics were administered at the end of the intervention in order to examine the effectiveness of each of the two approaches, in relation to teacher assessment skills and student achievement. The results of the second phase of the study showed that teachers, participating in each intervention group, managed to improve their assessment skills; however, the DIA had bigger impact on teacher assessment skills than the CBA. In addition, only the DIA was found to have an impact on student achievement in mathematics.

Contribution to the Theory

This study contributes to educational theory in four ways. First, it aids to the further development of the theory of educational assessment, since for the first time a dynamic framework that enables the definition and measurement of classroom assessment skills is developed. Until now, attempts to define what teachers should know and be able to do in relation to assessment have not managed to address assessment skills in a systematic way. Researchers have long recognized assessment skills as a crucial element of effective teaching practice (Smith, Silverman & Borg, 1980; Gullickson, 1986; Schafer, 1991). As a result, various lists outlining basic assessment competencies have been developed (American

Federation of Teachers, National Council on Measurement in Education National Education Association; 1990; Schafer, 1991; Stiggins, 1995; 1999). These lists describe assessment competencies in relation to general standards of assessment practice, without however providing details on the specific skills involved. In addition, these lists have not been associated with a specific theoretical background and empirical evidence supporting their validity has not been provided. Furthermore, recent conceptions of formative assessment are not effectively addressed (Brookhart, 2011). Recognizing the need for a comprehensive framework based on which skills associated with classroom assessment can be defined and measured, a dynamic framework of teacher assessment skills is proposed. Its dynamic nature can be attributed to the fact that the skills examined are based on a more comprehensive view of the assessment process. Thus, besides the main phases of the assessment process, assessment skills are defined and measured in relation to specific assessment techniques; whereas, both quantitative and qualitative characteristics of the assessment process are taken into consideration.

Given that assessment skills identified through the dynamic framework, as described above, will be examined in relation to student outcomes, it can be argued that this study also contributes to the identification of assessment skills that have a positive impact on student achievement. Most studies come from the area of formative assessment. Although research in that area suggests that the general practices associated with formative assessment can facilitate learning (Black & Wiliam, 1998; Wiliam, Lee, Harrison & Black, 2004), commonly made quantitative claims for effectiveness have been questioned (Bennett, 2011). This calls for more high quality studies to further strengthen the research base of formative assessment and assessment in general, in relation to their impact on learning.

The third contribution of this study to theory relates to the identification of developmental stages of teachers' assessment skills. Although attempts to classify teachers in

relation to the adoption of assessment strategies can be found in the literature (see Black et al., 2003; Wiliam, Lee, Harrison & Black, 2004), one can identify a lack of systematic research on teachers developmental stages, in relation to their assessment knowledge and skills. This study identifies, for the first time, types of assessment behavior that can stand as stages of assessment skill development. These stages move from relatively easy to more advanced types of behavior and are described in a distinctive way; thus, addressing a weakness of previous stage related studies to provide a clear picture of what each stage entails (Dall' Alba & Sandberg, 2006). In addition, the developmental scale proposed was identified in two measurement periods; thus, addressing another serious weaknesses of previous studies to investigate stage identification over a period of time (ibid, 2006).

Finally, this study contributes to the theory of teacher professional development. The review of the literature recognizes a lack of experimental studies investigating effective ways to improve teachers' skills (Demetriou, 2009; Tymms & Merrell, 2009). At the same time, the importance of experiments in educational research is highlighted (Slavin, 2010). For the first time, an experimental study will be conducted in order to investigate ways to achieve improvements in both assessment skills and student achievement. The competency-based approach, examined in this study, is by no means new; however, systematic empirical evidence on its effectiveness is lacking (Cohen & Ball, 1999; Richardson & Anders, 1994). On the other hand, the dynamic-integrated approach has been recently introduced and thus further investigation of its impact is needed. The results of this study could help identify effective ways to improve teachers' assessment skills and student achievement.

Significance of the Study: Implications for Policy and Practice

Educational Effectiveness Research (EER) addresses questions on what works in education and why. Over the last three decades, research into educational effectiveness has improved considerably, showing both methodological (Goldstein, 2003; Creemers, Kyriakides

& Sammons, 2010) as well as theoretical advances (Levine & Lezotte, 1990; Scheerens & Bosker, 1997). However, recognizing the constraints of existing approaches to contribute to the improvement of teaching practice, current approaches of EER examine how its findings can also be used for improvement purposes. Indeed, a major objective of educational science is to contribute to the effectiveness and the improvement of education by providing a knowledge base for practice and by helping schools develop effective intervention programs (Creemers & Kyriakides, 2006). In this context, the significance and importance of this research can be found in the way that its results can be used to yield improvement in the field of classroom assessment and teacher education at both policy and practice level.

In particular, the results of this study can be used by higher institutions, providing initial and in-service training to adjust their curriculum in order to provide adequate and appropriate assessment training to prospective and in service-teachers. Research in the area shows that although teachers spent a large amount of teaching time in assessment related activities (Crooks, 1988; Herman & Dorr-Bremme, 1982; Stiggins, 1991; Stiggins & Conklin, 1992), they however lack the necessary knowledge and skills to effectively assess their students (Lukin at.al., 2004; Schafer, 1993; Stiggins, 2002). If this study provides evidence in relation to which assessment skills have a positive impact on student learning, then institutions' responsible for teacher training can adjust the training offered to address these skills.

As stated earlier, this study further aims to develop a tool for measuring teachers' assessment skills, based on the theoretical framework proposed. Given that evidence supporting its validity is provided, this tool can be used to perform an initial evaluation of teachers' assessment skills. This evaluation can serve as a starting point for improvement and further professional development based on the needs identified. At the same time, this tool can be used to evaluate the impact of teachers' professional development programs upon teachers'

assessment skills. This could help determine the effectiveness of professional development programs in educational assessment, while at the same time provide information that can be used to improve their quality.

Moreover, if this study provides empirical justification of teachers' developmental stages in assessment, then policy should move on to establish mechanisms that allow stage identification in order to provide appropriate assistance to teachers. Research on the development of expertise suggests that teachers at different stages of development have differentiated needs. This suggests that professional development programs should adjust their content and structure to address these needs in order for improvement to be achieved.

Finally, this study aims to compare the Competency-based approach and the Dynamic Integrated approach, in relation to their impact on teachers' assessment skills and student outcomes. If the findings support one approach to be more effective, then policy should be directed towards the development of equivalent programs in order to train teachers in assessment. Subsequently, the results of this study can contribute to the development of evidence-based educational policy and practice, related to teacher training and professional development in assessment.

Thesis Structure

The complete thesis consists of five chapters. The first chapter is introductory and presents the research background, the research problem addressed as well as the research questions this study aims to answer. The first chapter also points out the scientific and practical relevance of the study. The second chapter is a literature review providing a theoretical framework of the fundamental concepts and issues related to the purpose of the study. It also examines the concepts of educational assessment, identifying the different phases and main purposes of the assessment process as indicated by the literature. In addition, Chapter 2 reviews literature in relation to teacher skills in assessment and it further moves on

to present a review of available literature and research concerning teacher professional development with particular emphasis on the two main professional development approaches. A brief overview of educational effectiveness research and how recent developments in the field can contribute to the purpose of this study are also discussed. Then, the theoretical framework proposed and used in this study is described in detail. The chapter ends with a presentation of the research agenda. Continuing, Chapter 3 describes the research methodology. In this chapter, research strategy, design and procedures are discussed, providing information on the sample population and the statistical techniques used. Finally it discusses the recognized limitations of the study. The next chapter, Chapter 4, presents the analysis of the data collected during the study. The analysis is made based on the research questions, presented in Chapter 1. Finally, Chapter 5, the last chapter of the study, presents a discussion on the outcomes of the study, in accordance to each research question and to the overall research problem. It also discusses possible consequences for theory, policy and practice recognized through the study. It ends with suggestions for further research.

CHAPTER 2

LITERATURE REVIEW

The literature review, presented in this chapter, aims to provide a theoretical framework of the fundamental concepts and issues related to the purpose of the study. It places the research problem in the wider theoretical context, recognizing possible connections and relationships. Through a critical literature review, it creates a frame of reference for the examination of the research problem and questions stated in the previous chapter. Therefore, this chapter concentrates on providing a review of the available literature within and across the fields of educational assessment, teacher education and educational effectiveness. Specifically, in the first section, literature on educational assessment is discussed. First, the purposes of educational assessment are identified. Then, the different phases of the assessment process are examined and research on assessment skills is reviewed. In the next section, the lack of research on professional development, specifically focused on assessment, is recognized. For these reasons, this section draws on general literature on teacher professional development, which is mainly concerned with generic teaching skills. A critical review of the main approaches to professional development is presented. Particularly, emphasis is given to the two dominant approaches: the competency-based approach and the holistic approach. Recognizing the strengths and weaknesses of the two approaches, the third section presents the dynamic integrated approach as an alternative, giving particular emphasis to its theoretical and methodological assumptions. The fifth section presents and describes in detail the theoretical framework proposed and used in this study. Finally, the last section summarizes the main conclusions drawn from the literature review and describes the research agenda for the present study.

Educational Assessment

Education is a highly charged evaluative setting (Broadfoot, 1996). Educational processes are purpose-oriented, and therefore carry with them the need to be evaluative at some point, whether this purpose has been achieved or not. In the history of education, assessment was developed as an antidote to the use of social criteria which for many years had determined educational provision. The need for more fair and accurate ways for selecting and classifying students elevated the notion of intelligence, with 19th century psychologists eagerly researching the determinants of various personal characteristics which could provide evidence of one's ability to be educated. According to Kamin (1974), the belief in the existence of an innate and fixed quality in students resulted in the extensive view of intelligence testing. Assessment was therefore used as a means of social control and reproduction. Indeed, as Broadfoot (1996) argues "the use of formalized procedures is an extremely powerful policy mechanism for exerting control over the education system" (p. 8) and intelligence tests were able to equate academic failure with inbuilt inadequacy.

However, in recent decades there is a growing dissatisfaction regarding traditional psychometric approaches, since a substantial body of research has shown that these tests cannot provide objective and reliable evidence of the attainments measured (IngenKamp, 1977; Raven, 1991; Satterley, 1994). In addition, educational literature has recognized that identifying differentiation and reliability as the most important features of assessment has led to emphasis being paid in quick economic and often in multiple-choice achievement tests rather than assessment procedures that provide a useful picture of what students can do (Broadfoot, 1996). The era we are living in today, has being characterized by many as an era of assessment reform, with assessment serving various purposes, using a number of methods and tools. This section presents a review of the recent literature on educational assessment, identifying its main purposes and phases. Questions of effectiveness are further examined.

Defining the Purpose of Assessment

The literature confirms that any designer or user of assessment should be familiar with the purpose an assessment wishes to serve (Airasian, 2001; Van der Horst & McDonald, 1997; Killen, 2003). Teacher assessment can serve a variety of purposes (Broadfoot, 1992; Brookhart, 2003; Gipps, 1994; Pellegrino et al., 2001; Torrance & Pryor 1998). Clarifying the purpose an assessment aims to serve is necessary in order for the appropriate procedures, methods and tools to be used. The review of the literature reveals two main purposes of classroom assessment: the summative and formative purposes of assessment.

Summative Assessment

Summative assessment is used for the recording of the overall achievement of a pupil in a systematic way (DES/WO, 1988). It aims at describing attainment, achieved at certain time, in order for comparisons to be made according to students' level of performance. Brookhart (1997) describes the main characteristics of summative assessment. First, summative assessment takes place at certain intervals, when achievement has to be reported, and it also relates to progression in learning against public criteria. In addition, the results of different pupils may be combined for various purposes because they are based on the same criteria, whereas decisions should be based on evidence from the full range of performance that is relevant to the criteria being used. Finally, summative assessment requires methods which are as reliable as possible without endangering validity; thus, it involves some quality assurance procedures.

Harlen (2005) further argues that the summative uses of assessment can be grouped into 'internal' and 'external' to the school community. Internal uses include the use of regular grading for recordkeeping, informing decisions about courses available within the school, and reporting to parents and students themselves. Teachers' judgments, often based on teacher-made tests or examinations, are commonly used in these ways. External uses include

certification by examination bodies or for vocational qualifications, selection for employment or higher education, monitoring the school's performance and school accountability, often based on the results of externally created tests or examinations.

When information about students' achievement is used for decisions that are considered important, not only for the individual student but also for the teachers and school, the results acquire a 'high stakes' character (Harlen, 2005). Indeed, according to Madaus (1988) a test is high-stakes when its results are used to make important decisions that affect students, teachers, administrators, communities, schools, and districts. The summative purpose of assessment has been associated with high-stakes assessment and accountability procedures (Harlen & Deakin Crick, 2002). However, high stakes assessment has been widely criticized for its negative impact on students, teachers and the curriculum. In particular, measurement specialists oppose high-stakes testing arguing that the use of a single indicator of competence to make important decisions about individuals or schools violates the professional standards of the measurement community (AERA, 1999). Other critics worry that the pressure of doing well, compromises instruction practice, since teachers tend to adopt a teaching style which emphasizes in knowledge transmission (Harlen & Deakin Crick, 2002). Moreover, it is argued that high-stakes testing undermines education because it narrows the curriculum (Nichols & Berliner, 2007; Watanabe, 2007), since it results in emphasizing on subjects tested at the expense of creativity and personal and social development. Finally, research shows that the increase in test scores, found on the introduction of tests is due to familiarity with the particular test content and not to increased achievement (Linn, 2000), thus posing questions for their effectiveness.

Formative assessment

The growing international dissatisfaction with high stakes assessment gave rise to a large body of literature, arguing in favor of assessment that not only evaluates but also promotes learning. Definitions and basic principles of formative assessment are discussed next.

Defining Formative Assessment

Scriven (1967) is the first to use the adjective "formative" to describe the evaluation of educational programs. Always in relation to his curriculum and teaching research, Scriven proposes two different uses of educational evaluation. According to Popham (2006) the need to satisfy the requirements of a U.S federal law stimulated Scriven's distinction. Indeed, the Elementary and Secondary Education Act of 1965 (ESEA) provided findings to US educators and required evaluative judgments about their programs to be made, in order for the funding to continue. Given the then lack of knowledge concerning program evaluation amongst the educators, the interest of educational researches and academics shifted to promoting a better understanding of the evaluative process.

Scriven's distinction is based on the differences in the purpose of these two types of evaluation; whereas, both evaluations are defined as appraisal for an educational program's worth or merit, in the case of formative evaluation, there is still time for the program's staff to make modifications that will aim at improving the program. On the contrary, summative evaluation addresses a mature, final-version program, intending to provide relevant decision makers with the information they need in order to decide whether the specific program should be continued or terminated. Scriven's distinction was the beginning of a new era in the field of educational evaluation. However, the word in its current meaning was first used in 1971 in the work of Bloom, Hastings and Madaus. The Learning for Mastery model presumes that all children can learn if they are provided with the appropriate learning conditions. According to

the model, in order for students to progress to the next learning objective, they must first master the current one. The method used to examine whether mastery has occurred is formative assessment, which is then used to draw information on students' specific weaknesses in order for corrective action to take place. Therefore, the adjective formative is now used to describe in- classroom assessment. Some years later, Sadler (1989) also addressed the notion of formative assessment, arguing that "formative assessment is concerned with how judgments about the quality of student responses (performance, pieces, or works) can be used to shape and improve the students' competence by short-circuiting the randomness and inefficiency of trial-and-error learning" (Sadler, 1989, p. 120). Thus, Sadler's approach places the weight on the use of qualitative judgments.

However, it can be argued that the work of Black and Wiliam in 1998 was the starting point for the notion of formative assessment to begin to flourishing, expanding the definition of formative assessment beyond the one Bloom described. In their article "Inside the black box" the two researchers bring forward, for the first time, through a meta-analysis, the instructional payoffs of classroom formative assessments. Therefore, they move formative assessment a step forward; from a promising concept to a research-proven path towards enhancing students' learning. In contrast to Bloom, they argue that a formative assessment does not need to be part of the everyday teaching process and provide immediate feedback. Instead, any assessment that aids in identifying and providing information that will be afterwards used effectively in order to adapt the teaching process to meet students' needs is recognized as formative. In their article (Black & Wiliam, 1998), they provide a definition according to which formative assessment refers to "all those activities undertaken by teachers and by their students in assessing themselves, which provide information to be used as feedback to modify the teaching and learning activities, in which they are engaged. Such

assessment becomes “formative assessment” when the evidence is actually used to adapt the teaching work to meet the needs” (p. 2).

The above definition requires that the information from a formative assessment must be in fact used to adjust instruction “to meet student needs”. However, Popham (2006) recognized this requirement as a flaw for the definition and went on to argue that “an assessment is formative to the extent that information from the assessment is used, during the instructional segment in which the assessment occurred, to adjust instruction with the intent of better meeting the needs of the students assessed” (p. 3). Popham (2006) agrees with Black and Wiliam (1998) that adjustments in instruction must occur with the intention to better meet students’ needs; however, he further highlights that these adjustments may not necessary be successful in order for assessment to be described formative. Indeed, the requirement set by Black and Wiliam (1998) subordinates the importance of formative intentions and moves all the weight on successful results. In this way, unsuccessful applications of formative assessment are completely ignored, along with the work of the teachers who attempted it. Acknowledging formative assessment as a risk technique discourages its use; therefore, limiting even more the possibility for assessment to be used in support of learning. The recognition of formative assessment as a type of assessment that aims at, but nonetheless cannot guarantee improvement of the learning and teaching process, is important for someone who wishes to encourage and support its widespread use amongst educators.

Indeed, by recognizing such a challenge found in their 1998 definition, Black and Wiliam (2009) re-state their definition, by taking in mind Pophams’ criticism, and further state that “practice in a classroom is formative to the extent that evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers, to make decision about the next steps in instruction that are likely to be better, or better founded, than the decision they would have taken in the absence of the evidence that was elicited” (p.9).

They argue that locating the focus on formative intentions would be unfortunate, since it will imply that "a situation in which evidence was collected, but not used would be formative" (Black & Wiliam, p.10). Additionally, they consider that locating the focus on the "resulting action", as they did in their 1998 definition, is perhaps to strict, since learning is unpredictable and formative actions may not always result in the improvement of learning. Therefore, in their new definition they appear to accept the fact that "even the best designed interventions will not *always* result in better learning for *all* students" (ibid, p.10). Black and Wiliam (2009) move on to clarify several features of their new definition. They argue that the term "instruction" (ibid, p. 9) represents a combination of teaching and learning and not a didactic approach to teaching. Furthermore, they acknowledge not only the teacher but also the individual learner and peers as agents of assessment. Finally, they clarify the requirement included in their definition for decisions "better or better founded" (ibid, p. 9) after the elicitation of evidence. This requirement accepts that some decisions indicated, may have been teacher's intentions prior to the elicitation of evidence.

Other definitions of formative assessment can also be found. The Assessment Reform Group defines formative assessment under the name of assessment for learning and states that "assessment for learning is the process of seeking and interpreting evidence for use by learners and their teachers to decide where the learners are in their learning, where they need to go and how best to get there" (Assessment Reform Group, 2002). Airasian (2001) has defined formative assessment as "the process of collecting, synthesizing, and interpreting information for the purpose of improving student learning while instruction is taking place" (p. 421), while Gronlund (2006) argues that formative assessment is intended "to monitor student progress during instruction […] to identify the students' learning successes and failures so that adjustments in instruction and learning can be made" (p. 6).

The absence of a commonly accepted definition is evident. Dunn and Mulvenon (2009) remark formative assessment as an "ethereal construct […] perpetuated in the literature due to a lack of agreed upon definition" (p. 2). Therefore, in order to better understand the concept of formative assessment, further investigation of its basic principles is presented next.

Basic Principles of Formative Assessment

The literature and research community of assessment has made many attempts to list the basic principles that underlie formative assessment in order to provide guidelines that will enable effective formative assessment practice to take place.

Stiggins (2001) describes students-involved classroom assessment and recognizes nine principles. He argues that in order for an assessment to be formative, the teacher must understand and articulate, prior to teaching, the achievement targets the students aim to fulfill. The teacher must regularly inform students about those targets, in a comprehensible way, and further transform these targets into dependable assessments that yield accurate information. Moreover, the teacher must have a clear understanding of the relationship between assessment and student motivation and subsequently use assessment to build students' confidence. Stiggins (2001) also recognizes as vital the use of assessment information to inform and revise instruction as well as the use of frequent and descriptive feedback. The last three principles concern the role of the student in the assessment process and highlight the need of students to be actively involved in their assessment, for them to actively communicate their achievement status and improvement and finally the need of students to be able to describe what targets they are aiming to fulfill and what comes next in their learning.

The Assessment Reform Group (2002) concluded ten principles which are considered essential in characterizing an assessment for learning. In order for an assessment to promote learning, it must be part of effective planning of teaching and learning and focus on how students learn. It must also be recognized as central to classroom practice and also be regarded

as a key professional skill for teachers. Furthermore, recognizing that any assessment has an emotional impact, the Group considers it vital for the assessment to be sensitive and constructive, while at the same time it takes into consideration the importance of learner motivation. Assessment for learning must promote commitment to learning goals and a shared understanding of the criteria by which students are assessed. It must also enable learners to receive constructive guidance about how to improve and develop their capacity for self-assessment in order for them to become reflective and self-managing. Finally, the last principle presents the need to recognize the full range of achievements of all students. (Assessment Reform Group, 2002)

The following twelve formative assessment principles were developed through the Re-engineering Assessment Practices (REAP) project, at the University of Strathclyde, by the Assessment Working Group, in order to provide guidance to teachers interested in improving the quality of the learning experience of students in higher education. The principles are based on recent research on assessment (Boud, 2000; Knight, 2002; Knight &Yorke, 2003) as well as the QAA guidelines on assessment of student learning (QAA, 2006). The first principle is to clarify what is considered a good performance by clarifying goals, criteria and standards. The second principle is to encourage time and effort on challenging learning tasks; thus, promoting deep rather than surface learning. The next two principles concern the provision of feedback and state the need of delivering high quality feedback information, helping learners to self-correct as well as the need of providing opportunities to act on feedback. The fifth principle is to ensure that formative assessment has a positive impact on learning, while the sixth is to encourage interaction and dialogue around learning. The REAP project has also highlighted that formative assessment must facilitate the development of self-assessment and reflection in learning and give choices to students regarding the topic, method, criteria as well as the weighting of timing in assessment. Formative assessment must also involve students in

decision-making about assessment policy and practice. Finally, it must support the development of learning communities, encourage positive motivational beliefs and self-esteem and provide information to teachers that can be used to help shape teaching itself.

Heritage (2007) is another assessment expert who attempts to clarify the formative assessment field by recognizing four core elements of formative assessment. The first element is learning progressions that helps teachers locate students' current learning status on the continuum along with students who are expected to progress. The second element is setting learning goals for students; such goals must be attainable from their current status in learning. The next element is feedback that provides students clear, descriptive criteria –based on information, indicating where they are in a learning progression, how their understanding differs from their desired goal and how they can move forward. Finally, the last element identified by Heritage (2007) is students' involvement through self- and peer-assessment and collaboration with the teacher in determining their current learning status and what they need to do to move forward in their learning.

Black et al (2003) conceptualize formative assessment in terms of some general learning principles. Therefore, they highlight the need to start from a learner's existing understanding as well as the need to involve the learner actively in the learning process. They also highlight the importance of meta-cognition "which calls both for a judgment of one's present understanding and for a clear view of the purpose of the learning and of the criteria for judging achievement of that purpose" (p. 78). Finally, they emphasize on the importance of the social element of learning which is made effective through interaction in discussion.

The discussion of the principles, defining formative assessment, is ongoing and despite the obvious overlaps, a clear and unified set of formative assessment principles is yet to be defined. Indeed, the principles of formative assessment are not prescriptive structures because each classroom has unique conditions to which the principles are applied as each practitioner

deems appropriate (Clark, 2011). What it is important to note is that the two main purposes of assessment, summative and formative, rest on different value assumption, and thus result in different practical implementation. The main phases that describe the practical implementation of the assessment process are presented next.

The Phases of Assessment

Student assessment is considered an integral part of teaching (Broadfoot & Black, 2004; Delandshere, 2002; Gipps, 1994; Harlen & James, 1997; Linn, 1993). It is defined as the systematic process of gathering information about students' learning (Shepard, 2000). It involves making our expectations explicit and public; setting appropriate criteria and high standards for learning quality; systematically gathering, analysing, and interpreting evidence to determine how well performance matches those expectations and standards; and using the resulting information to document, explain, and improve performance (Angelo, 1995).

Classroom assessment is frequently presented in the literature as a cycle subdivided into a number of phases. The Assessment Standards for school mathematics (National Council of Teachers of Mathematics [NCTM], 1995) describe the assessment process as four interrelated phases that highlight principal points, at which critical decisions need to be made. These phases are: a) planning, b) gathering evidence, c) interpreting the evidence and d) using the results. Each phase of the assessment process can be characterized by the decisions and actions that occur within that phase. Thus, during the planning phase, teachers make decisions concerning the purpose of the assessment, the framework of the activities, the methods for gathering and interpreting evidence, the criteria for judging performance and the formats used for summarizing judgments and reporting results. During the second phase, teachers make decisions concerning the creation or selection of tasks, the selection of the procedures for engaging students as well as the methods for creating and preserving evidence of the performances to be judged. The third phase, involves teachers' decisions concerning the

interpretation of the evidence. During this phase, teachers make decisions related to the determination of the evidence quality, the specific criteria to be applied for judging the performance, whether the criteria are applied appropriately and how the judgments will be summarized in the form of results. During the fourth and final phase of the assessment process, teachers make decisions concerning the use of the results. These decisions refer to the ways results will be reported, how inferences from the results will be made, which actions will be taken, and how to ensure that these results will be incorporated in subsequent instruction and assessment. It is highlighted that assessment does not proceed through these phases in a neat, linear fashion, and thus the phases should not be seen as necessarily sequential. Rea-Dickins (2001) also identifies four similar main decision-making stages in the assessment process. In the first stage, known as the planning stage, teachers consider the purpose and the procedures they will follow. This stage incorporates both the design and the operationalization phases of test development, since planning is being undertaken and the materials are being prepared. In the next stage, known as the implementation stage, teachers introduce the tasks to students and engage in scaffolding as required. In the third stage, known as the monitoring stage, teachers revise their teaching plans, share findings with other teachers, give learners detailed feedback and record evidence. In the final stage of the cycle, known as recording and dissemination, teachers record, report and make plans for the dissemination of the findings of their assessment procedures. Birenbaum et al. (2009) also present phases of the assessment process; however, this time, the phases are related to the formative purpose of assessment. They argue that an optimal formative assessment cycle consists of five phases: a) planning, b) evidence collection, c) interpretation, d) utilization and e) evaluation. During the planning phase, teachers set goals, define objectives and intended outcomes that will be used as evidence of students' performance, whereas during the interpretation phase, teachers estimate gaps between intended and obtained outcomes. During the utilization phase, teachers

implement interventions in order to narrow the gaps between intended and obtained outcomes. Lastly, during the evaluation phase, teachers assess the effectiveness of these interventions as far as narrowing the identified gaps is concerned.

As it is evident, various conceptualizations regarding the process of assessment can be identified. Based on the available literature, the four main phases describing the process of assessment design and use and the skills associated with each phase (i.e. planning/construction, administration, recording and reporting) are described next in more detail.

Planning and Construction of Assessment Tools

This phase includes skills that refer to the planning and designing of assessment as well as to the construction of the assessment tools. As mentioned above, it is necessary to clarify the purpose an assessment aims to serve in order for the appropriate procedures, methods and tools to be chosen. Thus, in this phase, teachers are expected to decide whether they are planning an assessment to achieve summative or formative purposes. Research, so far, has shown that achieving both purposes with one mechanism is not possible (Harlen & James, 1997; Black & Wiliam, 1998; Kyriakides & Campbell, 2003; Kyriakides, Demetriou & Charalampous, 2006). After deciding the purpose of assessment, it is necessary to define the learning goals based on which a student will be assessed. Herman et al. (2006) consider learning goals as the starting, ending, and recycling points in the selection and implementation of quality assessment tools, in the interpretation and analysis of student work, and in the use of results in order to provide informative feedback and take action that will further students' progress. Indeed, their role appears critical throughout the process of assessment. Goals provide a framework for interpreting and responding events that occur (Yorke, 2003) and affect performance by directing attention, mobilizing effort, increasing persistence, and motivating strategy development (Locke et al., 1981). Finally, this phase

includes the selection or construction of the necessary assessment instruments. Assessment methods hold an important role in ensuring the quality and effectiveness of assessment. Accepting Boud's (1988) argument that assessment methods probably have a greater influence on how and what students learn than any other single factor, we come to agree with Broan, Bull and Pendlebury (1997) that if we want to change student leaning, we must first change the methods of assessment. Stiggins (1992) argues that although we have many assessment tools at our disposal, they are not interchangeable. Choosing an assessment method depends on the achievement target to be assessed since "certain targets match up with certain assessment methods" (p.213). He further moves on to recognize the importance of the context in choosing an assessment method, arguing that "depending on the context the user is more or less able to take advantage of the strengths of a particular method and/or overcome its limitations" (ibid).

Summarily, the skills required in this phase cover decisions concerning the purpose that an assessment wishes to serve (Brookhart, 2003; Gipps, 1994; Pellegrino et al., 2001; Torrance & Pryor, 1998), the definition of learning goals based on which a student will be assessed (Herman et al. 2006; Sadler 1989); as well as the selection or/and development of quality assessment tools through which the purpose and goals of the assessment will be achieved (Green & Mantz, 2002; Shepard, 2000).

Administration of Assessment Instruments

The second phase includes skills associated with the implementation of assessment. Whereas external assessments are typically more standardized in terms of timing, setting and teacher support, the administration of classroom assessment rests mainly on teacher's decisions. As Black and Wiliam (2006) argue effective assessment relies heavily upon the adaptations teachers make; these adaptations vary in terms of both scope and time-scale. Skills included in this phase refer to decisions concerning the timing of an assessment, the

assessment's link to instruction, the variety of techniques used as well as the teachers' role during assessment administration (Anderson, 2003; Black & Wiliam, 1998; Shepard, 2007).

Recording and Analysing Data

The third phase refers to skills associated with the recording and analysis of data derived from the assessment process. Recording assessment information is necessary in order for information to be effectively used to inform learning and teaching. Unfortunately, as Schmoker (2006) points out, an enormous proportion of daily assessment is in fact never assessed, since no evidence is recorded, therefore, leaving no information to be further used or reported. A number of teachers experience difficulties in documenting, due to their limited understanding of the purpose, importance, process and effective use of documentation, in addition to the lack of resources and predetermined curricular guidelines (Kroeger & Cardy, 2006). Rinaldi (2006) views documentation as a "visible trace and a procedure that supports learning and teaching, making them reciprocal because they are visible and sharable" (p. 100). Gandini and Goldhaber (2001) correlate documentation with the cycle of inquiry, suggesting that teachers use documentation to explore questions and examine children's thinking as well as plan, project and respond to situation and ideas. Documentation is also seen as the representation of students' and teachers' communicative abilities (Abramson & Atwal, 2003; Abramson, 2006) and an excellent tool for communication with parents (Goldhaber & Smith, 2002). Documentation allows evidence of performance to be available for future use, interpretation and revision and it also aids in the identification of gaps in students' learning (Goldhaber & Smith, 2002). Effective documentation requires keeping regularly updated records of students' progress and involving students in record keeping (Harlen et al., 1992). Indeed, student-involved record keeping can be a powerful confidence builder as well as a mirror permitting students to watch themselves grow (Stiggins & Chappuis, 2005). However, there is no point in gathering information unless it can be acted upon (Black, 1993; Black &

Wiliam, 1998). Teachers are considered the primary users of information gathered in classroom assessments. Thus, after recording assessment information, teachers need to make decisions on how this information will be used. In practice, teachers must use assessment results to make responsive changes to instruction and learning (Popham, 2006); these changes must be early enough in the decision-making process, in order to actually influence student learning (Stiggins & Chappuis, 2008).

Summarily, skills included in this phase refer to skills associated with documentation of assessment results (Kroeger & Cardy, 2006), the eliciting information (Schmoker, 2006) as well as how this information is used (Stiggins & DuFour, 2009).

Reporting Results to Students and Parents

The last phase refers to skills related to the communication of assessment results to intended users. The communication of assessment results bridges the gap between the recorded data and their actual interpretation and use by the involved participants. In order for intended users to actually act upon assessment information, they must first be made aware of such information. The process of communicating or reporting assessment results entails two basic decisions: the first being what purpose is intended to be served through the assessment and the second being which are the best reporting methods or tools to fulfill this purpose. Reporting procedures deliver assessment results into the hands of the various intended users of the information in a timely and understandable manner (Stiggins, 2004; Roeber, 2003) and enhance the continuity and quality of students' learning experience (Berry, 2008). They also provide all intended users of assessment with knowledge of results that can be later used to make adjustments to support learning. Various methods can be used to report students' learning progress. The method or methods selected must be in alignment with the purpose the assessment wishes to serve and must be used appropriately to serve this purpose. In addition, Stiggins (2004) suggests that effective communication of results occurs when everyone

understands the meaning of the achievement target and the symbols used to convey information, when the information underpinning the communication is accurate and finally when the communication is tailored to the intended audience in the aspects of timing, detail and format (p. 17).

Summarily, skills included in this phase refer to decisions concerning the purpose of reporting (Guskey & Bailey, 2001; Harlen & James, 1997), the audience of reporting (Stiggins, 2004), the instruments used to report data (Guskey & Bailey, 2001) as well as the quality of teacher communication with parents and students (Stiggins, 2004).

