Teacher professional development: a cross-national analysis of quality features associated with teaching practices and student achievement

Submitted by

Fabian Barrera Pedemonte

for the Degree of Doctor of Philosophy

of the

UCL, Institute of Education

2016

Declaration

I, Fabian Barrera Pedemonte, hereby declare that this thesis and the work presented in it is entirely my own. Where I have consulted the work of others, this is always clearly stated.

Word Count (exclusive of appendices, list of reference and bibliography): 47,120

Signed ……. (Fabian Barrera-Pedemonte)

Date: London, 08/11/2016

Abstract

This thesis ponders three theory-based aspects of the relationship between quality features of teacher professional development (TPD) and national educational outcomes by using comparable data from the United States (US), England, Japan and Finland. Studies carried out in the US and England have suggested that TPD delivered with *content focus, coherence, active learning, collective participation* and *longer duration* is linked with better teaching practices and student achievement. However, there has been no systematic examination of the generalisability of this association into different contexts, thus data from Japan and Finland is used here to explore this aspect. Firstly, I analyse whether student achievement in mathematics is associated with TPD which is either *focused on content* (Chapter 2) or managed *coherently* by head-teachers (Chapter 3). Then, I examine whether active teaching practices are associated to TPD with greater degrees of *active learning, collective participation* and extended *duration* (Chapter 4).

I find that *active learning* is positively associated in Japan with all the teaching practices examined, whereas in Finland it is only related to project-based learning. *Collective participation* is also positively associated with project-based learning in Japan, but it is particularly detrimental in Finland, also for the use of information and communication technologies (ICT). TPD with longer *duration* increases the likelihood of using ICT in the US, cooperative learning in England and project-based learning in Finland. Contrary to expectations, I find that the achievement of students in the English-speaking countries seems to slightly decrease insofar as the *coherence* of TPD improves. Likewise, I also find a slight negative association for English and Japanese students in relation to the engagement of their teachers in mathematics *content-focused* TPD. These results suggest, contrary to current theory, that the relationships between the quality features of TPD and educational outcomes are country specific. What is more important, they cannot be accepted in all cases as a panacea for rasing the quality of education.

Acknowledgements

I would like to thank my supervisors, Dr. John Jerrim and Prof. Dick Wiggins, for their constant encouragement, impeccable professionalism and challenging feedback. I have benefited greatly from their thoughtful advice, and detailed comments and suggestions throughout all my PhD, and, certainly, it would not have been possible to complete this work without their intellectual support.

Most importantly, there are no words to thank my family enough. To my loving wife, Alejandra, for enduring together the emotional and physical ups and downs of this long journey, and for making me feel that everything would be alright. To my children, Camilo and Sofia, for giving in some of their precious time and much of their overwhelming joy when it seemed that this thesis would never be written. I also wish to thank my family in Chile, that supported this endeavour in every moment and enjoyed each of my small achievements.

Finally, I would like to thank the National Commission for Scientific and Technological Research (CONICYT) of Chile for providing the necessary financial support to undertake my PhD in the UCL Institute of Education.

"Man is a god when he dreams, a beggar when he reflects"

- Holderlin

Contents

List of figures

List of tables

Chapter 1

Introduction

The majority of teachers in developed countries and emerging economies are now expected to engage in activities of professional development. In fact, in many nations, the participation in continuing training has become a compulsory requirement to maintain employment, as well as a necessary component to obtain promotion and salary upgrades (European Commission/EACEA/Eurydice, 2013; OECD, 2012a). According to data from recent international teacher surveys (Mullis *et al.*, 2012d; OECD, 2009a; OECD, 2014d) more than 86% of teachers attend professional development activities every year. In the 2013 cycle of the Teaching and Learning International Survey (TALIS) the country with the lowest participation rate was Chile, with 72%, whereas in Australia, Croatia, Latvia, Malaysia, Mexico, Singapore and Alberta (Canada), practically all of the teachers in post were engaged in these type of activities.

As countries are increasingly challenged to improve the outcomes of their educational systems, a healthy scepticism has gained currency in relation to the actual contribution of these activities to the improvement of teaching and learning. Evidence about the effectiveness of the professional development of teachers is of great interest, in particular the identification of the quality features that are associated with better educational outcomes. In this regard, recent literature has remarked that teacher professional development (TPD) that is *focused on content* knowledge, delivered *coherently*, and with greater degrees of *active learning*, *collective*

participation and longer *duration*, is consistently associated with better teaching practices and student achievement (Caena, 2011; Desimone, 2009). Large-scale studies carried out in the United States (US) (Garet *et al.*, 2001) and England (Opfer and Pedder, 2011b) have used national probability samples and concluded that these five quality features of TPD are good predictors of educational outcomes.

This thesis attempts to replicate the positive results reported in these two countries in order to compare estimates with those obtained with data from Japan and Finland, which happen to be two high performing countries recognised as with excellent systems of TPD (Robinson, 2014; Stewart, 2011; Williams, 2013). Accordingly, the aim of this investigation is to compare across all these four countries different aspects of the association between the five quality features of TPD (*content focus*, *coherence*, *active learning*, *collective participation* and *duration*), classroom teaching practices and student achievement. Detailed multivariate and multiple regression analyses of national samples of teachers and students collected from recent rounds of international large-scale assessments are presented to estimate the relationship between each feature and specific outcomes.

The investigation that follows is divided into five chapters with Chapter 1 being an overall introduction to the research work presented in this thesis, in terms of its background, the motivation for adopting a cross-national approach and the rationale of each chapter. Chapter 2 presents and discusses findings on the association between *content-focused* TPD and students' achievement in mathematics, as measured in the 2011 Trends in International Mathematics and Science Study (TIMSS). In Chapter 3, a cross-nationally equivalent measure of the *coherence* of TPD is developed using data from the 2012 Programme for International Student Assessment (PISA) to discuss results on its relationship with students' achievement in mathematics. In Chapter 4 the relative contribution of the features *active learning, collective participation,* and *extended duration* on the odds of using specific instructional methods (small groups cooperative learning, project-based learning and information and communication technologies (ICT)) is evaluated using data from the 2013 Teaching and Learning International Survey (TALIS). Conclusions about the limitations of these findings and implications for the design of national and global strategies for TPD are provided in Chapter 5.

1.1. Research background

In the last fifteen years, research has started to empirically examine measurable dimensions of the delivery of TPD that might explain improvements in this area for learning. These efforts began in 1999 when a group of researchers of the Eisenhower Professional Development Program conducted the first large-scale comparison in the field of mathematics and science education in the US (Birman et al., 2000; Garet et al., 2001). In this noteworthy investigation, the researchers evaluated two sets of features related to variations in teachers' knowledge, skills and practices: (1) structural features (type of activity, duration and type of participation); and (2) core features (active learning, coherence and content focus).

Using a national probability sample and regression analyses, the study concluded that the core features produced significant and positive associations with teachers' self-reported classroom practices and skills, holding constant variables at the school and teacher level. Such results were in general confirmed in follow-up analyses (Desimone et al., 2002) and in a similar large-scale design carried out in Australia (Ingvarson, Meiers and Beavis, 2005)[1].

Based on the consistency of these findings with other strands of research in the area of TPD, Desimone (2009) proposed a set of five core features as consensual domains for future research and practice. What follows is a brief description of each domain:

- **Content focus:** Instead of focus on generic behaviours of teachers, effective TPD programmes focus their content either on subject knowledge, the curriculum or the way students learn about the subject matter.
- **Coherence:** Effective TPD programmes are logically aligned to the goals of the educational policies that support them, as well as to the knowledge and beliefs of teachers.

[1] More recently in England, Opfer and Pedder (2011b), also using a national probability sample, reported that secondary teachers from schools with higher achievement participated in TPD activities with greater degrees of active learning, collective participation and longer duration.

- **Active learning:** Effective TPD programmes provide opportunities orientated to observe, design, perform or expose teaching practices, as a manner to engage teachers in inquiry-based learning experiences.
- **Collective participation:** This feature refers to the necessary interaction of groups of teachers from the same school to develop collaborative and meaningful learning amongst peers.
- **Duration:** Although research has not yet defined a specific time span, it is argued that longer term TPD programmes are more effective, both with regard to the overall amount of time that the activity takes and the total amount of hours spent.

An assumption, implicit in the argument of Desimone (2009), is that these five core features of TPD have a similar potential to influence outcomes at the teacher and student level. Indeed, a closer examination of the corresponding operational theory presented by the author reveals no hierarchical relationship among them. Recent experimental studies carried out in the US (Greenleaf et al., 2011; Heller et al., 2012; Penuel, Gallagher and Moorthy, 2011; Walker et al., 2012) have consequently used all these indicators as measures of the quality of the TPD delivered, regardless the type of outcome measured (teachers' or students' learning). In order to evaluate distinctive attributes of TPD programmes, the activities implemented both in the intervention and control groups have been delivered with equivalent levels of *content focus*, *coherence*, *active learning*, *collective participation* and *duration* as a means to control for their effect. As a result, most of this research has failed to provide a precise estimation of the relative importance of each feature on specific educational outcomes.

Furthermore, there has been no systematic examination of the generalisability (Campbell and Stanley, 1963) of the contribution of these five features of TPD into different contexts to date. Thus the evidence presented here is unique. Considering that the quality of TPD is sensitive to teaching and learning environments (Desimone, 2009; Villegas-Reimers, 2003; Wayne et al., 2008), it is vital to compare data from diverse sites to shed light on this aspect (Borko, 2004). Unfortunately,

studies implemented in multiple settings[2] are still difficult to find in the specialised literature, certainly because they involve greater costs than research focused on the performance of single TPD programmes. Consequently, judgements on the potential influence of system conditions have been disregarded in favour of estimating the net effect of the interventions. In this sense, research in the field of TPD is still insufficient to warrant that the suggested five core features can be applied to contexts other than those in which they were originally tested.

Noting these flaws in the approach suggested by Desimone (2009), Wayne et al. (2008) were guardedly optimistic about the idea of consensus and suggested to conduct more empirical studies orientated to estimate the specific role of each feature either at the teacher or student level. In this respect, one of the main contributions of this thesis is to offer a comprehensive framework to model outcomes at each of these planes, based on the corresponding theories that nurture this area of research[3]. The three pieces of work here presented are clearly targeted towards the main domain of influence of each quality feature of TPD. Student achievement is analysed in relation to the features *content focus* (Chapter 2) and *coherence* (Chapter 3), as suggested by the theories of instruction and context, respectively. In turn, teaching practices are examined as an outcome in relation to the features *active learning*, *collective participation* and *duration*, as it is indicated in the theory of teacher change.

In addition, this thesis assumes that the mechanisms by which the quality features of TPD influence educational outcomes are context-specific and merit examination in the light of variations across countries in order to enhance their external validity (Borko, 2004). Considering the lack of analyses on the generalisability of these indicators, the thesis contributes to the literature by (1) using data that is representative from diverse country populations, and (2) analysing variations in standardised measures of student achievement and comparable survey

[2] Only Heller *et al.* (2012) have implemented a national trial in the US to evaluate the impact of three TPD courses on science scores across multiple federal states. However, even in this study the researchers disregarded the inclusion of distinctive characteristics of the participant states and their schools in the analysis.

[3] A detailed literature review is presented in Appendix A to describe the theories underlying the influence of the quality of TPD on educational outcomes.

data on teachers' and head-teachers' practices. The reasons for adopting a cross-national approach to the analysis are discussed in the following section.

1.2. Motivation for cross-national analysis

National school systems are deemed in this thesis as the natural contexts in which the contribution of the quality features of TPD must be necessarily analysed. Such approach is adopted for a number of reasons. Firstly, it is well known that TPD is often implemented by national bodies in the context of education reforms, which normally use this mechanism as a mean to engage teachers into the pursued innovations (Little, 1993). Policy makers proceed in this way because TPD enhances the control over the contents and methods employed to disseminate the reforms and because the expected outcomes could be achieved in a shorter period of time compared to using initial teacher education. On the other hand, the selection procedures established to participate in TPD are relatively well defined at the national level. Indeed, countries can be classified according to whether such activities are a duty of the teaching profession, a condition for upgrades in the career or an optional complement of teachers' work. All these conditions are defined at the country level and must be taken into account to understand the role of TPD and its potential influence on national educational outcomes.

TPD is firmly embedded within the national organisation of schools and its influence on educational outcomes occurs within them and through teachers' work. A valid point in this regard is that educational policies and practices of TPD are culturally bounded and represent idiosyncratic models of organisation derived from particular lines of historical and social development (Hardy, 2012; Hardy et al., 2010). Although the high rates of participation in TPD activities across countries could be interpreted as a proof of a fully globalised model of teacher learning (Baker and Letendre, 2005; Baker et al., 2005), national differences are still substantial in relation to the way teachers teach in the classroom (OECD, 2009a; OECD, 2014a; OECD, 2014d; Vieluf *et al.*, 2012) and to the level of achievement of their students (Mullis *et al.*, 2012a; OECD, 2014b). Therefore, if educational outcomes differ

across countries with similar selection procedures to engage in TPD, then it is worth assessing whether features of the quality of TPD are related to this variation.

The empirical analysis in this thesis focuses on the cases of the US, England, Japan and Finland, which represent four developed countries where TPD is a compulsory requirement to maintain employment (OECD, 2012a). As stated earlier, studies implemented in the US and England have shown that greater degrees of exposure to the quality features of TPD can be associated to better national educational outcomes (Garet *et al.*, 2001; Opfer and Pedder, 2011b). In this sense, an obvious starting point is verifying whether similar findings are obtained in both nations when a similar method of analysis is performed with current data. Whereas this aspect could be revealing of the stability of the association between the quality of TPD and national educational outcomes, the replication of estimates with those obtained with data from most diverse countries could certainly shed light about the generalisability of the relationship.