Educational Assessment: Issues of Effectiveness

Research supports that teacher effectiveness relates to the extent to which teachers use assessment for formative rather than summative purposes (Creemers & Kyriakides, 2008; Hattie & Temperley, 2007; Wiliam et al., 2004). International research supports the idea that tracking a student's progress toward objective learning goals is more effective than its comparison with peers' progress (Cameron & Pierce, 1994; Kluger & DeNisi, 1996). For example, in their study Hattie and Temperley (2007) obtained high effect sizes when students were given 'formative feedback'; feedback on how to perform on a task more effectively. On the other hand, they obtained far lower effect sizes when students were given praise, rewards or punishment. In Bangert-Drowns, Kulik and Kulik (1991) metanalysis study, the relationship between the frequency of assessment and student achievement was investigated and it was found that the systematic use of assessment as a form of feedback has a positive effect on students' outcomes.

Formative practices have also been associated with school effectiveness. Particularly, research shows that schools with an assessment policy focused on the formative purposes of assessment are more effective (e.g., de Jong, Westerhof, & Kruiter, 2004; Kyriakides, 2004). Formative assessment has been recognized as a determining factor of educational

effectiveness at both classroom and school level. Especially, in the context of Cyprus, teachers' emphasis on formative assessment regarding classroom level was found to be associated with teacher effectiveness (Kyriakides & Creemers, 2008a), whereas regarding school level, studies revealed that schools are more effective when the school policy promotes the formative purpose of classroom assessment (Kyriakides, 2005; Kyriakides, Campbell & Gagatsis, 2000). Similar results have also been reported in different contexts (Teddlie & Reynolds, 2000), as well as through metanalyses of school effectiveness studies (Kyriakides et al., 2010).

Teachers' Skills in Assessment

The growing accountability framework, the standard-based movement as well as the emphasis on effective classroom assessment practices have resulted in an increased need for teacher competency in the area of student assessment and evaluation. The review of the literature reveals that although teachers spent a large amount of teaching time in assessment related activities (Herman & Dorr-Bremme, 1982; Crooks, 1988; Stiggins, 1991; Stiggins & Conklin, 1992), they lack the necessary knowledge and skills to effectively fulfill their role as assessors during everyday classroom assessment (Lukin at.al., 2004; Schafer, 1993; Stiggins, 2002).

The concept of "assessment literacy" was introduced, while recognizing the need for assessment informed teachers, and it has been generally defined as an understanding of the principles of sound assessment (Popham, 2004; Stiggins, 2002). Teachers' ability to develop action plans, alter instruction and other factors in order to improve student learning is also recognized as an important component of assessment literacy (Fullan, 2000). The concept of assessment literacy entails that teachers must possess a set of assessment knowledge and skills. Acknowledging the need for clear standards of assessment literacy, the American Federation of Teachers, the National Council on Measurement in Education, and the National

Education Association (1990) developed a set of seven core competencies in assessment that teachers must possess. According to these standards, teachers should be skilled in choosing and developing assessment methods appropriate for instructional decisions, as well as in administrating, scoring and interpreting the results of both externally-produced and teacher produced assessment methods. In addition, teachers should be skilled in using assessment results, while making decisions about individual students, planning teaching, developing curriculum and school improvement. The fifth standard requires teachers to be skilled in developing valid pupil grading, while the sixth standard requires teachers to be skilled in communicating assessment results to students, parents, other lay audiences and other educators. Finally the seventh standard requires teachers to be skilled in recognizing unethical, illegal and otherwise inappropriate assessment methods and uses of assessment information.

Following the 1990 standards, other attempts to define criteria for assessment literacy were made. Schafer (1991) specified eight content areas in which teachers need to develop assessment skills; these content areas include basic concepts and terminology of assessment, uses of assessment, assessment planning and development, interpretation of assessments, description of assessment results, evaluation and improvement of assessments, feedback and grading and finally ethics of assessment. In an article, published in 1995, Stiggins also emphasizes the importance of having clear standards to define teacher assessment literacy and proposed a set of five standards to describe the concept of assessment literacy. He further argues that in order for teachers to be able to assist their students attain higher standards of academic achievement, they must be able to identify a clear purpose of assessment, focus on achievement targets, select proper assessment methods, sample student achievement and finally avoid bias and distortion. Stiggins (1998) also identifies gaps in the 1990 Assessment Competencies Statement, arguing that these statements ignore the use of assessment as an instructional intervention, the potential of student involvement as well as the connection

between assessment and student motivation. In 1999, Stiggins adds that assessment literate teachers should be able to understand what assessment methods to use, and when to use them in order to gather dependable information about student achievement. He further states that teachers should be able to communicate assessment results effectively to all intended users – including principals, other teachers, parents and students – whether they are using report card grades, test scores, portfolios, or conferences - and finally be able to understand how to use assessment to maximize student motivation and learning by involving students as full partners in assessment, record-keeping and communication.

Twenty one years after the publication of the 1990 Standards for Teacher Competence in Educational Assessment of Students (AFT, NCME, & NEA, 1990), Brookhart (2011) suggests an updated list of knowledge and skills that teachers need to apply on the assessment-related aspects of their work in a competent and professional manner. She argues that an update is necessary, since the two main developments in the area of educational assessment over the past years, the formative assessment and the standards-based accountability, have not influenced the content of the 1990 Standards. The researcher argues that although there are not any specific statements about "formative assessment" or "standards-based accountability", in the standards list she suggests, knowledge and skills in several of the statements apply to these areas. The updated list includes 11 standards statements that describe the necessary assessment knowledge and skills of teachers. First, teachers should understand learning in the content area they teach. Secondly, teachers should be able to articulate clear learning intentions that are congruent with both the content and depth of thinking, implied by standards and curriculum goals, in such a way that they are attainable and assessable. Thirdly, teachers should have a repertoire of strategies for communicating to students what an achievement of a learning intention looks like. Fourthly, teachers should understand the purposes and uses of the range of available assessment options and be skilled in using them. Fifth, teachers should

have the skills to analyze classroom questions, test items and performance assessment tasks in order to ascertain the specific knowledge and thinking skills, required for students to undertake. Sixth, teachers should have the skills to provide effective, useful feedback on student work. The seventh standard states that teachers should be able to construct scoring schemes that quantify student performance on classroom assessments into useful information for decisions about students, classrooms, schools, and districts. By following these decisions, it is expected to lead to improved student learning, growth, or development.

Next, the eighth standard states that teachers should be able to administer external assessments and interpret their results for decisions about students, classrooms, schools, and districts. The ninth standard states that teachers should be able to articulate their interpretations of assessment results and their reasoning concerning the educational decisions, based on the assessment results of the educational populations they serve (student and his/her family, class, school, community). The tenth standard requires teachers to be able to help students use assessment information in order to make sound educational decisions. Finally, the last standard states that teachers should understand and carry out their legal and ethical responsibilities in assessment, while conducting their work.

The detailed lists of teachers' assessment competencies, created under the assessment literacy movement, rest on the assumption that teachers must hold a set of competencies in order for them to be able to effectively assess their students. The lists' focus relies mainly on describing what teachers should be able to do in their classrooms, during assessment. However, assessment is an on-going process, integrated with teaching and infused in the everyday classroom life (Birenbaum, 2003; Cowie & Bell, 1999; Guskey, 2003). Presenting effective assessment practice as a list of isolated skills encourages a view of assessment as a process independent from teaching and breaks down what it is essentially; a continuous process. In addition, despite the argument that assessment literacy is an important competency

for teachers, the review of the literature highlights the lack of research on teachers' assessment skills and how these can be improved. Most research examines assessment literacy at pre-service level (see for example DeLuca & Klinger, 2010; Volante & Fazio, 2007), recognizing the inadequacy of teacher degree programs to sufficiently prepare future teachers in assessment issues. Furthermore, research investigating in-service teachers' assessment literacy (see for example Dekker & Feijs, 2005; Plake, Impara & Fager, 1993) is limited and provides no empirical evidence on how teachers' competency in assessment may be improved.

Finally, although the formative purpose of assessment has been widely promoted (Gipps, 1994; Stiggins, 1999; Shepard, 2000; Stobart, 2004; Popham, 2006) and the need for assessment literate teachers, able to design and administer more than summative end-of-unit tests (Green & Mantz, 2002; Shepard, 2000) has been highlighted, assessment research literature has failed to impact teachers' everyday assessment practice, which still appears to be outcome - oriented (Earl & Katz, 2000; Lock & Munby, 2000). Most attempts that aim for improvement in teachers' assessment practice focus on training teachers in the use of assessment strategies, recognized as beneficial to students' outcomes (e.g., Black & Wiliam 2005; Black et al. 2006). However, until today, there has been no systematic empirical evidence to describe in detail the skills related to effective assessment practice and how these can be developed.

As this section has shown, achieving effective assessment practice appears as a controversial issue in the literature. The need to define assessment skills, taking in mind the process and the purposes of assessment, as described in the literature, is highlighted. Defining the skills necessary for effective assessment practice to be achieved, could help us design professional development programs, aiming for improving assessment practice. The present study recognizes the importance of a professional development in enhancing the quality of

teaching and learning in schools. Thus, the next section presents a review of the literature on teacher professional development.

Teacher Training and Professional Development

Teacher training and professional development are considered important components of any effort to create effective schools (Smith & O'Day, 1991). Even more today, there is a wider recognition of the importance of professional development in equipping teachers to meet the numerous challenges, faced by our educational systems and education in general (Darling-Hammond, 2000; Guskey, 2003). Given the lack of research on professional development with particular reference to assessment, this section will draw on literature concerning teacher professional development in general, presenting the main approaches as these are recognized in the literature.

Main Approaches to Teacher Professional Development

Teacher professional development is considered an essential mechanism for deepening teachers' content knowledge and developing their teaching practices in order to teach to high standards (Borko, 2004). As a result, various systems and paradigms of professional development appear in the literature.

Ingvarson (1998) refers to the traditional system of professional development, mostly known as "in-service training", by comparing it with the "standard-base system". The first system incorporates models of professional development in which an outsider - usually the government, through its educational authorities- holds the control, defines the goals and provides short period workshops. On the other hand, the second system incorporates models in which professional bodies are in control and in which opportunities are clearly oriented to real needs, identified by teachers themselves. Cochran-Smith and Lytle (2001) describe three systems of professional development, which as they argue, co-exist in the educational setting; each one of them representing a different theorization of how improvement in learning can be

achieved. The first system, referred as “knowledge-for-practice”, is based on the assumption that university-based researchers generate knowledge and theory for teachers to use in order to improve their practice. The second system, “knowledge-in-practice”, rejects the idea of formal knowledge and recognizes practical knowledge, suggesting that most essential teaching knowledge is embedded in practice. The third system, “knowledge-of-practice”, also sees practice as central but only when teachers reflect on it to learn more on effective teaching.

Among the various systems and paradigms employed in teacher education and development, the Holistic or Reflective Approach (HA) and the Competency-Based Approach (CBA) are the dominant approaches to teacher professional development (Zeichner, 1983). Their main theoretical assumptions as well as their strengths and weaknesses are described below.

The Holistic/Reflective Approach to Teacher Professional Development

The dominant approach in teacher professional development today is focused on encouraging reflection of teaching practices, experiences, and beliefs (Golby & Viant, 2007). The main argument of the reflective paradigm is that “theory often fails to inform practice because the problems that arise in practice are generally neither caused by nor the result of teachers’ lack of knowledge about theory” (Johnson, 1996, p. 766). Based on this argument, reflection is seen as a way for teachers to develop informed practice while critically examining their practices in the classroom.

A large part of the literature identifies the origins of the term to the work of Dewey (1933) and Schon (1983; 1987; 1991). For Dewey (1933), reflection is seen as an active and deliberate cognitive process, which incorporates underlying beliefs and knowledge. He defines it as an action based on “the active, persistent and careful consideration of any belief or supposed form of knowledge in the light of the grounds that support it” (p. 9). For Dewey, reflection is associated with teacher professionalism, since through reflection teachers can

replace mechanistic and routine action with scientifically approved alternatives. On the other hand, for Schon (1983) reflection is an intuitive, personal, non-rational activity. Schon (1983) suggests that we can engage in reflection in either by 'reflecting on action', after the experience, or by 'reflecting in action', during the experience with the latter implying conscious thinking and subsequent modification while still in action. His conceptualization refers to a reflection which is inseparably associated with action. Indeed, for Schon (1987) it is through this interaction of thinking and doing that teachers improve their skills. Thus, knowledge is seen as the direct result of practice and not the type of knowledge that has been based on scientific approaches, as Dewey argues. As Fendler (2003) sums up "these days the meaning of professional reflection is riddled with tensions between Schon's notion of practitioner-based intuition, on the one hand, and Dewey's notion of rational and scientific thinking, on the other" (p. 19).

In this context, different attempts to define the concept of reflective practice based on its components can also be recognized. van Manen (1977) viewed reflection as a comprised of three elements: technical rationality, practical reflection, and critical reflection, while Korthagen (2001) regards reflection as consisting of organized, rational, language-based decision making processes that also include non-rational, gestalt type operations. Jay and Johnson (2002) regard reflective practice as consisting of the three crucial steps of description, comparison, and criticism.

Professional development in the context of reflective practice rests upon constructivist theorizations of learning. Learning is seen as individually and socially constructed in an integration of theory and action, whereas reality is seen as complex, multi-dimensional and multi-faceted. The idea of a single set of correct and effective teaching practices is rejected and professional development is directed towards more inclusive and teacher-involved strategies. Teachers are encouraged to develop their personal theories, theorize what they

practice and practice what they theorize (Kumaravadivelu, 2001). Emphasis is given on approaches that involve reflective capabilities of observation, analysis, interpretation, and decision-making, which enable teachers to critically review their teaching practice (Schon, 1983; Zeichner, 1987). Professional development takes amongst others the form of action research (Sparks-Langer & Colton, 1991; Pugach & Johnson, 1990; Carr & Kemmis, 1986; Zeichner, 1986) and professional learning communities (McLaughlin & Talbert, 2006). In addition, it involves making use of readings of journal writings, observation notes, transcribed conversations, videotaped analyses, and self-regulations (Cornford, 2002).

However, the holistic/reflective approach to teacher professional development has received extensive criticism. First, it has been argued that it is very difficult to define the concept and the practice of reflection (Hatton & Smith, 1995) since contradictory interpretations of the term exist (Cornford, 2002; Fendler, 2003). This has resulted in a confusion on what teachers are supposed to reflect on and how an effective reflection can be achieved. Another issue raised is that of problem identification. Effective reflective practice as described in the literature is bound up with problem identification, since as Schon (1983) states, "professional practice has at least as much to do with finding the problem as with solving the problem found" (p. 18). However, in many cases teachers do not have the ability to recognize what is wrong with their performance in the classroom, and use reflection to justify their actions instead of critically examining them. In addition, it is argued that the over-theorization of reflective practice as well as the assumption that teachers do not reflect unless someone teaches them how, has reduced reflection to a set of techniques (Jay & Johnson, 2002). Thus, whereas teachers are encouraged to reflect on their practice, this reflection is considered appropriate only when it satisfies the specific criteria set. Finally, whereas reflection is promoted as a way to improve the quality of teacher performance in the classroom there has been no clear connection between reflection and teacher effectiveness. Research

shows that reflective practices can have a positive impact on teachers' job satisfaction, teachers' interpersonal relationships and sense of self-efficacy (Braun & Crumpler, 2004); however, there is little empirical evidence to show that engagement in reflection will result in better student outcomes (Cornford, 2002; Korthagen & Wubbles, 1995).

Reflection and critical thinking are, or should be, important elements in all aspects of learning and performance. Yet critical thinking is necessary, but not sufficient (Ottesen, 2007), since it needs to be based on a combination of both knowledge and skills. Knowledge and skills well as the ability to act upon critical thinking are considered prerequisites of effective reflective practice (Cornford, 2002).

Competency-Based Approach (CBA)

Concerns about teacher quality in light of new student demands, the changed nature of the knowledge needed by teachers, and the balance between accountability and professional autonomy (Korthagen &Wubbels, 1995; Day, 2002) have given rise to the question of using competencies as a basis of teachers' education.

Competency-based approaches to teacher education and professional development are by no means recent (Whitty & Willmott, 1991). The development of competency-based education can be traced back to the 1920s, when the drive for technical and rational management systems first came into focus. However, the beginning of the 'modern competency movement' can be located in the late 1960s and early 1970s (Adams, 1996), as a result of various publications on competency-based organisational training and competency-based teacher education in the United States (Popham, 1984; Biemans, Nieuwenhuis, Poell, Mulder, & Wesselink, 2004). Back then, competency-based teacher education relied on a behaviourist model of training and learning with an emphasis on the "ability to demonstrate knowledge". The idea was that concrete, observable, behavioral criteria could serve as the basis for teacher training (Korthagen, 2004). However, many have criticized the focus on

teacher competencies understood as behaviours for favoring those instrumental aspects of teaching that can be subjected to tests of immediate use and applicability. In the recent competency-based movement, a holistic approach is put forward. Competence is regarded as the possession and development of integrated skills, knowledge, appropriate attitudes and experience for the successful performance of one's life roles (Popham, 1984; Korthagen, 2004).

The Competency-based approach (CBA) to professional development recognizes teachers' competencies as a determining factor of effective teaching and learning. In such a paradigm, teacher professional development takes the form of professional skills development. The basic assumption rests on the belief that if a teacher holds adequate knowledge and skills, this adequacy will be reflected on his or her practice and subsequently on the effectiveness of his or her teaching. Sparks and Loucks-Horsley (1990) identified a number of important assumptions, inherent in the training model. Two of these assumptions are: 1) there are behaviors and techniques worthy of replication by teachers in the classroom, and 2) teacher-education students and teachers can learn or change their behaviors to replicate these techniques in their classrooms. The argument in favor of competency-based approaches to professional development focuses on the validity and reliability of assessment practices, on quality assurance, transparency, accountability and also on addressing skills shortages in the workforce (Oberski & McNally, 2007). It is argued that competency statements improve standards by outlining the requirements of teachers through specific statements of required knowledge, skills and attitudes. According to Whitty and Willmott (1991) competency-based approaches to education help to understand better teacher education, provide a clearer role for schools/ colleges in the training process, clarify what beginning teachers can do, and finally enable teacher students to set clearer goals.

Professional development programs, deriving from the competency-based approach on professional development, tend to have a training character (Whitty & Whillmott, 1991). These programs are relatively short term, involving teachers in several hours or several days of workshops, with limited follow-up activities. They are underlined by the assumption that an effective teacher must possess a list of competencies, which can be obtained by his or her participation in training courses. Since the middle of the 20th century, the competency-based approach has resulted in numerous training sessions, aiming at improving teachers' knowledge and skills. This kind of professional development views learning as knowledge transmission and assumes a linear effect of professional competencies on professional practice. In other words, it argues that there are behaviors and techniques that are worthy of replication by teachers in the classroom (Sparks & Loucks-Horsley, 1990). Thus, various lists of strategies have been developed by experts (Sprinthall, Reiman, & Thies-Sprinthall, 1996) which are highly explicit (e.g., how to greet students / praise / ask high level questions) and teachers are expected to master these skills. These lists of competencies or standards of teaching seem to be supported by policy-makers (Becker, Kennedy & Hundersmarck, 2003).

However, strong arguments against the use of competence statements have been put forward (Barnett, 1994; Stronach et al, 1994; Humes, 2001). Essentially, the arguments against the use of competencies focus on the idea that in an attempt to ensure sufficient validity and reliability in the assessment of the teachers, the long detailed lists of skills, which were formulated gradually, resulted in a kind of fragmentation of teacher's role (Hattam & Smyth, 1995; Louden & Wallace, 1993). The competencies set, breaks down what is essentially a continuous process; whereas, putting the bits together does not necessarily describe what it means to teach or be a teacher. Thus, using competencies to describe teachers' work is possible to encourage reproductive rather than transformative teaching (e.g. Porter, Rizvi, Knight, & Lingard, 1992). Another argument is that the lists include a variety of

isolated skills, which cannot be covered thoroughly, no matter how long the training program is. In addition, the view of practice as a "container" (Lave, 1993) has been criticized for separating practice from content (McDermott, 1993), therefore, assuming that content, meaning professional skill, is de-contextualized and can be taught over professional development courses and can later on be applied on the appropriate "container". In more simplistic terms, it is assumed that teachers can be taught the necessary teaching skills outside their workplace and return afterwards to successfully implement them; this way ignoring the school-context in which these skills are to be applied. However, not taking into consideration the specific educational context of teachers may reduce the interest and affect the will and the efforts of the teachers to improve.

Another argument against the competency-based approach is the rather mechanistic procedure of implementing the prescribed guidance for each kind of teacher behavior, which does not allow teachers' critical and creative thinking to be expanded. Moreover, evaluations of the CBA approach suggest that although this approach has always acknowledged the need for flexibility in how the methods are applied in the classroom, yet the training itself does not encourage such flexibility (Sprinthall, Reiman, & Thies-Sprinthall, 1996). Finally, another argument against competency-based approaches is the fact that their effectiveness, in relation to student achievement, has not been scientifically supported. Whereas, research on the short-term impact of CBA has shown that student achievement does improve (Walberg, 1986), the long-term effects of this approach are rather questionable (Cohen & Ball, 1999; Richardson & Anders, 1994). For example, in a four-year study of a very popular staff development program, developed and conducted by Madeleine Hunter, which trained teachers in a structured approach to instruction, Stallings & Krasavage (1986) found that the program's effects on student achievement were minimal.

A distinctive feature of the competency-based professional development theory is the stages through which skills develop. Indeed, the process of learning to teach, the skills involved as well as the changes that a teacher experiences during this process are a common theme of discussion in the area of professional development. A brief review of the associated theories of professional stage development is presented next.

Developmental Stages and Professional Development

Dominant models of how people develop expert skill in professions (see Berliner, 1994; Billet, 2001; Sternberg et al., 2000) share some common characteristics. Based on a traditional view of professional skills as a set of fixed attributes, these models suggest fixed sequences of stages, representing successively higher levels of knowledge and skill acquisition. Even though there are differences in the number as well as the nature of each stage across the various models, as Feiman-Nemser and Remillard (1996) argue the tendency is to suggest " an initial stage of survival and discovery, a second stage of experimentation and consolidation, and a third stage of mastery and stabilization" (p.66). For example, Katz's model (1972) identifies four in-service teachers' development stages. During the first stage of Survival, teachers question their personal and professional competence as well as their desire to teach and have little understanding of their students' needs. The second stage is that of consolidation, during which teachers begin to focus on instruction and the needs of individual students. It is not until the third stage, the renewal stage, where teachers become competent in the practice of teaching students and begin to search for alternative and more effective teaching practices. The final stage of maturity occurs when teachers begin to ask deeper and more abstract questions about the philosophy of teaching and the impact it may have on both in and out the school settings. Similar models of pre- and in-service developmental stages also exist (i.e., Fuller & Bown, 1975; Kagan, 1992).

However, the conceptualization of skill development, as an accumulation of a defined body of knowledge and skills, has received extensive questioning (e.g., Ball & Cohen, 1999; Billet, 2001; Borko & Putman, 1996; Dall' Alba & Sandberg, 1996; 2006). The first critique is that the existence of strong generalized processing ability alone is not a sufficient quality for successful performance. Indeed, various studies (Ceci & Liker, 1986; Schraagen, 1993; Stevenson, 1996; Voss, Tyler, & Yengo, 1983) have provided evidence of the significant role of domain-specific knowledge in complex thinking, rather than general procedures. Billet (2001) attempts to illustrate the interdependence in the relationship between cognitive activities and the social world and identifies six bases from the cognitive literature. First, the domain-specificity of expertise is associated with social practice. Second, the knowledge constructed through problem solving is focused on overcoming barriers set in the social world. Third, the accumulation of procedures and concepts is the result of ongoing engagement with socially-determined tasks. Fourth, transfer is socially and culturally constructed. Fifth, individuals' efforts are relational to social practice, with some tasks demanding more effort than others. And finally, socially determined dispositional factors are related to cognitive structures and activities. Indeed, the container view of practice (Lave, 1993) decontextualizes professional practice and separates professionals from their activities and the situation in which they practice. However, as Dall' Alba and Sandberg (2006) argue practice is not a fixed or static container but has a rather a dynamic nature. As a result, practice varies across contexts, as does what is recognized as skillful performance in each context (Billet, 2001; Borko et al., 1997).

Another critique of stage models refers to the stepwise character of stage progression. As the Dall' Alba and Sandberg's (2006) meta-analysis illustrates, stepwise progression has been assumed without the support of empirical evidence obtained over extended periods of time. In addition, Huberman's (1989) study of teachers' professional life cycle demonstrates a

range of development trajectories, which challenges the possibility of a fixed sequence of professional development. Other studies (eg. Dall' Alba, 2004; Ollis, Macpherson & Collins, 2006; Sanberg, 1994) also raise questions concerning the stepwise development of professional skills. Furthermore, it is argued that stage models lack clarity about what is being developed, at each developmental stage (Dall' Alba & Sandberg, 2006). Indeed, stage models define developmental stages without making clear what each stage entails; this making the promotion of skill development even more difficult, since skillful performance is not defined and thus is difficult to be encouraged.

When compared with previous models, the Dreyfus model (1986) advances our understanding of skill development and addresses some of the critiques mentioned above. Indeed, the Dreyfus model is considered as one of the most advanced and influential models of skill acquisition. The model was proposed by Hubert and Stuart Dreyfus and it was based on their study of chess players, air force pilots, and army tank drivers and commanders (Dreyfus & Dreyfus, 1986). Working in the field of artificial intelligence, the two professors challenged the dominant view of human skill development as explicit rule-following in order to perform a task. On the contrary, their model is developmental and based on situated performance and experiential learning (Benner, 2004). According to the Dreyfus model, practitioners learn within the context of practice and develop their skills according to a progression through five skill levels: a) novice, b) advanced beginner, c) competent, d) proficient, and finally e) expert. At the first level, novices apply explicit context-free rules, instructed by others in order to respond to a given situation or objective. As they proceed to the advanced beginner level, practitioners have already acquired practical experience that allows them to also apply context-specific rules. At the competent level, practitioners are able to choose a plan, goals and strategies for when and how to apply rules and procedures; however still in a detached and deliberate way. This changes as the practitioners reach the proficient level; at this level,

practitioners use previous experience to intuitively assess each new situation. Practitioners, who reach the final level, are considered experts. Their skills are based on deep situational experience, acquired through involvement in a specific skill domain for extended periods. However, not all practitioners manage to reach this level of expertise.

As Dall' Alba and Sandberg (2006) argue the model extends previous models as it recognizes the importance of context for professional skills as well as the importance of experience in practical work situations, for advanced skill levels to be achieved. They continue by adding that in the Dreyfus model only those at lower skill levels approach each situation in a detached and deliberate way, whereas practitioners at more advanced levels approach each new situation intuitively. Although extremely influential, the Dreyfus model has also been widely criticized mainly for its inadequacy to reveal differentiation within one stage (Borko et al., 2000; Sandberg, 1994; 2000; Dall' Alba & Sandberg, 2006), even though empirical evidences support such a claim (Livingston & Borko, 1989; Engeström & Miettinen, 1999). Another inadequacy recognized is that the model fails to provide empirical evidence of the stepwise progression of skill development (Dall' Alba & Sandberg, 2006).

As the review of the literature has shown, the inability of both the holistic and the competency-based approaches to provide adequate evidence of their positive effect on teaching and learning has turned the attention of the professional development community to alternative theoretical paths. In this context, the Dynamic Integrated Approach was proposed.

The Dynamic Integrated Approach (DIA) to Teacher Professional Development

Both prevalent paradigms of professional development presented above have been criticized extensively. There is little empirical evidence to support the effectiveness of reflective or competency based approaches in promoting effective teaching, with advantages and disadvantages recognized in both approaches. In addition, research on teacher training and EER has been conducted apart from and without much reference to one another. Few

researchers of teacher training methods rationalize their selection of teaching skills in terms of EER, and very few evaluate the impact of teacher professional development on student learning. Taking this into consideration a Dynamic Integrated Approach (DIA) was recently proposed (Creemers & Kyriakides, 2012).

Theoretical and methodological advances in Educational Effectiveness Research (EER) and more specifically, the dynamic model of educational effectiveness (Creemers & Kyriakides, 2008) set the framework upon which the DIA is based. The dynamic model is established in a way that helps policy makers and practitioners improve educational practice by encouraging rational decision-making, concerning the optimal fit of the factors within the model and the present situation of the factors in the schools or educational systems. The DIA was, therefore, developed in an effort to facilitate the use of the model for improvement purposes.

The Dynamic Integrated Approach is based on the assumption that teacher improvement efforts should aim at the development of teaching skills, which relate to positive student outcomes. It is argued that teacher training and professional development should be focused on how to address specific groupings of teacher factors in relation to student learning rather than to an isolated teaching factor (as proposed by the CBA) or to the whole range of teacher factors (as implied by the HA), without considering the professional needs of student teachers and teachers. Therefore, the DIA lies between the two dominant approaches (i.e., the CBA and the HA) and aims to overcome their main weaknesses. Particularly, the dynamic dimension of this approach is attributed to the fact that its content derives from the grouping of teaching skills included in the dynamic model, and it is differentiated to meet the needs and priorities of teachers at each developmental stage. The integrated dimension of this approach is also attributed to the fact that, although the content of DIA refers to teaching skills that were found to be positively related with student achievement, the participants are also engaged into

systematic and guided critical reflection on their teaching practices. The main steps of the DIA, as well as the assumptions upon which each step is based on are presented next.

The Main Steps of the DIA

This section demonstrates the basic steps needed in order to develop a DIA to teacher professional development. During this process each teacher is expected to develop his/her own strategies and action plans for improvement. Therefore, a teacher is treated as being responsible for designing and implementing his/her own improvement strategies and action plans. However, at the same time the DIA acknowledges that support needs to be provided in order for teachers to be able to achieve improvement. Therefore, an external team, the Advisory and Research Team (A&RTeam), as Creemers, Kyriakides and Antoniou (2013) name it, is necessary in order to provide technical expertise as well as available knowledge-base on improvement of teaching factors. In addition, teachers are encouraged to use other available resources within and outside the school.

Step 1: Identify needs and priorities for improvement through empirical investigation. The first step of the proposed approach is based on the assumption that teacher improvement efforts should refer to the development of teaching skills related to student outcomes. The DIA suggests that evaluation data are required in order to identify the needs of each teacher participating in the improvement project. Thus, in any effort to train teachers, an initial evaluation of their teaching skills should be conducted in order to investigate the extent to which they possess certain teaching skills, whilst identifying their needs and priorities for improvement. The results of the initial evaluation will provide suggestions for the content of training, which are required for different groups of teachers, in order for the training to correspond to the professional needs and proximal development of each group of teachers, as denoted by their own stage of teaching skills.

Step 2: Provide guidelines for improvement: The role of the A&RTeam. The second step relates to the provision of appropriate material and specific guidelines for designing their improvement action plans. The A&RTeam is expected to support and guide teachers' improvement efforts by providing supporting literature, research findings as well as clear instructions, related to the area on which each group should concentrate for improvement. Teachers are expected to adopt and customise the provided guidelines to the specific context of their classroom and develop their own action plan for improvement, following the guidelines provided by the A&RTeam.

Step 3: Establish formative evaluation mechanism. The next step of the teacher professional development programme, based on the grouping of the factors of the dynamic model, comprises the establishment of formative evaluation procedures. The formative evaluation procedures involve: the identification of the learning goals, intentions or outcomes, and criteria for achieving them; the provision of effective, timely feedback to enable teachers advance their learning; the active involvement of teachers in their own learning, and lastly teachers responding to identified learning needs and priorities by improving their teaching skills.

Step 4: Establish summative evaluation mechanism. The final step of a DIA professional development programme is the establishment of a summative evaluation mechanism. The DIA suggests that summative evaluation is necessary in order to identify the overall impact of the programme on the development of teachers' skills and its indirect effect on student learning. The results of the summative evaluation will assist in measuring the effectiveness of this approach and allow subsequent decisions to be made regarding the continuity of the programme. This implies that at the end of the school year, teaching skills and student outcomes should be measured.

As the above description of the DIA suggests a measurement of teachers' skills is necessary in order for appropriate professional development programmes to be designed. The dynamic model of EER, based on which the DIA was developed, acknowledges effectiveness factors as multidimensional constructs and proposes a measurement framework upon which each effectiveness factor can be measured. Given the focus of this study on classroom assessment, a description of how the measurement framework proposed in the dynamic model applies for the factor of classroom assessment follows.

The model suggests the measurement of each factor by using five dimensions of frequency, focus, stage, quality and differentiation. The frequency dimension is a quantitative way to measure the functioning of each factor and refers to the quantity of a classroom assessment related activity that is present in a classroom. For example, frequency can be measured in terms of the number of assessment tasks teachers administer to students. On the other hand, the remaining four dimensions examine qualitative characteristics of classroom assessment. More specifically, focus is measured by looking at the ability of a teacher to use different ways of measuring student skills rather than using only one technique (Rao, Collins & DiCarlo, 2002). It also is important to examine whether the teacher makes more than one use of the information she/he collects (e.g., identify needs of students, conducting self-evaluation, adopting his/her long-term planning, using evaluation tasks as a starting point for teaching) (Black & Wiliam, 1998). The stage dimension is measured by investigating the period at which the evaluation tasks take place (e.g., at the beginning, during, and at the end of a lesson/unit of lessons) and the time lapse between collecting information, recording the results, reporting the results to students and parents, and using them for planning lessons. Quality is measured by looking at the properties of the evaluation instruments used by the teacher, such as the different forms of validity, the internal and external reliability, the practicality, and the extent to which the instruments cover the teaching content (Cronbach,

1990). The type of feedback the teacher gives to his/her students and the way students use the teacher feedback is also examined. Finally, differentiation in relation to the extent to which teachers use different techniques for measuring student needs and/or different ways to provide feedback to different groups of students by taking into account their background and personal characteristics is examined. Using this measurement framework implies that the factor of classroom assessment should not only be examined by measuring how frequently the factor is present (i.e., through a quantitative perspective) but also by investigating specific aspects of the way the factor is functioning (i.e., looking at qualitative characteristics of the functioning of the factor).