The cases of Japan and Finland are intrinsically interesting and worthy of study in this regard. Hanushek, Piopiunik and Wiederhold (2014) have recently shown that their teachers possessed the best cognitive skills among the 23 countries assessed in the recent cycle of the Programme for the International Assessment of Adult Competencies. In the context of a wide cross-national variation in numeracy and literacy skills, the authors were able to comment, for instance, that Japanese and Finnish teachers outperformed the abilities of a Canadian professional with a master or doctoral qualification[4]. In addition, Japan has regularly headed the league tables of international large-scale assessments along with the East Asian nations, whereas Finnish students have regularly produced the highest scores among their European counterparts[5] (Barber and Mourshed, 2007; Mullis *et al.*, 2012b; OECD, 2014b). All in all, key aspects of the teaching profession in these two countries are fairly dissimilar to the cases of the US and England. In particular, patterns in the demand of teachers, the stringency of requirements to become teacher, the utilisation of teachers' workload, the trajectories of career structure, and the salary comparison

[4] In contrast, teachers in the US and England demonstrated an average level of mastery in both tests compared to the international sample analysed.

[5] The scores of the US and England are usually close to the international average in these evaluations.

with similar professions, show that TPD may play a very different role when compared to the US and England[6].

A cross-national analysis on the association between the quality features of TPD and national educational outcomes is also nowadays workable considering the availability of information produced by a number of international large-scale assessments. My review of the most recently accessible data revealed that variables that describe each of the quality features of TPD can be adequately taken from the information reported by teachers in TIMSS 2011 (Mullis *et al.*, 2012d) and TALIS 2013 (OECD, 2014e), and by head-teachers in PISA 2012 (PISA Consortium, 2011). One of the advantages of using this information is that the link between the quality features of TPD and teacher and student learning can be statistically analysed within and between countries. At the student level, TIMSS 2011 and PISA 2012 provide high-quality assessments of students' achievement in mathematics, whereas at the teacher level, TALIS 2013 includes detailed self-reports of instructional practices.

1.3. Issues of causality in the use of observational data

The statistical analysis of the association between these educational outcomes and the quality features of TPD seeks to estimate the magnitude, direction and significance of such relationship at the country level, as it is observed in the data reported by their participants (i.e. head-teachers, teachers and students). For example, this study can estimate whether students taught by teachers that participated in mathematics content-focused TPD outperformed or not (i.e. direction) students taught by teachers that reported participation in any other focus of TPD. In addition, this study is able to reasonably determine the amount of points in test scores that represent this difference (i.e. magnitude) and whether such estimates can be statistically inferred to the national target populations (i.e. significance). The empirical chapters of this thesis examine via regression analyses whether the corresponding estimates of association produce consistent results insofar as different empirical conditions are applied to the models employed. In addition, the theoretical

[6] Detailed information on these systems conditions is discussed in Appendix B.

underpinnings of these statistical models lend validity to their specification and provide a meaningful interpretation to compare estimates across all the four selected educational systems.

Nonetheless, it is worth underlining that this study makes no claims of causality between the quality features of TPD and the educational outcomes chosen for each analysis. Indeed, the findings derived from the statistical analyses of this thesis cannot rule out the complementary contribution of factors different from the quality features of TPD to the differences observed in the outcome variables. To do so, participants should have had equivalent chances to be exposed (or not) to different quality levels of TPD, as measured by the selected features, which in turn supposed their random assignment to such levels before the TPD programmes started (Bando, 2013; Campbell and Stanley, 1963; Gertler *et al.*, 2011).

This is not the case of the data used in this thesis. Unfortunately, the international large-scale assessments utilised in this study are not fully suitable to discover cause-and-effect links because they are based on observational data (Kaplan, 2016; Rutkowski, 2016). This means that they were collected without an explicit randomisation scheme through which participants could have been allocated to different levels of the predictors. Following the previous example, teachers were not randomly assigned either to mathematics-focused TPD or programmes with any other focus, hence factors underlying such selection may bias the causal contribution of content-focused TPD to students' outcomes. Under such circumstances, the effect of the quality features of TPD may be confounded by the mechanisms used by participants or their environment to select into TPD programmes with lower or higher quality. In sum, selection bias (Campbell and Stanley, 1963; Wayne *et al.*, 2008) might affect the proper causal estimation because both the exposure to high/low quality TPD and the educational outcomes might emerge from their correlation with such mechanisms (Pokropek, 2016).

Selection bias can be formally described in the context of multiple regression analysis, which happens to be one of the main statistical tools used in this thesis. This strategy is well equipped to estimate the association between the quality features of TPD and the outcome variable, provided that a portion of the variance in the outcomes is assumed to remain unexplained. The error term (or residual) represents

such portion, thus it can be conceptualised as the contribution of other variables that are not specified in the model which may bias the causal link with the outcomes (i.e. confounders).

Random assignment of participants to different quality levels of TPD would override the correlation between the error term and the exposure to the quality features of TPD. To put it another way, under random assignment of individuals confounder variables are assumed to be constant in the population and, therefore, they cannot be correlated with the key explanatory variables of the regression models (i.e. zero conditional mean assumption (Wooldrigde, 2003)). Hence, the outcome measures becomes independent from selection mechanisms and causal effect can be fairly established.

From this follows a cautionary note because the use of observational data in this thesis cannot warrant that the correlation between the key predictors and the error is cancelled out. In this study, for example, some degree of correlation between the quality features of TPD and the error term exists in the regression models presented in the empirical chapters. In view of this, this thesis uses methods oriented to diminish the magnitude of such correlation as best as possible. These methods consider that the error term can be decomposed into observable and unobservable parts (Winship and Morgan, 1999), whereby selection bias can be specifically approached in each case.

Selection bias on the observable part of the error supposes that variables containing potential confounders from participants are available in the datasets and that the error term correlates with them, but not with its unobserved part. In this context, the "control function approach" (Winship and Morgan, 1999, p. 672) proceeds by including a number of such variables in the regression model in an attempt to remove the correlation between the key explanatory variables and the error.

This is implemented in all the empirical chapters of this thesis. For example, Chapter 2 includes in the regression analyses blocks of variables from teachers (gender, experience, specialisation in mathematics teaching, teaching hours, teacher shortage and satisfaction) and students (gender and cultural capital at home). Chapter 2 also details the type of teachers that attend mathematics content-focused TPD in

each of the four key countries using all these variables. In Chapter 3, students' gender, and their immigrant and socioeconomic status are used as controls along with key characteristics of schools, such as type of administration (public/private), location, average class size and school size. Finally, Chapter 4 adjusts for by teachers' gender, their experience, the completion of initial training, and their attitudes towards teaching and learning. Despite the careful inclusion of these control variables, these analyses recognise that there could be still variables omitted in the models that might be associated either to the outcome measures or the exposure of teachers to TPD with different levels of quality.

This lends weight to the argument that the unobservable part of the error should also be approached in order to decrease selection bias. "Selection on the unobservables" (Winship and Morgan, 1999, p. 669) assumes that the error term is correlated with its unobserved part. This is a worst-case scenario because the chances of assignment to the different levels of the quality features of TPD become a function of variables that are not necessarily available to the researcher. This is addressed in Chapter 4 taking advantage of the clustered structure of the data (i.e. teachers nested within schools). In this case, the regression is restricted to a school fixed-effects model (Clarke *et al.*, 2010; Snijders, 2005) which removes all between-schools variables (both, observed and unobserved) that could bias the estimates of association between the quality features of TPD and the outcome measures. However, although this specification cancels out the correlation between the key explanatory variables and the school-level part of the error term, teacher-level omitted and unobservables variables may still play a confounding role.

1.4. Research question and overview of chapters

The study of the association between the quality features of TPD and national educational outcomes is certainly important because the implementation of each of these dimensions would imply major efforts for schools and their national systems. Furthermore, supra-national bodies are aware that specific aspects of the delivery of these activities should be monitored in order to evaluate the trade-offs between the

universal provision of TPD and the quality of its implementation[7]. Hence, it is crucial to identify which features are more associated to better school outcomes, as well as to estimate the extent of their relative contribution in order to guide policy decisions.

In order to examine the potential for any generalisability of cross-national comparisons about the strength of any association between the quality features of TPD and educational outcomes, a series of statistical analyses are developed with data from the US, England, Japan and Finland. The overarching research question to be answered is:

> **Are there differences in teachers' exposure to the quality features of teacher professional development that might be associated with differences in national educational outcomes at the student and teacher level?**

Three pieces of empirical secondary analysis are developed to address specific aspects of this question. Each of these works is individually organised, including an introduction that states its relevance and contribution, the characteristics of the datasets analysed, the planned statistical analysis and their corresponding results and conclusions.

Chapter 2 aims to estimate the statistical association between teachers' engagement in TPD activities that *focus* on mathematics content knowledge and the achievement of their students in this subject. This research builds upon conflicting evidence reported by Telese (2012) from his analysis of data from the National Assessment of Educational Progress in the US. Contrary to what meta-analyses in this area have regularly indicated (Blank and de las Alas, 2009; Kennedy, 1998; Salinas, 2010; Scher and O'Reilly, 2009), the author suggested that mathematics *content-focused* TPD would be negatively associated to student achievement in this subject. In this regard, the multiple regression analysis (Tabachnick and Fidell, 2001; Wooldrigde, 2003) applied in this chapter allowed estimating differences in this outcome (as measured in TIMSS 2011) that were associated with teachers' exposure to this quality feature of TPD within each country of interest. In order to examine the

[7] The policy background provided in Appendix C demonstrates that this is recurrent topic of debate.

consistency of these results, national estimates were gradually fitted while a number of characteristics of teachers and students were held constant in successive regression models. Further, a series of additional analyses allowed to examine whether results differed when the nested structure of the data was taken into account and when the focus of TPD was either on the pedagogy or the curriculum.

The aim of Chapter 3 is two-fold: whereas it puts forward a novel approach to examine the mechanism through which the *coherence* of TPD is enhanced within schools, it estimates the association between this quality feature and students' achievement in mathematics (as measured in PISA 2012). Firstly, attention is paid to operationalising the concept of *coherence* as an attribute of the leadership practices of head-teachers, as reported by themselves in the school questionnaire of PISA 2012. The potential of a set of indicators included in this instrument both to detect a latent construct related to this feature and to operate equivalently across countries, is evaluated using appropriate techniques of factor analysis (Brown, 2006). Secondly, a Hierarchical Linear Modelling (Raudenbush and Bryk, 2002; Snijders and Bosker, 1999) is adopted to model the expected positive association between country-specific measurement models of the *coherence* of TPD and student achievement (taking account of clustering within schools). As in Chapter 1, a number of background characteristics at the level of students and schools is gradually included as statistical controls in order to examine whether the *coherence* of TPD made a consistent difference to the average achievement of students within each country.

Unlike the previous two chapters, Chapter 4 concerns the way teachers teach in the classroom and how this is associated with their exposure to the remaining three quality features of TPD (*active learning, collective participation* and *duration*). Data from TALIS 2013 is used to extend the results of the official report (OECD, 2014d), which –interestingly- included no analyses on the quality features of TPD. Therefore, this chapter primarily aims to fill this gap by analysing whether TPD with greater degrees of *active learning, collective participation* and longer *duration* increases the chances of using three instructional methods: small groups cooperative learning, project-based learning and ICT.

Furthermore, this piece of work takes advantage of recent findings indicating that the daily experiences of professional co-operation carried out by teachers within schools are also consistently related with their classroom practices (de Vries, Jansen and van de Grift, 2013; de Vries, van de Grift and Jansen, 2013; Opfer, Pedder and Lavicza, 2011a). In this context, this chapter additionally examines the relative contribution of each quality feature of TPD after controlling for the engagement of teachers in such practices of professional collaboration. Accordingly, by means of an Ordinal Regression Model (Long and Freese, 2006; Winship and Mare, 1984), the analysis is able to answer whether –and to what extent- the three quality features of TPD, contributed to the odds of using each of these three instructional methods. School fixed-effects and teachers' attitudes towards teaching and learning are added in successive models in order to assess the consistency of estimates.

In sum, the following three secondary analyses of data from the US, England, Japan and Finland were conducted using different multivariate modelling strategies for the analysis of cross-sectional data. Taking into account the complex survey design of each dataset, these techniques allowed estimations on the extent to which measures of educational outcomes varied according to their association with each of the quality features of TPD examined in the corresponding statistical models.

Chapter 2

Mathematics content-focused professional development and student achievement: a cross-national analysis of TIMSS 2011

2.1. Introduction

A number of educational policy documents pose that initial teacher education is insufficient to support teachers within the ever changing context to which education is exposed (Coolahan, 2002; Musset, 2010; OECD, 1998; OECD, 2005). In particular, the obsolescence of teachers' knowledge becomes a cause of concern given the accelerated progresses of science and technology in the current age of information (Jarvis, 2007). As a result, activities orientated to keep teacher knowledge 'up-to-date' are regularly promoted as a key strategy to improve the quality of education. Any national reform would require the implementation of this type of support mechanism to be successful.

The case of the US provides an exemplary illustration because its reform in subject matter teaching has put a constant strain on its educational system in the last twenty years (Little, 1993). In addition to the expected changes in knowledge *per se*, teachers in the US have had to enact new practices given the transformation of

curriculum and pedagogy promoted by the agenda of standards (Darling-Hammond and Ball, 1998; Darling-Hammond *et al.*, 2009). As a result, the area of TPD has become a relevant issue in the field of teacher education in this country. By promoting on-the-job teacher learning throughout teachers' career and encouraging a specific research domain focused on the effectiveness of these activities, the quality of TPD provision has been acknowledged as a key factor leading to quality of education overall (Department of Education. United States of America, 2011).

Research aimed to analyse the effectiveness of TPD is particularly profuse in the US and can be classified according to whether the main outcome of interest is at the teacher or student level (Supovitz, 2001; Wayne et al., 2008). The literature accounting for outcomes at the teacher level has proposed a number of core features of TPD activities that impact on teacher knowledge, beliefs and practices[8] (Garet *et al.*, 2001). However, when student achievement is considered as the main outcome variable, meta-analyses[9] have consistently remarked that *content focus*, i.e. TPD activities focused on subject matter knowledge, is the most important of such dimensions (Blank and de las Alas, 2009; Kennedy, 1998; Salinas, 2010; Scher and O'Reilly, 2009).