Summarily, the DIA rests on the assertion that neither competence nor reflection should be ignored if effective professional development is to be achieved. The potential of overcoming the apparent disadvantages, resulting from the competency-holistic dichotomy as well as research findings that support the effectiveness of such an approach in relation to student outcomes (Antoniou, 2009), lie behind the decision to further examine the effectiveness of the DIA. Support of its effectiveness on the development of general teaching skills has been provided (Creemers, Kyriakides & Antoniou, 2013), however, given the focus of the present study on assessment, it was considered necessary to examine whether the DIA can also have a positive impact when assessment skills are in focus. Although the DIA has been found more effective than the holistic approach, more research is required in order to examine its effectiveness in comparison to the competency-based approach. In order to do so, a theoretical framework for measuring teacher assessment skills was developed and is presented next.

A framework for Investigating Classroom Assessment

Recognizing the need for a comprehensive framework based on which skills associated with classroom assessment can be defined and measured, a framework based on the process

and the purposes of assessment, as these are described in the literature, was developed. The proposed framework takes into account the dynamic nature of assessment; thereby, skills that are associated with each phase of assessment are examined. In addition, assessment skills are defined and measured in relation to teachers' ability to use specific assessment techniques in order to measure different learning outcomes in mathematics. Traditional as well as alternative assessment techniques are taken into consideration, since the literature supports the use of a combination of assessment techniques to assess student learning (Suurtamm et al., 2010). Moreover, a measurement framework developed within the field of Educational Effectiveness Research (EER) is adopted and both quantitative and qualitative characteristics of the assessment process are taken into account. Finally, teachers' skills to use assessment results for formative purposes are taken into consideration. Each aspect of the framework is described below.

a) Main Phases of the Assessment Process

As previously mentioned, classroom assessment is frequently presented in the literature as a cycle, subdivided into a number of phases (e.g., Calfee & Masuda, 1997; National Council of Teachers of Mathematics [NCTM], 1995), most commonly these being: planning, gathering and interpreting evidence and use of results. Other important and distinctive aspects of the process, such as the construction of assessment tools (De Lange, 1993), assessment administration (Shepard, 2007), recording of assessment information (Kroeger & Cardy, 2006) and communicating assessment results (Stiggins, 2004) are also discussed in the literature. The literature highlights the dynamic relationship among the various phases of the assessment process (Black &Wiliam, 2009). In order to measure teachers' assessment skills, this study takes into account the four main phases of the assessment cycle as these have been presented earlier (see Figure 1). The division of the assessment process in particular phases is done in order to make sure that each aspect of

assessment practice is taken into account in measuring teacher skills. This division also helps us test the validity of the instrument, measuring assessment skills.

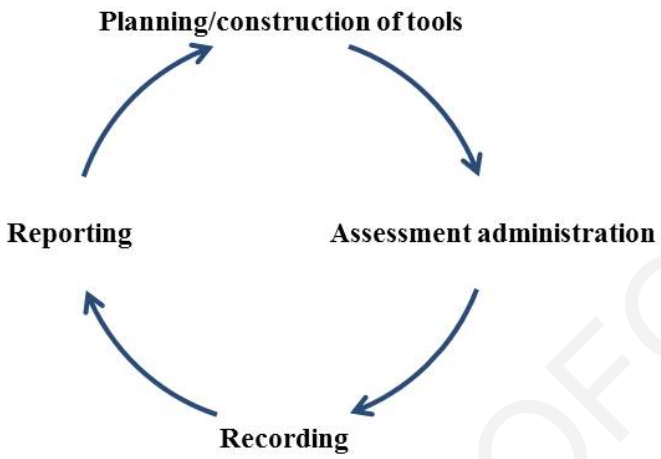


Figure 1. The assessment cycle illustrating the phases of assessment

These four phases are based on the assumption that effective teachers should make sure that: (i) appropriate assessment instruments are used to collect valid and reliable data,(ii) appropriate procedures in administering these instruments are followed, (iii) the data emerged from assessment are analyzed and recorded in an efficient way and without losing important information, and (iv) the results of assessment are reported to parents and students and help them take decisions on how support to students can be provided in order to improve their learning outcomes.

b) Assessment Techniques

Assessment techniques hold an important role in ensuring the quality and effectiveness of assessment, since they usually have an influence on how and what students learn. Current thinking in assessment also recognizes that a variety of assessment strategies needs to be employed, as learning is multidimensional and cannot be adequately measured by one instrument (Brookhart, 2003; Gipps, 1994). Therefore, teachers are encouraged to use a

variety of assessment strategies to provide students with multiple opportunities to show what they know and can do and to further provide insights into students' thinking (Moss, 2003; Shepard, 2001).

Choosing an assessment technique depends on the target to be assessed, since student achievement in relation to certain targets can be more appropriately measured by using specific techniques (Stiggins, 1992). For example, assessment of students' skills in oral communication requires the use of assessment techniques rather than the use of written tests. In addition, the use of a variety of techniques allows students to demonstrate different types of learning. This stands especially in the case of mathematics, since current views of effective mathematic instruction value the complexity of mathematics (Boaler, 2008) and require teachers to be able to use a variety of techniques to assess students' conceptual understanding as well as their problem-solving and reasoning abilities (Suurtamm et al., 2010). Given the development of alternative assessment methods as well as the re-conceptualization of existing traditional methods (Green & Mantz, 2002), it was considered necessary to examine assessment skills in relation to the four most common types of assessment techniques: a) written assessment, b) oral assessment, c) observation and d) performance assessment. For example, it was examined whether different types of written questions were included in teacher tests in order to examine the quality dimension of written assessment. The frequency of the use of formal and/or informal oral assessment to measure student achievement in mathematics was also examined.

c) Measurement Dimensions

Given that the Dynamic model of EER described treats teacher assessment as a factor associated with student achievement, it was considered relevant to take into account the measurement framework, proposed in measuring assessment skills. Specifically, the following

five dimensions used in the model to measure the functioning of each classroom factor were used: a) frequency, b) focus, c) stage, d) quality and e) differentiation. A description of the five dimensions and how these apply to the factor of assessment was provided earlier. It is important to note that these dimensions contribute to the effects that a characteristic of an effective teacher is expected to have on student outcome measures (Creemers & Kyriakides, 2008). Moreover, they help us describe, in a better way, the functioning of each characteristic of effective teachers. Frequency is a quantitative way to measure the functioning of each effectiveness characteristic, whereas the other four dimensions examine qualitative aspects of each characteristic. In addition, the dimensions are not only important from a measurement perspective but also, and even more, from a theoretical point of view. Actions of teachers, associated with each effectiveness characteristic, can be understood from different perspectives and not only by giving emphasis on the number of cases the actions occur in teaching and assessing their students. Moreover, the use of these measurement dimensions may help us develop strategies for improving assessment, since the feedback given to teachers could refer not only to the quantitative but also to the qualitative characteristics of their assessment practice.

Figure 2 shows the theoretical framework that was used in measuring teacher assessment skills. Specifically, each of the four assessment phases was defined based on the assessment knowledge and skills involved across the five dimensions of the dynamic model and in relation to the four most common assessment techniques.

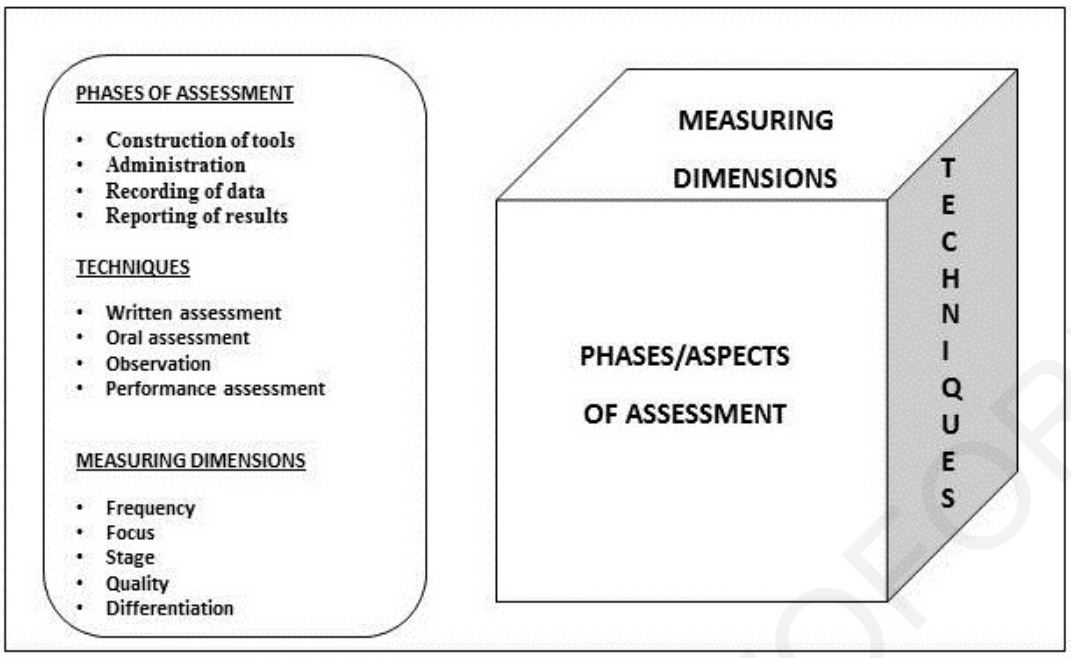


Figure 2. A framework for measuring teacher assessment skills

Research Agenda

The review of the literature presented in this chapter suggests that assessment can be a powerful force in supporting learning and also a mechanism for individual empowerment (Broadfoot & Black, 2004). The recognition of assessment as a key lever for promoting effective education has led to classroom assessment, being a centrepiece of various educational improvement efforts. Although the formative purpose of assessment has been widely promoted (Gipps, 1994; Popham, 2006; Shepard, 2000; Stiggins, 1999; Stobart, 2004) and even though the need for assessment literate teachers, who are able to design and administer more than summative end-of-unit tests (Green & Mantz, 2002; Shepard, 2000) is further highlighted, assessment research literature has failed to impact teachers' everyday assessment practice, which still appears to be outcome - oriented (Earl & Katz, 2000; Herman, Osmundson, Ayala, Schneider, & Timms, 2006; Lock & Munby, 2000).

Changing assessment practices is not a simple process, and some argue that it is more difficult than changing other teaching practices (Borko et al., 2000; Earl & Katz, 2000). The literature highlights the role of teacher training and professional development in any attempt to change teachers' classroom practices. Even though, a direct and clear cut relationship between professional development and student improvement has been assumed, there is little empirical evidence that supports the effectiveness of professional development approaches in promoting effective teaching (Cochran-Smyth & Zeichner, 2005; Guskey & Sparks, 2004). Even more, in the case of assessment targeted professional development there is inadequate solid empirical evidence to describe the change in teachers' actual assessment practice resulting from the received professional development (Kyriakides & Kelly, 2003).

This can be partly attributed to the fact that teachers' skills in assessment and how these can be developed were not taken into consideration. The literature investigating professional development in generic teaching skills reveals two dominant approaches; the competency-based and the holistic approach, with advantages and disadvantages recognized in both. Opposed to this classical dichotomy, the Dynamic Integrated Approach suggests a combination of the two approaches in an attempt to overcome their main disadvantages. Particularly, the DIA recognizes the need for a body of knowledge and skills, upon which teachers can critically reflect. It also incorporates recent findings revealing the grouping of effectiveness factors.

Based on the review of the literature, the present study makes the argument that teacher skills in assessment must be examined, prior to any attempt to improve classroom assessment practice. Drawing on research on classroom assessment and teacher developmental theory (Berliner, 1994; Dall' Alba & Sandberg, 2006), it examines whether developmental stages can be identified when teachers' assessment skills are under investigation. In addition, given the evidence supporting that the Dynamic Integrated Approach is more effective than

the Holistic Approach, in terms of bringing about improvement in teachers generic teaching skills, this study attempts to move a step forward. Particularly, it examines whether the DIA can also improve assessment skills and compares the effectiveness of the DIA approach, this time, in relation to the Competency-based approach.

Summing up, this study concentrates on questions that still require further investigation. Particularly, this research aims to investigate the following questions: (a) Can teachers be classified in distinctive developmental stages based on their assessment skills? (b) How can these stages be defined? (c) Do these stages describe overall assessment practice across the four aspects of the assessment process or are there differentiations between each aspect of assessment? (d) To which extent can teachers' stages in assessment be associated with student achievement? (e) Are teachers, who use more advanced types of behaviour, more effective than those demonstrating the relatively easy types? And finally (f) Which of the two professional development approaches has greater impact on the improvement of teachers' assessment skills and on the learning outcomes of their students?

In order to provide answers to the questions stated, a two-phase research study was conducted. The research design, the participants and the research methods for both phases of the study are presented in the following chapter.

CHAPTER 3

METHODOLOGY

This chapter presents the methodology used to examine the research questions set. It describes the methodology adopted, detailing and justifying the research design employed. It then describes in detail the processes of sampling and data collection with particular reference to the data collection instruments used. The different phases of the study are presented and the statistical techniques employed, during the analysis process are described. Finally, possible limitations, related to the methodological design of the study, are discussed.

Research Design and Justification of the Methods Chosen

The study included two main phases. The first phase examined teacher skills in using different techniques of assessment in mathematics and investigated whether teacher assessment skills can be grouped into different developmental levels. The second phase of the study aimed to examine the impact of two different professional development approaches on teachers' assessment skills and students' achievement. Table 1 presents the study timeframe of the study.

Phase 1

This study argues that prior to any attempt to improve classroom assessment practice teacher skills in assessment must be examined. Thus, the first phase of the study investigated the extent to which assessment skills can be grouped into different developmental stages. By taking into account the theoretical framework and its dimensions presented in Chapter 2, a teacher questionnaire was developed in order to measure teachers' assessment skills. In addition, the questionnaire was designed to measure explanatory variables, such as gender and years of experience. Details on how the questionnaire was developed are presented in the next section.

Table 1.

Study timeframe

Phase 1	May 2011	• Questionnaire Pilot study
	June 2011	• Final version of the questionnaire
	September 2011	• Teacher Questionnaire Administration • Teacher Interviews • Stage investigation/identification
Phase 2	October 2011	• Open invitation sent to participants of the 1st phase • Formation of control group (n= 102) and experimental group (n=76) • Random assignment of teachers of the experimental group into two groups
	November 2011	• Workshop session 1 • Student pre-test administration
	December 2011	• Workshop session 2
	January 2012	• Workshop session 3
	February 2012	• Workshop session 4
	March 2012	• Workshop session 5
	April 2012	• Workshop session 6
	May 2012	• Workshop session 7
	May 2012	• Teacher Questionnaire Administration • Teacher Interviews • Student post-test administration

It is acknowledged that the choice of a questionnaire as the prime method of investigation of teachers' assessment skills raises questions concerning the validity of the data gathered. As the framework described in Chapter 2 shows, the assessment process is conceptualized as a four phased process. Given that assessment is not a one instance process but an integral part of the teaching process, the use of a questionnaire was considered more appropriate for measuring a wide range of assessment skills situated at different phases of teacher's practice. For example, skills related to the planning and construction phase cannot be measured during a class observation, since teachers usually construct their assessment instruments outside the classroom; perhaps even at home. In addition, whereas assessment administration takes place during classroom instruction, classroom observation would have given us just a part of the picture; for example, a teacher may use performance assessment to assess his/her students, but he/she may not have used it on that particular day. Furthermore, using classroom observation would have not allowed us to measure skills related to the recording or the reporting of data, since once again these phases usually take place outside the classroom. Moreover, to measure teacher skills in administering assessment tasks, a large number of lessons per teacher have to be observed, since a significant percentage of teachers offer assessment tasks only at the end of a unit or series of lessons; it is therefore unlikely to collect data on teacher skills in assessment, unless a variety of lessons of each teacher is observed. Although the limitations of collecting data through teacher self-reports is acknowledged, it was not feasible to conduct a very big number of observations to ensure generalisability of the data. Document analysis could have been used to overcome some of the limitations mentioned; however, acquiring access to teacher and student data is not an easy task and confidentiality issues are raised. Finally, the large sample population needed, acted as a restriction for the use of teacher interviews as a primary method of data collection.

For the reasons described above, a questionnaire was used as the primary source of data, whereas interviews were used as a secondary source in order to examine the internal validity of the study. In order to minimize potential reporting bias, a big population was used to gather data. In addition, a pilot study (see next section) was conducted in order to ensure that the terminology used in the measures was clear and understandable and that teachers were able to consistently interpret what information the measures requested (Ball & Rowan, 2004). At this point, it is important to note that teachers were asked to complete the questionnaire eponymously. Whereas this might be considered as a thread to the authenticity of the responses, it was considered necessary in order to be able to identify teachers' needs and adjust the professional development program accordingly. In addition, anonymous administration of the questionnaire would not have allowed us to compare initial and final teacher data in order to examine improvement of teacher assessment skills. The fact that the participants were teachers who showed a special interest in improving their assessment skills and also that they were informed that the questionnaire data will be the base for the design of the professional development program to follow, increases the possibility that teachers were sincere in their responses.

Phase 2

The analysis of the data deriving from the first phase of the study showed that when teacher skills in assessment are measured, four developmental stages can be identified. Details on the analysis of the data from the first phase are provided later. Based on this finding, a decision was taken to investigate the effectiveness of different professional development approaches in bringing about improvement in teachers' assessment skills and students' achievement. The second phase of the study adopted an experimental design based on a multiple-treatment research methodology. This second phase served two main purposes: a) to test the generezability of the results of the first phase and b) to examine the effectiveness of

two different approaches to professional development in yielding improvement in teachers' assessment skills and students' achievement.

Experimental research is considered appropriate when searching for cause-effect relationships, where changes in an independent variable produce changes in dependent variables (Cohen, Manion & Morrison, 2007). Especially in the field of EER, it is argued that when correctly implemented, the randomized controlled experiment is a powerful design for detecting treatment effects of interventions (Creemers, Kyriakides & Sammons, 2010). During the second phase of the study, the professional development offered was considered as the main independent variable and it was examined whether this variable had an effect on a) teachers' assessment practice and b) student achievement.

Given that the first phase of the study involved the evaluation of teachers' assessment skills, during the second phase of the study, the four developmental stages identified are used to randomly allocate teachers, who agreed to participate in the second phase of the study, into two even intervention groups. The first group employed the Dynamic Integrated Approach (DIA), whereas the second group employed the Competency-Based Approach (CBA) (see Chapter 2). Teachers of the DIA group received differentiated training on specific assessment skills, related to their developmental stage. On the other hand, in the CBA group, no differentiation of content based on teachers' developmental stage was applied. As a result, all teachers of this group, despite their developmental stage, received the same training in classroom assessment; this training addressing assessment skills recognized across all four stages. Teachers, who did not attend any INSET course, were treated as members of the control group. A detailed description of each intervention is presented later in this chapter. Figure 3 presents the research design employed.

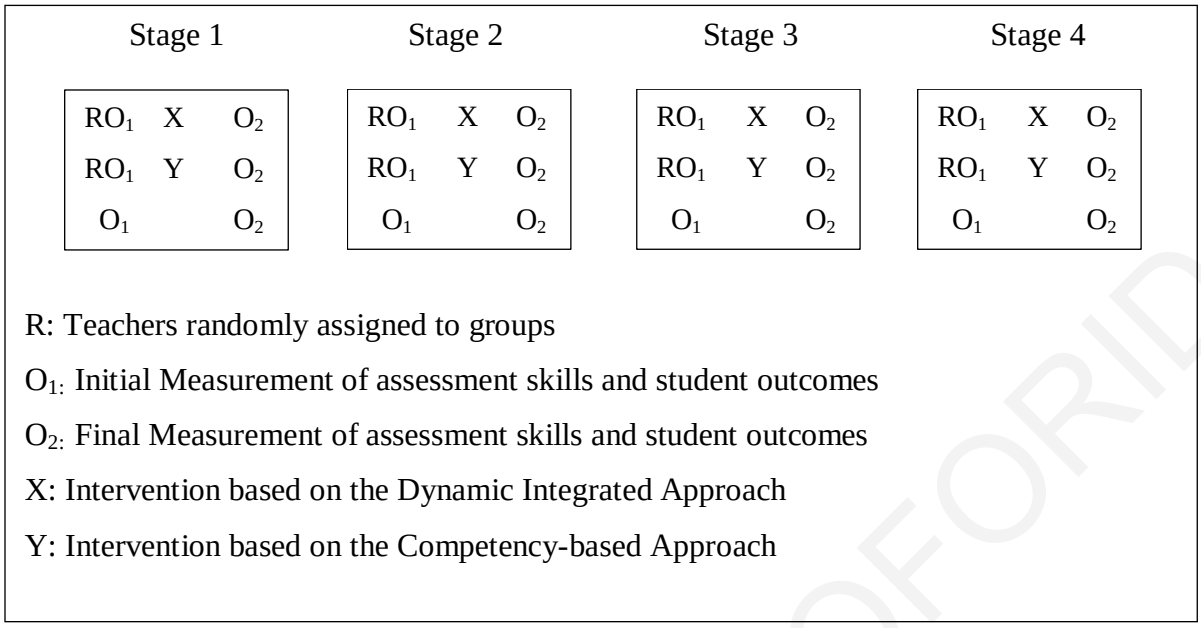


Figure 3. True experimental pretest-post-test control group design

As the figure above shows, participating teachers were assigned to developmental stages, according to the results of their assessment skills evaluation (Stage 1, Stage 2, Stage 3 and Stage 4). Then, the teachers of each stage were randomly assigned into the two groups (R). The teachers of DIA group received differentiated training, corresponding to their developmental stage (X). The teachers of the CBA group received training, covering assessment skills recognized across all four stages (Y). Teachers of the control group received no training. For all three groups, two outcome measures were used: assessment skills and student learning outcomes in mathematics. These were evaluated both at the beginning (O_1) and at the end (O_2) of the interventions.

Research Instruments

In order to examine the research questions, data concerning teachers' assessment skills as well as student performance in mathematics were collected. The instruments used were a) teacher questionnaire, b) teacher interviews and c) student tests in mathematics. Particularly, the same teacher questionnaire was used during Phase 1 as well as during the pre and post

measurement periods of Phase 2 of the study in order to investigate teachers' assessment skills. Teacher interviews were conducted both in phase 1 and 2 of the study in order to examine the validity of the questionnaire data. Finally, student tests were used during phase 2 of the study in order to provide data concerning the impact of the intervention on student achievement. The instruments used are described in more detail next.

Teacher Questionnaire

By taking into account the theoretical framework and its dimensions presented in Chapter 2, a teacher questionnaire was developed in order to measure teachers' assessment skills (see Appendix A). In order to construct the questionnaire, a specification table was first created (see Appendix B). The specification table was used in order to help us define how the various aspects of the assessment process (assessment phases, assessment techniques and measurement dimensions) will be addressed in order for skills, across all aspects to be measured. For example, in order to examine the focus dimension of the administration aspect for all four assessment techniques, items concerning the purposes of record keeping, whether the recording was for an individual student or a group of students, as well as the type of recording notes (specific/general/general trend) were created. The specification table was also used in order to help us categorize the questionnaire items. The items created guided the questionnaire development and later the compilation of the final version of the questionnaire.

The questionnaire consisted of five parts. In part A, teachers were asked to provide information related to their background characteristics (e.g. gender, position, years of experience etc.). In the next four parts (i.e. B, C, D and E), teachers were asked to indicate the extent to which they behave in a certain way, during mathematics teaching in their classroom. A Likert type scale was used to collect data. Teachers were asked to indicate the degree to which they behave in a certain way, by indicating a number from 1 (rarely/never) to 5(very often/always). In more detail, part B consisted of 14 statements, examining teacher behaviour

in relation to the use of written assessment. For example, teachers were asked to indicate how frequently they use ready-made tests and whether they take in mind their students' abilities when constructing written assessment items. Part C consisted of 8 statements, examining teacher behaviour in relation to the use of oral assessment. For example, in this section teachers were asked to indicate how frequently they use oral assessment in order to assess their students in mathematics as well as how they behave when a student has difficulty in answering a question. Part D consisted of 9 statements, which examined teacher behaviour in relation to the use of observation and performance assessment. For example, teachers were asked to indicate how often they decide in advance which students to assess through systematic observation, how often they use performance assessment to assess students' skills, such as the use of the compass, and how often they use observation to assess the procedure a student follows to solve a problem. The last part of the questionnaire, part E addressed specifically the recording and reporting of assessment results. This section included 22 statements. For example, teachers were asked to indicate how often they keep records for each assessment technique and how often during reporting to parents they refer to the child's performance, in relation to his/her classroom's level.

In more detail, each assessment technique was examined in relation to the four aspects of the assessment process (construction, administration, recording and reporting). For each aspect of the assessment process, each of the five dimensions (frequency, focus, stage, quality and differentiation) was applied. For example, in order to measure the quality dimension of the construction of written assessment, question 3 in part B asked teachers to indicate whether they include process questions in their written tests; whereas, question 8 in the same part asked whether a specification table is created, before developing a written test. Similarly, question 7 in part B item was used to measure the differentiation dimension of oral

assessment administration: “All students have the same amount of time available to answer the oral question I ask”.

A pilot study was conducted in June 2010 in order to to ensure that the terminology used in the measures was clear and understandable (Ball & Rowan, 2004). The pilot study involved the administration of the questionnaire to 12 teachers, followed by a personal interview by the researcher. The pilot study revealed only minor remarks concerning the layout of the instrument (i.e. spacing, indentation), as well as difficulties in understanding some statements such as the following: “The written tests I use include questions / activities that are related with each other” (B7). In order to improve this item, the following sentence was added “i.e. students are required to do operations and then create a graph based on their answers”. All remarks were taken into consideration and the final form of the questionnaire was developed.

Teacher Interviews

In addition to the questionnaire described above, semi- structured interviews were used in order to match responses with the questionnaire and ensure the internal validity of the results. It is generally acknowledged that interviews provide contexts where participants can ask for clarification, elaborate on ideas, and explain perspectives in their own words. Thus, by using interviews it was feasible to examine whether there was consistency between teachers’ responses to the questionnaire items and their responses to the interview questions. Semi-structured interviews were considered more appropriate, since this type of interviews allows us to collect detailed information concerning specific topics (Cohen, Manion & Morrison, 2007).

First an interview guide was created (see Appendix C). The interview guide included a list of key themes, issues, and questions to be covered. The first two questions were general. Question A asked teachers to share their opinion on student assessment, to identify prospects

for improvement and more specifically to share their expectations for a professional development program on assessment. Question B asked teachers to indicate the techniques they usually use to assess students in mathematics. Then next four questions were open-ended and were related to the four phases of the assessment process (i.e. construction, administration, recording and reporting). These questions allowed the interviewees to talk freely in relation to their assessment practice across the four assessment phases. Thus, it was feasible to get the responses the individuals gave spontaneously, while avoiding the bias that may result from close –ended questions (Foddy, 1993). A checklist of key themes was available for each question and the interviewer checked each theme covered in the interviewee's response. In case that the interviewee did not refer to some of the key themes included in the checklist, supplementary questions were available for each of the four questions. These were used to draw more information from the respondent in cases where further elaboration was considered necessary. In cases where the responses to the open-ended and supplementary questions did not address all key themes, close-ended questions were used in order to cover all themes on the checklist. This procedure allowed us to gather in-depth evidence of teacher assessment practice. A more specific description of the interview guide follows.

Question 1 asked teacher which procedures they usually follow to construct an assessment tool. The checklist for this question included the following key issues: time spent in the construction, type of questions (product / process), type of questions (objective / multiple choice / fill in / short answer / true-false /open-ended / coupling/ interpretive / layout), construction period, level of difficulty of questions, content representativeness and differentiation based on students and/or objective. In addition, two supplementary questions were provided. Question 1a asked teachers to describe in detail the way they work in order to construct an assessment tool, whereas question 2a asked them whether they consider written assessment to be inappropriate for the assessment of specific students or specific objectives

and finally, how they behave in such occasions. Question 2 asked teachers to describe their behavior during the administration of an assessment. The checklist for this question included the following key issues: frequency of administration of each technique, queries of students / clarifications, individual / group administration, type of guidelines (general / specific), administration period (beginning of a unit / end of a unit / when necessary), set / keep time limits, appropriateness of guidelines (students comprehend what they must do) and time differentiation based on students / objective. Supplementary question 2a asked teachers whether students usually ask clarifying questions during the administration of a written assessment and how they act in response. Supplementary question 2b asked teachers whether there are situations in which they give certain students more time in order to complete an assessment and if so, to comment on the reasons and the ways this is done. Next, question 3 asked teachers to describe the procedure they usually follow to record the results of an assessment by referring in particular for which assessment techniques they usually keep a record and what is usually the form of such a record. The checklist for this question included: frequency for recording results for each type, use of different recording tools, marking (symbolic/numeric), individual / group recording, comments (per student / per class as a total / per objective / per exercise), time gap between administration-recording, formative / comparative use of recorded data and recording differentiation based on students / objective. Supplementary question 3a asked teachers why they record data and describe how they usually go about. The next supplementary question, 3b, aimed to clarify the content of teachers' records and therefore asked teachers to mention in detail the information they include in data recording. Finally, the last question of the interview guide, question 4, asked teachers to describe the procedure they use to report assessment results. In particular, teachers were asked to comment to whom they usually report to, in what ways and during what period. The checklist for this question included the following key themes: reporting users, frequency of

reporting per user, purpose of reporting per user, type of reporting (general / specific / focused), reporting period per user, identification / definition of next steps, eliciting information from users / type of information, communication quality between teacher – user, differentiation of reporting context based on user / purpose and differentiation of communication language based on user. Once again two supplementary questions followed. In question 4a teachers were asked to comment in particular on the purposes of reporting to parents, as well as on the type of information this reporting must include. Question 4b asked teachers to comment on the way they report the results of a written assessment to their students.

Using the interview guide described above, interviews with 8 teachers were conducted in September 2010. These teachers were randomly selected out of the 178 teachers. The interviews took place at the interviewees' school at a time of their convenience. Note taking in addition to tape recording was used to document all interviews. In order to analyze the interview data, each interview was transcribed. The date, time, length and location of the interview was recorded on the transcripts. Details on how the interview data were analyzed are provided in the Analysis of Data section. After creating the profile of each interviewee, it was possible to match teachers' responses from the interviews with the questionnaire data. This procedure provided support to the internal validity of the study. In particular, consistency was identified between the way teachers responded to the two research instruments (i.e. questionnaire and interview). For example, teacher 5 circled number 5 (i.e. very often/ always) on the Likert scale for the statement B12 of the questionnaire "All students have the same amount of time to complete the written test". The same teacher during the interview stated: "I never allow extra time to students to complete a written test. I am very strict about it. Forty minutes is forty minutes". Likewise, teacher 8 stated: "I always correct homework. What is the point to assign homework if you are not willing to correct it? And I provide feedback to the

students so that they know what they did right and what they need to improve", her statement being consistent with her response to the item E12 of the questionnaire. For each teacher, the responses to all questionnaire items were compared to his/her responses to the interview questions. This comparison allowed us to identify consistency in the way teachers responded to the two research instruments.

In addition, the analysis of the interview data provided support to the grouping of assessment skills into levels of difficulty. As already mentioned, when the scaling and developmental structure of teachers' abilities was examined, using the Rasch and Saltus models, four developmental stages were identified (see Chapter 4 for details on the analysis). By examining the interview data, it was found that skills situated at the lowest level (i.e., those with the negative logit scores in the Rasch scale, see table 5 in Chapter 4) were considered by teachers to be easy, whereas skills situated at higher levels were considered to be more difficult. For example, skills related to the differentiation of assessment were mentioned by the teachers as difficult to be achieved. In particular, teacher 2 who was found to be situated at level 1, stated: "Differentiation sounds good. I agree that it is necessary, especially nowadays that we have a lot of immigrant students and students with learning difficulties in our classrooms. However, I don't do it. It is very difficult for me to prepare different tests or exercises according to their needs. If there were ready made tests for each group, then I would use them. I only use ready-made tests to assess my students". Similarly, teacher 4 who was found to be situated at level 2, commented on the reporting of assessment information for formative reasons. She stated that: "I am sure it would help my students more if I provided feedback for their assessment results. I write on the tests comments like "well done" or "you need to work harder', but that's about it. You cannot write specific feedback for each student. It is very difficult and it takes time". The same teacher also stated "I use oral assessment all

the time. I ask a lot of questions to my students. However, I do not prepare them in advance. I find it easier to do it spontaneously during teaching".

As it is evident from the above, interviews gave the opportunity to teachers to elaborate on their answers, thus providing us with more in-depth information concerning the questionnaire data.

Student Written Tests

Since one of the aims of the second phase of the study was to examine the impact of the interventions on student outcomes, student tests were necessary in order to measure student achievement. Given that our teacher sample included teachers, who taught mathematics in various grades, it was decided that a battery of mathematics criterion-referenced equated tests were to be used in order to assess students' achievement at the beginning and at the end of the intervention. Criterion-referenced tests enabled the evaluative description of the qualities to be assessed, without reference to the performance of others. The tests were administered to all students of the 178 teachers at the beginning and at the end of school year 2010-2011.The tests used were developed and validated in other studies conducted in Cyprus (Kyriakides, 2005; Kyriakides & Creemers, 2008a; Antoniou, 2009). The tests were designed to assess knowledge and skills in mathematics in accordance to the Cyprus Curriculum. Students were asked to answer at least two different tasks, related to the objective in the curriculum of mathematics for their group year. Particularly, seven tests were used. At the first phase, the pre-test for each school year was used, covering in this way the initial assessment of students' mathematics achievement for school years 1, 2, 3, 4, 5, and 6. At the fourth phase of the study, the pre-test for each year was used as a post test for the previous year, covering in this way the final assessment of students' mathematics achievement for school years 1, 2, 3, 4, 5, and 6. The pre-test for year 1 (i.e. Test 0) was a performance-based

test, since students at the beginning of year 1 are not expected to have reading and writing skills. Table 2 presents the administration procedure.

Table 2

Pre and Post- student test administration

School Year (Y)	Y1	Y2	Y3	Y4	Y5	Y6
Pre-test	Test 0	Test 1	Test 2	Test 3	Test 4	Test 5
Post-test	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6

Written tests used were subject to control for reliability and validity. None of the respondents achieved full score, and none showed full zero performance. Moreover, less than 5% of the students achieved over 80% of the maximum score, and less than 10% of the students achieved over 70% of the maximum score. Based on the range of the results, the ceiling and floor effects in the attainment data were not observed.

Due to the fact that the test, taken by Grade 6 students was administered when they were at the end of the school year, was obviously more difficult than the test administered to Grade 2 students when they were at the beginning of the school year, it was considered necessary that the scores were made comparable. Equating was done using Item Response Theory (IRT) modeling. The method of equating follows the same procedure as that used in the Programme for International Student Assessment (PISA) studies. However, in PISA, equating is horizontal (equating the different versions of tests), whereas in this study the equating was vertical. Specifically, the scores were transformed into the same scale on the basis of characteristics of IRT models that students' latent level of ability (y) and difficulty level of an item (b) are identical, when certain preconditions are fulfilled (Bond & Fox, 2001). The latent ability level for each student can be determined in every version, given that there

are so-called anchoring items connecting the versions. For the purposes of this study, the tests used had enough common items (i.e., approximately 8% of anchoring items across the tests) with representative content to be measured (Kolen & Brennan, 1995). Estimation was made by the Extended Logistic Model of Rasch (Andrich, 1988), which revealed that each scale had satisfactory psychometric properties (see Kyriakides & Creemers, 2008a). Thus, for each assessment period, achievement in mathematics was estimated by calculating the Rasch person estimates.

Research Sample

Phase 1

The teacher sample of the first phase of the study consisted of 178 teachers. Specifically, the questionnaire was administered to a random sample of 10% of primary Cypriot teachers in early September 2010. Out of the 240 teachers that were approached, 178 responded; a response rate of 74.2%. Particularly, the research sample for the first phase consists of 69 men (38.8%) and 109 women (61.2%). Out of the 178 participants, 28 were assistant head teachers (15.7%), whereas the rest 150 (84.3%) were teachers. Although the sample included head teachers, none of them responded to the questionnaire invitation. Participating teachers' years of experience ranged from 3 to 24 years, with the mean and median of their teaching experience estimated at 10.8 years and 10.5 years respectively. The standard deviation of the variable was 5.3.The teacher-sample was found to be representative of the teacher population of Cyprus in terms of gender (X^2=0.81, d.f. =1, p=0.42) and years of experience (t=1.21, d.f. =1278, p=0.22).

Phase 2

In mid-October 2010 an open invitation was sent to all 178 teachers participating in the first phase of the study. The invitation asked teachers to participate in a professional development program in order to improve their skills in classroom assessment during the

school year 2010-2011. The program was to be completed through seven three-hour meetings from November 2010 to May 2011. All meetings were scheduled in non-working time and volunteer participation applied. Out of the 178 teachers, 76 teachers agreed to use their free time to attend this course, a response rate of 42.5%. The fact that more than 2 out of 5 of the teachers, who were invited, accepted to participate in the program and spend their own free time for professional development reasons seems to reveal the interest that these teachers had in improving their assessment skills. Teachers from 58 different schools, representing the four districts of Cyprus, volunteered to participate. Teachers who did not attend any INSET course (n=102) were treated as members of the control group. The two intervention groups created did not differ from the control group in terms of their general characteristics (i.e., gender and years of experience). However, it must be acknowledged that the 76 teachers of the intervention groups showed personal interest in developing their assessment skills, something that cannot be generalized to the 102 teachers of the control group.

As mentioned earlier, during the first phase of the study teachers' assessment skills were measured and four stages of development were identified. Based on the results of stage allocation, teachers of each stage were randomly assigned into two groups, resulting into a total of 38 teachers in each group. Further details on group formation are given in the next section. The student sample consisted of 2358 students (51.4 % boys and 48.6% girls). The student sample derived from 178 classrooms in which the participating teachers, from both intervention and control groups, taught Mathematics during the academic year of 2010-2011. Out of the 2358 students, 267 attended grade 1 (11.3%), 183 attended grade 2 (7.8%), 554 attended grade 3 (23.5%), 534 attended grade 4 (22.6%), 499 attended grade 5 (21.2%) and finally the rest 321 students attended grade 6 (13.6%). It is important to note that although the teacher sample was not randomly selected; the student sample consisted of students from all 6 grades of primary school education. Student data were collected both at the beginning and at

the end of the intervention. Less than 5% of the original student sample was excluded from the analysis, due to missing prior or post attainment data. A detailed description of the intervention is described in the next section.

The Intervention

The intervention consisted of three steps described below.

Step 1: Initial evaluation of teachers' assessment skills and student outcomes and allocation of teachers into treatment groups

Teacher data collected during the first phase of the study were used as the initial evaluation of teachers' assessment skills. In addition, data on student achievement were collected using external written forms of assessment designed to assess knowledge and skills in mathematics. A detailed description of the instruments used was provided in the previous section. During this step, the analysis of teacher questionnaire data was conducted and teachers' developmental stages were identified (see Chapter 4 for the analysis of data). Thus, based on this analysis, the two intervention groups were formed. Particularly, the teachers, who according to the evaluation of their assessment skills were found to be in a certain developmental stage, were randomly allocated evenly into two groups. For example, the 10 teachers that were found to be situated at stage 4 were randomly allocated into the two experimental groups, each one consisting of 5 teachers. Random assignment of teachers into groups was considered necessary in order to satisfy the experimental research design employed. Indeed, randomization is considered an essential element of a true experimental design in order to minimize selection bias threats (Slavin, 2010). Since the two experimental groups involved training associated with skills of one or all development stages identified, it was considered necessary to use random assignment from within a selected pool; this being the stage in which teachers were situated. As a result, group randomization at the student level was also applied. Randomization at group level is important when investigating the effects of

teacher level factors (i.e., classroom assessment) on student achievement (Creemers, Kyriakides & Sammon, 2010). Since the interventions employed aimed at changing teacher behavior in order to affect the behavior of interrelated people (i.e., students in their classroom), random assignment of each student to each group was not feasible. In addition, breakdowns in the randomization of students were minimized, since the students involved came from different schools; therefore, resulting in high isolation of units. The number of teachers in each group according to their developmental stage is shown in Table 3.

Table 3

Teachers' allocation into groups

	Stage 1		**Stage 2**		**Stage 3**		**Stage 4**		**Total**	
	N	**F**	**N**	**F**	**N**	**F**	**N**	**F**	**N**	**F**
Group A	13	34.2%	10	26.3%	9	23.7%	6	15.8%	38	100%
Group B	13	34.2%	10	26.3%	9	23.7%	6	15.8%	38	100%
Control Group	30	29.4%	28	27.5%	29	26.4%	15	15.2%	102	100%
Total	56	31.4%	48	27%	47	26.4%	27	15.2%	178	100%

Each group employed a different professional development approach in order to improve participating teachers' assessment skills. The first treatment group, referred to as "experimental group A", employed the Dynamic Integrated Approach (DIA), whereas the second treatment group, referred to as "experimental group B" employed the Competency Based Approach (CBA). Therefore, teachers of experimental group A received training only on the assessment skills associated with their developmental stage, whereas teachers of

experimental group B received training on assessment skills associated with all four developmental stages.

Step 2: Training sessions

The second step of the intervention took place from November 2010 to May 2011. During this time, teachers participated in a series of seven training sessions; one session per month. Each session had a three-hour duration and aimed at improving teachers' assessment skills, through the use of the professional development approach employed. The first session was common for both groups, whereas sessions 2-7 were held separately for each group. A description of the sessions for each experimental group is provided below.

Session 1 (Experimental Group A+B): The first session was common for both groups and therefore all 76 teachers attended. It served as an introductory session, aiming at presenting the overall scope and procedures of the program to the participants. Thus, during this session the rational of the professional development program was presented and the main goals set were illustrated. In addition, administrative issues were discussed. Particular emphasis was also given on the program's evaluation procedures. Teachers were informed that the focus of the evaluation was going to be on the impact of the program on teacher behavior and student outcomes. Accordingly, the relevant procedures of teacher questionnaire administration, interviews as well as student test administration, at both the beginning and at the end of the program, were justified.

Furthermore, the action research methodology was presented. The presentation included definitions, basic value assumptions and a description of the four-step process of action research project development. Then, an action plan example, relevant to classroom assessment, was given to each participant. Participating teachers were asked to discuss the action plan with the research team, recognize advantages and disadvantages and finally

provide suggestions for its improvement. This example was planned to be used, throughout the program, as a guide for teachers to develop their own action plans.

Since sessions 2 to 7 were held separately for each group, a description of the sessions is presented separately for experimental group A and B accordingly.

Experimental Group A: DIA approach. The first experimental group consisted of 38 teachers. This group employed the Dynamic Integrated Approach (DIA). As already mentioned in Chapter 2, the DIA is based on the assumption that the content of the professional development program should address and therefore be differentiated to meet the needs and priorities of teachers, at each developmental stage. The grouping of skills derived from the Rasch and Saltus analyses on teacher questionnaire data (see table 5 in chapter 4). Therefore, four focus areas were created, with each focus area addressing assessment skills found to be situated at the same level. Detailed description of the skills associated with each focus area is presented in the next section. Moreover, another basic assumption of the DIA is that teachers should be engaged into systematic and guided critical reflection on their teaching practices. Thus, the research team provided opportunities for teachers, employing the DIA, to engage in reflection on their assessment practices throughout the sessions

Session 2. During this second session, teachers were distributed into four smaller groups, each group consisting of teachers of the same developmental stage. The working groups established remained the same during all sessions until the end of the program. The members of the research team provided an overall description of the focus area of each working group, making the skills on which each team had to work to improve clear. Specific areas of activity were identified for each team. At the same time, supporting material related to these areas was provided. The four focus areas identified, as well as the areas of activity for each group are presented next.

Focus Area 1- Working group A (Stage 1): The first working group focused on basic assessment skills. Classroom assessment is an integral part of the teaching process and thus teachers are expected to use effectively everyday assessment routines as part of their teaching. The role of assessment as a means, not only to evaluate but also to achieve learning, was justified and teachers were asked to reflect on their own everyday practice and especially how they use assessment and what they are trying to achieve. The areas of activity for this group included:

a) Enrichment or alteration of ready-made written tests (i.e., add their own questions; remove items that are not in line with the goals set or items of bad quality).

b) Using different types of written questions to assess students' performance (i.e., use multiple choice, matching, completion questions; include both process and product questions).

c) Using oral assessment and observation but not in a systematic way (i.e., understand the basic principles of oral assessment and observation; recognize when it is more appropriate to use these techniques)

d) Assessing group work, based on more than just the overall result (i.e., assess students' contribution to the team).

e) Being consistent with checking homework (i.e., inform students if homework is not going to be checked the same day; assign homework that can be assessed).

f) Keeping records for written assessment (i.e., keep data on all students' written assessment; register data in ways that can be used to inform learning; avoid keeping only overall results).

g) Reporting assessment results in a summative way (i.e., report results to both parents and students; understand the basic principles of reporting).

Focus Area 2- Working group B (Stage 2): The second working group moved beyond basic knowledge and skills, towards skills associated with the use of assessment for improvement purposes. The formative purpose of assessment has been widely supported in the literature as means of achieving improvement in students' learning (Black & Wiliam, 1998; Shepard, 2000). Therefore, teachers of this working group focused on the following areas of activity:

a) Construction on quality written assessments (i.e., develop a specification table in order to construct written tests ; use test items which not only ask for the final product of a task, but also the process used to reach this outcome)

b) Using oral assessment and observation (i.e., use of these techniques in a planned and systematic way)

c) Quality administration of written tests (i.e. provide clarification comments)

d) Marking homework for formative reasons (i.e., use data from homework check to improve learning)

e) Keeping records, using descriptive comments (i.e., avoid numerical or letter recording of results; record comments that can be used to inform learning)

f) Reporting assessment results to parents (i.e., report results to parents to achieve formative purposes, at least when concerning written assessment results)

Focus Area 3- Working group C (Stage 3): Teachers in this working group worked towards improving their assessment skills in order to measure more complex educational objectives. Therefore, skills involved move beyond the assessment of knowledge towards the assessment of skills and abilities of students. The areas of activity for the third group involved:

a) Developing relevant observation tools (i.e., setting specific goals; developing observational tools in line with the goals set).

b) Assessing group work (i.e. focus on students' contribution to the team, instead of overall performance).

c) Keeping records for the performance of students (i.e., keeping records for each exercise/goal included in the specification table of the assessment instrument).

d) Reporting results (i.e., reporting results deriving from all assessment techniques; reporting to both parents and students).

Focus Area 4- Working group D (Stage 4): The fourth working group worked towards improving their skills, associated with the differentiation of assessment. The need to differentiate assessment procedures and tools, based on students' needs, has been recognized as an essential element of effective learning (Chapman & King, 2005; Koutselini, 2008; Tomlinson, 1999). Therefore, teachers of this working group focused on differentiation concerning the following areas of activity:

a) Construction and administration of written assessment (i.e., construction and use of differentiated written assessments; extra tasks to those who finish earlier; extra time to slow learners)

b) Construction and administration of oral assessment (i.e., use of questions of differentiated difficulty; differentiation of wait time)

c) Reporting to parents and students (i.e., more often to those needed, adjusting forms/ language that are in line with the educational level of parents)

During the second session, instructions on how to develop their own action plan were also given to teachers. Then, under the supervision of the research team, each teacher developed his/her own action plan according to the focus area of his/her working group (see Appendix D for the action plan template). Since sessions were distributed once each month, sufficient time was available for teachers in order to pursue the goals set in their action plan.

Teachers were specifically asked to implement the activities included in their plan, and be ready to reflect on their experiences in the next meeting.

Sessions 3-6. During the next sessions each working group was working separately. With the support of the research team, teachers of each group were asked to reflect on their experiences and identify effective or non-effective practices, share comments on the activities implemented and receive and provide feedback. Additional material was also provided in each session. Furthermore, teachers were asked to complete exercises in the areas of activity of their focus area. These exercises had as a purpose to encourage collaboration within the team, while providing practical examples of new knowledge and skill application. Examples of these exercises are presented in Appendix E. Teachers were also encouraged to revise their action plans, based on their own and others' experiences and on the material provided; this was done always under the support and guidance of the research team.

Experimental Group B: CBA approach. The second experimental group consisted of 38 teachers. This group employed the Competency Based Approach (CBA), based on which teachers were given training in skills related to all four developmental stages identified. As mentioned in chapter 2, the CBA is based on the assumption that effective teachers must possess a list of competencies, which can be obtained by their participation in training courses. This approach is in line with the assessment literacy movement (Popham, 2004; Schafer, 1993), which argues the need for teachers to possess a set of assessment knowledge and skills in order for them to be effective.

The CBA does not address specific groupings of assessment skills, as the DIA does. The primary aim of CBA sessions was to improve teachers' competence in assessment by providing the necessary knowledge, associated with all the identified assessment skills. In particular, teachers received training in each skill separately. Initially, the programme was concerned with the easiest assessment skills (i.e., those with the negative logit scores in the

Rasch scale, see table 5 in chapter 4) and gradually moved on to the most difficult ones. Therefore, during the first training session, teachers received training in easier skills, such as the frequency of constructing and administrating written assessments; whereas, in their last training session, teachers were trained in more difficult skills, such as the differentiation in reporting and recording observation data. In this way, all of the skills in the four focus groups were covered, and it was expected that every teacher could master all the assessment skills.

Another difference between the two experimental groups is the focus on reflection. Whereas, for the DIA, reflection is considered an essential component of teachers' professional development and was therefore encouraged, for the CBA group, reflection is not considered necessary, since the basic assumption of this approach is that training in the assessment skills is identified as important and therefore adequate for achieving improvement.

Each session for the CBA group included three parts. During the first part, the experience on working on the action plans was discussed in a whole-group discussion. Teachers were asked to recognize possible limitations of the activities they had tried; provide suggestions for improvement and comment on other teachers' experience. The second part of the session included training on the skills associated with the focus area under study, at the particular session. More specifically, the research team presented the skills providing supporting material from the literature. During the third part of the session, teachers had to work in their working groups to discuss the material given and work on the application tasks given by the research team. In the CBA mixed ability grouping was employed, whereas in the DIA ability grouping was employed; since teachers, who were situated at the same level, were working together towards improvement, In particular, teachers of the CBA group were given the opportunity to form their working groups, as they wished. Therefore, groups formulated included teachers found to be situated at different developmental stages. Teachers worked in their working groups, only during the practical part of the session. All groups were given the

same training, the same material and the same application activities in each session. All sessions for experimental group B followed this three-part course.

Session 2. During the second session, the members of the research team provided an overall description of the skills to be addressed over the remaining sessions. At the same time, it was noted that supporting material, related to these skills, was going to be provided. Instructions on how to develop their own action plan were also given to teachers. Then, the more easy skills, as these were identified by the Rasch and Saltus analyses, were presented (see table 5 in chapter 4). The research team presented the skills, providing supporting material from the literature. Under the supervision of the research team, each teacher developed his/her own action plan for those skills (see Appendix D for the action plan template). As in experimental group A, teachers were specifically asked to implement the activities included in their plan until the next meeting. Finally, as in experimental group A, teachers were distributed into four smaller groups. The working groups established were used for all sessions, until the end of the program. However, as mentioned above, all groups were given the same training and the same material in each session and were put in operation only during the practical part of the session. Each group was asked to work on the application activities given. These activities were common for all working groups.

Sessions 3-6. During the next sessions, training following the three-part course concerning the remaining skills, as described above, was provided to all members of experimental group B. Following the list of assessment skills identified, enabled the research team to provide training, starting from basic assessment routines and gradually move on to more advance assessment skills. The competency based approach adopted aimed to improve teachers' competency in assessment, by providing the necessary knowledge associated with all assessment skills identified. Opportunities for application of this knowledge were also given, in the practical part of the session. Teachers were also expected to create a new action plan for

the skills presented in each session. The one month break between sessions gave teachers the opportunity to identify areas of improvement in relation to the skills presented and implement activities in order to improve these skills.

Session 7 (Experimental Group A+B). The final session of the professional development program was common for both experimental groups. During this final meeting, the research team explained the two different approaches employed. Teachers were asked to express their comments in relation to the approach used for their group. Positive and negative aspects regarding both approaches were identified and suggestions were made. In particular, teachers of experimental group A (DIA) recognized as a positive aspect the fact that the training offered was focused and thus provided them with a more comprehensive view of the skills involved. Opportunities to examine the skills in depth were provided, as well as the time to put them in practice. However, teachers of this group felt that with this approach they had missed the opportunity to receive training on other skills, not included in their focus area. On the other hand, teachers of experimental group B (CBA) recognized as a positive aspect the fact that their training provided them with an overall view of assessment skills. Starting from basic and moving to more advance skills helped them understand better what effective assessment practice entails. However, they also recognized as a negative aspect the fact that due to the large number of skills involved, it was difficult to attempt their application in their classroom practice. A one month focus for each group of skills was considered inadequate for changes in the classroom practice to be achieved. The general impression was that they managed to get a glimpse of everything, without in-depth enactment. Teachers from both groups expressed their wish for a follow-up professional development program during the next school year.

Step 3: Final evaluation of teachers' assessment skills and student outcomes

During the third and final step of the intervention, teachers' assessment skills and student outcomes in mathematics were measured by using the same procedures and

instruments as in phase one. In particular, teachers' assessment skills were measured by using the same questionnaire and interviews. Student outcomes in mathematics were measured by using the same pool of written assessment instruments. The results of the evaluation were presented to the participants by the research team on an individual basis.

Analysis of Data

In this study, various methods were used to analyze data, elicited from all research instruments at both phases of the study. The methods used to investigate teachers' skills in assessment, the qualitative analysis used to analyze the interview data and finally multilevel analysis used to examine the impact of the interventions are presented.

The Rasch and Saltus models

The Rasch model provides a mathematical framework against which test developers can compare their data. The model is based on the idea that useful measurement involves examination of only one human attribute at a time (i.e. unidimensionality), on a hierarchical "more than/less than" line of inquiry (Bond & Fox, 2001). The basic assumption of the model is that all persons have a higher probability of correctly answering easier items and a lower probability of correctly answering more difficult items. The extended logistic model of Rasch (Andrich, 1988) is an extension of the dichotomous model, in the case where items have more than two response categories and is, therefore, used to analyze the data emerging from teachers' responses to the questionnaire items. Since each item of the questionnaire had five response choices (1 = rarely/never, 5 = very often/always), it could be modeled as having four thresholds. Each threshold (k) has its own difficulty estimate (F), and this estimate is modeled as the threshold at which a person has a 50/50 chance of choosing one category over another. For example, the first threshold is modeled as the probability of choosing a response of 2 instead of a response of 1, and it is estimated with the following formula:

$$P_{ni1}(X = 1 / B_n, D_i, F_l) = \frac{e^{\{B_n - (D_i + F_1)\}}}{1 + e^{\{B_n - (D_i + F_1)\}}}.$$

In the above equation, F_1 is the difficulty of the first threshold, and this difficulty calibration is estimated only once for this threshold, across the entire set of items in the rating scale. The threshold difficulty F_1 is added to the item difficulty D_i (i.e., $D_i + F_1$) to indicate the difficulty of the threshold 1 on item i. Modeling subsequent thresholds in the rating scale follows the same logic. Thus, the general form of the rating scale model expresses the probability of any person choosing any given category on any item as a function of the agreeability of the person n (Bn) and the endorsability of the entire item i (Di) at the given threshold k (F_k) and is as follows:

$$P_{nik} = \frac{e^{\{B_n - (D_i + F_k)\}}}{1 + e^{\{B_n - (D_i + F_k)\}}}.$$

It is also important to note that the natural log of the odds of these probabilities result in the direct comparison between a person's ability and the difficulty of threshold k on item i {i.e., $\ln \frac{P_{nik}}{1 - P_{nik}} = B_n - (D_i + F_k)$}. The ability of the Rasch model to compare persons and items directly, means that we have created person-free measures and item-free calibrations. In addition, the Rasch model provides indices that help the investigator to determine whether there are enough items spread along the continuum, as opposed to clumps of them, and enough spread of ability among persons. Thus, reliability is estimated for both persons and items. Specifically, the person reliability index indicates the replicability of person ordering that can be expected, if this sample of persons were given another set of items, measuring the same construct (Wright & Masters, 1982). Person reliability is enhanced by small error in ability estimates, which in turn is affected by the number of targeted items. Person reliability does not only require ability estimates well-targeted by a suitable pool of items, but also a large-enough spread of ability across the sample, so that the measures demonstrate a hierarchy of ability

(i.e., person separation) on this construct (Fox & Jones, 1998). Therefore, high person reliability means that a line of inquiry in which some persons score higher and some score lower is developed, and that confidence in the consistency of these inferences can be placed. Similarly, the item reliability index indicates the replicability of item placements along the pathway, if these same items were given to another sample with comparable ability levels. Thus, from high item reliability, it can be inferred that a line of inquiry in which some items are more difficult and some items are easier is developed, and that confidence in the consistency of these inferences can be placed. The estimate of both the person separation reliability and the item separation reliability are based on the same concept as Cronbach's alpha. Thus, separation indices represent the proportion of the observed variance considered to be true. A value of 1 represents high separability in which errors are low and item difficulties and person abilities are well separated along the scale (Wright &Masters, 1981).

In this study, the model was used in order to identify the extent to which the assessment skills, measured by the questionnaire, could be reducible to a common unidimensional scale. As mentioned above, the Rasch model does not test only the unidimensionality of the scale but it is also able to find out whether the tasks can be ordered, according to the degree of their difficulty. At the same time, the people, who carry out these tasks, can be ordered according to their performance in the construct under investigation. This procedure is justified theoretically and is used in studies on teacher evaluation (e.g., Burry & Shaw, 1988; Wang & Cheng, 2001; Wright & Linacre, 1989). For this study, specifying the position of one assessment skill on the scale provides exact information about the individuals (teachers), who can perform sufficiently (i.e., those scoring higher than the position of this assessment skill on the scale) or insufficiently (those scoring lower than the position of this assessment skill). This analysis also makes it possible to make statements about the relative difficulty of each assessment skill. Similarly, specifying individual teacher's position on this

continuum provides information about the probability of this teacher to show assessment competence below or above this position (Bond & Fox, 2001).

Next, the procedure for detecting pattern clustering in measurement designs, developed by Marcoulides and Drezner (1999), was used in order to examine whether assessment skills are grouped into different levels of difficulty; that may be taken to stand for types of teacher assessment behavior. The procedure for detecting pattern clustering in measurement designs (MD) is an extension to G theory and can be considered as a special type of item response theory, capable of estimating latent traits, such as examinee ability estimates, rater severity and item difficulties. This procedure enabled us to segment the observed measurements into constituent groups (or clusters), so that the members of any group are similar to each other, according to selected criterion that stands for difficulty.

The Saltus model (Mislevy & Wilson, 1996; Wilson, 1989), a variant of the Rasch model developed by Wilson, was also used to differentiate between different types of developmental stages. The Saltus model allows the researcher to differentiate between major and less pervasive changes in development, without sacrificing the idea of one common underlying continuum. Formally, when comparing two groups of persons, the Saltus model states that the difficulty parameter changes by a certain amount for a subtest of items in one group, $P(X_{ij}=1) = f(\theta_j - \beta_i + \tau)$, in which t denotes the change in difficulty, also called the Saltus parameter, and f is the logistic distribution function (Wilson,1999). A positive value of t implies that all of the items to which the Saltus parameter pertains, become easier to the same extent in that group. For the other items in this group, and for all the items in the other group, the difficulty remains unchanged, and the equation of the Rasch model holds. Hence, when two groups of people who are assumed to be at different stages of development are compared; a positive and significant value of the t parameter for a subset of items in the more developed group may reflect a discontinuity in development that may reflect some kind of qualitative

change. The change consists in part of the items being easier. It is situated on the same dimension and supplements a progression along the same latent scale.

In most theories of development, developmental sequences involve stages that are qualitatively discrete from each other and follow a constant order of succession. In the Saltus model, these two aspects of development are summarized by the twin concepts of 'gappiness' and 'rigidity', respectively. Gappiness is indicated in the Saltus model by segmentation, which is specified as the distance between the most difficult item of Level 1 and the easiest item of Level 2. Segmentation is measured through two segmentation indices, one for each person group. The difference between these two segmentation indices is called the asymmetry index. When the asymmetry index is zero, the Saltus model is equivalent to the Rasch model, which can be interpreted to mean that the difference, as far as difficulty is concerned, between the two item types is the same for both participant groups. On the other hand, if the asymmetry index is positive, the Level 1 students perform on the items as being further apart in difficulty than students in Level 2 do. This pattern indicates rigidity and is typical of hierarchical development. That is, the upper stage items are near to impossible for persons at the lower stage, whereas persons at the upper stage can solve items of both stages; although facing some difficulty in dealing with the upper stage items and making some errors in dealing with the lower stage items is reasonable. This diminishes the observed difference in difficulty of the item types. This pattern is also manifested in a jump in the predicted probability of success at the border between the two groups that is not present, when the asymmetry index is zero.

Qualitative Analysis

Qualitative data which emerged from the interviews were analysed by using the constant comparative method (Maykut & Morehouse, 1994). As Strauss and Corbin (1998) describe the constant comparative method consists of looking for patterns, within and across data, in order to define emerging constructs, phenomena and relationships. In particular, the

constant comparative method involves breaking down the data into discrete 'units' (Lincoln & Guba, 1985) or 'incidents' (Glaser & Strauss, 1967) and coding them to categories. "Within-case analysis" (Denzin & Lincoln, 1998) of each teacher's responses to the interview were conducted in order to link them with his/her responses to the questionnaire. For this reason, transcripts were read with the intention of identifying integrating themes, foci, frequently used metaphors and possible incongruities. Matching teachers' responses from the interviews with the questionnaire data provided support to the internal validity of the study.

Multilevel Analysis

It is now generally accepted that a satisfactory approach to school effectiveness modeling requires the deployment of multilevel analysis techniques (Goldstein, 1997). Multi-level analysis is a methodology for the analysis of data with complex patterns of variability. It is based on the notion that individuals or any other type of objects are naturally nested in groups, with membership in the same group leading to a possible correlation between the individuals (de Leeuw & Meijer, 2008). Therefore, students are nested within classrooms, classrooms are nested within schools and schools are nested within educational districts, systems or nations.

In this study, multilevel analysis was considered appropriate, since the data set had a hierarchical structure in which students were nested within classrooms. Multilevel analysis was conducted using the MLwiN software and was used to measure the impact of the independent variables (DIA and CBA approaches to professional development) on student achievement. Separate multilevel analysis for each dependent variable was performed. The first step in each analysis was to determine the variance at individual, class, and school level without explanatory variables (i.e., baseline model). In subsequent steps, explanatory variables at different levels were added. Explanatory variables, except grouping variables, were centered

as Z-scores with a mean of 0 and a standard deviation of 1. Grouping variables were entered as dummies with one of the groups as baseline (e.g., boys = 0).

Research Limitations

As the literature indicates, several conditions may put at risk the validity of a research (Bracht & Glass, 1968; Campbell & Stanley, 1963). As described above, systematic and consistent efforts have been made in order to eliminate to the best possible extent threats to the internal and external validity of the study. However, some possible limitations can be acknowledged.

First, the use of a questionnaire as the primary method for investigating teachers' assessment skills may be questioned. Reporting bias is a threat commonly associated with the use of self-reported data, since it is acknowledged that there is no way to tell for sure how truthful a respondent may be. However, as already mentioned, the proposed framework for measuring assessment skills refers to all four phases of assessment. Therefore, it was practically not possible to measure teacher skills in assessment by using external observation. More specifically, observation of teacher behavior in the classroom may not allow us to measure teacher skills in assessment tool construction, recording and reporting of data, since these tasks may take place outside classroom. Moreover, to measure teacher skills in administering assessment tasks, a large number of lessons per teacher have to be observed, especially since a significant percentage of teachers offer assessment tasks only at the end of a unit or series of lessons; therefore it is very unlikely to be able to get data on teacher skills in assessment, unless a variety of lessons of each teacher is observed. Although the limitations of collecting data through teacher self-reports are acknowledged, it was not feasible to conduct a very big number of observations in order to ensure generalisability of the data. Recognizing this limitation, this study addresses validity issues in two ways. First, the analysis of the semi-structured interviews provides evidence supporting the internal validity of the study. Second,

the results of the Rasch analysis show that teacher assessment skills can be measured by using the three aspects of the proposed framework, used in developing the questionnaire. Thus, data analysis provides support for the construct validity of the questionnaire. Finally, the fact that teachers were informed that their responses to the questionnaire would be used to design the professional development program to follow, suggests that it is more possible that they responded with honesty to the questionnaire administered.

Another limitation of the study has to do with the sampling procedure employed during the experimental phase of the study. Random assignment of participants is considered to be an important characteristic of experimental studies, since random assignment to treatment conditions assures that treatment group assignment is independent of the pre- treatment characteristics of group members (Creemers, Kyriakides & Sammons, 2010). In the present study, the teacher sample of the intervention groups was not randomly selected, since the 76 teachers volunteered to participate in the professional development program, whereas the rest 102 teachers formed the control group. It is acknowledged that this poses questions concerning the external validity of the study and whether its results can be generalized to a wider population. It is important to note that this study recognizes the fact that the control group cannot be comparable to the two intervention groups, since teachers who agreed to participate in the training seem to show a special interest in improving their assessment skills. The use of the control group was decided in order to help us identify differences between the two intervention groups and not differences between each intervention group with the control group. In addition, since participating teachers were randomly assigned to two experimental groups, it is argued that homogeneous experimental groups were formed; thus, allowing the comparison between the two groups in terms of outcomes measures and teacher assessment skills.

Finally, another limitation recognized is the small sample used during the experimental phase of the study. Volunteer participation in professional development programs, which take place at non-work time, is difficult to be achieved. However, each treatment group consisted of 38 teachers, a number considered acceptable in experimental research designs. In addition, given the teacher sample used, the research team was able to provide one to one support to all participating teachers throughout the intervention.

This chapter has outlined the research design and methods used in this study. The next section presents how data, derived from both phases of the study, were analyzed in order to address the research questions set.

CHAPTER 4

RESEARCH RESULTS

This chapter presents the analysis of the data collected throughout the study. Research results are presented addressing the research questions set in Chapter 1. Particularly, the first section presents the results deriving from the first phase of the study, which examines the identification of developmental stages of teacher assessment skills. The second section presents the results of the second phase of the study and in particular, the impact of the two interventions upon teacher assessment skills and student achievement.

Searching for Stages of Teacher Skills in Assessment

This section presents the results of the first phase of the study. Teachers' responses to the questionnaire, measuring assessment skills were analysed by a number of different methods in order to provide answers regarding the scaling and developmental structure of teachers' abilities in assessment. Specifically, the Rasch and Saltus models were used.

Using the Rasch Model to Specify the Hierarchy of Item Difficulty

The extended logistic model of Rasch (Andrich, 1988) was used in order to identify the extent to which the assessment skills measured by the questionnaire could be reducible to a common scale.

The extended logistic model of Rasch was applied to the whole sample of teachers and all 87 measures concerned with their assessment skills together, using the computer program Quest (Adams & Khoo, 1996). This model (Andersen, 1977; Wright, 1985) is an extension of the dichotomous Rasch model to the case in which items have more than two response categories and it was therefore used to analyze the data that emerged from teachers' responses to each questionnaire item. Since each item has five responses, it can be modeled as having four thresholds. Each threshold has its own difficulty estimate, and this estimate is modeled as

the threshold at which a person has a 50% chance of choosing one category over another (Andersen, 1977). These thresholds are calculated in log odds (otherwise called logits) and should be ordered to represent decreasing probability of each assessment behaviour occurring. Thresholds that do not increase monotonically are considered disordered. The magnitudes of the distances between the threshold estimates are also important. Threshold distances should indicate that each step defines a distinct position on the variable and thereby they should be neither too close together nor too far apart on the logit scale (Bond & Fox, 2001). Specifically, guidelines indicate that thresholds should increase by at least 1.4 logits (i.e. to show distinction between categories) but no more than 5 logits (i.e. to avoid large gaps in the variable; Linacre, 1999).