For example, Scher and O'Reilly (2009) reported that students of teachers engaging in mathematics content-focused TPD obtained an estimated .38 standard deviations higher score on mathematics evaluations than their counterparts whose teachers did not engaged in this type of TPD. Blank and de las Alas (2009) reported similar positive results for TPD focused on content knowledge. By synthesising 16 empirical studies, the effect sizes[10] on student mathematics achievement were .21 for pre-post measures and .13 for only post measures.

[8] See page 14.

[9] Meta-analysis is one form of systematic synthesis of research in which "a large collection of analysis results from individual studies [are statistically analysed] for purposes of integrating the findings" (Glass, 1976).

[10] Effect size refers to "the standardised difference between two means" (Howell, 2007, p. 229). This is calculated by taking away the mean parameter of a baseline group – μ_0 (e.g. control group, before intervention measure, etc.)- from the mean parameter of an intervention group –μ_1-, and dividing this

In contrast to these studies, Telese (2012) has suggested that content-focused TPD is rather negatively associated with student achievement in mathematics, as measured in the US's National Assessment of Educational Progress (NAEP). This is a nationally representative study carried out annually to evaluate students' performance in several subjects. Using data from 8th grade schools in 2005, the author found that students of teachers participating in TPD that included a moderate or large focus of mathematics content had lower scores in this subject than their counterparts that were taught by teachers reporting a null level of this dimension of the quality of TPD. Whilst small effect sizes were reported (.06 and .08 in relation to moderate and large extents, respectively), these results suggested that teachers' participation in TPD activities with more than a small dose of mathematics knowledge would result in lower scores in standardised tests of student achievement in mathematics.

Unfortunately, current experimental research (Greenleaf *et al.*, 2011; Heller *et al.*, 2012; Penuel, Gallagher and Moorthy, 2011; Walker *et al.*, 2012) have failed to clarify this issue. As these studies have been aimed to estimate the effect of distinctive characteristics of TPD programmes, the differential contribution of *content focus* on students' outcomes has remained unexplored. Only Walker et al. (2012) have attempted to test the effect at the student level of variations in the focus of TPD across several programmes. However, as the outcome variable in this study was based on self-reported student gains, the effect of this feature on standardised measures of student achievement is still largely unknown. Furthermore, most of these findings proceed from Randomised Controlled Trials, with samples that are generally small and that in no case represent information about the whole country. Sample sizes usually involved the participation of less than twenty-five teachers per study and they were regularly drawn from limited regions within some of the federal states.

In contrast, when evidence comes from observational data –as in Telese (2012)-, studies analyse what actually happen with TPD under "normal" (status quo) conditions. In these cases, sample sizes include up to thousands of teachers per study,

result by the standard deviation of the parent population – σ ($d = \mu 1 - \mu 0 / \sigma$). Cohen (1988) suggests the following levels in order to interpret results: *d*<.2, Small; *d*<.5, Medium; *d*<.8, Large.

and several intervening variables are embedded to the analysis as controls. Due to these features, observational evidence is likely to be more generalizable because it includes more representative samples of target populations and external validity is enhanced. Nevertheless, as the assignment of TPD activities is unlikely to be random with respect to student's outcomes, one cannot rule out the presence of possible unobserved (and uncontrolled) confounding factors.

In sum, more research is needed to understand the individual contribution of TPD focused on content given its expected association with student achievement. Considering that this feature is sensitive to characteristics of context (Desimone, 2009), researching its degree of generalisability into different teaching and learning environments becomes crucial. The issue challenges *per se* the debate on effective TPD by carrying the attention to the contextual conditions that might constrain its link with student achievement (Opfer and Pedder, 2011a). Larger scale and comparative designs are particularly fitted for this purpose (Borko, 2004), and a cross-national analisys can shed light about the external validity of its contribution at the macro level.

This study is intended to contribute to this literature with comparative evidence about the relationship between student achievement in mathematics and the participation in TPD activities that focus specifically on mathematics content knowledge. The analysis is aimed to answer what is the specific contribution of this variable after controlling for contextual characteristics at the student and teacher level. By statistically modelling this relationship in recent international large-scale data produced by the 2011 Trends in International Mathematics and Science Study (TIMSS), this chapter compares the role of TPD in the US to three other developed countries (England, Japan and Finland).

I focus particularly on the US in this chapter due to its contribution to the evidence on the topic (Blank and de las Alas, 2009; Yoon et al., 2007), its current policy debate on the issue (Department of Education. United States of America, 2010; Department of Education. United States of America, 2011) and its high teacher participation in TPD activities focused on mathematics content (IEA, 2012). Such level of participation might be partially explained by the regular shortage of teachers in this subject and the comparatively lower requirements to become teacher (see

Appendix B). In this context, the US system seems to particularly concentrate the opportunities for TPD in the delivery of pure knowledge of mathematics.

The case of England is similar to the US regarding the high attrition of mathematics' teachers and the low requirements to obtain a teaching qualification. In contrast, teacher training is highly demanded in Japan and Finland and their selection of candidates is very competitive. These two countries have consistently shown good results in international tests of mathematics (Mullis *et al.*, 2012a; OECD, 2014b), thus I use their cases in this chapter in order to understand the results given such diverse conditions. Recall that TPD activities are compulsory in all the four countries examined in this thesis, however they differ in other teacher policy areas and such divergence is useful to contextualise the findings of this analysis.

Having these national contexts in mind, the main question to be answered is: does mathematics content-focused TPD relate to student achievement in this subject, controlling for characteristics of students and teachers? This question partially contributes to answer the overaching question of the thesis, in terms of examining whether differences in teachers' exposure to this kind of TPD might be associated with variations in student achievement within countries[11].

This secondary analysis of data collected from 8th grade teachers and students is conducted following an Ordinary Least Squares (OLS) regression. In the next section I describe the empirical methodology implemented in order to answer this question, as well as relevant features of the dataset analysed, i.e. Grade 8 TIMSS 2011. Section 2.3 provides estimates of association between student achievement in mathematics and TPD focused on mathematics knowledge in the four countries of interest. This is followed in section 2.4 by a discussion of findings and conclusions.

[11] See page 23.

2.2. Data sources and methodological strategy

2.2.1. Survey design

In this study, I use data drawn from the 2011 round of the TIMSS; an international large-scale assessment conducted every four years by the International Association for the Evaluation of Educational Achievement. I focus on the 8th grade target population of the assessment, which corresponds to all students enrolled in the eighth year of schooling in each country (13/14 years old). In order to collect data from this population, TIMSS uses a two-stage stratified cluster sampling strategy: in the first stage, each country is expected to randomly select at least 150 schools from their national frames, with probability proportional to their size. In the second stage, one or two intact classes are sampled within each school, therefore students belonging to them are those finally tested. The procedure usually involves the participation of more than 4,000 students from each educational system. In TIMSS 2011, 45 countries and 14 benchmarking participants (usually, states within countries) took part of the Grade 8 assessment in mathematics (Mullis *et al.*, 2012c).

The survey design of TIMSS is complex and its precision is actually jeopardised by nonparticipation of schools and absence of students on the day of the test. For this reason, the organisers of the assessment set minima of school, classroom and student rates participation, regarding original sample sizes. In Grade 8 TIMSS 2011 average response rates of both schools (95%) and pupils (96%) were high, though England just satisfied guidelines for the sample of schools (75%) (Mullis *et al.*, 2012c). In order to correct for the classrooms and students non-response a set of sampling weights (Rust, 2013) are calculated in each round of the assessment by the organisers. As a result, for each student there is a specific weighting factor that informs his or her inverse probability of selection, with the necessary adjustments for nonresponse. Then, by multiplying this number with measures of interest, estimates become representative of the national target population. In this paper I consider these features of the sampling strategy, thus weighting factors at the student level are part of every analysis here developed.

2.2.2. Student achievement

The outcome variable "student achievement" is defined in this study as the overall mean score in mathematics as measured in Grade 8 TIMSS 2011. This variable has been transformed to have a mean of 500 points and a standard deviation of 100 across all participant countries. Data on "student achievement" are also complex due to assessment design characteristics that derive from the curricular framework of TIMSS. Broad subject content domains in mathematics are expected to be covered in each round of the test, thus in order to yield accurate results of domain proficiencies at the national level, TIMSS follows a multiple-matrix sampling strategy based on Item Response Theory (Embretson and Reise, 2000; Mirazchiyski, 2013). Thereby, a huge amount of items are elaborated to validly measure specific domains such as Number, Algebra, Geometry, and Data and Chance, and Mathematics as a composite of all of them. Unlike a normal test not all items are administered to each student as it would take a considerable amount of time to answer the whole battery. But responses of all students to items assigned are used, thus combined results of the test are regarded as student achievement in TIMSS.

By the same token, estimations of proficiency for each student are actually unknown. However, they can be calculated from a hypothetical distribution based on the individual responses to the assigned set of items. In order to report the performance that a student might reasonably have in the overall score of mathematics, TIMSS uses a fixed number of five plausible values as random draws from this distribution. Rutkowski et al. (2010) remark the threats to statistical analysis when plausible values are not adequately considered for estimates of student achievement in international large-scale assessment such as TIMSS. One of these threats is the calculation of national estimates using the mean of plausible values as a single parameter of achievement for students. In this case, the authors comment the mistake of calculating the average of the five plausible values for each student and then the mean of every student score to obtain the national estimation. In this case, standard errors are dramatically underestimated.

In this study, the TIMSS assessment design is taken into account, but only the first plausible value is analysed as dependent variable. This alternative procedure has been suggested by the PISA organisers as an efficient method to obtain

equivalent mean and regression estimates, but with standard errors that might slightly differ from estimates calculated with the five plausible values (OECD, 2009b). In line with other studies that have used this method (Gilleece, 2015; Grilli *et al.*, 2014; Jerrim, 2011), this chapter applies the first plausible value drawn from the populations of interest as the main outcome variable.

2.2.3. Key explanatory variable

Data about the key explanatory variable "mathematics content-focused TPD" is drawn from one of the items included in the teacher questionnaire of TIMSS 2011. This instrument is administered to every mathematics teacher of the sample of students and contains questions about their background, the school where they work and their teaching practices. Among the questions which are designed to collect information about their preparation to teach mathematics, they are requested to indicate the main focus of recent experiences of TPD. The item content is taken from the instrument as follows (TIMSS & PIRLS International Study Center, 2011, p. 14):

"In the past two years, have you participated in professional development in any of the following? Check one circle for each line.

a) Mathematics content (yes/no)"

It is probable that the self-report procedure that is used in the instrument might become an important source of measurement error[12] for this item. Problems such as confusion of teachers trying to comprehend the substantial content of every single alternative of the item and memory imprecision according to the timeframe might yield inaccurate information on the variable. In addition, there is no possibility within the instrument to contrast the precision of the answers against actual data or other variables. It is inevitable the presence of this type of deficiencies in survey data, but the point is whether the extent of their influence on the quality of information collected is tolerable.

[12] The term *measurement error* denotes "unexplained variation in a measurement" (Hutchison, 2008, p. 444). It can be observed when a replication of the same measurement process produces a different value.

Desimone (2009) comments that teacher survey items eliciting factual data about TPD experiences show acceptable indicators of reliability and validity. In this sense, as long as the information requested about the focus of TPD is descriptive (about facts) and no evaluative (personal judgements about facts), self-reported data are well supported, as in this case. In addition, after reading more carefully every alternative of response, it becomes easier to realise that contents do not overlap and, on the contrary, they refer to very clear and probable topics addressed in TPD activities. The fact that every item is binary also helps teachers as they do not need to rate or compare across points on semantic scales or between multiple choices.

Furthermore, the clarity and substantive independence of the topics addressed in each of the alternatives of the question allows the isolation of different foci of TPD experiences. For the purposes of this study, this aspect becomes relevant because among these possible foci, the focus on mathematics content (e.g. alternative (a)) can be analysed separately from the rest of topics. This study aims to estimate the specific contribution of this feature to student achievement, considering that specialised literature has been suggesting its critical role in order to influence student achievement in mathematics (Blank and de las Alas, 2009; Kennedy, 1998). Therefore, the analyses in this chapter only uses the responses of teachers to the first alternative of the question. Table 2.1 indicates the relative frequency of this item in every educational system that took part in TIMSS 2011, as well as the percentage of teachers who received TPD in any of the foci. Countries of interest are highlighted in grey.

Table 2.1 Percentage of teachers attending TPD focused on mathematics content and TPD in general. TIMSS 2011 8th Grade Mathematics

	TPDCont		TPD			TPDCont		TPD	
	%	(SE)	%	(SE)		%	(SE)	%	(SE)
Thailand	75	(3.5)	91	(2.3)	**United States**	57	(2.4)	96	(1.0)
Lithuania	75	(3.4)	97	(1.5)	Alabama, US	57	(6.1)	100	(0.0)
Ukraine	75	(3.8)	93	(2.2)	Saudi Arabia	55	(4.2)	81	(3.3)
Alberta, CAN	75	(3.4)	96	(1.7)	**England**	55	(4.1)	90	(2.8)
Chinese Taipei	73	(3.6)	94	(2.0)	Honduras	55	(4.4)	81	(3.7)
Kazakhstan	72	(3.7)	95	(1.8)	Lebanon	54	(4.1)	84	(3.0)
Indonesia	71	(4.2)	86	(3.5)	Georgia	52	(3.9)	85	(2.9)
Israel	70	(2.7)	93	(1.4)	Iran	51	(3.5)	85	(2.4)
Romania	69	(3.7)	93	(2.2)	Quebec, CAN	51	(4.1)	87	(3.0)
Hong Kong	69	(4.1)	92	(2.3)	Korea, Rep. of	49	(3.0)	77	(2.5)
Tunisia	68	(3.6)	89	(2.3)	Oman	47	(3.3)	80	(2.7)
Qatar	68	(4.6)	92	(2.6)	Abu Dhabi, UAE	44	(4.3)	88	(2.8)
Minnesota, US	67	(5.3)	96	(3.0)	Dubai, UAE	44	(4.3)	89	(3.4)
Russian Feder.	66	(3.2)	92	(1.8)	United Arab Emir	44	(2.5)	86	(1.9)
Singapore	66	(3.1)	93	(1.5)	California, US	43	(5.5)	90	(4.4)
Ghana	66	(3.9)	84	(3.1)	Malaysia	39	(3.7)	70	(3.5)
Japan	65	(3.9)	84	(2.9)	Australia	38	(3.6)	92	(2.4)
South Africa	65	(3.6)	88	(2.8)	Morocco	36	(3.2)	81	(2.7)
Massachus..., US	65	(6.0)	96	(2.6)	Hungary	33	(3.6)	81	(3.0)
Armenia	64	(3.6)	96	(1.6)	Palestina	30	(3.8)	79	(3.4)
North Carol., US	64	(6.0)	96	(3.3)	Turkey	30	(3.2)	67	(3.2)
Florida, US	63	(5.8)	100	(0.0)	Bahrain	29	(4.9)	78	(4.8)
Colorado, US	62	(6.1)	98	(1.3)	Sweden	28	(2.8)	77	(3.1)
Indiana, US	61	(5.8)	97	(1.7)	Syria	25	(3.6)	78	(3.9)
New Zealand	60	(3.7)	89	(2.2)	Jordan	23	(3.3)	60	(3.9)
Macedonia	59	(3.9)	100	(0.0)	Italy	21	(3.0)	81	(3.0)
Chile	59	(3.9)	78	(3.4)	Botswana	21	(3.4)	64	(4.3)
Ontario, Canada	58	(3.8)	91	(2.6)	Norway	21	(3.3)	52	(4.3)
Connecticut, US	58	(6.1)	98	(1.6)	**Finland**	9	(1.9)	51	(3.9)
Slovenia	57	(2.9)	93	(1.7)	**Total**	57	(1.0)	86	(0.8)

Source: TIMSS 2011 database

Notes: TPDCont = participation in mathematics content-focused TPD.

It is worth noting from this table the important contribution of TPD focused on mathematics content to the high implementation of TPD in general. For example, 86% of teachers in the total sample participated in some type of TPD in the last couple of years. Even in the countries with the lowest rate (Finland and Norway), TPD involved the participation of more than a half of the sample, whereas in a number of educational systems practically all teachers participated in some type of TPD (Macedonia; Alabama, US; and Florida, US, 100%).

The partial proportion of teachers who reported in average that they had attended TPD focused on mathematics content is not negligible (57%), however the rate is highly dispersed across countries. In a quarter of countries more than two thirds of teachers attended this specific aspect of TPD. The range of variation goes mainly from 75% in Thailand, Lithuania, Ukraine and Alberta, Canada to 9% in Finland. The rest of the countries of interest for this study shows percentages that are close to the international proportion: England (55%), the US (57%) and Japan (65%). Among these, I would highlight the low percentage of participation in Finland in TPD in general (51%) and particularly in TPD focused on mathematics content (9%), which is less than a half of the proportion of its predecessor in the list (Norway, 21%)[13].

2.2.4. Analytic strategy

I use an OLS approach to answer the research question of this study. This strategy allows estimating how a continuous outcome variable varies with changes in one or more predictor variables (Howell, 2007; Wooldrigde, 2003). In this case, I am interested in how within each country student achievement in mathematics varies with changes in the participation of their teachers in mathematics content-focused TPD. The analysis proceeds examining how this relationship changes once blocks of control variables are gradually included in successive statistical models which build up on the number of controls. In the context of OLS, a simple bivariate regression analysis firstly informs about the magnitude, direction and statistical significance of the unconditional association between the outcome and the key explanatory variable (Model 0). Then, a multiple regression analysis evaluates how this relationship fluctuates holding constant the rest of predictors in two nested successive models. In this sense, the OLS approach indicates the conditional expectation of the association by using control variables grouped into three thematic blocks.

[13] The percentage of missing data in the key explanatory variable was 1% for Japan, 4% for Finland, 9% for England and 23% for the US.

Block 1 and 2 includes background characteristics of students and teachers, respectively, such as student and teacher gender, student cultural capital at home[14], teaching experience and specialisation in mathematics teaching. Because it is possible that teachers in different types of schools or teachers with different characteristics might experience different types of TPD, I include these student and teacher characteristics as control variables in Models 1 and 2. Block 3 adds in organisational variables to the second model (Model 2) (teaching hours, teacher shortage and teacher satisfaction). It is suggested that teachers with different levels of satisfaction or schools with different number of classroom hours and levels of teacher shortage, might experience different types of TPD. This draws on evidence for the US where the most qualified teachers are the ones who attend content-focused TPD (Desimone, 2009). Block 3 variables attempt to control for these teacher characteristics within each country [15].

The final form of the model is:

$$A_{ijk} = \alpha + \beta_{1.} TPDContent_j + \beta_{2.} Block1\&2_i + \beta_{3.} Block3_j + \varepsilon_{ij} \qquad \forall k$$

Where:

A = Student achievement measured as the Overall Mean Score in Grade 8 TIMSS 2011.

TPDContent = A binary variable indicating the participation of teachers in mathematics content-focused TPD (1=yes, 0=no).

Block 1&2 = A set of six variables about student and teacher background characteristics.

[14] The *highest parental education level* is included as indicator of student family background and the *number of books at home* is used as indicator of family scholarly culture (Evans *et al.*, 2010).

[15] The list of variables included in the analysis is provided in Appendix E. In general, the typical percentages of missing data were approximately 2% in Japan and Finland, 9% in England, and 20% in the US. In order to maximise the amount of information available and boost sample sizes, item mean substitution was applied (Eekhout, 2014; Hawthorne and Elliott, 2005). This method involves replacing the missing values for a case on one variable with the weighted mean value of all other participants that have valid values for that variable. Dummy variables were generated for each predictor to indicate those cases where the missing values were substituted. These variables were also included in the corresponding regression models.

Block 3 = A set of three variables about organisational features of teachers work.

ε = Error term.

i = Student i.

j = School j.

k = Country k.

A potential threat to efficiency in OLS is the lack of independence amongst predictor variables. For instance, when two predictors are highly correlated the information of one of them is rather redundant and it might indicate that they are actually measuring the same construct. A polychoric correlational matrix method has been used to examine the relationship among the predictors for each country of interest[16]. The results yielded none correlation higher than .5 between pairs of predictors in each of the four key countries, which indicates no high multicollinearity for the purposes of the analysis.

Furthermore, in order to ensure accuracy of standard errors, the error variance in OLS has to be constant for each level of the observed predictors, i.e. homoscedasticity (Howell, 2007). For example, considering student gender as predictor, the error variance in student achievement should be statistically similar either for male or female students. Concerning this aspect in the context of complex survey data, all the analyses here presented have been executed using the "svy" command in STATA 12©, which adjusts estimations for potential departures from homoscedasticity (Reale, 2006; StataCorp, 2011a).

From OLS estimates it is possible to gain knowledge on the association between the key explanatory variable and the outcome. However, the causality of this link cannot be claimed from this analysis. To do so, extraneous variables should be assumed to be constant in the population and, therefore, not correlated with the predictors, i.e. zero conditional mean assumption (Wooldrigde, 2003), which is a strong assumption to make. From this follows a cautionary note because controlling for the effect of all the potential unobserved variables is difficult to be accomplished by conducting a secondary analysis of cross sectional data. There could be still

[16] See Appendix F.

variables that are actually associated either with student achievement or the participation of teachers in mathematics content-focused TPD, and not included in the models here presented.

The point is worth to mention as it constrains causal interpretations from results yielded by the statistical analysis followed in this study. In this sense, for example, it must be noted that TPD is not randomly assigned in this study, therefore self-selection of teachers to TPD might work to bias estimates (Desimone, 2009). In order to describe which type of teachers attend mathematics content-focused TPD in each of the four selected countries, characteristics of teachers and their students are presented in the results section using some of the available variables.

Finally, in the analysis I consider mathematics content-focused TPD as a teacher-level variable interpreted at the level of each student (Rutkowski et al., 2010). Recall that students are randomly selected in TIMSS, not teachers, thus every interpretation of results has to proceed as a student-level analysis. Nevertheless, the design of TIMSS also randomly selects the schools from which students were sampled, therefore teacher-level variables might be also interpreted as an attribute of schools. In this case, a Hierarchical Linear Modelling (HLM) strategy (Raudenbush and Bryk, 2002; Snijders and Bosker, 1999) would be suitable, especially in the US and England where the proportion of between-classes variance in student achievement (e.g. intra-class correlation, 56% and 67%, respectively) is higher than the within-classes[17]. In this regard, I conducted HLM for each of the models as part of a set of additional analyses in order to know whether findings using this approach differ from OLS models. Instead of content, these tests also include the use of pedagogy or curriculum as key explanatory variables in order to evaluate whether such foci might alternatively work as better predictors of student achievement than mathematics content-focused TPD.

In the following section the main statistical results are reported. Estimates are presented from a comparative perspective that ultimately focuses upon the US, England, Japan and Finland. Firstly, a macro level analysis of the relationship between the key explanatory variable and the outcome is illustrated in the context of the 59 educational systems that took part in mathematics Grade 8 TIMSS 2011. The

[17] The intra-class correlation in Finland is 40% and in Japan 30%.

coefficients informing their unconditional (bivariate) association (Model 0) are then presented for the 15 participant OECD countries, paying special attention to the results observed in the four key countries. Secondly, variations in the estimates from Model 0 to 2 are examined in detail for the US, England, Japan and Finland. In addition, characteristics of teachers engaging in mathematics-focused TPD in these countries are described using a number of available variables, whereas the cross country variation between the estimates in Model 2 is assessed using an Independent Samples t-test technique[18]. Finally, a set of additional analyses are used to evaluate how OLS results might vary either using an HLM approach or alternative foci (pedagogy or curriculum) of TPD as key explanatory variables.

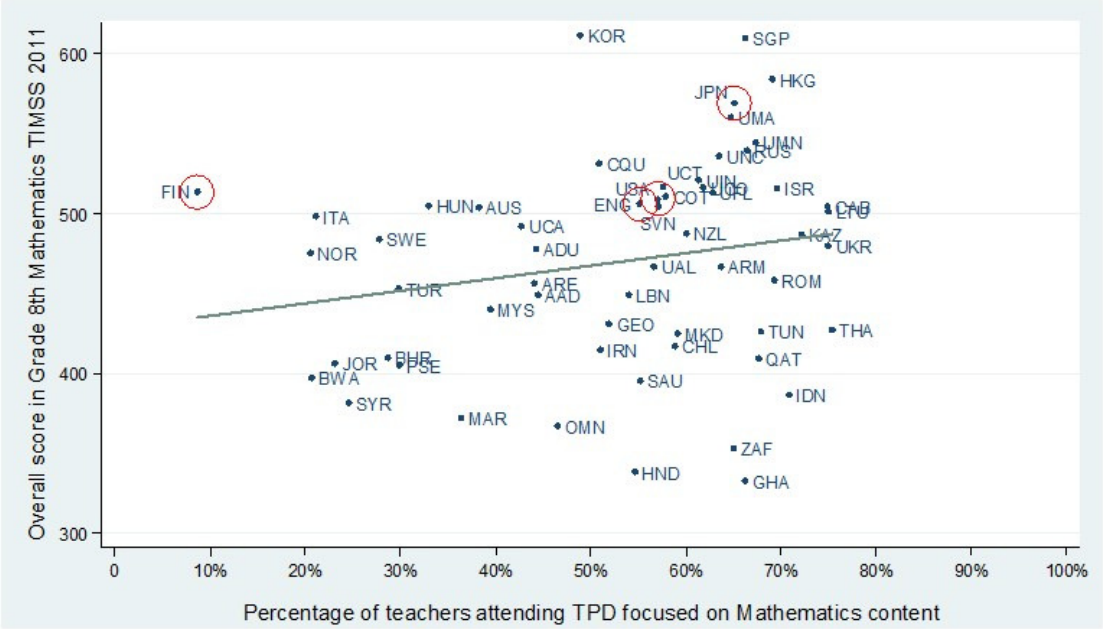
2.3. Results

2.3.1. OLS results

This section presents the main findings for the research question stated above on page 30. Figure 2.1 illustrates the relationship between the proportion of teachers attending mathematics content-focused TPD and student achievement across the 59 educational systems that took part in Grade 8 TIMSS 2011. The dots in the diagram indicate each country, whereas the four selected countries for detailed analysis are indicated within red circles.

[18] As at this point the individual hypothesis stating potential differences between pairs of countries will be tested in multiple occasions, the 95% level of statistical confidence of each comparison might no longer represent the error rate of the set of comparisons among countries as a whole. Nevertheless, as the analysis involves only a small number of simultaneous planned hypotheses -just six, one for each comparison between countries-, I considered this not being a substantial issue and I preferred do not execute a multiple hypothesis testing approach using Bonferroni correction (Shaffer, 1995).