Figure 4 illustrates the scale for the 87 measures of assessment skills with item difficulties and teacher measures calibrated on the same scale. The item threshold values were found to be ordered from low to high, indicating that the teachers answered consistently with the ordered response format of our Likert scale. The threshold distances range from 1.7 to 2.5 logits. Figure 4 also shows that the 87 items of the questionnaire, measuring teacher assessment skills, have a good fit to the measurement model, indicating a strong agreement among the 178 teachers located at different positions on the scale, across all 87 items. Moreover, the questionnaire items are well targeted against the teachers' measures since teachers' scores range from -3.14 to 3.11 logits and item difficulties range from -3.11 to 3.34 logits.

Furthermore, Table 4 provides a summary of the scale statistics for the whole sample and the two subgroups (female and male teachers). Reliability is calculated by the Item Separation Index and the Person Separation Index. Separation indices represent the proportion of the observed variance considered to be true. A value of 1 represents high separability in which errors are low and item difficulties and students' measures are well separated along the

scale (Wright & Masters, 1981). It can be observed that for the whole sample and each subgroup the indices of cases and item separation are higher than 0.92, indicating that the separability of the scale is satisfactory (Wright, 1985). In addition, the infit mean squares and the outfit mean squares were found to be near one and the values of the infit t-scores and the outfit t-scores are approximately zero.

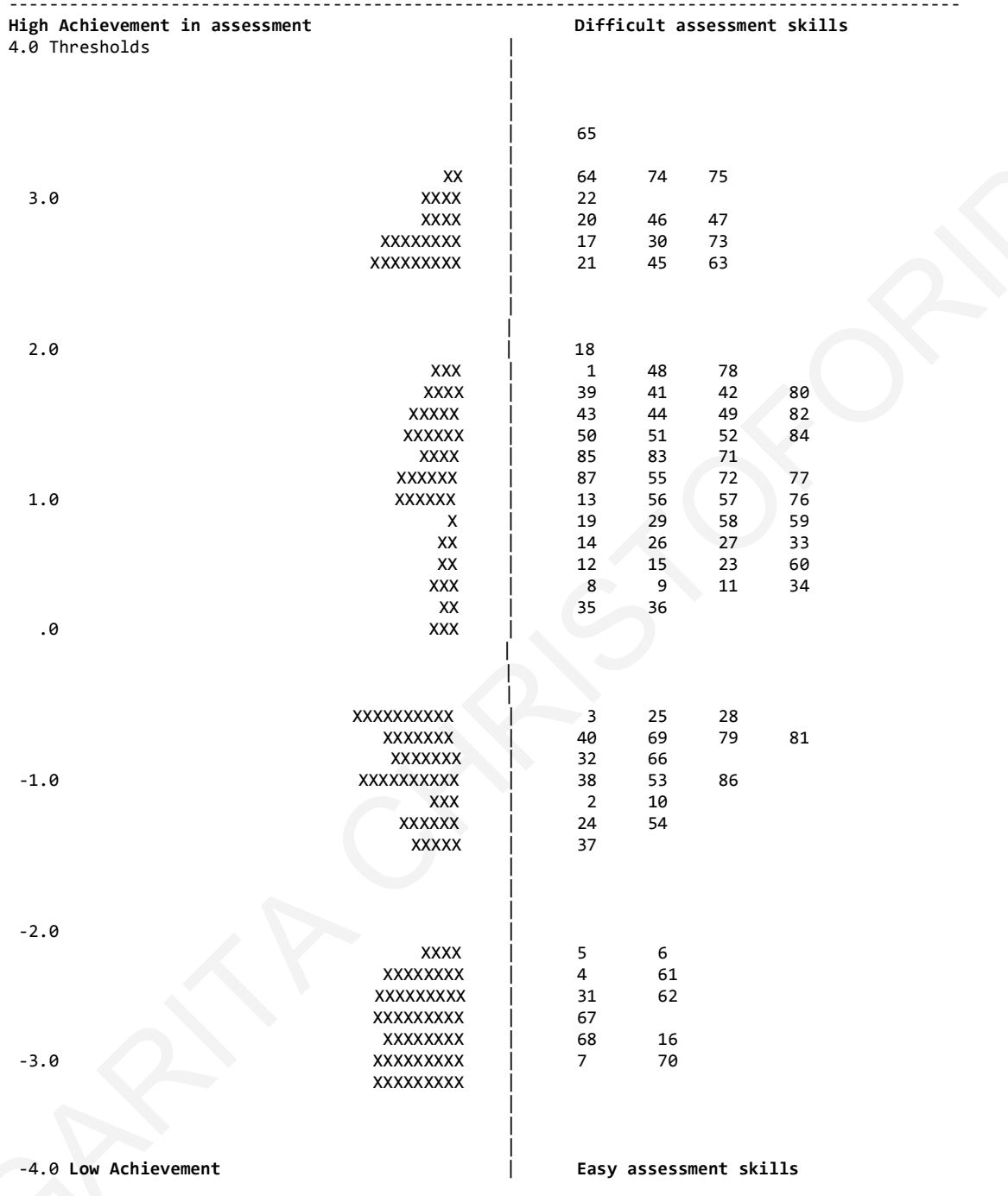
The results of the various approaches used to test the fitting of the Rasch model to our data also revealed that there was a good fit to the model when teachers' performance in these assessment skills was analyzed. More specific, all assessment skills were found to have item infit with the range of 0.85–1.16, and item outfit with the range of 0.76–1.40. All the values of infit t for both persons and assessment skills were greater than - 2.00 and smaller than 2.00. Finally, the procedure proposed by Yen (1993) was used to test for local independence; and local independence was found not generally violated.

Table 4

Statistics relating the questionnaire measuring assessment skills, based on the Rasch analysis of the whole sample and of Rasch analysis of each gender group separately.

Statistic	Whole sample (n=178)	Female (n=109)	Male (n=69)
Mean (items*)	0.00	0.00	0.00
(persons)	0.08	0.10	0.07
Standard deviation (items)	1.12	1.02	1.05
(persons)	1.02	0.96	0.93
Separability (items)	0.99	0.98	0.99
(persons)	0.95	0.94	0.93
Mean Infit mean square (items)	0.99	1.00	0.99
(persons)	1.00	1.00	1.00
Mean Outfit mean square (items)	1.03	1.02	1.03
(persons)	1.04	1.07	1.03
Infit t (items)	0.04	0.05	-0.01
(persons)	0.02	-0.04	-0.03
Outfit t (items)	0.01	0.03	-0.05
(persons)	0.06	0.05	0.04

* L=87 items



Note: Each X represents 1 teacher

Figure 4. Rasch Scale of teacher's skills in assessment (N = 178 teachers; L = 87 skills)

Using Cluster Analysis to Specify Levels of Difficulty

Having established the reliability of the scale, the procedure for detecting pattern clustering in measurement designs, developed by Marcoulides and Drezner (1999), was used to find out whether assessment skills are grouped into levels of difficulty that may be taken to stand for types of teachers' behaviour in evaluating student achievement in mathematics, which move from relatively easy to more difficult. Applying this method to segment, the assessment skills on the basis of their difficulties that emerged from the Rasch model, showed that they are optimally clustered into four clusters. Specifically, the cumulative D for the four-cluster solution was 59%, whereas the fifth gap adds only 2%. According to the literature in cluster analysis, the four-cluster solution explaining 59% of the observed variance is considered satisfactory (Romesburg, 1984). These four clusters are further explored and specified by using the Saltus model, as described below.

Using the Saltus model to Specify the Developmental Structure of Assessment Skills

How deep is the divide separating the four types of teacher behavior in assessment that emerged from cluster analysis and which can be ordered into different levels according to their difficulty? The Rasch model and the clustering method used so far, cannot answer this question and therefore the Saltus model was used.

To apply the Saltus model, it is assumed that the 87 questionnaire items are developmentally structured in the four levels identified through the cluster analysis. The Saltus solution was found to represent a better fit to the actual data rather than the Rasch model, and offers a statistically significant improvement over the Rasch model which is equal to 899,5 chi-squared units at the cost of 20 additional parameters (i.e. 9 ts, four means and standard deviations and three independent proportions). Table 5 presents the item difficulty parameters for teachers of Level 1 (i.e. column 3) and the implied within stage difficulty (i.e. columns 4, 5, and 6). The Saltus parameter estimates (i.e. t values) are shown in the bottom of the table.

Table 5

Rasch and Saltus parameter estimates for the 87 items of the teacher questionnaire grouped into four levels of assessment skills

Assessment skills by measurement dimension	Rasch	Level 1	Level 2	Level 3	Level 4
Frequency: Construction of written tests	-3,11	-3,41	-3,41	-3,41	-3,41
Frequency: Administration of written tests	-3,09	-3,39	-3,39	-3,39	-3,39
Frequency: Recording of homework (check for summative reasons)	-3,07	-3,37	-3,37	-3,37	-3,37
Frequency: Recording of written tests	-3,05	-3,35	-3,35	-3,35	-3,35
Focus: Construction of written tests (various types of written questions)	-3,01	-3,31	-3,31	-3,31	-3,31
Focus: Recording of written tests (single score)	-2,97	-3,27	-3,27	-3,27	-3,27
Focus: Reporting of written tests (parents only)	-2,95	-3,29	-3,29	-3,29	-3,29
Stage: Construction of written tests	-2,93	-3,23	-3,23	-3,23	-3,23
Frequency: Reporting of written tests	-2,91	-3,21	-3,21	-3,21	-3,21
Focus: Construction of written tests (basic skills)	-2,89	-3,19	-3,19	-3,19	-3,19
Stage: Recording of written tests	-2,87	-3,27	-3,27	-3,27	-3,27
Focus: Construction of written tests (product questions)	-2,86	-3,16	-3,16	-3,16	-3,16
Quality: recording of written tests (for summative reasons)	-2,85	-3,22	-3,22	-3,22	-3,22
Stage: Reporting of written tests (parents only)	-2,84	-3,14	-3,14	-3,14	-3,14
Quality: Construction of written tests (only for summative reasons)	-2,83	-3,13	-3,13	-3,13	-3,13
Quality: Administration of written tests (for summative reasons)	-2,82	-3,21	-3,21	-3,21	-3,21
Quality: Reporting of written tests (for summative reasons)	-2,82	-3,12	-3,12	-3,12	-3,12
Frequency: Administration of oral assessment	-1,1	-1,01	-2,84	-2,8	-2,89
Frequency: Construction of oral assessment (systematic)	-1,09	-1	-2,83	-2,79	-2,88
Focus: Construction of written tests (process questions)	-1,08	-0,98	-2,81	-2,77	-2,86
Quality: Construction of written tests (specification table)	-1,07	-0,95	-2,78	-2,74	-2,83
Quality: Administration written tests (clarifications)	-1,05	-0,91	-2,74	-2,7	-2,79
Frequency: Reporting of homework (for measuring basic skills)	-1,03	-0,93	-2,76	-2,72	-2,81
Frequency Reporting of oral assessment	-1,00	-0,9	-2,73	-2,69	-2,78
Quality: Construction of written tests (representative)	-0,98	-0,88	-2,71	-2,67	-2,76
Frequency: Construction of performance test	-0,95	-0,81	-2,64	-2,6	-2,69
Frequency: Administration of performance test	-0,89	-0,79	-2,62	-2,58	-2,67
Stage: Recording written tests (value added)	-0,88	-0,82	-2,65	-2,61	-2,7
Focus: Administration of other forms (homework basic skills)	-0,88	-0,78	-2,61	-2,57	-2,66
Frequency: Recording oral assessment	-0,87	-0,75	-2,58	-2,54	-2,63
Frequency: Reporting observation (non-systematic)	-0,84	-0,74	-2,57	-2,53	-2,62
Stage: Reporting written tests to parents only	-0,83	-0,84	-2,67	-2,63	-2,72

Focus: Construction of performance tests (basic skills only)	-0,81	-0,71	-2,54	-2,5	-2,59
Quality: Construction of written tests (take into account learning needs of students)	-0,78	-0,8	-2,63	-2,59	-2,68
Frequency: Reporting of performance tests	-0,71	-0,65	-2,48	-2,44	-2,53
Stage: Administration of oral assessment	-0,66	-0,63	-2,46	-2,42	-2,51
Focus: Recording of written tests (descriptive comments)	-0,62	-0,58	-2,41	-2,37	-2,46
Frequency: Recording of performance tests	-0,59	-0,49	-2,32	-2,28	-2,37
Frequency: Construction observation	-0,54	-0,44	-2,27	-2,23	-2,32
Focus: Construction of oral tests (basic skills)	-0,52	-0,41	-2,24	-2,2	-2,29
Focus: Administration of oral tests (clarifications)	-0,45	-0,38	-2,21	-2,17	-2,26
Frequency: Administration of observation	-0,38	-0,3	-2,13	-2,09	-2,18
Focus: construction of observation (basic skills only)	-0,32	-0,27	-2,1	-2,06	-2,15
Focus: construction of performance tests (basic skills)	-0,30	-0,2	-2,03	-1,99	-2,08
Stage: construction of oral assessment	-0,27	-0,23	-2,06	-2,02	-2,11
Stage: construction of performance test	-0,23	-0,18	-2,01	-1,97	-2,06
Stage: administration of oral assessment	-0,22	-0,12	-1,95	-1,91	-2,00
Stage: administration of observation	-0,20	-0,1	-1,93	-1,89	-1,98
Stage: construction of observation	-0,17	-0,04	-1,87	-1,83	-1,92
Focus: construction of oral assessment (basic skills)	-0,15	-0,05	-1,88	-1,84	-1,93
Stage: recording of oral assessment	-0,09	-0,01	-1,84	-1,8	-1,89
Stage: administration of performance test	0,05	0,08	-1,75	-1,71	-1,8
Quality: reporting of written tests (formative)	0,95	0,99	0,13	-1,66	-1,5
Stage: administration of observation (systematic and continuous)	0,97	1,09	0,23	-1,56	-1,4
Quality: construction of observation tools	0,98	1,08	0,22	-1,57	-1,41
Quality: recording of performance tests (formative reasons)	1,00	1,1	0,24	-1,55	-1,39
Focus: construction of written tests (complex objectives)	1,03	1,11	0,25	-1,54	-1,38
Focus: construction of oral assessment (mathematics communication)	1,11	1	0,14	-1,65	-1,49
Focus: construction of performance test (complex objectives)	1,14	1,06	0,2	-1,59	-1,43
Quality: reporting of performance tests (formative)	1,2	1,07	0,21	-1,58	-1,42
Quality: recording of oral assessment (multi-dimensional)	1,21	1,11	0,25	-1,54	-1,38
Quality: recording of performance test (formative – comments on a variety of skills)	1,24	1,17	0,31	-1,48	-1,32
Focus: administration of observation (complex skills such as communication)	1,25	1,21	0,35	-1,44	-1,28
Focus: administration oral assessment (maths communication)	1,28	1,19	0,33	-1,46	-1,3
Focus: administration observation (low-inference observation tools)	1,34	1,21	0,35	-1,44	-1,28
Focus: construction of observation (focused on specific skills/objectives)	1,35	1,28	0,42	-1,37	-1,21
Focus: reporting written tests (parents and pupils)	1,37	1,32	0,46	-1,33	-1,17
Focus: reporting performance tests (parents and pupils)	1,38	1,34	0,48	-1,31	-1,15
Focus: reporting oral assessment (parents and pupils)	1,42	1,39	0,53	-1,26	-1,1

Focus: reporting observation (parents and pupils)	1,43	1,33	0,47	-1,32	-1,16
Stage: construction of observation (systematic and continuous)	1,45	1,43	0,57	-1,22	-1,06
Differentiation recording written tests	2,6	2,91	2,1	1,98	-0,86
Differentiation: construction of written tests	2,68	2,96	2,15	2,03	-0,81
Differentiation: construction of oral assessment	2,78	2,94	2,13	2,01	-0,83
Differentiation: administration of written tests	2,88	2,97	2,16	2,04	-0,8
Differentiation: construction of performance tests	2,95	3,00	2,19	2,07	-0,77
Differentiation recording performance tests	3,00	3,08	2,27	2,15	-0,69
Differentiation: administration of performance test	3,08	3,12	2,31	2,19	-0,65
Differentiation: reporting of written tests	3,12	3,21	2,4	2,28	-0,56
Differentiation: construction of oral assessment	3,16	3,29	2,48	2,36	-0,48
Differentiation: administration of oral assessment	3,19	3,38	2,57	2,45	-0,39
Differentiation: administration of observation	3,23	3,42	2,61	2,49	-0,35
Differentiation: reporting of performance tests	3,27	3,50	2,69	2,57	-0,27
Differentiation: reporting of observation	3,29	3,56	2,75	2,63	-0,21
Differentiation recording of oral assessment	3,31	3,64	2,83	2,71	-0,13
Differentiation recording of observation	3,33	3,69	2,88	2,76	-0,08
Differentiation: reporting oral assessment	3,34	3,73	2,92	2,8	-0,04

Note 1: Lines in the body of the table above separate the four stages of assessment as indicated by cluster analysis.
Note 2: The Saltus parameter estimates (i.e. τ values) are shown below.

	Examinee Stage			
Item Class	1	2	3	4
1	0.00*	0.00*	0.00*	0.00*
2	0.00*	1.83	1.79	1.88
3	0.00*	0.86	2.65	2.49
4	0.00*	0.81	0.93	3.77

*Fixed at zero for model identification

The following observations arise from table 5. First, item difficulty parameters for teachers in Level 1 are more spread out than those of the Rasch model, exhibiting a large gap between the items of Level 1 and the items in Levels 2, 3, and 4. The gap between the items of Level 1 and the items of Level 2 closes considerably when we look at the difficulty estimates that pertain to Level 2 teachers. Specifically, for teachers who belong to Level 2, items of Level 2 are almost as easy as items of Level 1. As far as the difficulties of items of Level 3 are concerned, these items are relatively difficult for Level 2 teachers, whereas for Level 3

teachers these items are almost as easy as Level 2 items. Similar observations can be made in relation to items of Level 4.

Second, a comparison of the segmentation indices reveals that all of them are very large. Similarly, all the asymmetry indices were relatively large. However, the asymmetry index between Levels 3 and 4 is extremely high. This implies that the transition from one level to the other is not linear; specifically, the transition from level 3 to 4 is much more difficult than any transition among the first three levels. Thus, the development of teacher skills in assessment is discontinuous rather than continuous. In addition, the discontinuity in development is much more obvious for teachers moving from Level 3 to 4. A description of the different levels of teacher behaviour in assessment is given below.

Type 1: Using written tests to measure basic skills in mathematics for summative reasons (-3.10 up to -2.20 logits). The assessment skills included in this stage reveal that teachers (n=56) demonstrating this type of behaviour use everyday assessment routines. Type 1 teachers enrich or alter ready-made written tests and use a variety of types of written questions to assess students' performance. However, they do not use the oral assessment and/or observation, in a systematic way, in order to assess their students' performance. Records are kept only in relation to written assessment results, whereas results are reported to parents only for summative purposes. Finally, type 1 teachers appear to be consistent to homework check.

Type 2: Using different techniques of assessment to measure basic skills in mathematics (-1.40 up to 0.50 logits). The assessment skills included in this stage reveal that teachers (n=48) demonstrating this type of behaviour are able to use the various techniques of assessment appropriately in order to measure basic skills in mathematics. Specifically, type 2 teachers create a specification table before developing their written tests. In this way, they try to ensure that their tests are representative to what has been taught in the classroom. They also

include test items which measure students' ability to give a correct answer to a task and items which investigate the process that was used by each student in his/her attempt to find an answer to a problem (i.e., process questions are included). In designing test items, they also take into consideration their students' abilities. In addition, they reported that they offer clarification comments to students during assessment administration and that oral assessment and observation are planned in advance. Furthermore, teachers of this stage move beyond homework check and use homework information to assess the basic skills of their students in mathematics. As far as recording of assessment data is concerned, they use descriptive comments to give feedback to their students. Finally, they report to parents on their students' assessment results.

Type 3: Using assessment techniques to measure more complex educational objectives for formative reasons (0.20 up to 1.95 logits). Teachers demonstrating this type of behaviour (n=47) are able to use assessment techniques to measure more complex educational objectives in mathematics such as their ability to communicate by using mathematics. Thus, observation is used in a systematic way, by setting specific goals and creating observation tools related to these goals. Recording is done for data that derive from all assessment techniques and not only from written assessment (as in type 2 teachers) and takes the form of goal and/or exercise specific documentation. In addition, reporting is done for formative reasons and it is expanded to cover all assessment techniques. Furthermore, teachers of this stage report assessment information not only to parents but to their students as well. Finally, group assessment is used in a systematic way and is primarily concerned with each student's contribution to the team work rather than the team's overall performance.

Type 4: Differentiation in Assessment: Applying assessment in and for different occasions and students (2.60 up to 3.35 logits). Based on the assessment skills included in this type of behavior, it appears that type 4 teachers (n=27) are able to differentiate assessment

procedures and tools based on their students' needs. Therefore, teachers of this stage do not use the same written tests to measure the achievement of different groups of students and they are also more flexible during the administration process (e.g., they give extra tasks to those who finish earlier and more time to slow learners). They also differentiate reporting of assessment information to both parents and students (e.g., reporting is done more frequently to those that need it the most; they use different forms/language that are in line with the educational level of parents) and pursue teacher-parent communication especially when the last rarely or never visit the school.

The Impact of the Interventions on Teacher Assessment Skills and Student Achievement

This section presents the results of the second phase of the study. Specifically, the results of the analysis related to the impact of the interventions on a) teacher assessment skills and b) student achievement are discussed.

Impact on Teacher Assessment Skills

The analysis procedure described in the previous section was also used to analyze teacher data from the second phase of the study. Thus, the questionnaire data were analyzed in order to validate the identification of the four types of assessment behavior.

Data from the initial and final measurement of participating teachers' assessment skills were analysed. All assessment skills were found to have item infit with the range 0.88–1.15, and item outfit with the range of 0.79–1.38. All the values of infit t for both persons and assessment skills were greater than - 2.00 and smaller than 2.00. By comparing the difficulty index of all items of the scales which emerged from the two data collection phases (i.e., at the beginning and at the end of the school year), it was found that there are difficulties that could be considered invariant across the two administration periods, within measurement error (0.13). This implies that person estimates that emerged from the two Rasch analyses could be

considered as comparable. By applying the procedure for detecting pattern clustering in measurement designs, it was found out that assessment skills can be grouped into the same four levels of difficulty, as identified through the analysis of data that emerged from the first measurement. Specifically, the cumulative D for the four-cluster solution was 64%, whereas the fifth gap adds only 2%.

In order to measure the impact of the two professional development programs upon teachers' skills in assessment, the Rasch person estimates of each group were compared. Table 6 presents the means and standard deviations of teacher scores of each experimental group and the control group, which emerged by measuring assessment skills at the beginning and at the end of the intervention.

Table 6.

Means and standard deviations of teacher scores measuring assessment skills of the control and the experimental group before and after the intervention

Group	Before		After	
	Mean	S.D.	Mean	S.D.
Control Group (n=98)	-0.05*	1.00	-0.04	0.97
Employing DIA (n=36)	-0.05	1.03	0.43	0.99
Employing CBA (n=36)	-0.06	0.97	0.17	0.88

*Rasch person estimates in logits

First, it can be observed that the initial mean scores of the three groups were almost the same. One-way analysis of variance revealed that was no statistically significant difference among the three groups in regard to the initial Rasch person estimates (F=0.011, p=.989). Second, the final score of teachers employing the DIA (Mean=0.43, SD=0.99) was bigger than their initial score (Mean=-0.05, SD=1.03) and the t-test paired sample revealed that this difference was statistically significant (t=7.81, df=35, p=.001). This finding reveals that

teachers, employing the DIA, managed to improve their assessment skills. On the other hand, the mean final and initial scores of the control group were almost the same and the t-test paired test reveals that teachers of the control group did not manage to improve their assessment skills (t=0.103, df=97, p=.92). Third, the t-test paired sample test reveals that teachers, employing the CBA, also managed to improve their assessment skills (t=3.89, df=35, p=.001).

In order to identify whether each intervention had an impact on the assessment skills of teachers, a regression analysis was also employed. The final score of teachers was treated as dependent variable, whereas the initial score as well as two dummy variables representing each intervention were treated as independent variables. In this way, the control group was treated as the reference group. The model found to fit better the data, was able to explain a very large percentage of the variance of the final score of teachers' skills in assessment (82%) and all three variables were entered to the equation that emerged which is given below:

Post score = -.002 + 0.868 * pre score + 0.474 * DIA + 0.216 * CBA + r

The equation above presents the unstandardized solution and helps us identify the impact of each approach on teacher' post scores in comparison to the control group. The equation suggests that a teacher employing the DIA will differ by 0.474, in terms of the post score, compared with a teacher of the control group with the same pre score. In addition, when a teacher, employing the CBA approach, is compared with a teacher of the control group, the difference will be 0.216 in favor of the DIA teacher. Finally, comparing the impact of the DIA (.474) with the impact of CBA (0.216) on teachers' post scores, it can be observed that there is a difference of .258 in favor of the DIA. The conclusion that teachers employing the DIA manage to improve their skills at a statistically higher level than teachers employing the CBA, is also supported by comparing the standardized beta coefficients. In particular, it can be observed that the impact of the DIA (.200) was twice as big than the impact of CBA (.091).

Furthermore, by comparing the classification of teachers into different stages at the beginning and at the end of the intervention, it was found that 13 out of 36 teachers of the group employing the DIA, managed to move to the next more demanding stage, whereas the other 23 teachers remained at the same stage. Specifically, four teachers of this group moved from stage one to stage two, six teachers of stage two managed to move to stage three and three teachers situated at stage three were found to be at stage four at the end of the intervention. On the other hand, only five teachers of the group employing the CBA, managed to move to the next most demanding stage, whereas almost all teachers of this group (i.e., 31 out of 36) remained at the same stage. More specifically, four teachers managed to move from stage one to stage two and one teacher moved from stage two to stage three. Finally, by using the t-test paired sample, it was found that teachers situated at stages three and four who made use of CBA, did not make any statistically significant progress in their skills ($t=1.13$, $df=13$. $p=.279$), whereas teachers of these two stages, employing the DIA managed to improve at a statistically significant level ($t=6.05$, $df=18$, $p=0.001$).

Impact on Student Achievement

The results of the multilevel analysis, conducted in order to measure the impact of each of the two approaches on teacher professional development on student achievement, are presented next.

Empty models with all possible combinations of the levels of analysis (i.e., student, teacher, and school) were established and the likelihood statistics of each model were compared (Snijders & Bosker, 1999). An empty model consisting of student, teacher, and school level represented the best solution. Statistical power is also an issue that has to be taken into account in using multilevel modeling approaches to analyze nested data (Cools, De Fraine, Van den Noortgate, & Onghena, 2009). It is typically recommended that at least 40 higher level units must be sampled in order to tap sufficient variance. In this study, the sample

consisted of 174 teachers, appointed at 62 different schools and thereby the three-level model was considered appropriate. The empty model revealed that 74.3% of the total variance was situated at the student level, 16.7% of the variance was at the classroom level and 9.0% was at the school level. In subsequent steps, explanatory variables at different levels were added, starting at the student level. Explanatory variables, except grouping variables, were entered as Z-scores with a mean of 0 and a standard deviation of 1. This is a way of centering around the grand mean (Bryk & Raudenbush, 1992) and yields effects that are comparable. Grouping variables were entered as dummies with one of the groups as baseline (e.g., girls=0). The models presented in Table 7 were estimated without the variables that did not have a statistically significant effect at level .05.

In model 1 the context variables at each level and the teacher background information were added to the empty model. The following observations arise from the figures of the third column in Table 2. First, model 1 explained 33.0% of the variance, most of which was attributed to the student level. Second, all student background variables had statistically significant effects on student achievement. Prior knowledge has the strongest effect in predicting student achievement at the end of the school year. In addition, prior knowledge was the only contextual variable which had a consistent effect on achievement when aggregated either at the teacher or the school level. Finally, length of teaching experience was the only teacher background factor, which had a statistically significant effect on student achievement.

In model 2, the impact of teacher assessment upon student achievement was investigated. Since teachers were assigned to four developmental stages according to their assessment skills, the extent to which the classification of teachers into these four stages could explain variation in student achievement was examined. Thus, teachers at stage 3 were treated as a reference (or baseline) group and three dummy variables were entered in model 1. The developmental stage, at which a teacher is situated, was found to have a statistically significant

effect on student achievement. Specifically, students of teachers at stage 1 had the lowest achievement, whereas students of teachers at level 4 had higher achievement than students of the first three levels.

Finally, in model 3 the effect of each approach employed towards teacher professional development in assessment was investigated. Thus, teachers of the control group were treated as a reference (or baseline) group and two dummy variables, indicating the teacher professional approach employed (i.e., DIA and CBA), were entered into model 2. Only the effect of the dummy variable measuring the impact of the DIA was found to be statistically significant at .05 level.

The results of the multilevel analysis provide evidence that only the DIA yields better results in student achievement than the control group. However, it is not clear whether this approach is equally effective for teachers situated at different levels. To test this assumption, four separate multilevel analyses were conducted. Each analysis engaged only the teachers of the same stage and not the overall teacher sample. In this way, we could compare the effect size of the variable, concerned with the use of the DIA and the use of the CBA, upon achievement of students who were taught by teachers situated at different stages of teaching competences.

Table 8 illustrates the figures of the final model of each of the four separate multilevel analyses that were conducted. None of the teacher background factors was found to have a statistically significant effect on student achievement and, therefore, teacher background factors are not presented in the table.

Table 7

Parameter Estimates and (Standard Errors) for the analysis of student achievement in mathematics (Students within classes, within schools)

Factors	Model 0	Model 1	Model 2	Model 3
Fixed part (Intercept)	2.19 (0.40)	1.20 (0.12)	0.66 (0.10)	0.34 (0.10)
Student Level				
Context				
Prior achievement in maths		0.64 (.12)	0.64 (.11)	0.64 (.12)
SES		0.41 (.14)	0.41 (.14)	0.40 (.14)
Gender (0=boy, 1=girl)		0.12 (.04)	0.11 (.03)	0.11 (.03)
Classroom Level				
Context				
Average achievement		0.40 (.10)	0.40 (.10)	0.40 (.10)
Average SES		0.21 (.10)	0.21 (.10)	0.21 (.10)
Percentage of girls		N.S.S.	N.S.S.	N.S.S.
Teacher background				
Gender (0=male, 1=female)		N.S.S.	N.S.S.	N.S.S.
Years of experience		0.14 (.04)	0.10 (.04)	0.10 (.04)
Position (0=teacher, 1=deputy head)		N.S.S.	N.S.S.	N.S.S.
Quality of Assessment				
Stage 1			-.34 (.07)	-.33 (.07)
Stage 2			-.19 (.07)	-.18 (.07)
Stage 4			.18 (.07)	.17 (.07)
DIA group				.16 (.06)
CBA group				N.S.S.
School Level				
Context				
Average achievement		0.10 (.04)	0.10 (.04)	0.09 (.04)
Average SES		N.S.S.	N.S.S.	N.S.S.
Percentage of girls		N.S.S,	N.S.S,	N.S.S,
Variance components				
School	9.0%	7.8%	7.1%	6.9%
Class	16.7%	14.2%	10.5%	9.2%
Student	74.3%	45.0%	44.1%	44.0%
Explained		33.0%	38.3%	39.9%
Significance test				
X^2	1033.4	810.1	705.0	651.3
Reduction		223.3	105.1	53.7
Degrees of freedom		7	3	1
p-value		.001	.001	.001

N.S.S. = No statistically significant effect at level .05.

Table 8

Parameter Estimates and (Standard Errors) emerged from separately analyzing achievement of students taught by teachers situated at the same level

Factors	Stage 1	Stage 2	Stage 3	Stage 4
Fixed part (Intercept)	0.65 (.20)	0.58 (.20)	0.62 (.10)	0.63 (0.08)
Student Level				
Context				
Prior achievement in maths	0.64 (.12)	0.65 (.12)	0.68 (.11)	0.63 (.11)
Sex (0=Girls, 1=Boys)	0.10 (.04)	0.10 (.04)	0.11 (.04)	0.10 (.04)
SES	0.33 (.11)	0.30 (.12)	0.35 (.11)	0.31 (.12)
Classroom Level				
Context				
Average achievement	0.35 (.09)	0.37 (.09)	0.35 (.09)	0.36 (.09)
Average SES	0.21 (.09)	0.22 (.09)	0.21 (.09)	0.20 (.09)
Percentage of girls	N.S.S.	N.S.S.	N.S.S.	N.S.S.
Intervention				
DIA	0.11 (.05)	0.15 (.05)	0.19 (.08)	0.18 (.05)
CBA	0.10 (.05)	N.S.S.	N.S.S.	N.S.S.
School Level				
Context				
Average achievement	0.08 (.04)	0.08 (.03)	0.07 (.03)	0.07 (.03)
Average SES	N.S.S.	N.S.S.	N.S.S.	N.S.S.
Percentage of girls	N.S.S.	N.S.S.	N.S.S.	N.S.S.
Variance components				
School	7.1%	7.2%	6.8%	6.7%
Class	8.3%	9.5%	8.7%	9.4%
Student	44.5%	44.3%	44.6%	44.0%
Explained	40.1%	39.0%	39.9%	39.9%

N.S.S. = No statistically significant effect at level .05.