Figure 2.1 Mathematics content-focused TPD and student achievement in Grade 8 TIMSS 2011 for 59 participating countries



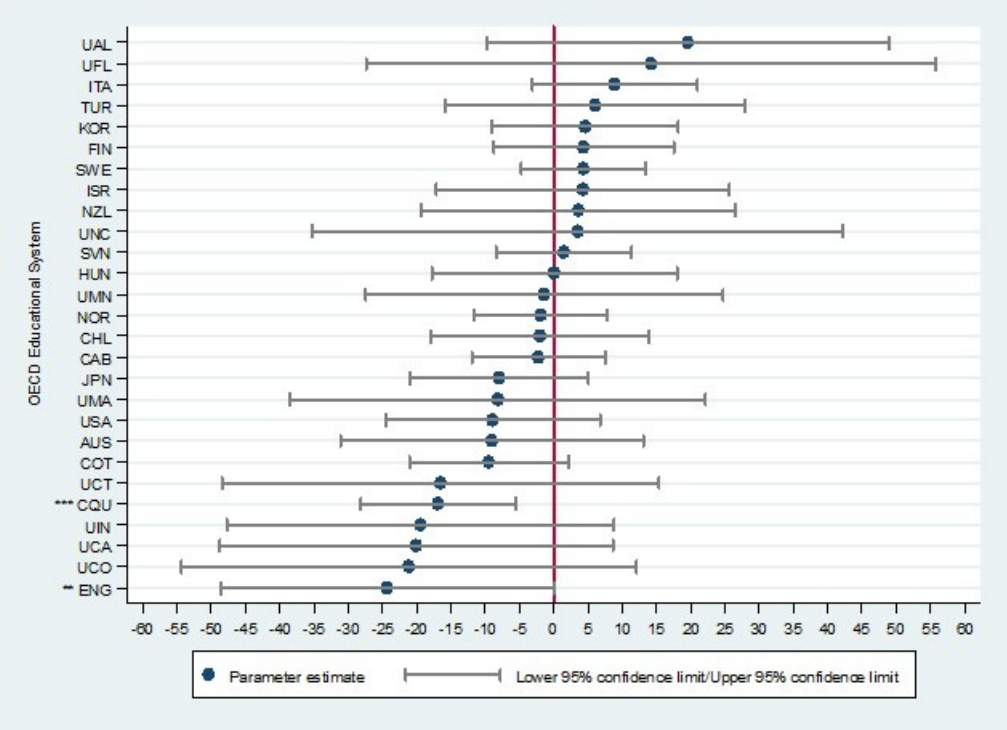
Source: TIMSS 2011 database

In general, there is a slight positive correlation between student achievement and the participation in mathematics content-focused TPD across all countries. In other words, as the percentage of teachers that attended this type of TPD increases, mathematics student achievement increases as well. The association is weak (Pearson's $r = .16$), with only 4% of the variance in student achievement explained by for this predictor at the macro level. In this sense, the evidence is poor at the country level for the link between mathematics content-focused TPD and student mathematics achievement. Among the key countries of this study, Japanese students showed high performance in the subject, whereas 65% of their teachers attended this type of TPD. On the other hand, student achievement in the US (508 points), England (507 points) and Finland (513 points) was close to the international mean score (500 points), but unlike US (57%) and English (55%) teachers, only 9% of Finnish teachers attended mathematics content-focused TPD.

Figure 2.2 shows the magnitude, direction and statistical significance of the unconditional association between the key explanatory variable and student achievement (Model 0) for each of the OECD systems participating in the assessment. Quantity of score points in the assessment associated to the participation

of teachers in this type of TPD is indicated on the horizontal axis. The direction is referenced using a red line starting from zero and statistical significance with intervals in grey colour indicating a range of 95% of confidence.

Figure 2.2 Unconditional association between student achievement and mathematics content-focused TPD (Model 0) across the 15 OECD educational systems participating in Grade 8 TIMSS 2011



Source: TIMSS 2011 database

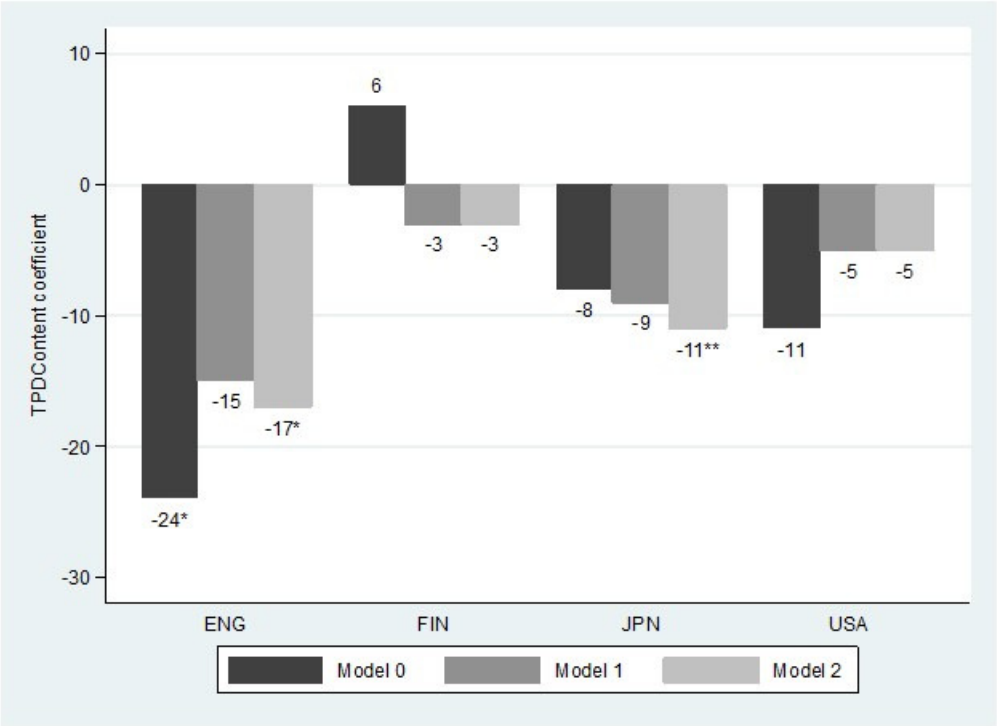
Notes: *p < .1. **p < .05. ***p < .01.

In general, the magnitude of the unconditional association between these variables is small across this pool of countries. Further, the direction of the relationship is negative in the only two countries where this is statistically significant. This is the case of Quebec, Canada, where the parameter estimate is -17 points (0.17 international standard deviations). And it is also the case of one of the key countries of the study, e.g. England, where students taught by teachers that attended mathematics content-focused TPD scored 24 points less in the assessment (0.24 international standard deviations). England yielded the highest negative estimate, whereas Alabama, US yielded the highest positive value (20 points). These ranges were substantially small taking into account that they represented less than a quarter

of one standard deviation of the international score (100 points) and they were centred on zero; indeed, in only half of the educational systems the estimation was positive. Among the other key countries of this study, the estimates in the sample were positive for Finland (6 points) and negative for Japan (-11 points) and the US (-8 points).

Figure 2.3 below shows how the parameters in Model 0 (unconditional association) change once blocks of control variables are added to the models in the four countries of interest. Bars indicate the magnitude, direction and statistical significance of the association and they are grouped by country. Detailed parameters on each predictor included in the models is provided in Appendix G.

Figure 2.3 Variation of the conditional association between student achievement and mathematics content-focused TPD across models for England, Finland, Japan and the US



Source: TIMSS 2011 database

Notes: *p < .1. **p < .05. ***p < .01.

As in the general pattern found in Model 0, estimates were typically small and negative in these samples. In the US, the association decreased from -11 points to -5 points once background variables were added in Model 1 and it remained the

same value in Model 2. However, as these values were not statistically significant this study does not provide evidence of a consistent association between mathematics content-focused TPD and student achievement in the US population. Against conventional wisdom and even against existing literature supporting this association in the country, this type of TPD seemed not to be linked to the performance of students in this country (possible explanations for this will follow in the next section). This point was also valid for Finland, where the association became negative in the sample once blocks of predictors were included in successive models.

Nevertheless, the case of England and Japan is different as the association was consistently negative once controlling by for background and organisational variables. Though in England, the -24 points unconditionally associated to the explanatory variable decreased to -15 points in Model 1, the estimate was again statistically significant and its magnitude slightly increased to -17 once organisational variables were added in Model 2. In other words, mathematics content-focused TPD was associated in this country to -17 points in the overall mean score of students in the TIMSS assessment, while background and organisational variables were held constant. On the other hand, in Japan, the estimate was significant only in Model 2, therefore Japanese students with teachers attending mathematics content-focused TPD achieved 11 points less in the assessment, taking into account the same conditions. Nonetheless, it is worth recalling that these values represented a relatively small difference between students with and without teachers' attendance for mathematics-content focused TPD, as they featured only 17% and 11% of the international standard deviation, respectively.

At this point, it is worth highlighting that when estimates of the association between mathematics content-focused TPD and student achievement became statistically significant (e.g. Model 2 for Japan and England), the parameters were small and negative. These findings are relevant because they indicate that when there is evidence of a link between the key explanatory variable and the outcome, the direction of the conditional association is the opposite to the expected according to the literature.

The use of an international large-scale assessment such as TIMSS provides statistically powerful support for this result due to its large sample sizes. However,

it might fail to take into account the bias of teacher self-selection. In other words, the negative associations obtained in this analysis might be not related to mathematics content-focused TPD itself, but to the conditions that made teachers to participate in this type of TPD that are, in addition, potentially related to the level of achievement of their students. Harnessing the set of available variables for this secondary analysis, this potential limitation becomes productive as it provokes a serious challenge to know more about which type of teachers attend mathematics content-focused TPD in each of the four countries of interest.

For this purpose, tables 2.2 and 2.3 break down student and teacher characteristics, respectively. Table 2.2 presents for each country of interest the percentage of students whose teachers engaged (or not) in mathematics content-focused TPD in relation to students' gender, their number of books at home and the highest level of education of their parents. Numbers in bold indicate a significant association (95% statistical confidence) between the key explanatory variable and the student variables.

Table 2.2 Percentage of students whose teachers attended mathematics content-focused TPD or not, by student characteristics in Finland, Japan, US and England

	FIN			JPN			US			ENG		
	yes	no	tot	yes	no	tot	yes	no	tot	yes	no	tot
Student gender												
Female	50	48	48	50	49	49	52	50	51	50	47	49
Books at home												
One or less bookcases	53	60	60	68	72	69	69	64	67	70	63	67
Two or more bookcases	47	40	40	32	28	31	31	36	33	30	37	33
Parental education												
Some or No School	0	1	1	0	0	0	**3**	**2**	**2**	**4**	**2**	**3**
Secondary	38	41	41	37	34	36	**34**	**27**	**32**	**46**	**42**	**44**
Post-Secondary	62	58	58	63	66	64	**63**	**71**	**66**	**50**	**56**	**53**

Source: TIMSS 2011 database.

Notes: numbers in bold indicate $p < .05$ (Chi-squared tests)

According to these data, students from the US and England whose parents had less educational attainment were more likely to be taught by teachers that participated in mathematics content-focused TPD. As an illustration, among US students whose teachers did not engage in this type of TPD, 27% had parents that only attained secondary education and 71% had parents with higher qualifications. In contrast, teachers that attended mathematics content-focused TPD taught 7% more students whose parents had only secondary education (34%) and 8% less students whose parents attained post-secondary degrees (63%). Likewise, English teachers that took part in mathematics content-focused TPD taught 4% more students whose parents had only secondary education (46%) and 6% less students whose parents had higher educational attainment, when compared with their colleagues that did not attend this type of TPD. There were no significant difference in relation to the gender of students or the amount of books in their home. In addition, teachers in Finland and Japan that attended mathematics content-focused TPD seemed to teach similar proportions of students than their counterparts that did not attend this type of TPD according to these three student variables.

Table 2.3 displays for each country the percentage of teachers that attended (or not) mathematics content-focused TPD in relation to a number of their own characteristics. These variables include their gender, years of teaching experience, whether they were majored in mathematics, the extent to which they considered that

the amount of teaching loads were problematic in their schools and their level of job satisfaction. As in the previous table, figures in bold indicate a significant association (95% statistical confidence) between the key explanatory variable (participation in mathematics content-focused TPD) and the teacher variables.

Table 2.3 Percentage of teachers that attended mathematics content-focused TPD or not, by teacher characteristics in Finland, Japan, US and England

	FIN			**JPN**			**US**			**ENG**		
	Yes	no	tot	yes	no	tot	yes	no	tot	yes	no	tot
Teacher gender												
Female	71	49	51	24	26	25	73	64	70	48	57	52
Male	29	51	49	76	74	75	27	36	30	52	43	48
Years of experience												
0-10	22	37	36	42	28	37	50	48	50	55	56	55
11-20	23	27	26	20	16	19	27	23	27	30	18	25
>=21	55	36	38	38	56	44	23	24	23	15	26	20
Majored in mathematics												
Yes	**94**	**70**	**72**	**87**	**69**	**81**	47	38	45	80	71	76
No	**6**	**30**	**28**	**13**	**31**	**19**	53	62	55	20	29	24
Teaching hours												
Serious problem	0	1	1	18	24	20	3	7	4	1	7	4
Moderate problem	9	8	8	36	40	37	10	11	11	15	23	18
Minor problem	44	39	39	23	24	24	24	18	22	34	37	35
Not a problem	47	52	52	23	12	19	63	64	63	50	33	43
Teacher satisfaction												
Disagree a lot	3	1	2	1	4	2	3	9	5	4	9	6
Disagree a little	7	7	7	12	13	13	7	7	7	5	9	7
Agree a little	39	48	47	53	50	52	24	29	25	28	24	26
Agree a lot	51	44	44	34	33	33	66	55	63	63	58	61

Source: TIMSS 2011 database.

Notes: numbers in bold indicate $p < .05$ (Chi-squared tests)

In general, regarding teacher participation in mathematics content-focused TPD, consistent differences were observed in Finland and Japan in relation to the specialisation in this subject. For instance, Finnish teachers that took part in mathematics content-focused TPD included 24% more staff majored in mathematics (94%) when compared with their counterparts that did not attend this type of TPD (70%). Similarly, among Japanese teachers majored in mathematics, 12% more engaged in content-focused TPD (87%) when compared to their colleagues without

such specialisation. Interestingly, there were no association between this variable and the rest of teacher characteristics. In addition, no association was found in the US or England.

To recap, there was no evidence of a link between mathematics content-focused TPD and student achievement in the US sample, even when certain characteristics of teachers and students were taken into account. These results were shared with Finland, but not with England and Japan. In these latter two countries, though the association was still small, it showed that students taught by teachers who attended this type of TPD performed relatively less in mathematics Grade 8 TIMSS 2011, while teacher and student characteristics were considered. These findings are worth to be interpreted in the context of the conditions of participation in this type of TPD exposed in the previous two paragraphs (e.g. level of parental education and teacher specialisation in mathematics).