By comparing the four models, the following observations arise. First, all four separate models explained approximately 40.0% of the variance, most of which was attributed to the student level. Second, in all models student background variables were found to have statistically significant effects on student achievement, with prior knowledge having the strongest effect in predicting student achievement at the end of the school year. Third, in analysing the data which emerged from teachers of stage 1, it can be observed that for teachers

of this stage both the DIA (0.11) and the CBA (0.10) had similar impact on student achievement. In all other models (i.e. Stage 2, Stage 3, Stage 4) only the DIA was found to have a statistically significant impact on student achievement. In particular, for teachers of stage 2 the DIA had an impact of .15, whereas for teachers of stage 3 the impact of the DIA was found to be 0.19. The impact of the DIA for teachers situated at stage 4 was found to be 0.18.

It is also important to note that each analysis did not reveal similar effects of the dummy variable concerned with the use of the DIA upon student achievement. The fixed effects obtained with multilevel analysis can readily be converted into standardized effects or 'Cohen's d' by dividing them by the standard deviations in the "treatment groups". Thus, the relative strength of the effects can be more easily compared among the four groups of teachers who are situated at different stages. When the effects of the DIA, presented in Tables 7 (whole sample effects) and 8 (effects per stage), are expressed in this way (see Table 9); they do not turn out to be at the same level.

Table 9

Effect of employing each approach expressed as Cohen's d per group of students taught by teachers situated at the same stage and for the whole sample

Stage	Effect	Pooled S.D.	Cohen's d
Employing CBA			
Teachers at stage 1	0.10	0.76	0.13
Employing DIA			
Teachers at stage 1	0.11	0.77	0.14
Teachers at stage 2	0.15	0.74	0.20
Teachers at stage 3	0.19	0.73	0.26
Teachers at stage 4	0.18	0.72	0.25
Whole sample	0.16	0.96	0.17

The impact of the DIA on student achievement was found to be small (0.14 and 0.20), in cases where teachers of the first two stages were taken into account, whereas relatively higher effect sizes (0.25 and 0.26) were identified in cases where teachers of stages 3 and 4 were taken into account (see Cohen, 1988, p. 19-27). In addition, the two approaches were found to have almost the same effect size (0.13 and 0.14), in cases where data from teachers of stage 1 were taken into account. This implies that the DIA was equally beneficial to the CBA for teachers situated at level 1, however, for teachers situated at the higher stages, only the DIA was beneficial. The overall effect of the DIA was found to be .17.

This chapter presented the analysis of the data collected in order to provide answers to the research questions set. The first section investigated teacher skills in using different techniques of assessment in mathematics. The analysis of the data provided support to the scaling and developmental structure of teachers' abilities in assessment. In particular, it was found that assessment skills can be grouped into four types of assessment behavior, which are discerned in a distinctive way and move gradually from easier to more advanced skills. The second section examines the impact of two professional development approaches (i.e. DIA and CBA) upon teacher skills in assessment as well as student outcomes. The various analysis procedures employed showed that teachers employing the DIA managed to improve their assessment skills more than teachers employing the CBA. The DIA approach was found to be particularly effective for teachers situated at higher stages. Finally, taking student outcomes as criteria of effectiveness, it was found that teachers who use more advanced types of assessment behaviour were more effective than those demonstrating the relatively easy types. Further discussion of the results is presented in the next chapter.

CHAPTER 5

DISCUSSION AND SUGGESTIONS FOR FURTHER RESEARCH

This chapter builds on the findings of the two phases of the study and draws conclusions in relation to the research questions set. First, findings in relation to the measurement of teacher assessment skills are discussed. Next, conclusions in relation to the teacher professional development in assessment are drawn. Finally, suggestions for further research are made.

Measurement of Teacher Assessment Skills

Current teaching practices emphasize in the integration of teaching and assessment, with assessment becoming an on-going process, infused in the everyday teaching practice; therefore, demanding changes in teachers' practice in order for effective learning to take place (Gardner et al., 2010; Wylie & Lyon, 2009). This re-conceptualization of assessment has led to an understanding of the importance of teachers' role in assessment, bringing forward the need for teachers to effectively implement assessment practices. In this context, classroom assessment research appears to be of high priority in the field of education. However, despite the numerous attempts for establishing a theoretical base for classroom assessment (Black & Wiliam, 2006; 2009; Brookhart, 2004; Gipps, 1994; Pryor & Crossouard, 2008; Sadler, 1989), a research gap still exists on what constitutes effective assessment (Perrenoud, 1998; Yorke, 2004) and how it translates into action (Wiliam, Lee, Harrison & Black, 2004). In addition, there is little research investigating teachers' assessment skills either for formative or summative purposes (Mok, 2010; Wiliam, Lee, Harrison & Black, 2004).

In this study, a specific measurement framework was used. The proposed framework directly associated classroom assessment with specific aspects. In particular, the framework took into account the dynamic nature of assessment and thereby the skills associated with each

phase of assessment were examined. In addition, assessment skills were defined and measured in relation to teacher's ability to use specific assessment techniques in order to measure different learning outcomes. Finally, a measurement framework developed, within the field of Educational Effectiveness Research (EER), was adopted and both quantitative and qualitative characteristics of the assessment process were taken into account. Based on the aspects presented above, the framework allowed us to define and measure specific skills associated with assessment practice. This is important because, even though a number of essential assessment concepts, principles, techniques and procedures that teachers need to know have been identified (i.e., Calfee & Masuda, 1997; Fullan, 2000; Popham, 2004), teachers' assessment skills have not until now been systematically addressed (Brookhart, 2011). Moreover, the framework incorporates skills related to the assessment process for both summative and formative reasons, making it possible to examine effective assessment practice irrespective of its purpose orientation. Thus, the framework enables the measurement of classroom assessment's effectiveness not only in terms of its formative purpose; but also in terms of all aspects of the assessment process.

What is also important to note is the fact that even though the main phases of the assessment process were considered as one of the three aspects based on which the framework was developed, this does not imply a view of assessment as a step-by-step model that is "done" by the teacher. On the contrary, the framework is based on current thinking in assessment that views assessment as an on-going, iterative, dynamic process that engages both teacher and learner in the assessment process (Shepard, 2000; Gardner et al., 2010; Wiliam et al., 2004). Without neglecting the sequential character of the four phases in the process of the design and implementation of assessment, this study considers all phases as interrelated and interchangeable. In addition, although the framework focuses on the role of the teacher and how he/she interacts with his/her students and their parents, during the different phases of the

assessment process, this does not however mean that the role of students in assessment is neglected. Effective assessment requires for the student to have ownership of his/her learning (Black & Wiliam 1998; 2004; Tunstall & Gipps, 1996; Harlen, 2005), to be actively involved in the processes of assessment (Lieberman, 1991; Stiggins, 2001; Heritage, 2007) and hold responsibility over his/her actions (Boud & Falchikov, 1989; Black & Harrison, 2001; Nicol &McFarlane- Dick, 2006). Therefore, students assume a more active and responsible role in teaching and learning. Their involvement in negotiating goals and criteria (Boud, 1995; Orsmond et al., 2000; Falchikov, 1995) as well as in assessment methods (Harris & Bell, 1994) facilitates students' responsible and active engagement with assessment. In this context, the framework acknowledges that both teachers and students should have a strong voice in the assessment process (Brookhart, 2011; Suurtamm et al., 2010). For instance, in using the quality and the differentiation dimension to measure assessment skills, the role of students was seriously taken into account. More specifically, the quality dimension is concerned with teachers' skills in providing meaningful opportunities for students to take actions about their own learning based on assessment information (Brookhart, 2011). This was done not only through looking at teachers' skills in reporting assessment results but also in measuring their skills in constructing, administering assessment instruments, and recording assessment data. Therefore, whereas the framework identifies specific teacher skills, these skills are defined in a way that acknowledges the important role that students hold in the assessment process (Stiggins & Chappuis, 2006).

In order to examine whether the proposed framework could be used to design instruments that examine assessment knowledge and skills in relation to actual assessment practice, a teacher questionnaire was developed. As already mentioned, examining teachers' assessment skills though observations was not possible, given the continuous and complex nature of the assessment practice. Thus, the questionnaire was considered as an appropriate

tool for measuring a wide range of assessment skills situated at different phases of teacher's practice. By using the questionnaire, we were able to examine each assessment technique in relation to the four aspects of the assessment process (construction, administration, recording and reporting), whereas for each aspect of the assessment process, each of the five dimensions (frequency, focus, stage, quality and differentiation) was applied. Given that the questionnaire provided self-reported data, issues regarding the validity of the study could be raised. This study addressed validity issues in two ways. First, the analysis of the semi-structured interviews provided evidence supporting the internal validity of the study. Second, the results of the Rasch analysis showed that teacher assessment skills can be measured by using the three aspects of the proposed framework, used in developing the questionnaire. Thus, data analysis provided support for the validity of the proposed framework, as well as the construct validity of the questionnaire.

By using the questionnaire developed, we were able not only to define and measure specific assessment skills, but also to examine whether stage classification can be identified when examining teachers' assessment skills. The results of the study provided support to the initial assumption that teacher assessment skills can be grouped into different developmental levels. In particular, the use of specific measurement dimensions to describe the factor of classroom assessment helped us to group assessment skills into four types of assessment behavior. The developmental scale was consistently identified in both measurement periods (at the beginning and at the end of the intervention) and thereby the generalizability of the results in the context of Cyprus was demonstrated twice. The four stages of teacher assessment behavior identified were described in a distinctive way; thus, addressing a weakness in previous stage related studies to provide a clear picture of what each stage entails (Dall' Alba & Sandberg, 2006). The content of each stage was specifically determined, whereas previous stage models had suffered from vagueness and lack of clarity on what could actually constitute

each developmental stage (Dall' Alba & Sandberg, 2006). In addition, moving away from the commonly applied summative-formative distinction, the four stages identified represented an integrated approach to assessment practice, including various functions and purposes of assessment. The findings described above suggest that the questionnaire developed can be used by institutions offering both pre- and in-service teacher education in order to perform initial evaluation of teachers' assessment skills. This evaluation can serve as a starting point for improvement and further professional development based on the needs identified. At the same time, this tool can be used to evaluate the impact of teachers' professional development programs upon teachers' assessment skills. This could help determine the effectiveness of professional development programs in educational assessment, while at the same time provide information that can be used to improve their quality.

Looking at the description of these four stages, a movement from relatively easy towards more advanced types of teacher behaviour in assessing student knowledge and skills in mathematics can be observed. Starting from skills associated with everyday classroom routines with a mainly summative orientation, a gradual movement towards skills associated with the use of assessment for formative purposes can be observed. This is in line with recent literature supporting that effective teachers use formative–oriented assessment in everyday classroom practice (Black & Wiliam, 1998; Creemers & Kyriakides, 2009; Hattie & Temperley, 2007; Kyriakides & Creemers, 2008a; Wiliam, Lee, Harrison, & Black, 2004). Important conclusions also arise when examining in more detail the content of the four stages. First, the stages appear to provide support to arguments concerning the dynamic nature of the assessment process. The four phases of assessment process, which were used to measure teachers' skills, do not stand independent but on the contrary they are found to coexist in all four stages. This implies that teachers of all four stages are involved in the cycle of assessment, with their skills differentiated in terms of their complexity in each phase.

In particular, teachers of stage 1 and 2 differ in relation to the techniques they use during student assessment. As already mentioned in Chapter 2, learning is multidimensional and cannot be adequately measured by one instrument; therefore, a variety of assessment techniques needs to be employed by the teachers (Brookhart, 2003; Gipps, 1994). In addition, differences in learner characteristics imply that over-reliance on one form of assessment disadvantages students who are able to display their knowledge, skills or abilities more effectively through other methods (e.g., Leder et al., 1999). As the analysis of data shows, stage 1 teachers rely only on the use of written tests, whereas stage 2 teachers use a variety of assessment techniques. This is in line with research, suggesting that especially in the subject of mathematics, teachers feel more comfortable about assessing through a reproduction of a broad range of facts and mathematical operations in timed pencil-and-paper tests (e.g., Firestone et al., 2000; Goodlad, 1984; Powell et al., 1985). On the other hand, teachers of stage 2 are able to use, in an appropriate way, the various techniques of assessment in order to measure basic skills in mathematics. However, teachers of both first two stages appear to use assessment, only in order to achieve summative purposes and they also attempt to measure only basic skills in mathematics. This contradicts the view that assessments need to move beyond the scope of basic skills and be aligned with and support the new ideas of effective mathematics instruction and assessment (Suurtamm, 2004). Moving on, differences between stage 2 and stage 3 teachers are found in terms of the purpose, as well as the content of assessment in mathematics. In particular, stage 3 teachers use assessment to achieve formative purposes and in addition expand the content of their assessment to include more complex educational tasks. It is evident that teachers situated at a higher levels have developed the skills necessary to assess students in a more comprehensive and learning-centred way. Finally, it is important to note that the dimension of differentiation is only present in the last stage. This implies that differentiating assessment across the different phases of the assessment

process and in relation to different techniques is more difficult to be achieved, as the analysis of data using the Rasch and Saltus models has shown. This finding is in line with previous studies that found the differentiation of instruction situated at higher levels of teacher development (Kyriakides, Creemers & Antoniou, 2009; Kyriakides, Archambaul & Janosz, 2013)

Classifying teacher skills into levels of difficulty is only meaningful if these can be related to student achievement (Kyriakides, Creemers & Antoniou, 2009). Thus, a question that may rise here is related to the extent to which the classification of teachers into these four stages explains variation in student achievement. This study stresses the need to identify those activities associated with classroom assessment, which have positive impact on student outcomes. Therefore, taking student outcomes as criteria of effectiveness it was examined whether teachers who use more advanced types of behaviour were more effective than those who demonstrate the relatively easy types. Multileveled analysis showed that the developmental stage, at which a teacher is situated, had a statistically significant effect on student achievement. Specifically, students of teachers at stage 1 had the lowest achievement, whereas students of teachers at level 4 had higher achievement than students of the first three levels. Therefore, it was found that teachers exercising more advanced types of assessment behaviour had better student outcomes. Thus, this study contributes to the identification of assessment skills that have a positive impact on student achievement. This is important since, as already mentioned, most studies investigating the impact of assessment practice on student achievement come from the area of formative assessment (Black & Wiliam, 1998; Wiliam, Lee, Harrison & Black, 2004) and their findings have been questioned (Bennett, 2011). In this study effectiveness was investigated not only in terms of the formative purpose of assessment but across all aspects of the proposed framework. Thus, the findings can be used to determine what constitutes effective assessment and how it translates into action.

The fact that teacher assessment skills were identified and that they were found to be related to student achievement outcomes implies that effective assessment practices can be defined and promoted through teacher professional development programs. Therefore, the results of this study can be used by higher institutions, providing initial and in-service training to adjust their curriculum in order to provide adequate and appropriate assessment training to prospective and in service-teachers. Furthermore, stage identification suggests that policy should move on to establish mechanisms that allow stage examination in order to provide appropriate assistance to teachers. Research on the development of expertise suggests that teachers at different stages of development have differentiated needs (Fullan, 1990, Hargreaves, 1994), while professional development opportunities which are planned and focused upon teachers' needs, are more likely to be effective (Eraut, 1995; Duncombe & Armour, 2004). Thus, for teachers situated at stage 1, professional development programs could include training on everyday assessment routines such as the enrichment or alteration of ready-made written tests and the use of different types of written questions to assess students' performance. For teachers at stage 2, professional development could address the use of different assessment techniques, the use of both product and process questions; as well as the use of descriptive comments to give feedback to their students. Training for teachers of stage 3 could have as a basic aim the development of assessment for formative purposes. Thus, development programs could address issues related to the use of various assessment techniques to measure more complex educational objectives in ways that promote learning and the recording and reporting of the assessment information to both students and parents. Finally, professional development for stage 4 teachers could focus on developing skills for the differentiation of assessment in and for different occasions and students.

Teacher Professional Development in Assessment

The literature highlights the growing evidence and recognition of the importance of professional development in equipping teachers to meet contemporary challenges (Darling-Hammond, 2000; Guskey, 2003). However, it also highlights the failure of professional development to adequately address assessment (Stiggins, 2002; Popham, 2004). It is argued that although teachers spend a large amount of teaching time in assessment related activities (Crooks, 1988; Herman & Dorr-Bremme, 1982; Stiggins, 1991; Stiggins & Conklin, 1992), they still lack the necessary skills to effectively assess their students in the everyday classroom (Lukin et al., 2004; Stiggins, 2002). In addition, given the popularity of formative assessment, the fact that a lot is done under its name and not all of it is helpful, testifies to a need for improved formative assessment knowledge and skills (Brookhart, 2011).

Professional development in assessment appears as a controversial issue in the literature. One line of research recognizing the inadequate assessment training at both pre- and in-service teacher education (Popham, 2004; Stiggins, 1991) shifts the attention to the need for teachers to understand the principles of sound assessment in order for effective practice to be achieved. On the other hand, another line of research brings forward other factors, besides teacher competence, that impact the effectiveness of assessment practice such as the role of the classroom assessment culture (Shepard, 2000), teachers' perceptions and beliefs (Brown, 2004; Pajares, 1992), as well as the formative function of assessment (Black & Wiliam, 1998; Creemers & Kyriakides, 2008). As a result, teacher professional development efforts to address assessment practice either adopt a competency based approach (e.g., Assessment Literacy Project [KSDE]) providing training sessions in order to enhance teachers' assessment literacy or alternatively adopt a more holistic approach by creating professional learning communities (Brookhart, Moss & Long, 2010). Taking the above into consideration, this study investigated in more detail the development of teachers' skills in assessment.

In particular, an experimental study was conducted in order to examine the impact of two different professional development approaches on teachers' assessment skills and students' achievement. Both approaches addressed teacher skills in classroom assessment, however with fundamental differences to their content and structure. The first approach was competency based and offered all teachers the same training on assessment skills in order for them to acquire the necessary competencies in assessment. The second approach, the dynamic integrated approach, adopted a theory driven evidenced based approach to professional development advocating the provision of professional development adjusted to teachers' developmental stage (Creemers & Kyriakides, 2010). Teachers of both groups attended 7 professional development session and were encouraged to get involved in action research in order to improve their assessment skills. Throughout the meetings, support was provided by the research team to teachers of both groups, recognizing that effective integration of new skills requires opportunities to practice the new skills and receive feedback (Ingvarson, Meiers & Beavis, 2005). The two approaches differ in relation to two important aspects. First, the DIA is based on the assumption that the content of the professional development program should address and therefore be differentiated to meet the needs and priorities of teachers at each developmental stage. Therefore, four focus areas were created, with each focus area addressing assessment skills found to be situated at the same level. On the other hand, the CBA does not address specific groupings of assessment skills, as the DIA does. The primary aim of CBA sessions was to improve teachers' competence in assessment by providing the necessary knowledge associated with all the identified assessment skills. Thus, all of the skills in the four focus groups were covered, and it was expected that every teacher could master all the assessment skills. The second basic difference between the two approaches refers to the use of reflection as an essential tool for improvement. Whereas, for teachers employing the DIA, opportunities to engage in reflection on their assessment practices throughout the

sessions were provided; for the CBA group, reflection was not considered necessary since the basic assumption of this approach is that training in the assessment skills identified as important is adequate to achieve improvement.

Data were analyzed in order to investigate which of the two professional development approaches had greater impact on teacher assessment skills and student learning outcomes. The analysis of the data showed that teachers employing the DIA managed to improve their assessment skills more than teachers employing the CBA. Particularly, 13 out of 36 teachers of the DIA group managed to move to the next, more demanding level of assessment competence, whereas the other 23 teachers remained at the same stage. Specifically, 4 teachers of this group moved from stage 1 to stage 2, 6 teachers of stage 2 managed to move to stage 3 and finally 3 teachers situated at stage 3 were found to be at stage 4 at the end of the intervention. On the other hand, only 5 teachers of the group employing the CBA, managed to move to the next stage, whereas the rest 31 teachers remained at the same stage. More specifically, 4 teachers managed to move from stage 1 to stage 2, 1 teacher moved from stage 2 to stage 3 whereas teachers situated at stages 3 and 4 did not make any statistically significant progress in their skills. The larger impact of the DIA on teacher assessment skills identified could be attributed to the way this approach differentiates from the CBA; since teachers employing the DIA were encouraged to reflect critically on specific assessment skills associated with their level of competence.

Furthermore, looking at the progress of teacher assessment skills, it can be observed that in the cases where change occurred, this change was towards the next demanding level. This implies that the improvement of assessment skill took place gradually, since all progress was done in a stepwise manner (i.e., from stage 1 to stage 2, from stage 2 to stage 3 etc.). It is however, important to note that teachers situated at stage 4, in both intervention groups, did not manage to move to the next development stage. This might imply that the upward

movement from stage 4 is more difficult to be achieved than the upward movement from stage 1, 2 and 3. This assumption is also supported by the symmetry indices estimated by the Saltus model. Another issue raised is that movement towards the next stage was identified to teachers of the DIA that were found to belong to stages 1, 2, and 3 at the beginning of the intervention. On the other hand, upward movement of teachers employing the CBA was identified only in stages 1 and 2. This could be attributed to the fact that teachers of higher level of competence, employing the CBA, spend a lot of time developing and applying action plans, referring to skills that they already possessed. For example, teachers situated at level 3, employing the CBA, developed action plans that aimed to encourage record keeping for data deriving from all assessment techniques. However, based on the evaluation of their skills, teachers of this stage were already skilled in record keeping and therefore this particular involvement in action research had no added value. In addition, the findings suggest that improvement of teacher skills takes place gradually. This implies that it is perhaps naive to believe that the methods of the experts can or should be taught directly to beginners (Combs et al., 1974), thus having important implications for the content of both pre- and in-service teacher education. Teacher education needs to be in line and to address teachers' needs. Trying to engage a non-experienced teacher in activities aiming at differentiated assessment, when he/she has not yet acquired basic assessment skills, may not be appropriate. The same applies when you invite an experienced teacher with advanced skills in test construction, to attend an introductory course in test development.

The research suggests that teachers can improve and ultimately progress to the next developmental stage of assessment skills, by undertaking appropriate interventions and participating in effective professional development programs (Creemers, Kyriakides & Antoniou, 2013). This argument is supported by the fact that teachers of the control group did not manage to improve their assessment skills, while all of them remained at the same stage

that they were found to be situated at the beginning of school year. On the other hand, teachers employing either the DIA or the CBA managed to improve their assessment skills. The added value of using the DIA rather than the CBA to design a teacher professional development programme was identified by comparing the progress in assessment skills that each intervention group managed to achieve, since teachers of the DIA managed to improve their skills more than those employing the CBA. This appears to provide support to the main assumption of the DIA that training initiatives are more effective when they are structured to correspond to the professional needs of teachers (Creemers, Kyriakides & Antoniou, 2013). Identifying individual teacher's needs and providing realistic and differentiated staff development allows teachers to build on their existing personal and professional strengths and grasp learning opportunities (Fullan, 1990, Hargreaves, 1994, Hargreaves & Hopkins, 1991, Newton &Tarrant, 1992). In addition, whereas the CBA focused only on competency development teachers employing the DIA were also engaged into systematic and guided critical reflection on their assessment practices, suggesting that a focus on both competence and reflection is necessary in order to achieve effective professional development. In this way, it is possible for professional development to result in changes in practice rather than merely changes in knowledge levels (Darling-Hammond, 1998; Lieberman, 1996; Wilson & Berne, 1999), since it is argued widely that what teachers know does not necessarily define what teachers practice (Elmore, 1996; Kauffman, 1996; Kennedy, 1997; Greenwood & Abbott, 2001; Gersten, Vaughn, Deshler, & Schiller, 1997).

Moving a step forward and taking student outcomes as criteria of effectiveness, it was found that when analyzing the whole sample effects, only the DIA yielded better results in student achievement than the control group, whereas the CBA was found not to have any significant effect on student achievement. However, examining further the impact of the DIA on student achievement, this time focusing on the effect per stage, it was found that its impact

was small in cases where teachers of the first two stages were taken into account, whereas relatively higher effect sizes were identified in cases where teachers of stages 3 and 4 were taken into account. In addition, the two approaches were found to have almost the same effect size, in cases where data from teachers of stage 1 were taken into account. This implies that whereas only the DIA was beneficial in relation to student outcomes for teachers situated at the higher stages; the DIA was equally beneficial to the CBA for teachers situated at stage 1. This could be attributed to the fact that teachers of stage 1 lack the necessary routines and basic skills which are easier to address whereas those situated at the higher stages need more specific interventions that are focused on their needs. What is important to note is that whereas positive changes in student outcomes are the ultimate measure of professional development's success, until now there is relatively little systematic research conducted on the effects of professional development on student outcomes (Garet et al., 2001). The positive impact of the DIA on student outcomes, found in this study, provides further support to the argument that professional development programs should employ the DIA in order for effectiveness to be achieved (Creemers, Kyriakides & Antoniou, 2013).

This study has showed that the DIA was more effective than the CBA in improving teachers' skills and student outcomes. This finding comes to agree with the argument that when traditional professional development approaches are employed, despite the good intentions, teachers' practice and students' chances for academic success remain unchanged (Tyack & Tobin, 1994). These findings also have important implications for policy and practice. Since the findings showed that the DIA is more effective in improving teachers' skills and students' achievement, perhaps policy should implement professional development programs employing the DIA in its official teacher training. This way, teachers will be engaged in professional development opportunities that encourage them to reflect critically on specific assessment skills associated with their level of competence. In order to do so,

evaluation mechanisms of teachers' assessment skills also need to be established. These can be used to evaluate teachers' needs in order to offer appropriate professional development as well as to examine the effectiveness of the program offered, based on the changes identified in relation to teachers' skills.

Summarily, the findings of this study, as these are discussed above, provide important information in relation to the measurement of teacher assessment skills, as well as the professional development in assessment. Given that researchers in the area of educational effectiveness highlight the need for establishing stronger links between educational effectiveness research and improvement of practice (Creemers & Kyriakides, 2008; Reynolds, Hopkins & Stoll, 1993) implication of the findings for policy and practice especially in the context of Cyprus are presented next.

Implications for Policy and Practice in the context of Cyprus

This research is in line with recent attempts to merge the findings of educational research with initiatives for improvement in educational policy and practice. This stands especially in the context of Cyprus, since from 2005 an ambitious educational reform attempt has been launched (Ministry of Education and Culture, 2008).

First, this study has identified assessment skills that have a positive impact on student achievement. In particular, this study provided further support to previous research suggesting that effective teachers use formative–oriented assessment in everyday classroom practice (Black & Wiliam, 1998; Hattie & Temperley, 2007; Wiliam, Lee, Harrison, & Black, 2004). The use of assessment for formative purposes is also of the main objectives of the Cyprus' educational reform initiative (Ministry of Education and Culture, 2004). However, as already mentioned, despite the positive attitudes of Cypriot teachers towards formative assessment, only a limited number of teachers actually implement it in their everyday practice (Creemers, Kyriakides & Antoniou, 2013). The results of this study can be used by higher institutions of

Cyprus providing teacher education degrees to adjust their curriculum in order to provide adequate and appropriate assessment training to prospective teachers. In addition, the results can be used to inform educational policy in order to move forward to the establishment of assessment targeted training and professional development opportunities for in- service teachers. Developing assessment skills thought well-targeted pre- and in-service teacher education could contribute to a more effective use of assessment in the everyday classroom.

In addition, given the support provided to the questionnaire developed to measure teacher assessment skills, the questionnaire could be used to examine assessment knowledge and skills in relation to actual assessment practice. Thus, the questionnaire could be used by educational institutions offering teacher education and professional development in order to identify their trainees' needs and subsequently adjust the content of the training provided accordingly. Moreover, the questionnaire could be used as part of the initial training for newly hired teachers. Its use could help identify beginning teachers' needs in relation to assessment in order for their mentors to provide appropriate and targeted support. Furthermore, with appropriate adjustments the questionnaire could be used as a teacher self-assessment instrument. Self-assessment is a powerful technique for self-improvement and professional growth (McDonald & Boud, 2003). Teachers could use it as learning tool in order to examine their practice and identify areas for improvement.

Finally, this study has provided support for the effectiveness of the DIA to teacher professional development. Particularly, the DIA was found more effective than the CBA in improving teachers' skills and student outcomes. Given that teacher training in Cyprus has mainly a competency-based orientation; policy should be directed towards the development of professional development opportunities that employ the DIA in order to improve teacher assessment skills and their effectiveness status.

Suggestions for Further Research

As many OECD countries are beginning to develop commonalities of understanding and practice in relation to classroom assessment (Sebba, 2006), the difficulties in effective implementation need to be identified and tackled by researchers and policy makers. As Torrance and Pryor (2001) argue, it is necessary to explore systematically teachers' daily practice in order to facilitate the firm grounding of future programmes of change. Without systematic analyses of classroom assessment, based on empirical research in classrooms, research evidence can only provide us with a limited understanding of the nature and process of assessment in the service of learning. This study used a specific framework to measure teacher skills in assessment and investigated the effectiveness of two different professional development approaches in improving teacher assessment skills and student achievement. The findings revealed that assessment skills can be grouped into 4 stages of assessment behaviour and that the Dynamic Integrated Approach is more effective than the Competency-Based Approach in improving assessment skills and student achievement.

However, further research is needed in order for more in-depth analysis of the findings. First, this study used a teacher questionnaire to measure teacher assessment skills. Thus, the use of other methods to measure teacher assessment skills also needs to be investigated. Using a variety of classroom data such as content analysis of teachers' tests and assessment records, student surveys, students' work samples as well as classroom videotaping could provide important insights in relation to teachers' skills assessment.

In addition, while this study has provided evidence that teacher skills in assessment can be grouped into certain types of assessment behaviour, more studies in this area are necessary. More specifically, given the fact that the study was conducted in a single country and was concerned with primary teachers' assessment skills in mathematics, further research is needed in order to test the generalisability of the findings of this study. Future studies will have to

face more issues in developing comparable questionnaires, since not only adaptation but also translation of questionnaire items might be needed. Whether the developmental stages of classroom assessment skills can also be identified in measuring skills of teachers in assessment of student achievement in various subjects (not only in mathematics) and in assessment of students at different phases of schooling (not only at primary school level) should be further investigated in order to test the generalisability of the findings.

Moreover, further research is needed in order to examine the generalizability of the results of the experimental study. Particularly, future research might investigate further the effectiveness of the Dynamic Integrated Approach in different educational contexts and settings. It is also important to recognize that the long-term effect of the interventions needs to be also investigated. This is important since, changes due to interventions to improve assessment skills may revert to baseline after the intervention stimulus ends (Hargreaves, 2002). Thus, a follow – up measurement could help us consider the dynamic character of effectiveness since not only teachers moving to a higher stage might be identified but also teachers who drop to a lower stage for several reasons (including burnout) could be identified. Longitudinal studies that expand for more than one year are therefore necessary in order to examine, not only the long term effect of the interventions, but also the sustainability of the effects.

Finally, it is necessary to acknowledge that both intervention groups provided professional development externally, since they took place on non-working hours and teacher participation was on a volunteer basis. Teachers of both groups were working towards improvement on an individual basis, irrespective of their colleagues at school or their school's policies. However, it is argued that teacher professional development is not sufficient without considering the larger system in which teachers find themselves (Wylie & Lyon, 2009) and thus the working context is considered as the most suitable place for professional development

(Hargreaves, 1997; Retallick, 1999; Scribner, 1999). According to Evans (1993) school-based in service education derives from the curriculum needs and plans of the school. It may concern the school as a whole or in part, as well as provide for the individual teacher's in service needs. Since, school-based teacher professional development is a growing trend in many European countries, it is necessary to investigate further a school-based professional development intervention and to examine the added value of using the DIA to develop school-based INSET courses. Comparing external and school-based professional development opportunities will help identify which professional development approach is more effective in bringing about improvement to both teachers and students. It will also help identify school factors that may influence the effectiveness of different professional development approaches, since it is argued that school wide influences are considered as an element that impacts teacher development (Joyce & Showers, 1995).

REFERENCES

Abramson, S. (2006). Documentation—communication—action: Co-inquiry meetings for facilitated interchange. *Co-Inquiry Journal, 1*(1), 1-14.

Abramson, S., & K. Atwal. (2003). Teachers as co-inquirers. In *Next steps in teaching the Reggio way,* J. Hendrick (Ed.),(pp. 86–95). Englewood Cliffs, NJ: Merrill.

Adams, K. (1996). S/NVQs and assessment. *Training Tomorrow,* 10(4), 14-15.

Adams, R. J., & Khoo, S.T (1996). *Quest - the interactive test analysis system*. Australian Council of Educational Research.

Airasian, P. W. (2001*). Classroom assessment* (4th ed.). Boston: McGraw-Hill.

American Educational Research Association. (1999). *Standards for educational and psychological testing*. Washington, D.C.: Author.

American Federation of Teachers, National Council on Measurement in Education, & National Education Association (AFT/NCME/NEA). (1990). Standards for teacher competence in educational assessment of students. *Educational Measurement: Issues and Practice, 9*(4), 30-32.

Andersen, E. B. (1977). Sufficient statistics and latent trait models, *Psychometrika*, 42, 69–81.

Anderson, L. (2003). *Classroom assessment: Enhancing the quality of teacher decision making.* Mahwah, NJ: Erlbaum.

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

Angelo, T. A. (1995). Reassessing and defining assessment. AAHE Bulletin (Nov.), 7-9.

Antoniou, P. (2009). *Using the Dynamic Model of Educational Effectiveness to Improve Teaching Practice: Building an Evaluation Model to Test the Impact of Teacher Professional Development Programs.* Unpublished Doctoral Dissertation, University of Cyprus, Cyprus.

Assessment Reform Group. (2002). *Assessment for Learning: 10 principles research-based principles to guide classroom practice*, Assessment Reform Group, London, United Kingdom.

Ball, D. L., & Cohen, D. K. (1999). Developing practice, developing practitioners: Toward a practice-based theory of professional education. In L. Darling-Hammond and G. Sykes (Eds.), *Teaching as the learning profession* (pp. 3-31). San Francisco, C A:Jossey-Bass.