Finally, in order to evaluate whether estimates of association in Model 2 were statistically different across key countries of this study, Independent Samples T-Tests analysis were conducted. Table 3.3 below reports the p-values yielded by the test for each pair of countries.

Table 2.4 P-levels based on t-tests for independent samples under Model 2 for each country pairing on mathematics achievement

p-value	JPN	US	ENG
FIN	0.352	0.758	0.377
JPN		0.765	0.422
US			0.398

Source: TIMSS 2011 database

There were no statistically significant differences among any pair of countries between their conditional associations of mathematics content-focused TPD and student achievement. In this sense, once controlling for background and organisational variables, the specific contribution of the key explanatory variable to the outcome is not different among the key countries.

2.3.2. Additional analyses

In order to explore whether results yielded by the OLS strategy might differ when the key explanatory variable (and the rest of teacher characteristics) is measured at the school level, I re-analysed Model 2 by applying HLM to take account of clustering. In addition, in order to explore whether the association changes when other types of TPD foci are considered as key explanatory variables, I conducted every OLS model using pedagogy focused and curriculum focused TPD as main predictors. These two additional analysis are presented in the following subsections.

2.3.2.1. HLM results

This section presents the results of Model 2 across the key countries by using HLM. Table 2.5 provides information on the conditional association of mathematics content-focused TPD with student achievement, considering this key explanatory variable as a school-level predictor. The first three rows present Block 1 predictors as student-level variables in the model; Block 2 and 3 predictors are shown below them as school-level variables of teacher characteristics. School average of each of the teacher variables were calculated to be included in the model; thus the key explanatory variable (TPD Content), now is considered as the proportion of teachers attending this type of TPD in each school. Below the key explanatory variable, between-school and within-school unexplained variance, and the number of observations (students) and clusters (schools) are reported.

The findings of this set of analysis suggested a similar pattern for the conditional association in Model 2 compared to the results yielded previously by using the OLS approach. The relationship between mathematics content-focused TPD (now measured as a school-level predictor) was still weak in the US sample and could not be inferred to the US population as it was not statistically significant. These results were shared with Finland and England, but not with Japan. Similarly to the OLS results, Japanese students that attended schools where teachers participated in this type of TPD scored 11 points less in TIMSS 2011.

Table 2.5 HLM of Model 2 by key countries

	FIN			JPN			US			ENG		
	coef	SE		coef	SE		coef	SE		coef	SE	
Student-Level Variables:												
Student gender	-2.6	1.8		-5.1	2.2	**	-7.3	1.0	***	-7.1	1.6	***
Books	12.2	0.8	***	13.2	0.9	***	7.3	0.4	***	9.9	0.7	***
Parental education	12.6	1.1	***	20.1	1.4	***	2.1	0.5	***	5.9	1.1	***
School-Level Variables:												
Teacher gender	-3.3	4.4		-1.6	5.8		2.8	6.1		1.0	14.7	
Teaching experience	0.1	0.2		0.2	0.2		0.6	0.3	**	0.1	0.7	
Math majored	5.5	5.5		6.6	6.8		-2.8	5.4		15.7	16.9	
Teaching hours	-3.2	3.4		7.9	2.5	***	2.7	3.2		0.5	9.0	
Teacher shortage	2.5	2.8		4.1	3.1		2.0	3.2		0.4	6.5	
Teacher satisfaction	5.5	3.4		11.0	3.3	***	-0.7	3.6		13.1	8.6	
TPD Content	1.7	8.2		-11.1	5.5	**	-2.8	6.5		-22.1	15.3	
Between-school variance	0.20			0.23			0.53			0.63		
Within-school variance	0.57			0.73			0.48			0.48		
N (students)	4286			4593			10477			4030		
n (classes)	145			138			501			118		

Source: TIMSS 2011 database

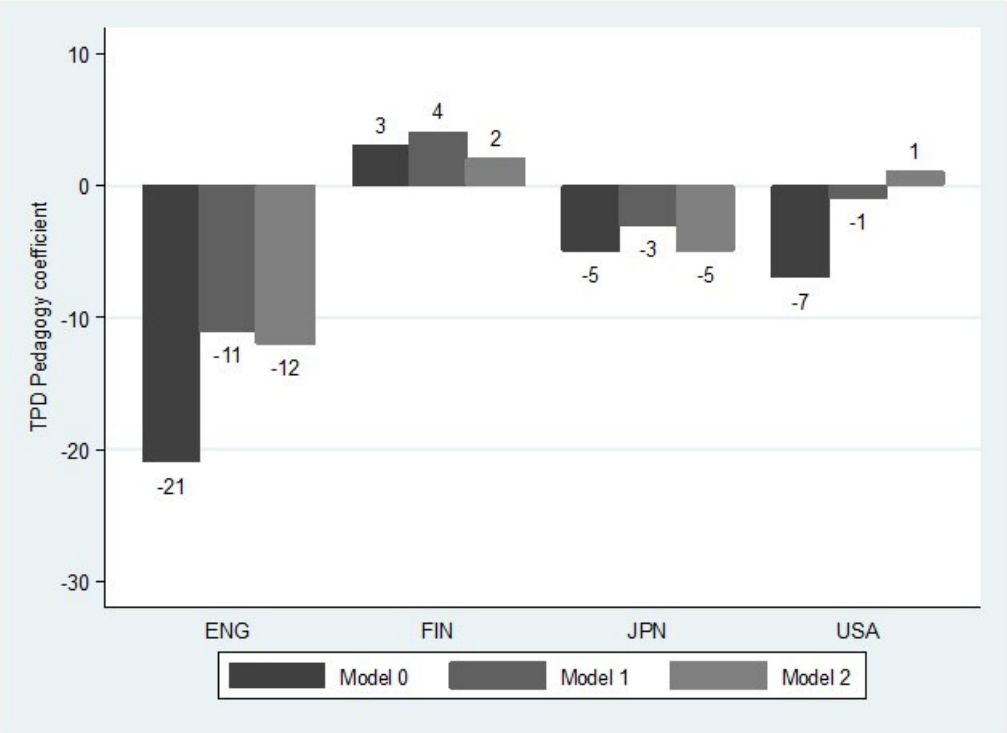
Notes: Outcome variable: Overall mean score Grade 8 mathematics TIMSS 2011; *p < .1. **p < .05. ***p < .01.

2.3.2.2. Pedagogy and Curriculum as key explanatory variables

This section presents results applying the same OLS modelling strategy, but using either the focus on pedagogy or the focus on curriculum as key explanatory variables. A detailed set of parameter estimates for the US, England, Japan and Finland are provided in appendices H and I.

As for Figure 2.3, Figure 2.4 illustrates how parameters of the unconditional association between pedagogy-focused TPD and student achievement (Model 0) change once blocks of control variables are added to the models. Bars indicate the magnitude, direction and statistical significance of the association and they are grouped by the key countries of the study.

Figure 2.4 Variation of the conditional association between student achievement and mathematics pedagogy-focused TPD across models for England, Finland, Japan and the US



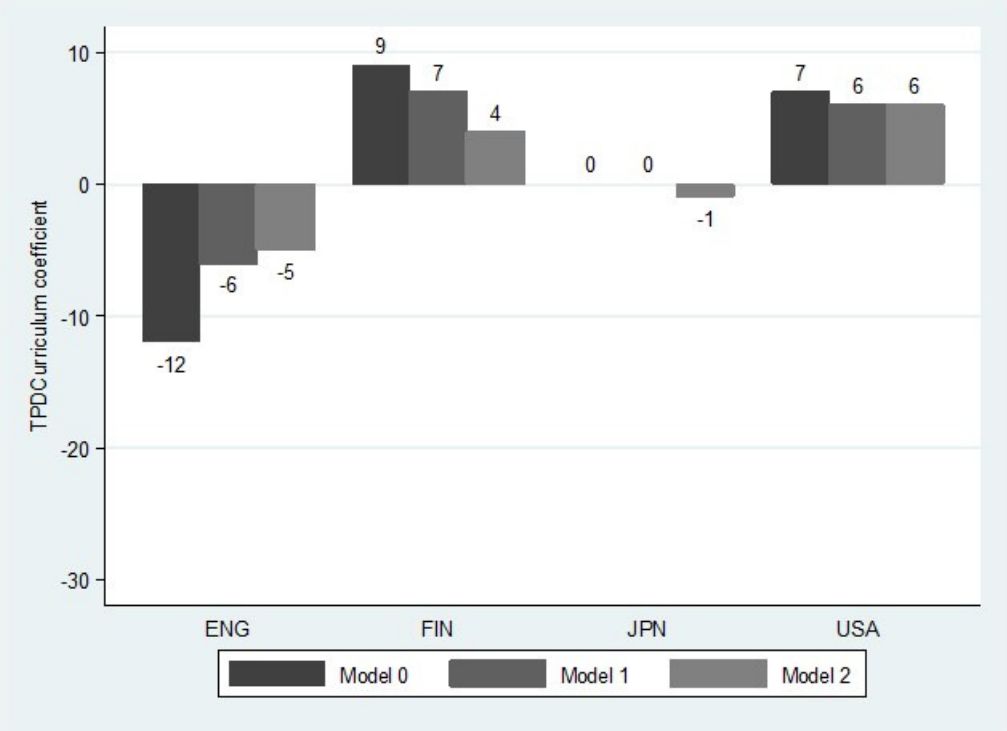
Source: TIMSS 2011 database

Notes: *p < .1. **p < .05. ***p < .01.

In general, the trend was similar to results analysed previously using mathematics content-focused TPD as key explanatory variable: estimates were small and regularly negative in the samples analysed. Once controlling variables in Model 2, the range of magnitude of estimates ranged between -12 points in England and 2 points in the US, which represented only 12% and 2% of the international standard deviation of TIMSS 2011, respectively. In addition, unlike previous findings, none of the estimates was statistically consistent, thus no evidence of association between pedagogy-focused TPD and student achievement can be supported in any country and model.

Figure 2.5 repeats the previous analysis, but using curriculum-focused TPD as key explanatory variable instead of content focused TPD.

Figure 2.5 Variation of the conditional association between student achievement and mathematics curriculum-focused TPD across models for England, Finland, Japan and the US



Source: TIMSS 2011 database

Notes: *p < .1. **p < .05. ***p < .01.

As in the case of mathematics content and pedagogy TPD foci, estimates associated to curriculum were also small, though in Finland and the US samples they were positively associated to student achievement. Once controlling for student and teacher variables in Model 2, the range of magnitude of estimates was between -5 points in England and 6 points in the US, which represented barely 5% and 6% of one standard deviation across the educational systems participating in the assessment, respectively. Moreover, none of these coefficients was statistically significant at any level of conventional confidence, thus no evidence of a relationship between curriculum-focused TPD and student achievement in mathematics was supported from this analysis.

To sum up, findings yielded by the set of analyses included in this chapter indicated little evidence of a link between mathematics content-focused TPD and student achievement in the pool of countries examined. Whereas the correlation between teacher participation in this type of TPD and student performance in mathematics was weak at the level of countries taking part in TIMSS 2011, estimates of the unconditional association in the OECD educational systems were also small and close to zero. Once controlling for student and teacher variables, coefficients remained small and among the key countries of the study, they were only consistent in the case of England and Japan; there was no evidence of association between the key explanatory variable and the outcome in the US sample.

Robustness tests used in this section supported the findings obtained through applying the OLS approach and using the focus on content knowledge as key explanatory variable. Therefore, these findings differed from the hypothesised positive direction of the association between mathematics content-focused TPD and student achievement and raise more questions than answers about the nature of the relationship between maths content TPD and student achievement.

2.4. Discussion and conclusion

The comparative analysis provided in this study has been aimed to estimate the statistical association between student achievement in mathematics and teacher participation in TPD activities that focus on mathematics content knowledge. Here one question was in focus: does this type of TPD relate to student achievement in our four selected countries? OLS regression was applied by introducing predictors in block sequences using data from the recent Trends in International Mathematics and Science Study (TIMSS) 2011 and a series of additional analyses confirmed the main findings.

The estimates derived from this cross-sectional study represent the actual implementation of mathematics content-focused TPD, as well as its contribution to student performance in an international assessment of mathematics achievement. They inform the relative prevalence of this feature among mathematics teachers in every country and they allow a valid analysis of its association with student outcomes using a comparable measure of achievement. In contrast, most of the evidence supporting a positive effect of the key explanatory variable has been produced in the context of Randomised Controlled Trials evaluating specific features of TPD programmes in the US. As I remarked above, these studies usually involved a limited number of participants, in no case they represented the population at the national level, and they regularly used their own measures of student outcomes. It must be acknowledged that due to random allocation of teachers to TPD interventions, they are closer to infer a causal relationship between mathematics content-focused TPD and the outcome variable. However, evidence provided in this study closely matched those obtained by Telese (2012) with observational data from the NAEP 2005: they show a non-existent even weakly negative association for this type of TPD experiences and student achievement.

If findings yielded by this study were to be trusted, the favourable effect of the key explanatory variable on student achievement is questioned through the analysis of observational data. Results yielded in this study refuted in two ways what specialised literature has been indicating about the relationship between mathematics content-focused TPD and student achievement. Firstly, they contradicted the expected positive association that has been regularly reported in the US (Blank and de las Alas, 2009; Kennedy, 1998; Salinas, 2010; Scher and O'Reilly, 2009), as the analysis showed no consistent link for this country in any model or robustness test. TIMSS 2011 data showed that in this country this

type of TPD is rather neutral to student achievement in mathematics. Secondly, it showed that in countries where there was empirical evidence of this relationship, this was rather small and negative, as it was in the case of England and Japan (when background and organisational variables were controlled).