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

Bangert-Drowns, R. L., Kulik J. A., & Chen-Lin C. Kulik, C.l.C. (1985). Effectiveness of Computer-Based Education in Secondary Schools. *Journal of Computer-Based Instruction*, 12(3), 59-68.

Barnett, R. (1994). *The Limits of Competence. Knowledge, Higher Education and Society*, Buckingham: Open University Press.

Becker, B. J., Kennedy, M. M., & Hundersmark, S. (2003, April). Communities of scholars, research, and debates about teacher quality. *Paper presented at the annual meeting of the American Educational Research Association*, Chicago.

Benner, P. (2004). Using the Dreyfus model of skill acquisition to describe and interpret skill acquisition and clinical judgment in nursing practice and education. *Bulletin of Science, Technology & Society*, 24(3),188.

Bennett, R. (2011). Formative assessment: a critical review. Assessment in Education, 18(1), 5-25.

Berliner, D. C. (1994). Expertise: The wonder of exemplary performances. In J.N. Mangieri & C.C. Block (Eds.), *Creating powerful thinking in teachers and students: Diverse perspectives* (pp. 161-186). Fort Worth, TX: Holt, Reinhardt, & Winston.

Berry, R. (2008). *Assessment for learning*. Hong Kong University Press.

Biemans, H., Nieuwenhuis, L., Poell, R., Mulder, M., & Wesselink, R. (2004). Competence-based VET in the Netherlands: background and pitfalls. *Journal of Vocational Education and Training*, 56, 523-538.

Billett, S. (2001). Knowing in practice: Re-conceptualising vocational expertisen. *Learning and Instruction*, 11(6), 431-452.

Birenbaum, M. (2003). New insights into learning and teaching and their implications for assessment. In Segers, M., Dochy, F. & Cascallar, E. (Eds.). Optimizing New Modes of Assessment: In Search of Qualities and Standards (pp. 13-36). Dordrecht: Kluwer Academic Publishers.

Birenbaum, M., Kimron, H., Shilton, H., & Shahaf-Barzilay, R. (2009). Cycles of inquiry: Formative assessment in service of learning in classrooms and in school-based professional communities. *Studies in Educational Evaluation* 35(4), 130–149.

Black, P. (1993). Formative and summative assessment by teachers. *Studies in Science Education*, 21, 49 – 97.

Black, P. & Harrison, C. (2001). Self and peer assessment and taking responsibility: the science student's role in formative assessment. *School Science Review*, 83(302), 43-49.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, 80(2), 139-148.

Black, P., & Wiliam, D. (2005). *Developing a theory of formative assessment*. In J. Gardner (Ed.), Assessment and Learning, London, UK: Sage.

Black, P., & Wiliam, D. (2006). Developing a theory of formative assessment. In J. Gardner (Ed.), *Assessment and learning* (pp. 81–100). London: Sage.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment Evaluation and Accountability*, 21(1), 5–31.

Black, P. J., Harrison, C., Lee, C., Marshall, B. & Wiliam, D. (2003). *Assessment for learning: putting it into practice.* Maidenhead: Open University Press.

Black, P., McCormick, R., James, M. & Pedder, D. (2006). Learning how to learn and Assessment for Learning: a theoretical inquiry. *Research Papers in Education,* 21(2), 119–132.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on formative and summative evaluation of student learning*. New York: McGaw-Hill.

Boaler, J. (2008). *What's math got to do with it?* London: Penguin Press.

Bond, T.G., & Fox, C.M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, N. J.: Lawrence Erlbaum Associates, Publishers.

Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. *Educational Researcher, 33*(8), 3–15.

Borko, H., & Putnam, R. (1996). Learning to teach. In D. Berliner & R. Calfee (Eds.), Handbook of educational psychology (pp. 673-708). New York: Macmillan.

Borko, H., Mayfield, V., Marion, S., Flexer, R. & Cumbo, K. (1997). Teachers' developing ideas and practices about mathematics performance assessment: successes, stumbling blocks, and implications for professional development. *Teaching and Teacher Education*, *13*, 259-278.

Borko, H., Peressini, D., Romagnano, L, Knuth, E., Yorker, C., Wooley, C., Hovermill, J., & Masarik, K. (2000). Teacher education does matter: A situative perspective of learning to teach secondary mathematics. *Educational Psychologist, 35*(3), 193-206.

Boud, D. (Ed.). (1988). *Developing Student Autonomy in Learning*, London: Kogan Page, Second Edition.

Boud, D. (1995). *Enhancing Learning through Self Assessment*. London: Kogan Page.

Boud, D.J. (2000). Sustainable assessment: rethinking assessment for the learning society. *Studies in Continuing Education*, 22(2), 151-167.

Boud, D. & Falchikov, N. (1989). Quantitative studies of student self assessment in higher education: a critical analysis of findings. *Higher Education*, 18, 529-549.

Bracht, G. H., & Glass, G. V. (1968). The external validity of experiments. *American Education Research Journal*, 5, 437-474.

Braun, J. A. & Crumpler, T.P. (2004). The social memoir:an analysis of developing reflective ability in pre-service methods course. *Teaching and Teacher Education*, 20, 59-75.

Broadfoot, P. (1992). Exploring the forgotten continent: a traveller's tale. *Scottish Educational Review.* 26(2), 88-96.

Broadfoot. P. (1996). *Education, assesment and socitey*. Open University Press.

Broadfoot, P., & Black, P. (2004). Redefining assessment? The first ten years of Assessment in Education. *Assessment in Education, 11*(1), 7-27.

Brookhart, S. M. (1997). A theoretical framework for the role of classroom assessment in motivating student effort and achievement. *Applied Measurement in Education*, 10, 161–180.

Brookhart, S. M. (2003). Developing Measurement Theory for Classroom Assessment Purposes and Uses. *Educational Measurement: Issues and Practice, 22*(4), 5–12.

Brookhart, S. M. (2004). Classroom assessment: tensions and intersections in theory and practice. *Teachers College Record*, 106(3), 429-458.

Brookhart, S.M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice*, 30(1), 3–12.

Brookhart, S. M., Moss, C. M. & Long, B. A. (2010). Teacher inquiry into formative assessment practices in remedial reading classrooms. *Assessment in Education: Principles, Policy & Practice*, 17(1), 41-58.

Brown, G. T. L. (2004). Teachers' conceptions of assessment: Implications for policy and professional development. *Assessment in Education: Policy, Principles and Practice*, 11(3), 305-322.

Brown, G., Bull, J., & Pendlebury, M. (1997). *Assessing student learning in higher education*. London: Routledge.

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

Burry, J., & Shaw, D. (1988). Teachers and administrators differ in assessing teacher effectiveness. *Journal of Personnel Evaluation in Education, 2*(1), 33-41.

Calfee, R. C., & Masuda, W. V. (1997). Classroom assessment as inquiry. In G. D. Phye (Ed.) *Handbook of classroom assessment: Learning, adjustment, and achievement*. NY: Academic Press.

Cameron, J. & Pierce, D.P. (1994). Reinforcement, reward, and intrinsic motivation: a meta-analysis, *Review of Educational Research*, 64, 363-423.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171–246). Chicago, IL: Rand McNally.

Carr, W., & Kemmis, S. (1986). *Becoming critical: Education, knowledge and action research*. Geelong: Deakin University Press.

Ceci, S. J., & Liker, J. K. (1986). A day at the races: A study of IQ, expertise and cognitive complexity. *Journal of Experimental Psychology: General*, 115, 225-266.

Chapman, C. & King, R. (2005). *Differentiated assessment strategies: One tool doesn't fit all*. Thousand Oaks: Corwin Press.

Clark, I. (2011). Formative assessment: Policy, perspectives and practice. *Florida Journal of Educational Administration & Policy*, 4(2), 158–180.

Cochran-Smith, M., & Lytle, S. (2001). Beyond certainty: Taking an inquiry stance on practice. In A. Lieberman & L. Miller (Eds.), *Teachers caught in the action: Professional development that matters* (pp. 45-58). New York: Teachers College Press.

Cochran-Smith, M. & Zeichner, K. (Eds.). (2005). *Studying teacher education.* New York: Routledge.

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

Cohen, D.K., & Ball, D.L. (1999). *Instruction, capacity, and improvement* (CPRE Research Report No. RR-043). Philadelphia: University of Pennsylvania, Consortium for Policy Research in Education.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Combs, A.W., Blume, R. A., Newman, A. J., & Wass, H. L. (1974). *The professional education of teachers: A humanistic approach to teacher preparation* (2nd ed.). Boston: Allyn and Bacon.

Cools, W., De Fraine, B., Van den Noortgate, W., & Onghena, P. (2009). Multilevel design efficiency in educational effectiveness research. *School Effectiveness and School Improvement*, *20*, 357–373.

Cornford, I.R. (2002). Reflective teaching: Empirical research findings and some implications for teacher education. *Journal of Vocational Education & Training, 54*, 219–236.

Cowie, B., & Bell, B. (1999). A model of formative assessment in science education. *Assessment in Education*, 6(1), 102–116.

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

Creemers, B.P.M., & Kyriakides, L. (2009). Situational effects of the school factors included in the dynamic model of educational effectiveness. *South African Journal of Education, 29*(3), 293-315.

Creemers, B. P. M. & Kyriakides, L. (2010). Using the Dynamic Model to Develop an Evidence-Based and Theory-Driven Approach to School Improvement. *Irish Educational Studies,* 29 (1), 5-23.

Creemers, B.P.M., & Kyriakides, L. (2012). *Improving Quality in Education: Dynamic Approaches to School Improvement*. London: Routledge.

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

Creemers, B.P.M, Kyriakides, L., & Sammons, P. (2010). *Methodological Advances in Educational Effectiveness Research.* London: Routledge Taylor Francis.

Cronbach L. J. (1990). *Essentials of Psychological Testing* (5th ed.). New York: Harper Collins.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58,* 438-481.

Dall'Alba, G. (2004). Understanding professional practice: investigations before and after an educational programme. *Studies in Higher Education,* 29(6), 679–692.

Dall'Alba, G. & Sandberg, J. (1996). Educating for competence in professional

Practice. *Instructional Science,* 24, 411-437.

Dall'Alba, G. & Sandberg, J. (2006). Unveiling professional development: a critical review of stage models. *Review of Educational Research,* 76, 383–412.

Darling-Hammond, L. (1998). Teachers and teaching: Testing policy hypotheses from a national commission report. *Educational Researcher, 27* (1), 5-15.

Darling-Hammond, L. (2000). Reforming teacher preparation and licensing: Debating the evidence. *Teachers College Record,* 102(1), 28-56.

Day, C., (2002). School reform and transitions in teacher professionalism and identity. *International Journal of Educational Research,* 37(8), 677-692.

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 first year of junior general education in the Netherlands. *School Effectiveness and School Improvement,* 15(1), 3-31.

De Lange, J. (1993). Assessment in problem-oriented curricula. In N.L. Webb & A.F. Coxford (Eds.), *Assessment in the mathematics classroom (NCTM Yearbook)* (pp. 197 – 208). Reston, VA: NCTM.

De Leeuw, J., & Meijer, E. (2008). Introduction to multilevel analysis. In J. de Leeuw & E. Meijer (Eds.), *Handbook of multilevel analysis* (pp. 1–75). New York: Springer.

Dekker, T. & Feijs, E. (2005) Scaling up strategies for change: Change in formative assessment practices. *Assessment in Education,* 12(3), 237–254.

Delandshere, G. (2002). Assessment as Inquiry. *Teachers College Record,* 104(7), 1461-1484.

DeLuca, C., & Klinger, D. A. (2010). Assessment literacy development: Identifying gaps in teacher candidates' learning. *Assessment in Education: Principles, Policy & Practice, 17,* 419-438.

Demetriou, D. (2009). *Using the dynamic model to improve educational practice.* Unpublished doctoral dissertation, University of Cyprus.

Denzin, N.K. & Lincoln, Y.S. (1998). *Collecting and interpreting qualitative materials.* Thousand Oaks, CA: Sage.

DES/WO (1988). *National Curriculum Task Group on Assessment and Testing—a report.* London: DES.

Dewey, J.(1933). *How we think*: *A Restatementof the relation of reflective thinking to the educative process.* Chicago: Henry Regnery.

Dreyfus, H. L., & Dreyfus, S. E. (1986). *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*, New York: Free Press.

Duncombe, R. & Armour, K. (2004). Collaborative Professional Learning: From Theory to Practice. *Journal of In-Service Education*, 30(1), 141-166.

Dunn, K. E., & Mulvenon, S. W. (2009). A critical review of research on formative assessments: The limited scientific evidence of the impact of formative assessments in education. *Practical Assessment & Research and Evaluation*, 14(7), 1-11.

Earl, L., & Katz, S. (2000). Changing classroom assessment: Teachers' struggles. In N. Bascia & A. Hargreaves (Eds.), *The sharp edge of educational change* (pp. 97–111). London: Routledge.

Elmore, R. (1996). Getting to Scale with Good Educational Practice. *Harvard Educational Review*, 66(1), 1-26.

Engeström, Y. & Miettinen, R. (1999). Introduction. In Y. Engeström., R. Miettinen & RL. Punamäki (1999) (Eds), *Perspectives on Activity theory.* Cambridge: Cambridge University Press.

Eraut, M. (1995). In Service Teacher Education. In L.W. Anderson (Ed.), *International Encylopedia of Teaching and Teacher Education.* Oxford: Pergamon Press.

Evans, T. K. (1993). *School Based Inservice Education*. Phaedon, Culemborg.

Falchikov, N. (1995). Peer feedback marking: developing peer-assessment. Innovations in *Education and Training International, 32*, 175-187.

Feiman-Nemser, S. & Remillard, J. (1996). Perspectives on learning to teach. In F. B. Murray (Ed.), *The teacher educator's handbook: Building a knowledge base for the preparation of teachers* (pp. 63-91). San Francisco: Jossey-Bass

Fendler, L. (2003). Teacher reflection in a hall of mirrors: Historical influences and political reverberations. *Educational Researcher*, 32(3), 16-25.

Firestone, W.A., Winter, J. & Fitz, J., (2000). Different assessments, common practice? Mathematics testing and teaching in the USA and England and Wales. *Assessment in Education*, 7(1), 13–37.

Foddy, W. (1993). *Constructing questions for interviews and questionnaires: Theory and practice*. Melbourne: Cambridge University.

Fox, C. M., & Jones, J. A. (1998). Uses of Rasch modeling in counseling psychology research. *Journal of Counseling Psychology*, 45, 30-45.

Fullan, M.G. (1990). Staff development, innovation, and institutional development. In *Changing school culture through staff development*, 1990 ASCD yearbook (3-25). Alexandria, VA: ASCD.

Fullan, M. (2000). The return of large-scale reform. *Journal of Educational Change*, 1(1), 5–27.

Fuller, F., & Bown, O. (1975). Becoming a teacher. In K. Ryan (Ed.), *Teacher education.* Chicago: University of Chicago Press.

Gandini, L., & J. Goldhaber. (2001). Two reflections about documentation. In , L. Gandini & C. Edwards (Eds.), *Bambini: The Italian approach to infant/toddler care*, (pp.124–45). New York: Teachers College Press.

Gardner, J., Harlen, W., Hayward, L., Stobart, G. & Montgomery, M., (2010). *Developing Teacher Assessment*. Maidenhead: Open University Press.

Garet, M.S., Porter, A.C., Desimone, L., Birman, P.F., & Yoon, K.S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal*, 38(4), 915-945.

Gersten, R., Vaughn, S., Deshler, D., & Schiller, E. (1997).What we know about using research findings: Implications for improving special education practice. *Journal of Learning Disabilities*, 30, 466–476.

Gipps, C. (1994). *Beyond Testing*. RoutledgeFalmer, London.

Glaser, B. G. & Stauss, A. L. (1967). *The Discovery of Grounded Theory*. Chicago: Aldine.

Golby, M., & Viant, R., (2007). Means and ends in professional development. *Teacher Development*, *11*(2), 237 – 243.

Goldhaber, J., & Smith, D. (2002). The development of documentation strategies to support Teacher reflection, inquiry, and collaboration. In V. Fu, A. Stremmel, & L. Hill (Eds.), *Teaching and learning*: *Collaborative exploration of the Reggio Emilia approach*, (pp. 147-160). Columbus, OH: Merrill/Prentice-Hall.

Goldstein, H. (1997). Methods in School Effectiveness Research. *School Effectiveness and School Improvement*, 8(4), 369-95.

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London: Edward Arnold.

Goodlad, J.A., (1984). *A Place Called School*. McGraw-Hill, New York.

Green, S. K., & Mantz, M. (2002 April). Classroom assessment practices: Examining impact on student learning. *Paper presented at the Annual Meeting of the American Educational Research Association*, New Orleans, LA.

Greenwood, C. R., & Abbott, M. (2001). The research to practice gap in special education. In C. R. Greenwood, (Ed.), Bridging the gap between research and practice in special education: Special issue. *Teacher Education and Special Education*, 24(4), 276-289.

Gronlund, N. E. (2006). *Assessment of student achievement* (8th ed.). Boston: Pearson.

Gullickson, A. R. (1986). Teacher education and teacher-perceived needs in educational measurement and evaluation. *Journal of Educational Measurement*, 23, 347-354.

Guskey, T. R. (2003). What makes professional development effective? *Phi Delta Kappan, 84* (10), 748-750.

Guskey, T.R. & Bailey, J.M. (2001). *Developing grading and reporting systems for student learning.* Thousand Oaks, CA: Corwin Press, Inc.

Guskey, T.R., & Sparks, D. (2004). Linking professional development to improvements in student learning. In E.M. Guyton & J.R. Dangel (Eds.), *Teacher education yearbook XII: Research linking teacher preparation and student performance* (pp. 11–21). Dubuque, IA: Kendall/Hunt.

Hargreaves, A. (1994). *Changing teachers, changing times: Teachers' work and culture in the post-modern age*. London: Cassell; New York: Teachers' College Press.

Hargreaves, A. (1997). Rethinking educational change: Going deeper and wider in the quest for success. In A. Hargreaves (Ed.), *Rethinking educational change with heart and mind: 1997 ASCD Yearbook* (pp. 1-26). Alexandria, Virginia: Association for Supervision and Curriculum Development.

Hargreaves, A. (2002). Sustainability of educational change: The role of social geographies. *Journal of Educational Change, 3*, 189-214.

Hargreaves, D. H. & Hopkins, D. (1991). *The Empowered School: The Management and Practice of Development Planning*, London: Cassell.

Harris, D. & Bell, C. (1994). *Evaluating and Assessing for Learning,* London: Kogan Page.

Harlen, W. (2005). Teachers' summative assessment practices and assessment for learning-tensions and synergies. *The Curriculum Journal, 16*(2), 207-223.

Harlen, W., & James, M. J. (1997). Assessment and learning: differences and relationship between formative and summative assessment. *Assessment in Education, 4*(3), 365-380.

Harlen, W. & Deakin Crick R. (2002) *A systematic review of the impact of summative assessment and testing on pupils' motivation for learning*, London, Evidence for Policy and Practice Co-ordinating Centre Department for Education and Skills.

Harlen, W., Gipps, C., Broadfoot, P. & Nutall, D. (1992). Assessment and the improvement of education. *The Curriculum Journal*, 3, 215-230.

Hattam, R., & Smyth, J. (1995). Ascertaining the nature of competent teaching. A discursive practice. *Critical Pedagogy Networker*, 8(3&4), 1-12.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112.

Hatton, N., & Smith, D. (1995). *Reflection in Teacher Education: Towards Definition and Implementation*. The University of Sydney: School of Teaching and Curriculum Studies.

Heritage, M. (2007). Formative assessment: What do teachers need to know and do? *Phi Delta Kappan, 89*, 140-145.

Herman, J., & Dorr-Bremme, D. W. (1982). *Assessing students: Teachers' routine practices and reasoning.* Paper presented at the annual meeting of the American Educational Research Association, New York.

Herman, J.L., Osmundson, E., Ayala, C., Schneider, S., & Timms, M. (2006). *The nature and impact of teachers' formative assessment practices. CSE Technical Report #703.*

National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Huberman, M. (1989). The professional life cycle of teachers. *Teachers College Record*, 91, 31-57.

Humes, W. (2001). Conditions for professional development. *Scottish Educational Review*, 33(1), 6-17.

Ingenkamp, K. (1977*). Educational assessment*. Windsor, UK: National Foundation for Educational Research.

Ingvarson, L. (1998). Professional development as the pursuit of professional standards: The standards-based professional development system. *Teaching and Teacher Education*, 14(1), 127–140.

Ingvarson, L., Meiers, M., & Beavis, A. (2005). Factors affecting the impact of professional development programs on teachers' knowledge, practice, student outcomes and efficacy. *Education Policy Analysis Archives*, 13(10), 1-28.

Jay, J. K., & Johnson, K. L. (2002). Capturing complexity: a typology of reflective practice for teacher education. *Teaching and Teacher Education*, 18, 73-85.

Johnson, K. E. (1996). The role of theory in second language teacher education. *TESOL Quarterly*, 30(4) 765-771.

Joyce, B., & B. Showers. (1980). Improving Inservice Training: The Messages of Research. *Educational Leadership*, 37(5), 379-385.

Kagan, D. M. (1992). Professional growth among preservice and beginning teachers. *Review of Educational Research*, 62(2), 129 –169.

Kamin, L. (1977). *The Science and Politics of I.Q*. Harmondsworth: Penguin.

Katz, L. (1972). Developmental stages of preschool teachers. *Elementary School Journal*, 73(1), 50-54.

Kauffman, J. M. (1996). Research to practice issues. *Behavioral Disorders*, 22(1), 55-60.

Kennedy, M. M. (1997). The connection between research and practice. *Educational Researcher*, 26(7), 4-12.

Killen. R. (2003). *Effective teaching strategies: Lessons from research and practice (*3rd Ed). Tuggerah, Australia: Social Science Press.

Kluger, A.N. & DeNisi, A. (1996). The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory, *Psychological Bulletin*, 119, 254-284.

Knight, P. (2002). Summative Assessment in Higher Education: practices in disarray. *Studies in Higher Education*, 27, 275-286.

Knight, P. & Yorke, M. (2003). *Assessment, Learning and Employability*. Maidenhead: Society for Research into Higher Education and Open University Press

Kolen, M.J., & Brennan, R.L. (1995). *Test equating: Methods and Practices.* New York: Springer-Verlag.

Korthagen, F. A. J (2001). *Linking practice and theory: The pedagogy of realistic teacher education*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Korthagen, F. (2004). In search of the essence of a good teacher: Towards a more holistic approach in teacher education. *Teaching and Teacher Education*, 20, 77-97.

Korthagen, F. A. J. & Wubbels, T. (1995). Characteristics of reflective practitioners: Towards an operationalization of the concept of reflection. *Teachers and Teaching: Theory and Practice*, 1(1), 51-72.

Koutselini, M. (2008). *Differentiation of Teaching and Learning*. Nicosia. (in Greek).

Kroeger, J., & Cardy, T. (2006). Documentation: A hard to reach place. *Early Childhood Education Journal*, *33*(6), 389–398.

Kumaravadivelu, B. (2001). Toward a postmethod pedagogy. *TESOL Quarterly*, 35, 537–560.

Kyriakides, L. (1997). Primary teachers' perceptions of policy for curriculum reform in Mathematics. *Educational Research and Evaluation, 3*(3), 214-242.

Kyriakides, L. (2004). Investigating Validity from Teachers' Perspective through their engagement in Large-Scale Assessment: the Emergent Literacy Baseline Assessment Project. *Assessment in Education: Principles, Policy and Practice*, 11(2), 143-165.

Kyriakides, L. (2005). Drawing from teacher effectiveness 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. *Journal of Classroom Interaction, 40*(2), 44-66.

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. (2008a). 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. (2008b). A longitudinal study on the stability over time of school and teacher effects on student outcomes. *Oxford Review of Education, 34,* 521-546.

Kyriakides, L., & Kelly, K. (2003). The impact of engagement in large scale assessment on teacher professional development: The Emergent Literacy Baseline Assessment project. *Journal of Research in Childhood Education, 18*(1), 38-56.

Kyriakides, L., Archambault, I., & Janosz, M. (to be accepted). Using Student Ratings to Identify Stages of Effective Teaching. Journal of Classroom Interaction.

Kyriakides, L., Campbell, R.J., & Gagatsis, A. (2000). The significance of the classroom effect in primary schools: An application of Creemers' comprehensive model of educational effectiveness. *School Effectiveness and School Improvement*, 11(4), 501-529.

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

Kyriakides, L., Creemers, B., Antoniou, P., & Demetriou, D. (2010). A synthesis of studies searching for school factors: Implications for theory and research. *British Educational Research Journal*, 36(5), 807-830.

Lave, J. (1993). The practice of learning. In Chaiklin, S. & Lave, J. (Eds.), *Understanding Practice.* Cambridge, Cambridge University Press.

Leder, G.C., Brew, C. & Rowley, G., (1999). Gender differences in mathematics achievement – Here today and gone tomorrow?. In G. Kaiser, E. Luna & I Huntley (Eds.), *International Comparisons in Mathematics Education* (pp.213-224), Falmer Press: London.

Levine, D.U., & Lezotte, L.W. (1990). *Unusually effective schools: A review and analysis of research and practice.* Madison, WI: National Center for Effective Schools Research and Development.

Lieberman, A. (1991). Accountability as a reform strategy. *Phi Delta Kappan*, 73, 219.

Lieberman, A., (1996). Practices that support teacher development: Transforming conceptions of professional learning. In M. W. McLaughlin & I. Oberman (Eds.), *Teacher*

learning: New policies, new practices (pp. 185–201). New York: Teachers College Press

Linacre, J. M. (1999). Understanding Rasch measurement: Estimation methods for Rasch measures. *Journal of Outcome Measurement*, 3(4), 382-405.

Lincoln, Y. & Guba, E. (1985). *Naturalistic Inquiry*. Beverly Hills, CA: Sage Publications Inc.

Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis*, 15, 1-16.

Linn, R. (2000). Assessments and accountability, *Educational Researcher*, 29(2), 4–16.

Livingston, C., & Borko, H.(1989). Expert-novice differences in teaching: A cognitive analysis and implications for teacher education. *Journal of Teacher Education*, 40(4), 36-42.

Lock, C.L., & Munby, H. (2000). Changing assessment practices in the classroom: A study of one teacher's change. *The Alberta Journal of Educational Research, 46*, 267-279.

Locke, E. A., Shaw, K.N., Saari, L. M., & Latham,G. P. (1981). Goal setting and task performance: 1969-1980. *Psychological Bulletin*, 90, 125-152.

Louden, W., & Wallace, J. (1993). Competency standards in teaching. *Unicorn*, 19(1), 45-53.

Lukin, L., et al. (2004). Facilitating the development of assessment literacy. *Educational Measurement: Issues and Practice*, 26-36.

Madaus, G. F. (1988). The influence of testing on the curriculum. In L. N. Tanner (Ed.), *Critical issues in curriculum: Eighty-seventh yearbook of the national society for the study of education* (pp. 83–121). Chicago: University of Chicago Press.

Marcoulides, G., & Drezner, Z. (1999). A procedure for detecting pattern clustering in measurement designs. In M.Wilson & G. Engelhard, Jr (Eds.), *Objective measurement: Theory into practice* (Vol. 5). New Jersey: Ablex.

Maykut, P. & Morehouse, R. (1994). *Beginning qualitative research: a philosophical and practical guide.* London: The Falmer Press.

McDermott, R. (1993). The acquisition of a child by a learning disability. In S. Chaiklin & J. Lave (Eds.), *Understanding practice: Perspectives on activity and context* (pp. 269–305). Cambridge, UK: Cambridge University Press.

McDonald, B., & Boud, D. (2003). The impact of self-assessment on achievement: The effects of self-assessment training on performance in external examinations. *Assessment in Education,* 10(2), 209-220.

McLaughlin, M. W., & Talbert, J. E. (2006). *Building school-based teacher learning communities: Professional strategies to improve student achievement*. New York: Teachers College Press

Ministry of Education and Culture, Republic of Cyprus (2004). *Manifest of school reform*. Lefkosia: Author.

Ministry of Education and Culture, Republic of Cyprus. (2008). *Inclusion in the Cyprus educational system at the beginning of the twenty first century: An overview - national report of Cyprus*. Nicosia, Cyprus.

Mislevy, R. J., & Wilson, M. (1996). Marginal maximum likelihood estimation for a psychometric model of discontinuous development. *Psychometrica,* 61, 41–71.

Mok, M. M. C. (2010). *Self-directed Learning Oriented Assessment: Assessment that Informs Learning & Empowers The Learner*. Hong Kong: Pace Publications Ltd.

Moss, P.A. (2003). Reconceptualizing validity for classroom assessment. *Educational Measurement: Issues and Practice*, 22(4), 13-21.

National Council of Teachers of Mathematics. (1995). *Assessment standards for school mathematics*. Reston, VA: Author.

Newton, C. & Tarrant, T. (1992). *Managing change in schools.* London: Routledge.

Nicol, D.J. & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2), 199-218.

Nichols, S. L., & Berliner, D. C. (2007). *Collateral damage: How highstakes testing corrupts America's schools*. Cambridge, MA: Harvard Education Press.

Oberski, I. & MacNally, J. (2007). Holism in teacher development: A Goethean perspective. *Teaching and Teacher Education*, 23, 935-943.

Ollis, S**.,** MacPherson, A. C.**,** & Collins, D. (2006). Expertise and talent development in rugby refereeing: An ethnographic enquiry. *Journal of Sports Sciences, 24*, 309-322.

Orsmond, P., Merry, S. & Reiling, K. (2000). The use of student-derived marking criteria in peer- and self-assessment. *Assessment and Evaluation in Higher Education*, 25(1), 23–39.

Ottesen, E. (2007). Reflection in teacher education. *Reflective practice*, 8, 31–46.

Pajares, F. M. (1992). Teachers' Beliefs and Educational Research: Cleaning Up a Messy Construct. *Review of Educational Research* , 62 (3), 307-332.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

Plake, B. S., Impara, J. C., & Fager, J. J. (1993). Assessment competencies of teachers: A national survey. *Educational Measurement: Issues and Practice, 12*(4), 10-12.

Perrenoud, P. (1998). From formative evaluation to a controlled regulation of learning processes: towards a wider conceptual field. *Assessment in Education*, 5(1), 85–102.

Popham, W.J. (1984). Teacher competency testing: The Devil's dilemma. *Teacher Education and Practice, 1*, 5-9.

Popham, W. J., (2004). *Classroom assessment: What teachers need to know (4th ed.)*. Upper Saddle River, NJ: Pearson Education.

Popham, W. J. (2006). All about accountability: Those [fill-in-the-blank] tests! *Educational Leadership*, 63(8), 85–86.

Popham, W. J. (2009). Assessing student affect. *Educational Leadership, 66*, 85-86.

Porter, P., Rizvi, F., Knight, J., & Lingard, R. (1992). Competencies for a clever country. Building a house of cards? Unicorn, 18(3), 50-58.

Powell, A.G., Farrar, E. & Cohen, D.K., (1985). *The Shopping Mall High School*, Houghton Mifflin, Boston MA.

Pryor, J. & Crossouard, B. (2008). A socio-cultural theorisation of formative assessment. *Oxford Review of Education*, 34(1),1-20.

Pugach, M.C.& Johnson, L.J. (1990). Fostering the continued democratization of consultation through action research.Teacher Education and Special Education, 13, 240-245.

Quality Assurance Agency for Higher Education, (2006). *Code of practice for the assurance of academic quality and standards in higher education, Section 6: Assessment of students,* Gloucester.

Rao, S., Collins, H.L., & DiCarlo, S.E. (2002). Collaborative testing enhances student learning. *Advances in Physiology Education*, *26*(1), 37-41.

Raven, J. (1991). *The tragic illusion: educational testing*. New York: Trillium Press.

Rea-Dickins, P. (2001). Mirror, mirror on the wall: identifying processes of classroom assessment. *Language Testing, 18*, 429-462.

Retallick, J. (1999). Teachers' workplace learning: towards legitimation and accreditation. *Teachers and Teaching*, 5(1), 33-50.

Reynolds, D., Hopkins, D.& Stoll, L.(1993). Linking School Effectiveness Knowledge & School Improvement Practice: Towards Synergy. *School Effectiveness & School Improvement*. 4(1), 37-58.

Richardson, V., & Anders, P.L. (1994). A theory of change. In V. Richardson (Ed.), *Teacher change and the staff development process: A case in reading instruction* (pp. 199–216). New York, NY: Teachers College Press.

Rinaldi, C. (2006). *In Dialogue with Reggio Emilia: Listening, Researching and Learning*. New York, USA: Routledge.

Roeber, E. D. (2003). Assessment models for No Child Left Behind. Denver, CO: Education Commission of the States.

Romesburg, H. C. (1984). *Cluster Analysis for Researchers*. Belmont, CA: Lifetime Learning Publication and Wadsworth Inc..

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*, 119-144.

Sandberg, J. (1994). *Human competence at work: An interpretative approach*. Goteborg, Sweden:Bas

Sandberg, J. (2000). Understanding human competency at work: An interpretive approach. *Academy of Management Journal*, 43(1), 9-25.

Satterley, D. (1994). The quality of external assessment. In Harlen, W. (Ed.). *Enhancing Quality in Assessment* . London: Paul Chapman.

Schafer, W. D. (1991). Essential assessment skills in professional education of teachers. *Educational Measurement: Issues and Practice,* 10(1), 3-6.

Schafer, W. D. (1993). Assessment literacy for teachers. *Theory Into Practice*, *32*(2), 118-126.

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

Schmoker, M. (2006). *Results now.* Alexandria, VA: Association for Supervision and Curriculum Development.

Schon, D.A. (1983). *The reflective practitioner: How professionals think in action.* New York, NY: Basic Books.

Schön, D. *(*1987*). Educating the Reflective Practitioner.* Jossey Bass, San Francisco.