Regarding the first of these points, it is worth considering the high level of participation in mathematics content-focused TPD which is normally reported in the US (IEA, 2012). Policies have promoted the idea that TPD programmes that make a difference in students' outcomes concentrate particularly in delivering knowledge about the subject area, with most of the research in this country supporting this view with evidence specificly produced from analysis of mathematics' achievement. This option of the US system can be well understood considering the serious shortage of high-quality teachers in this subject, which stems from the high attrition in this area and the comparatively low requirements to obtain teaching qualifications. In other words, the US system seems to address the shortage of good teachers in mathematics via delivering such knowledge throughout the teaching career. However, the findings from this chapter show that students taught by teachers that took part in mathematics content-focused TPD achieved equivalent scores in TIMSS 2011 than their conterparts taught by teachers that engaged in TPD focused in any other topic. In this sense, this type of TPD is neutral to this outcome.

I recognise that two caveats well deserve consideration for this point. On the one hand, discrepancy with previous research might be due to the lack in this study of the desired alignment between TPD experiences and outcome measures of achievement (Wayne et al., 2008). Blank and de las Alas (2009) commented that studies using student level measures that were more tuned to capture the areas that were expected to be improved by TPD, were those more likely to report larger effect sizes. Standardised assessments of achievement seem to be less sensitive to trace the effect of the key explanatory variable on student performance, hence an international test, such as TIMSS, would be even less useful for this purpose.

On the other hand, it can be argued that my analysis does not encompasses other variables that can be relevant in order to evaluate the contribution of mathematics content-focused TPD in the US. There could be still omitted variables that are actually associated either to student achievement or to the participation of teachers in this type of TPD that are not included in the models presented (see subsection 1.3). Taking into account the literature

accounting for the effect of TPD on teaching practices, I would mention in particular variables such as the rest of quality features of TPD (coherence, collective participation, duration and active learning) (Desimone, 2009; Garet et al., 2001). As these indicators have been successfully tested in large-scale designs in the US (Birman *et al.*, 2000; Desimone *et al.*, 2002; Garet *et al.*, 2001) and England (Opfer and Pedder, 2011b) it is thinkable that they might have some influence on the association between the focus of TPD and the outcome variable[19].

Nonetheless, one of the advantages of a cross-national design is that it allows contrasting this kind of considerations with data collected from different contexts, and, indeed, the second point of this discussion about the findings reported for England and Japan disputes the validity of the two aforementioned caveats. In fact, the measures of student achievement provided by TIMSS are useful in order to analyse the statistical association with mathematics content-focused TPD. Further, by utilising variables of the assessment that are available in the database it is possible to control for some characteristics of TPD national systems in order to estimate the degree of association between the key explanatory variable and student achievement in mathematics. These two aspects were solidly used in my OLS models, either in the case of England as in Japan.

The thing is that in these two educational systems the relationship was rather small and –against what common sense would indicate- negative. To be more precise, English and Japanese 8th grade students would obtain respectively 17 and 11 points less in the mathematics assessment of TIMSS when they are taught by teachers attending mathematics content-focused TPD. This is relatively a small size association, taking into account that one standard deviation in the test is 100 points across all the participant countries. However, this small figure is not produced by sampling variation, thus it is worth to provide further explanations for the unexpected direction of the outcome.

In the case of England, it is worth mentioning that teachers' shortage in mathematics and the comparatively low requirements to become teacher are -as in the US- also serious

[19] Unfortunately, a limitation of this secondary analysis is that it was not possible to control for this set of predictors as they were not included in the teacher questionnaire of TIMSS 2011. It would be useful that TIMSS could include this group of questions in the next rounds of the assessment. The Teaching and Learning International Survey (TALIS) 2013 conducted by the OECD has already included questions related to these features.

policy issues that may limit the contribution of mathematics content-focused TPD to students' achievement. It may be that mathematics teachers in this country need greater efforts and time (e.g. more than one year) to internalise new concepts and translate them into effective classroom practices. In other words, mathematics content-focused TPD in England would be counterproductive to improving students' outcomes because such focus may take longer time to be learnt and, consequently, delivered to students. Hence, while English teachers participate in mathematics content-focused TPD (even months after the programme has finished) they are still struggling to teach recently acquiered concepts, which would affect their mastery of the subject. This would explain the relatively lower test scores of their students in TIMSS 2011.

The case of Japan is different because this system selects the best candidates to fill the teaching positions and in-service teachers have strong numeracy skills (Hanushek, Piopiunik and Wiederhold, 2014). However, one alternative way of viewing the negative association of mathematics content-focused TPD with student outcomes is considering the high overloading experienced by teachers in this country. As commented in Appendix B, Japanese teachers work approximately ten hours per day (53 hours per week in average), with 19 hours of the week schedule occupied in tasks that are not fully related to their teaching duties. At this juncture, engaging in TPD activities may be experienced as a poorly attractive extra task, which in practice requires trading-off hours of teaching related work in favour of participating in such compulsory events. In addition, there is a high participation in mathematics content-focused TPD in Japan (67%, see page 35), which may reflect their involvement in the traditional Japanese form of TPD, i.e. lesson study (Lewis, 2009). This type of TPD is very demanding as it requires time to collaboratively plan, conduct, and evaluate a specially designed lesson focused in a particular mathematics content (a loop that can be repeated more than once considering the feedback received from colleagues). It is possible that engaging in such type of TPD may hinder their performance with the rest of their classes in a context of high overloading, which would reduce the quality of their teaching and the opportunities to learn of their students.

All in all, the possibility of reverse causality cannot be dismissed for these two countries (England and Japan): instead of a negative effect of the key explanatory variable on student achievement, it could rather be that teachers attending this type of TPD are those teaching in schools with low achiever students. In other words, self-selection of teachers to

content-focused TPD might work as a bias of estimates, so it is worth to manage some information of the conditions that make teachers to participate in this type of TPD.

That is why my analysis introduced further descriptive information about what type of teachers attend mathematics content-focused TPD in each of the four countries of interest. However, this information is insufficient to solve the problem of weak or negative association. Unfortunately, in this point the analysis confront a relevant limitation derived from the cross-sectional feature of TIMSS assessment. As Goldstein (2008) claimed, it is necessary that designs of international large-scale assessments such as TIMSS introduce a longitudinal component of prior achievement in order to fully contrast the variation which is attributable –in this case- to the participation of teachers in mathematics content-focused TPD. On the other hand, we need to know so much more about how teachers apply the knowledge acquired in these activities and their motivation to select into this type of TPD, as well as the varying readiness of their students to engage in learning activities delivered by this kind of teachers.

What the analysis does is to demonstrate that with the available data the current implementation of mathematics content-focused TPD cannot be accepted as positive investment or panacea for raising student achievement in mathematics. At least as it is measured in an international large-scale assessment like TIMSS 2011. Contrary to what specialised literature has been indicating, the participation of teachers in this kind of experiences is rather neutral to student performance in the US. Furthermore, by putting the analysis in a cross-national context, the relationship is rather small and negative, as it was exposed for the cases of England and Japan (once student and teacher variables are controlled).

The following chapter examines in the four countries of interest the statistical association between student achievement in mathematics and a second quality feature of TPD, i.e. *coherence*, using data from PISA 2012.

Chapter 3

Coherence of teachers' professional development and student achievement: a cross-national analysis of PISA 2012

3.1. Introduction

In this chapter, I set out to provide an empirical examination of the concept of coherence in TPD and its link with student achievement, drawing data from the 2012 round of the Programme for International Student Assessment (PISA). Coherence is defined as the extent to which TPD activities are actively managed to be consistent with the overall goals of schools, in particular with those related to students' learning[20]. Thereby, the concept is conceived in the context of the improvement of the capacity of teachers to enhance school achievement (Newmann, King and Youngs, 2000) and, therefore, assumed as a logical step for effective TPD (Darling-Hammond et al., 2009; Villegas-Reimers, 2003). In particular, it represents one of the key components of the quality of such activities (Desimone, 2009) and,

[20] The term is not used here in the conventional sense of TPD programmes perceived as internally well-structured, with adequate consistency between its focus, duration and the types of activities included (Firestone *et al.*, 2005); nor in terms of correct alignment with external educational policies and standards (DeMonte, 2013; Fuhrman, 1993; Hochberg and Desimone, 2010; O'Day and Smith, 1993).

as such, it has been found to be associated to better school outcomes at the national level in the US (Garet et al., 2001). The concept of coherence in TPD is usually measured using teachers' perception about recent experiences of in-service training (Desimone *et al.*, 2002; Newmann *et al.*, 2001a; Penuel *et al.*, 2007), however a number of studies have also described the concept in terms of how districts become able to organise a coherent system of teacher learning activities for schools' improvement (Borko, Elliott and Uchiyama, 2002; Elmore and Burney, 1997; Firestone et al., 2005).

The study of the coherence in TPD is important both for teachers and policy makers as it represents "the logical chain" by which the learning needs of the staff connect with the educational goals of the system (Ofsted, 2006). Indeed, the design of any strategy orientated to improve schools through TPD should be based on information that emerges from the schools themselves and, consequently, states clear learning objectives and methods that determine what must be accomplished in every TPD activity. If such aims are neither sufficiently supported by what is actually required to ameliorate each school, nor adequately specified in the programmes of TPD, then the benefits of reforms would be less probable to be observed (Borko, 2004). Having coherent TPD is also relevant for teachers because a consistent approach to such activities is likely to fulfil their demand for in-service training, a process which in turn contribute to encourage retention in the system. Most importantly, the quality of teachers' performance could be boosted by this feature, given that coherent TPD is also likely to develop teachers' skills according to what is specifically required in the school context in which they work. As a consequence, students' achievement might benefit as teachers become more skilled to synthesise their acquired knowledge according to the particular characteristics of their students.

Despite its importance, it is striking that very little is known about the process through which school goals and teacher learning might converge into TPD activities, which largely resides under the realm of head-teachers' management (Youngs and King, 2002). In this regard, Sebastian and Allensworth (2012) report that the influence of the leadership of head-teachers on teaching practices is importantly explained by the extent to which TPD is perceived by teachers as a coherent practice within schools. This aspect suggests that the role of head-teachers becomes particularly significant to teachers' work insofar as they put efforts in making of TPD a strategy in line with the goals of the school. Certainly, specific tasks undertaken daily by leaders determine this feature of the quality of TPD, so even the

most carefully designed strategy depends on that role. However, no study has directly examined self-reports from head-teachers about what they actually do to affect the coherence of TPD.

Furthermore, it is worth noting that this construct has been so far only studied in the US context, so a question remains in relation to its portability as a measure used to assess the quality of TPD in other countries. As national contrasts in the organisation of schools and the provision of TPD might affect this dimension, further research is necessary to establish how adequate using the coherence of TPD is for cross-national comparisons in this area. For instance, and unlike the US, results about schools' performance based on individual examinations of their students do not exist in all countries (e.g. Japan and Finland) and where they do, they are not always available to permit opportune decisions about the coherence of TPD activities (e.g. UK/England). This is relevant because school goals aimed to improve student learning require a precise knowledge of the subjects and curriculum areas that need to be supported on the basis of results from the assessment of students' skills. Therefore, it is quite possible that the extent of coherence in TPD within schools will vary across nations.

At this juncture, data gathered in PISA 2012 provide an interesting opportunity to examine this construct as a number of actions implemented by head-teachers aimed to make of TPD a coherent practice within their schools have been collected from 65 countries. Given that these data have not yet been analysed in detail nor in relation to the influence of TPD on national educational outcomes, relevant insights about the effectiveness of countries in developing coherent TPD could be obtained by means of describing the extent to which such actions are undertaken.

In this context, I examine the degree of coherence evident in TPD across the US, UK/England, Japan and Finland. The study of this construct has received special attention in the US, which is probably due to the critical shortage of high-quality teachers in key subject areas (see Appendix B) and the consequential concern on the efficient utilisation of the staff. Policy and research in this country usually promote that the leadership style of their head-teachers' should focus on improving students' learning and supporting teachers in instructional improvement, hence TPD must necessarily be a coherent practice within US schools. The situation is similar in England in terms of shortage of high-quality staff, whereby policy has explicitely indentified the coherence of the TPD as a "best practice" (Ofsted, 2006) that effectively raise learning standards.

By contrast, TPD plays a different role in Japan and Finland because these systems put greater efforts in selecting the best candidates for the teaching positions in a context of surplus of applicants. Hanushek, Piopiunik and Wiederhold (2014) have recently described the outstanding cognitive skills of teachers in these two countries, thus it is very likely that all their students have access to high-quality teaching and the coherence of TPD is not a major concern. However, TPD is also compulsory for Japanese and Finnish teachers (as in the US and England), hence it becomes relevant to examine what is sufficient for their head-teachers in order to make TPD a coherent practice within schools.

Considering these national differences, the aim of this chapter decomposes into three specifics questions: firstly, do the variables comprising a measuring instrument of the construct in PISA 2012 operate equivalently across these countries? Secondly, what is the performance of each nation in this dimension? A third research question examined in this chapter is whether the willingness of head-teachers to promote coherent TPD becomes an effective mechanism to improve school outcomes. More specifically: does a coherent approach to TPD in schools relate to student achievement? This specific question is important to answer the main question of the thesis, concerning whether variations in teachers' exposure to this feature of TPD might be associated with differences in student achievement within countries[21].

The chapter mainly aims to detect in each of the four countries of interest the presence of an unidimensional latent construct (Kline, 1994) related to the concept of coherence in TPD using a number of items included in the head-teacher's questionnaire contained in PISA 2012. This analysis is carried out following the application of Exploratory Factor Analysis (EFA) (Asparouhov and Muthén, 2009; Baglin, 2014; Browne, 2001) and Multiple-Group Confirmatory Factor Analysis (MG-CFA) (Brown, 2006; Byrne, 2012; Desa, 2014; Jöreskog, 1969), which is appropriate to evaluate the invariance of this measurement model across the four countries of interest. To examine whether the positive association between coherent TPD and school outcomes found in the US (Garet et al., 2001) is replicated with current data, and whether such results are also observed in the UK, Japan and Finland, a Hierarchical Linear Modelling (HLM) (Raudenbush and Bryk, 2002; Snijders and Bosker, 1999) analysis is applied. Any causal interpretation derived from results

[21] See page 23.

reported in this analysis must be taken with caution as attributes of schools that are actually related either to student achievement or to the coherence of TPD activities may not be covered or available in the existing data. Otherwise, the chapter provides statistical evidence to compare the role of this key component of the quality of TPD in all the four countries selected.