Schön, D. (1991). *The Reflective Practitioner.* Aldershot: Ashgate Publishing Ltd

Schraagen, J. M. (1993). How expert solve a novel problem in experimental design. *Cognitive Science*, 17, 285-309.

Scribner, J. P. (1999). Professional development: Untangling the influence of work context on teacher learning. *Educational Administration Quarterly*, 35(2), 238-266.

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation*, (pp. 39-83). Chicago, IL: Rand McNally.

Sebba, J. (2006). Policy and practice in assessment for learning: the Experience of selected OECD countries, in J. Gardner (ed.*), Assessment and Learning* (pp. 333-351). Sage Publications, London,

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational researcher*, 29(7), 4-14.

Shepard, L.A. (2001). The role of classroom assessment in teaching and learning. In V. Richardson (Ed.), *The Handbook of Research on Teaching*, (4th edition). Washington, DC: American Educational Research Association.

Shepard, L. A. (2007). Will commercialism enable or destroy formative assessment? In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning*. Mahwah, NJ: Lawrence Erlbaum Associates.

Slavin, E.R. (2010). Experimental studies in education.In Creemers, B.P.M, Kyriakides, L., & Sammons, P.(Eds.), *Methodological Advances in Educational Effectiveness Research, (pp. 102-114).* London: Routledge Taylor Francis.

Smith, B. O., Silverman, S. H., & Borg, J. M. (1980). *A design for a school of pedagogy* [Publication No. E-80-4200). Washington, DC: U.S. Department of Education.

Smith, M. & J. O'Day. (1991). *Putting the Pieces Together: Systemic School Reform*. CPRE Policy Brief. New Brunswick, NJ: Eagleton Institute of Politics.

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

Sparks, D., & Loucks-Horsley, S. (1990). Models of staff development. In W. R. Houston (Ed.), *Handbook of Research on Teacher Education* (pp. 234-250). New York: Macmillan.

Sparks-Langer, G. & Colton, A. (1991). Synthesis of Research on Teachers' Reflective Thinking. *Educational Leadership*, 37-44.

Sprinthall, N. A., Reiman, A. J., & Theis-Sprinthall, L. (1996). Teacher professional development. In J. Sikula, T.,J Buttery & E. Guyton (Eds.), *Handbook of research on teacher education* (pp. 666-703). New York: Simon-Schuster Macmillan.

Stallings, J., & Krasavage, E. (1986). Program implementation and student achievement in a four-year Madeline Hunter Follow-Through project. *Elementary School Journal*, 87(2), 117-138.

Sternberg, R. J., Forsythe, G. B., Hedlund, J., Horvath, J. A., Wagner, R. K., Williams, W. M., et al. (2000). *Practical intelligence in everyday life*. Cambridge: Cambridge University Press.

Stevenson, J. (Ed.). (1996). *Learning in the workplace: Tourism and Hospitality*. Brisbane, Australia: Centre for Skill Formation Research and Development, Griffith University.

Stiggins, R. J. (1991). Assessment Literacy. *The Phi Delta Kappan*, 534-539.

Stiggins, R. J. (1992). High quality classroom assessment: What does it really mean? *Educational Measurement: Issues and Practice*, 11 (2), 35–39.

Stiggins, R. J. (1995). Assessment literacy for the 21st century. *Phi Delta Kappan*, *77*(3), 238-245.

Stiggins, R. J. (1998). *Classroom assessment for student success*. National Education Association: Washington, DC.

Stiggins, R. J. (1999). Evaluating classroom assessment training in teacher education programs. *Educational Measurement: Issues and Practice*, *18*(1), 23-27.

Stiggins, R. (2001). *Student Involved Classroom Assessment* (3rd Edition). Columbus, OH: Merrill Publishing.

Stiggins, R. J. (2002). Assessment crisis: The absence of assessment FOR learning. *Phi Delta Kappan*, *83*, 758-765.

Stiggins, R. J., (2004). New assessment beliefs for a new school mission. *Phi Delta Kappan*, 86(1), 22–28.

Stiggins, R. J., & Conklin, N. F. (1992). *In teachers' hands: investigating the practices of classroom assessment*. Albany, NY: State University of New York Press.

Stiggins, R., & Chappuis, S. (2005). Putting testing in perspective: It's for learning. *Principal Leadership (High School Ed.)*, 6(2), 16-20.

Stiggins, R. J., & Chappuis, J. (2006). What a difference a word makes: Assessment FOR learning rather than assessment OF learning helps students succeed. *Journal of Staff Development*, *27*(1), 10–14.

Stiggins, R. J., & Chappuis, J., (2008). Enhancing student learning. *District Administration*, 43-44.

Stiggins, R.J., & DeFour, R., (2009). Maximizing the power of formative assessments. *Phi Delta Kappan, 90*(9), 640–644.

Stobart, G. (2004). *The formative use of summative assessment: possibilities and limits,* Philadelphia: 30th Annual IAEA Conference.

Strauss, A. & Corbin, J. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (2nd ed.). Thousand Oaks, CA: Sage.

Stronach, I., Cope, P., Inglis, W., & McNally, J. (1994). The SOED 'competence' guidelines for initial teacher training: issues of control, performance and relevance. *Scottish Educational Review,* 26(2), 118-133

Suurtamm, C. (2004). Developing authentic assessment: Case studies of secondary school mathematics teachers' experiences. *Canadian Journal of Science, Mathematics and Technology Education,* 4, 497–513.

Suurtamm, C., Koch, M., & Arden, A. (2010). Teachers' emerging assessment practices in mathematics: Classrooms in the context of reform. *Assessment in Education: Principles, Policy, and Practice, 17*(4), 399-417.

Teddlie, C., & Reynolds, D. (2000). *The international handbook of school effectiveness research*. London: Falmer Press.

Tomlinson, C. A. (1999). *The differentiated classroom: Responding to the needs of all learners*. New Jersey: Pearson Education.

Torrance, H. & Pryor, J. (1998). *Investigating Formative Assessment. Teaching, Learning and Assessment in the Classroom*. Buckingham, Open University Press.

Torrance, H., & Pryor, J. (2001). Developing formative assessment in the classroom: Using action research to explore and modify theory. *British Educational Research Journal, 27*, 615-631.

Tymms, P. & Merrell, C. (2009). Attainment, Standards and Quality. In *Childrlen, their World, their Education: Final report and recommendations of the Cambridge Primary Review.* Alexander, R. London: Routledge.

Tunstall, P., & Gipps, C. (1996).Teacher Feedback to Young Children in Formative Assessment: A Typology. *British Educational Research Journal*, 22(4), 389-404.

Tyack, D. & Tobin, W. (1994). The "Grammar" of Schooling: Why has it been so hard to change? *American Educational Research Journal*, 31(3), 453-479.

Van der Horst, H. & McDonald, R. (1997). *Outcomes-based education: a teacher's manual*. Pretoria: Kagiso Publishers.

Van Manen, M. (1977). Linking ways of knowing with ways of being practical. *Curriculum Inquiry*, 6, 205-228.

Volante, L., & Fazio, X. (2007). Exploring teacher candidates' assessment literacy: Implications for teacher education reform and professional development. *Canadian Journal of Education, 30*, 749-770.

Voss, J. F., Tyler, S. & Yengo, L. (1983). Individual differences in the solving of solving of social science problems. In R. Dillion & R. Schmeck (Eds.), *Individual Differences in Problem-Solving*. San Diego, CA: Academic Press.

Walberg, H. (1986). Synthesis of research on teaching. In M. Wittrock (Ed.)., *Handbook of research on teaching* (pp. 214-229). New York: Macmillan.

Wang, W.C., & Cheng, Y.Y. (2001). Measurement issues in screening outstanding teachers. *Journal of Applied Measurement*, *2*(2), 171–186.

Watanabe, M. (2007). Displaced teacher and state priorities in a high stakes accountability context. *Educational Policy Analysis,* 21(2). 311-368.

Whitty, G., & Willmott, E. (1991). Competence-based teacher education: Approaches and issues. *Cambridge Journal of Education,* 21(3), 309–319.

Wiliam, D., Lee, C., Harrison, C., & Black, P. J. (2004). Teachers developing assessment for learning: Impact on student achievement. *Assessment in Education: Principles Policy and Practice, 11*(1), 49-65.

Wilson, M. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin,* 105(2), 276–289.

Wilson, M. (1999). Measurement of developmental levels. In G. N. Masters & J. P. Keeves (Eds.), *Advances in measurement in educational research and assessment,* (pp. 151–163). Oxford: Pergamon.

Wilson, S. & Berne, J. (1998). Teacher learning and the acquisition of professional knowledge: An examination of research on contemporary professional development. In Iran Nejad & P. D. Parsons (Eds.), *Review of Research in Education,* 24 (pp. 173-209).Washington, DC: American Educational Research Association.

Wright, B.D. (1985). Additivity in psychological measurement. In E.E. Roskam (Ed.), *Measurement and personality assessment* (pp. 101-112). Amsterdam: Elsevier Science Publishers BV.

Wright, B.D., & Linacre, J.M. (1989). Observations are always ordinal: measurements, however, must be interval. *Archives of Physical Measurement and Rehabilitation*, *70*(12), 857–860.

Wright, B. D., & Masters, G. (1981). *The measurement of knowledge and attitude* (Research Memorandum 30). Chicago: University of Chicago, Department of Education, Statistical Laboratory.

Wright, B. D., & Masters, G. (1982). *Rating scale analysis*. Chicago: MESA Press

Wylie, C., & Lyon, C. (2009). *What schools and districts need to know to support teachers' use of formative assessment*. Teachers College Record.

Yen, W.M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187-213.

Yorke, M. (2003). Formative assessment in higher education: Moves towards theory and the enhancement of pedagogic practice. *Higher Educatio*n, 45(4), 477-501.

Yorke, M. (2004). Retention, persistence and success in on-campus higher education, and their enhancement in open and distance learning. *Open Learning*, 19(1), 19-32.

Zeichner, K. (1983). Alternative Paradigms of Teacher Education. *Journal of Teacher Education*, 34, 3-9.

Zeichner, K. M. (1986). Content and contexts: Neglected elements in studies of student-teaching as an occasion for learning to teach. *Journal of Education for Teaching, 12,* 5-24.

Zeichner, K.M. (1987). Preparing reflective teachers: An overview of instructional strategies which have been employed in preservice teacher education. *International Journal of Educational Research, 11,* 565–575.

APPENDIX A: Teacher Questionnaire

The aim of this research is to examine the needs of teachers, as far as assessment is concerned, in order to provide appropriate and focused professional development. The following statements concern the assessment of students in mathematics and refer to the four phases of the assessment process: a) planning and construction of assessment tools, b) administration of assessment instruments, c) recording and analyzing data and d) reporting results. This questionnaire will be used to design a professional development programme that meets the needs of each teacher. In addition, this questionnaire will allow us to assess the effectiveness of this programme after its completion. Therefore, providing your name upon completion of this questionnaire is required. You are reassured that all information found in this questionnaire will be strictly confidential. Finally, you are kindly requested to answer all questions in all honesty.

PART A

1) **Gender:** Male ☐ Female ☐

2) **Position:** Teacher ☐ Assistant head teacher ☐ Headteacher ☐

3) Years of Working Experience: …….

(Consider the current year as a full year)

4) Having in mind the way students are assessed in mathematics rank the following by using the numbers 1 to 6 in such a way that in **column A** number 1 refers to the most appropriate assessment technique, number 2 to the next more appropriate technique and so on, whereas number 6 is the least appropriate technique.
Then follow the same procedure for **column B**, so that number 1 refers to the technique you **use more often,** number 2 refers to the technique you use less often and so on, whereas number 6 refers to the technique you rarely or never use.

	COLUMN A Appropriateness	**COLUMN B** Usage Frequency
A. Written Test	☐	☐
B. Oral Assessment	☐	☐
C. Observation	☐	☐
D. Performance Assessment	☐	☐
E. Homework	☐	☐

5) Rank the following **assessment purposes** by using numbers 1 to 4, so that number 1 refers to the most important purpose for which you assess your students in mathematics, number 2 refers to a less important purpose and so on, whereas number 4 refers to the least important assessment purpose.

In mathematics I assess my students in order to:

A. assess myself ☐

B. be able to compare my students with each other ☐

C. identify the needs of my students and plan my teaching ☐

D. assess the effectiveness of the curriculum delivered to the students ☐

PART B: WRITTEN ASSESSMENT IN MATHEMATICS

Tick (√) the appropriate box:

1) When I assess my students in mathematics, I use a written test:

A. Never ☐

B. Once a semester ☐

C. At the end of each unit ☐

D. More than once for a unit ☐

At the previous question, if you ticked 'Never', do not answer questions 2-14 of Part B and proceed to Part C. If you ticked another choice, then answer regularly all the questions of Part B.
The statements 2-14 of Part B refer to the use of written tests in mathematics. Circle a number, from a scale 1-5, in order to show to what extend the following statements respond to what occurs during the mathematics assessment in your classroom. Number 1 refers to facts that occur very rarely or never, whereas number 5 refers to facts that occur very often (i.e. at least once during a unit).

1 means 'Very Rarely or Never' and 5 means 'Very Often or Always'

2) My written tests include only questions / activities found in the mathematics assessment book. 1 2 3 4 5

3) The questions / activities that I use require from students to explain the procedure they used in order to find the result. 1 2 3 4 5

4) All students are asked to answer exactly the same questions in every written test.	1	2	3	4	5
5) The written tests I use, usually include:					
A. Multiple Choice Questions	1	2	3	4	5
B. True / False questions	1	2	3	4	5
C. Fill-in exercises	1	2	3	4	5
D. Matching exercises	1	2	3	4	5
E. Short Answers	1	2	3	4	5
F. Open-ended questions (that do not have only one correct answer)	1	2	3	4	5
G. Problem solving	1	2	3	4	5
6) The written tests I use include questions / activities that are related with each other (i.e. students are required to do operations and then create a graph based on their answers)	1	2	3	4	5
7) Before creating a test, I write down the objectives I want to assess and further indicate which exercises of the test correspond to each objective.	1	2	3	4	5
8) I provide help to a student when I realize that she/he is having some difficulties during the written assessment.	1	2	3	4	5
9) When I construct the questions / activities for a written test, I take into consideration my students' abilities (i.e. in a class of lower-ability students I use easier exercises).	1	2	3	4	5
10) During written assessment administration, students require clarifications regarding the exercise instructions.	1	2	3	4	5
11) Once I recognize that a student has difficulties in comprehending the exercises, I provide clarification to that student.	1	2	3	4	5
12) All students have the same amount of time to complete the written test.	1	2	3	4	5
13) Before students begin to complete the written test:					
A. I provide a detailed explanation of the instructions of each question / activity.	1	2	3	4	5
B. I provide general instruction on how to complete the test.	1	2	3	4	5
C. I do not provide any instructions, because I consider the instructions of the test comprehensible.	1	2	3	4	5
14) When I recognize that a number of students have not fully comprehended a question / activity, I interrupt the test and provide further instructions for the whole class.	1	2	3	4	5

PART C: ORAL ASSESSMENT

Tick (√) the appropriate box:

1) When I assess my students in mathematics I use oral assessment:

A. Never ☐

B. Once a semester ☐

C. At the end of each unit ☐

D. More than once for a unit ☐

At the previous question, if you ticked 'Never', do not answer questions 2-8 of Part C and proceed to Part D. If you ticked another choice, then answer regularly all the questions of Part C.

In order to answer the questions of Part C, circle a number, from scale 1-5, in order to show to what extend the following statements respond to what occurs during mathematics assessment in your classroom.

1 means 'Very Rarely or Never' and 5 means 'Very Often or Always'

2) I orally assess my students in mathematics:					
A. during discussion in the classroom (randomly)	1	2	3	4	5
B. after planning and when students are aware of the assessment (formal)	1	2	3	4	5
C. after planning and when students are not aware of the assessment (informal)	1	2	3	4	5
3) Sometimes students do not seem to have understood the question asked and I am required to rephrase it.	1	2	3	4	5
4) I know in advance which students I am going to assess and which questions I am going to ask each student.	1	2	3	4	5
5) I orally assess students to check to what extend the results correspond to the results of the written test.	1	2	3	4	5
6) I take into consideration students' abilities when I ask questions (i.e. easier questions are addressed to lower-ability students)	1	2	3	4	5
7) All students have the same amount of time to orally answer the questions	1	2	3	4	5
8) When a student has difficulties in answering an oral question, then:					
A. I rephrase the question	1	2	3	4	5
B. I provide further clues	1	2	3	4	5
C. I provide a different question	1	2	3	4	5
D. I request other students to answer the same question	1	2	3	4	5

PART D: OBSERVATION/PERFORMANCE ASSESSMENT

Tick (√) the appropriate box:

1) I use observation / performance assessment (i.e. if students know how to use the compass) to assess my students in mathematics:

A. Never ☐

B. Once a semester ☐

C. At the end of each unit ☐

D. More than once for a unit ☐

At the previous question, if you ticked 'Never', do not answer questions 2-9 of Part D and proceed to Part E. If you ticked another choice, then answer regularly all the questions of Part D.

In order to answer the questions 2-9 of Part D, circle a number, from a scale 1-5, in order to show to what extend the following statements respond to what occurs during mathematics assessment in your classroom.

1 means 'Very Rarely or Never' and 5 means 'Very Often or Always'

2) In mathematics I observe my students randomly for assessment purposes (without doing any planning in advance)	1	2	3	4	5
3) I decide in advance which students to assess through systematic observation.	1	2	3	4	5
4) When students work in groups, I observe in order to assess to what extend each student cooperates well with others (in case you do not use group work activities, do <u>not</u> answer the question)	1	2	3	4	5
5) During mathematics instruction, I assess how a student performs an activity to check her/his skills (i.e. if s/he knows how to use the compass)	1	2	3	4	5
6) Before proceeding to observation, I write down the objectives I want to assess and how to achieve that.	1	2	3	4	5
7) I use observation to assess the procedure a student follows to solve a problem.	1	2	3	4	5
8) When I use observation to asses my students, I take interest in identifying each student's contribution to the team.	1	2	3	4	5
9) When students work in groups I use observation in order to assess only the final outcome of each group.	1	2	3	4	5

PART E: RECORDING AND REPORTING RESULTS

Part E refers to statements concerning the recording and reporting of assessment results. Circle a number, from a scale 1-5, in order to show to what extend the following statements respond to what occurs during mathematics assessment in your classroom.

1 means 'Very Rarely or Never' and 5 means 'Very Often or Always'

1) I keep a record of the results from:					
A. Written Test	1	2	3	4	5
B. Oral Assessment	1	2	3	4	5
C. Performance Assessment	1	2	3	4	5
D. Observation	1	2	3	4	5
2) When I record the results of an assessment I use:					
A. numeric rating scale	1	2	3	4	5
B. letter and symbol rating scale (i.e. A, B / ++)	1	2	3	4	5
C. comments regarding student's specific needs and requirements	1	2	3	4	5
D. general comments regarding student's performance and progress	1	2	3	4	5
3) The records I keep concern:					
A. each student's performance during the exercise	1	2	3	4	5
B. the student's general performance	1	2	3	4	5
C. the classroom's overall performance	1	2	3	4	5
D. the student's performance by objective	1	2	3	4	5
4) When students work in groups I record comments regarding:					
A. each team's overall performance	1	2	3	4	5
B. the contribution of each student to the team	1	2	3	4	5
C. each student's performance in relation to the other members of the team.	1	2	3	4	5
5) The results of the written assessments are given back to students in the form of:					
A. numeric rating scale	1	2	3	4	5
B. letter and symbol rating scale (i.e. A, B / ++)	1	2	3	4	5
C. general commentary (i.e. 'Very Good', 'You need to study harder', etc.)	1	2	3	4	5
D. specific commentary in relation to weaknesses that were identified	1	2	3	4	5
E. neither commentary nor rating	1	2	3	4	5
6) I inform the school's headteacher regarding the results of each assessment.	1	2	3	4	5
7) At the end of parents meeting, I usually have the impression that they did not understand the information I provided them regarding the results of their child's assessment.	1	2	3	4	5

8) When I inform parents I usually refer to their child's grades in each test.	1	2	3	4	5
9) During parents meeting, I ask parents to give me information regarding:					
A. the amount of time the student spends on studying at home	1	2	3	4	5
B. the student's behavior outside school	1	2	3	4	5
C. the student's family environment and situation	1	2	3	4	5
D. the ways in which they help their child at home	1	2	3	4	5
10) When I realize that there is a communication problem between me and the parents, I adapt the conversation to their level.	1	2	3	4	5
11) All students are informed about the assessment results in the same way.	1	2	3	4	5
12) I correct the homework and provide feedback to students.	1	2	3	4	5
13) When I inform parents I refer to the child's performance in relation to his/her classroom's level.	1	2	3	4	5
14) I discuss, on a personal level, with every student regarding his/her assessment results.	1	2	3	4	5
15) When I inform parents I advise them on how to help their child improve.	1	2	3	4	5
16) I inform <u>parents</u> regarding the results of:					
A. written assessment	1	2	3	4	5
B. oral assessment	1	2	3	4	5
C. observation/performance assessment	1	2	3	4	5
D. activities assigned in classroom	1	2	3	4	5
17) When I inform students regarding their assessment results I point out what they can do to improve themselves.	1	2	3	4	5
18) I inform <u>students</u> regarding the results of:					
A. written assessment	1	2	3	4	5
B. oral assessment	1	2	3	4	5
C. observation / performance assessment	1	2	3	4	5
D. activities assigned in classroom	1	2	3	4	5
19) In order to draw conclusions regarding students' performance I take into consideration the way they do their homework	1	2	3	4	5

20) Rank the following **purposes** that serve the way by which you record the assessment results of Maths by using numbers 1 to 6, so that number 1 refers to the most important purpose, number 2 refers to the less important purpose and so on, whereas number 6 refers to the least important purpose.

The way by which I record the assessment results allows me:

A. to know the level of each student ☐

B. to be able to inform parents any time I am asked ☐

C. to check the appropriateness of the objectives set ☐

D. to be able to inform school administration at any time I am asked ☐

E. to check the appropriateness of the questions / activities ☐

F. to have a better awareness of the level of my classroom ☐

Tick (√) the appropriate box:

21) When I realize that some parents do not ask to be informed systematically about their children's assessment, then:

A. I contact parents by phone and ask them to meet at school ☐

B. I send a letter to parents asking them to meet at school ☐

C. I only contact parents whose children face difficulties ☐

If you chose 'C' in question 21, do not answer question 22.

22) After contacting parents, either by phone or letter, the parents come at the school:

A. Never ☐

B. Rarely ☐

C. Often ☐

D. Always ☐

Thank you for your cooperation

APPENDIX B: Questionnaire Specification Table

ASSESSMENT SKILLS		FREQUENCY	FOCUS	STAGE	QUALITY	DIFFERENTIATION
CONSTRUCTION	WRITTEN ASSESSMENT ORAL ASSESSMENT OBSERVASION PERFOMANCE ASSESSMENT	• Number of techniques used • Frequency of the use of each technique • Frequency of each type of each technique • Time devoted to construction **A4, B5**	• Purposes to be achieved with each technique • Setting learning goals for each technique • Types of questions/tasks included (product or process) **B3, D7**	• Period of construction • Time lapse between construction and administration	• Instrument representative of the content • Instruments validity and reliability • Variety of techniques used to control quality **A5, B2, B6, B7, B9, C4, C5, D6**	• Differentiation of techniques according to the goals aimed • Differentiation of techniques according to students' needs **B4**
ADMINISTRATION	WRITTEN ASSESSMENT ORAL ASSESSMENT OBSERVASION PERFOMANCE ASSESSMENT	• Frequency for technique administration • Time devoted to complete assessment tasks • Frequency of clarification questions **B1, C1, C2, D1, , E19**	• Individual or group administration • Specificity of instructions (too general/too specific/ focused) **B13, D2, D4**	• Period of administration • Time gap between administration and recording **B1, C1, D1**	• On time administratio n (begin and finish on time) • Quality of instructions provided (accomplish intended use) **B10, B11, B14,C3,C8, D3, D5, D8, D9**	• Differentiation of time according to students needs • Differentiation of techniques according to students needs **B8, B12, C6, C7,**
RECORDING	WRITTEN ASSESSMENT ORAL ASSESSMENT OBSERVASION PERFOMANCE ASSESSMENT	• Frequency of record keeping • Time devoted to record keeping • Number of Instruments used for record keeping **E1**	• Purposes of record keeping • Individual or group recording • Recording notes (specific/gener al/general trend) **E3, E4**	• Period of recording • Time lapse between recording and reporting	• Recording instruments interpretative validity • Formative and summative use of records **E2, E5**	• Differentiation of record keeping according to purpose and student
REPORTING	WRITTEN ASSESSMENT ORAL ASSESSMENT OBSERVASION PERFOMANCE ASSESSMENT	• Frequency of reporting • Time devoted to reporting • Number of users of reporting (to whom?) **E16, E18**	• Purposes of reporting • Reporting information (general/ specific/ focused) **E13**	• Period of reporting **E6**	• Formative feedback • Receiving student information from users / type of information • Quality of communicati on between teacher and users **E7, E8, E9, E15, E17, E19**	• Differentiation of reporting according to users • Differentiation of reporting according to purposes • Differentiation of communication discourse according to users **E10, E11, E14**

APPENDIX C: Interview Guide

INTERVIEW FOR INVESTIGATING TEACHERS' SKILLS IN ASSESSMENT

Instructions for using the tool:

Every interviewee must provide information for each one of the four phases of assessment. Questions A and B are general and must be asked at the beginning of the interview. Four questions will follow, each one of them corresponding to one of the four phases, as well as a check list for each question. In case one or more topics is or are not mentioned by the interviewee, then the supplementary questions as well as clarification questions must be asked.

Question A:

First I would like to thank you for the time you took for this interview. This interview concerns the assessment of students in the mathematics classroom and it specifically focuses on the four phases of the assessment process: construction, administration, recording data and reporting results. The aim of this interview is to ensure that the professional development program to be offered will correspond to the real needs of each participant.

What is your opinion on the assessment of the student as it is? What are the prospects for improvement and more specifically, what do you expect to gain from the specific professional development program?

Note: In case the interviewee refers to issues relating to a specific phase, pass on to the specific question and after that, pass on to the questions concerning the other phases.

Question B:

Which assessment techniques do you usually use to assess your students in mathematics?

Check List

- *written assessment*
- *oral assessment*
- *observation (random – intentional)*
- *performance assessment*

Question 1:
Which procedures do you usually follow to construct an assessment tool?

Check List

- *time spent in the construction*
- *type of questions (product / process)*
- *type of questions (objective / multiple choice / fill in / short answer / true-false /open-ended / coupling/ interpretive / layout)*

- *construction period*
- *level of difficulty of questions*
- *content representativeness / Specifications Table / Comments Table (1a)*
- *differentiation based on students / objective (1b)*
-

Supplementary question (1a): Having in mind the last written assessment you constructed, could you describe how you worked on it? For example, did you use a ready-made test and if so how did you choose it? What do you take into consideration when choosing the exercises / activities for a written assessment?

Supplementary question (1b): Are there any situations where you consider written assessment inappropriate for the assessment of specific students or specific objectives? If so, what are the reasons and how would you deal with these kinds of situations?

Question 2:

We would like to know how you behave during the administration of an assessment. For example, what procedure do you usually follow during the administration? (i.e. do you provide clarification guidelines / solve queries / set time limits etc.)

Check List

- *frequency of administration of each technique*
- *queries of students / clarifications*
- *individual / group administration*
- *type of guidelines (general / specific)*
- *administration period (beginning of a unit / end of a unit / when necessary)*
- *set / keep time limits*
- *appropriateness of guidelines (students comprehend what they must do) (2a)*
- *time differentiation based on students / objective (2b)*

Supplementary question (2a): During the written assessment, do students usually ask questions? If so, how do you act in response?

Supplementary question (2b): Are there situations in which you give certain students more time in order to complete the assessment? If so, what are the reasons and in which ways this is done?

Question 3:

We would like you to describe the procedure you usually follow to record the results of an assessment. For example, for which assessment techniques do you usually keep a record? And what is usually the form of such a record?

Check List

- *frequency for recording results for each type*
- *use of different recording tools*
- *marking (symbolic/numeric)*
- *individual / group recording*
- *comments (per student / per class as a total / per objective / per exercise)*
- *time gap between administration-recording*
- *formative / comparative use of recorded data(3a)*
- *recording differentiation based on students / objective*

Supplementary question (3a): In your opinion, data recording serves in what? Can you describe how you usually use these records?

Check List 3a

- Purpose / Use of recording

a) *information on classroom level*
b) information on student's level
c) evaluation / revision of objectives / activities
d) information for reporting purposes

Supplementary question (3b): Can you mention in detail the information you include in data recording?

Check List 3b

- The recording includes:

a) *comments on student's general performance*
b) comments on the needs / weaknesses that aroused
c) comments on the progress shown
d) comments on the next improvement steps

Question 4:

We would appreciate it if you could describe the procedure you usually follow in order to report assessment results? For example, who do you report to, in what way and during what period?

Check List

- *reporting users*
- *frequency of reporting per user*
- *purpose of reporting per user*
- *type of reporting (general / specific / focused)*
- *reporting period per user*
- *identification / definition of next steps*
- *eliciting information from users / type of information*

- *communication quality between teacher - user*
- *differentiation of reporting context based on user / purpose*
- *differentiation of communication language based on user*

Supplementary question (4a): Having in mind your experiences of parents' visits, what purposes does reporting to parents serves and what type of information this reporting must include?

Supplementary question (4b): After your last written assessment, in which way did you report to your students about their performance? For example, when students received their tests what type of comments did you provide?

APPENDIX D: Action Plan Template

School:	**School Year:**	**Area of actions:**
Goal(s):		
Actions/Steps	**Time-frame**	**Resources**
Evaluation		

APPENDIX E: Examples of Application Activities

Application activity – Table of Specification

A) By using the written essay given, complete the following table of specification.

Course: Math **Class: A'** **Unit: 7** **Concepts /Skills**	**KNOWLEDGE**	**COMPREHENSION**	**APPLICATION**	**TOTAL**
1. Solid shapes (cube, cylinder, rectangular parallelepiped, sphere)				
2. Flat shapes (triangle, circle, square, rectangle)				
3. Completing the table				
4. Graph (completion / creation)				
5. Patterns (abc/abcd)				
6. Measurement				
7. Problem Solution				
TOTAL				

A) Do you consider that the distribution of the exercises fulfills the objectives adequately?

……
……
……
……
……
……

B) Would you remove any exercises in order to improve the test? If so, which ones and why?

……
……
……
……
……
……
……
……
……
……
……
……
……

C) Would you add any exercises in order to improve the test? If so, such changes would be in relation to which concepts / activities / dimensions?

……
……
……
……
……
……
……
……
……

Application activity – Use of multiple assessment techniques

- Complete the following table in order to show how different techniques can be used in order to assess an objective

Objective	Written Exercise	Oral Assessment	Observation / Performance Assessment
1. Addition of similar fractions			
2. Identify the properties of the operations when they are presented in natural numbers and symbols ➢ commutative property of addition/multiplication a+b=b+a, a.b=b.a ➢ associative *property of addition/multiplication* (a+b)+c= a+(b+c), (a.b).c= a.(b.c) ➢ distributive property of multiplication over deduction (b-c).a=(b.a)-(c.a) ➢ distributive property of division over addition (a+b):c= (a:c)+(b:c)			

Application activity – Group Work

- Complete the following table in order to illustrate how you would organize a group work to assess the following objective:

Calculating the area of plane figures

Group Composition	**Group Organization**	**Activities**	**Assessment**
Number of members: ☐	➢ role assignment by the teacher ☐ ➢ role assignment by the team ☐ ➢ no role assignment ☐	Suggestions for activities that could be used:	➢ individual ☐ ➢ team ☐
• Homogenious Ability grouping ☐ • Heterogeneous Ability Grouping ☐	➢ fixed timetable / schedule ☐ ➢ flexible timetable / schedule ☐		Assessment concerning ➢ team contribution ☐ ➢ the result ☐ ➢ The degree of cooperation ☐
➢ Only boys ☐ ➢ Only girls ☐ ➢ Both boys and girls ☐	➢ only group work ☐ ➢ combination of group / individual work ☐		Assessment technique(s): …… ……... …… …….

Application activity – Homework

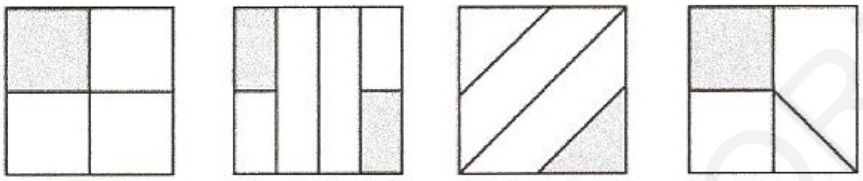
- **Math's Book, pg. 35-36 (problem solution).**
- **Math Worksheet (dividing fractions).**
- **Prepare for a test in Unit 5.**
- **Write down, 6 times, the multiplication table for the number 8.**

- The above are examples of homework. After examining them, discuss with your team:
  - Which objectives they serve?
  - What is their contribution to learning?
  - Do you consider the above homework appropriate for providing better learning in relation to the objectives mentioned?
- Which alternative homework or guidelines would you recommend? Take into consideration the characteristics of a constructive homework given to you.

……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……

Application activity – Example of Diagnostic Assessment

- Which of the following shapes has the ¼ of its area shaded? It may be more than one.



A B C D

- After examining the above exercise, complete the following tables.

Each alternative is aiming to check what?

A	B	C	D

How teaching should be organized for students who answered as follows:

A	A,B	A,D	A,C,D

Application activity – Differentiation of written assessment

- After examining the following objective, create exercises that could be used for the assessment of the particular objective in relation to the three levels.

Objective: Place value of digit / Analyze numbers into hundreds, tens and ones

Lower-level Exercise	Average-Level Exercise	Higher-level Exercise