In the next section I describe the methodological strategy implemented to address the research questions above presented, as well as relevant features of the dataset analysed, i.e. PISA 2012. Section 3.3 provides properties of the measurement model that I suggest to examine the key explanatory variable of this chapter, as well as estimates of association with student achievement. This is followed by a discussion of findings and conclusions in section 3.4.

3.2. Data sources and methodological strategy

3.2.1. Survey design

In this chapter, I use data drawn from the 2012 cycle of PISA, an international large-scale assessment conducted every three years by the Organization for Economic Cooperation and Development (OECD). The target population of the assessment are 15-year-old students (e.g. between 15 years 3 months and 16 years 2 months at the time of testing) with a minimum of six years of schooling. In order to gather representative data from this population, PISA implements a two-stage stratified cluster sampling procedure: in the first stage, each country is expected to randomly select at least 150 schools, with probability proportional to their size; in the second stage, 35 students are randomly sampled with equal probability within each sampled school. In PISA 2012, 65 nations took part of the assessment, which is the second study focused on mathematics literacy, after the 2003 round (OECD, 2014b).

The ideal situation is when in a country 100% of the originally sampled schools and their students take part of the assessment. However, this is accomplished in every round of PISA only by a small number of nations, thus the organisers set minimum standards of participation for the sampled schools (85%) and students within them (80%) in order to

preserve the desired representation of national target populations. Countries where schools' response rates are below the standard are allowed to improve this number by substituting with units that were not originally selected from the national list of schools. In PISA 2012, countries' response rates were largely accomplished after replacement (in average, 98% of schools and 92% of students), with the only exception of the US (77% of schools). In order to correct for the unit non-response at the school and student level, sampling weights are calculated in each cycle by the organisers, thus a specific weighting factor (Rust, 2013) informing the probability of selection and adjustments for nonparticipation is assigned to each school and student. Therefore, by applying these inverse probability weights to the indicators of interest, estimates are adjusted to be representative of the national target populations (OECD, 2012b).

Furthermore, PISA 2012 provides an additional weighting factor based on a proportional transformation of the design and response weights that rescales the sample size to be fixed to 1,000 cases at the country level. The use of this weighting factor is recommended in the instance that data from several school systems are simultaneously analysed as pooled datasets and/or when the different sizes of national samples might lead to overestimation of results (Stapleton, 2013). For instance, this occurs in this chapter when combined datasets are used to estimate indices of measurement invariance across the countries of interest. Thereby, appropriate weighting factors are employed in every analysis throughout the chapter.

The details of the achieved sample sizes used in the analyses and their corresponding weighted values are provided in Table 3.1 for the pooled dataset and each country. For instance, the first row (No weight) details the actual number of schools included in the combined dataset (1,060) and the contribution of each country to this number. The second row (W_FSCHWT, this is the actual name of the variable in the dataset) presents the school level adjustment for non response, which corresponds to the number of schools represented in each country and in the pooled dataset (43,400). The third row indicates the additional weighting factor (SENWGT_SCQ, also called senate weight) that rescales the countries sample size to approximately 1,000 cases each to facilitate cross-country comparisons. The following three rows present similar information but for the students included in the sample.

Table 3.1 Sample sizes from PISA 2012 data for countries of interest using two different types of sampling weights

	Weighting factor in PISA 2012 dataset	Key countries combined (pooled dataset)	US	UK[f]	JPN	FIN
Schools	No weight[a]	1,060	162	396	191	311
	W_FSCHWT[b]	43,399.5	31,091.1	4,410	7,041.4	857
	SENWGT_SCQ[c]	3,915.3	1,000	915.3	1,000	1,000
Students	No weight[a]	29,872	4,978	9,714	6,351	8,829
	W_FSTUWT[d]	5,361,348	3,538,783	634,338	1,128,179	60,047
	senwgt_STU[e]	3,922	1,000	922	1,000	1,000

Source: PISA 2012 database

Notes: [a] Actual number of observations in the dataset; [b] Grade nonresponse adjusted school base weight; [c] Schools' senate weight - sum of weight within the country is 1000; [d] Grade nonresponse adjusted student base weight (total weight); [e] Students' senate weight - sum of weight within the country is 1000; [f] Only England, Wales and Northern Ireland were included (Scotland was excluded).

3.2.2. Student achievement

The outcome variable "student achievement" is defined in this study as the overall score in the mathematics scale as measured in PISA 2012. The measure is transformed to have a mean of 500 points and a standard deviation of 100 across the OECD participant countries. The data on "student achievement" is complex given test features that derive from the PISA 2012 framework for assessing literacy in mathematics. Extensive categories related to contexts, contents and processes included in this subject are expected to be covered throughout the assessment. Thus in order to yield precise results of domain proficiencies at the country level, PISA follows a multiple-matrix sampling strategy based on Item Response Theory (Embretson and Reise, 2000; OECD, 2014b). A considerable number of items are elaborated to validly assess several content categories and mathematics literacy as a composite of all of them. Unlike a conventional test, not all questions are delivered to each student as responding the whole instrument would exceed the two hours defined for the administration of the test. Questions are randomly assigned to students to generate estimates of student achievement and these scores are combined as in other international large-scale assessments (e.g. TIMSS) to produce the final score for mathematics.

Estimations of proficiency for each student are actually unknown, though they can be calculated from a hypothetical distribution based on responses to the assigned set of items. In order to report the performance that a student might reasonably have in the overall score of mathematics, PISA employs a fixed number of five plausible values as random draws from this distribution. Rutkowski et al. (2010) mention the flaws of statistical analyses in which these plausible values are insufficiently used for estimates of student achievement. One example is the calculation of national estimates using the mean of plausible values as a single parameter of achievement for students. In this instance, calculating the average of the five plausible values for each student and then the mean of every student score to obtain the national estimation, might dramatically underestimate standard errors. In this study I follow recommended practise regarding how estimates are combined to calculate their corresponding standard errors. For instance, each HLM model involves running five regressions (one on each plausible

value) and then the calculation of the average coefficients and their errors (Macdonald, 2014)[22].

3.2.3. Key explanatory variable

The "coherence of TPD within schools" is operationalised by asking head-teachers to rate a group of items included in the PISA school questionnaire (PISA Consortium, 2011). This instrument is administered to every head-teacher of the sample of schools and contains questions about the organisation of the school and the learning environment. Among the questions orientated to gather data on climate, policies and practices, head-teachers are requested to indicate the presence of specific management strategies, in particular about their performance in relation to particular courses of action that might determine the level of coherence of TPD within the school. These are the five items that were administered:

- Item 1. A standardised policy for mathematics (i.e. school curriculum with shared instructional materials accompanied by staff development and training) is implemented for quality assurance and school improvement.[23]
- Item 2. Extent to which appraisals of and/or feedback to teachers have directly led to opportunities for TPD.[24]
- Item 3. Frequency head-teacher made sure that TPD activities were in accordance with the teaching goals of the school during the last year.
- Item 4. Frequency head-teacher led or attended in-service activities concerned with instruction during the last year.

[22] Unlike the analysis developed in Chapter 2, this chapter uses the five plausible values available in the PISA 2012 dataset in order to produce more precise standard errors.

[23] This item was recoded for scoring from its original values in the PISA 2012 dataset (1=yes/2=no) to 1=yes/0=no.

[24] 4-point Likert-type scale, consisting of the following: 1=No change, 2=A small change; 3=A moderate change; 4=A large change.

- Item 5. Frequency head-teacher set aside time at faculty meetings for teachers to share ideas or information from in-service activities during the last year.[25]

Table 3.2 describes how this group of items compares to those used in the literature to capture the coherence of TPD. The first column details the statements used in the school questionnaire of PISA 2012, whereas the second and third columns display the corresponding items used by Sebastian and Allensworth (2012) and Murray (2012), respectively.

[25] Items 3 to 5 use a 6-point Likert-type scale, consisting of the following: 1=Did not occur, 2=1-2 times during the year, 3=3-4 times during the year, 4=Once a month, 5=Once a week, 6=More than once a week.

Table 3.2 Comparison with other measures of coherence of TPD used in the literature

PISA 2012	Sebastian and Allensworth (2012)	Murray (2012)
Item 1. A standardised policy for mathematics (i.e. school curriculum with shared instructional materials accompanied by staff development and training) is implemented for quality assurance and school improvement.	Curriculum, instruction, and learning materials are well coordinated across the different grade levels at this school. There is consistency in curriculum, instruction, and learning materials among teachers in the same grade level at this school.	Professional development activities are aligned with the school curriculum.
Item 2. Extent to which appraisals of and/or feedback to teachers have directly led to opportunities for TPD.	Teachers are left completely on their own to seek out professional development	Specific teacher needs inform the design of our professional development activities.
Item 3. Frequency head-teacher made sure that TPD activities were in accordance with the teaching goals of the school during the last year.	Overall, my professional development experiences this year have been closely connected to my school's improvement plan.	Professional development activities relate directly to our institutional goals.
Item 4. Frequency head-teacher led or attended in-service activities concerned with instruction during the last year.	Once we start a new program, we follow up to make sure that it's working.	Teacher professional development is part of our school improvement plan.
Item 5. Frequency head-teacher set aside time at faculty meetings for teachers to share ideas or information from in-service activities during the last year.	We have so many different programs in this school that I can't keep track of them all. Many special programs come and go at this school. You can see real continuity from one program to another at this school.	Our personnel conduct our professional development activities. We involve teachers in designing the activities of our professional development program.

Sources: Murray (2012); PISA Consortium (2011); Sebastian and Allensworth (2012)

The first item refers to the concept of instructional coherence (Newmann *et al.*, 2001b), which is deemed as a favourable school condition for the coherence of TPD because it connects in-service training with other relevant resources used by teachers in the classroom. The measures of coherence in TPD developed by Sebastian and Allensworth (2012) and Murray (2012) also include this aspect, although here the standardisation feature assures that decisions about this link are consistent for all the teachers of mathematics. In addition, they are based on what the school needs to accomplish its learning goals in this subject.

The other four variables are directly related to the role of head-teachers and describe the extent to which leaders undertake specific tasks that reinforce the coherence of TPD activities within schools, either when they are presented to teachers or during their implementation. For instance, item 2 provides information on the degree to which teacher evaluation strategies includes opportunities for TPD as expected course of improvement. This variable provides information on the consistency between teacher learning activities and the needs of knowledge and skills required by the staff. Such aspect is also included in the aforementioned instruments, however the individual appraisal of the needs of TPD is highlighted here with the key role of head-teachers in this process of assessment and feedback. The rest of indicators illustrate the regularity of the supervision over TPD activities carried out in situ by head-teachers, in particular the frequency that they attend such events and check whether school goals are being accomplished. They are also mentioned by Sebastian and Allensworth (2012) as actions that support following-up new TPD programmes in the school, and by Murray (2012) as to the alignment with school improvement plan and the participation of teachers in the design of TPD.

These five observed indicators are not exempt of measurement error as translation issues and the self-reported nature of the data undoubtedly affect the validity of the information collected. Although standardised guidelines and double translation of items have been implemented by PISA organisers (PISA Consortium, 2010), the question remains as to whether specific and complex concepts originally conceived in one language can be fully interpreted in dissimilar cultures (Zhang, 2011). In this regard, it must be acknowledged that the term *coherence* is selected from a particular context (here the US), thus its original meaning or some of its attributes are constrained to the potential understanding of participants from different countries. This becomes clear when one

examines specific ideas mentioned in the questions (e.g. "standardised policy" and "appraisals and/or feedback to teachers") which are assumed as existing and signifying the same content in school systems with dissimilar cultural backgrounds and organisational characteristics. In the previous round of PISA (OECD, 2012b), important sources of measurement error derived from translation procedures were reported for versions of the test written in non-Indo-European languages (e.g. Middle-East and Asian countries). Although similar flaws were not observed in the current round of the assessment (OECD, 2014c), the official report does not provide specific information about the items included in the school questionnaire, thus it may well be the case that such measures lack to some extent of cross-cultural comparability.

Furthermore, it is possible that the self-report procedure that is utilised in the administration of the instrument might become an important source of measurement error for this group of items. Drawbacks such as uncertainty in the head-teachers' understanding when attempting to decipher long sentences including several objects, as well as memory imprecision for those question requesting information from past events, might yield inaccurate data on every item. Given that the precision of the information collected through this instrument cannot be verified with actual data or other variables, it is worthwhile inquiring about the extent to which such potential limitations actually affect the quality of the data collected. In this regard, Desimone (2009) remarks that teacher surveys in the field of TPD that elicit factual data (instead of evaluative data) show adequate estimates of reliability and validity. In other words, as long as the information requested about the coherence of TPD is descriptive (about facts) and not based on personal judgements about facts, the use of self-reported data is well supported. In this sense, the quality of these data could be judged as satisfactory as it seems evident the implementation of such principle in the item content and the structure of the questions posed.

Taking all these aspects together, it seems acceptable that the more that these five items are implemented in the school context, the greater the coherence of TPD. In other words, it is possible to argue that the consistency between TPD and curriculum materials, the extent to which teachers' appraisals lead to individualised TPD and/or the degree to which such activities are closely overseen, are aspects to a great extent determined by the coherence of TPD pursued by head-teachers. From a measurement perspective these five items are all potentially tapping the same latent construct, which is suggested in this

chapter as the extent to which TPD activities are actively managed by school leaders to be consistent with the improvement goals of schools. If this is the case, then the inter-tem correlations should yield positive and noticeable associations, in which case the indicators would clearly measure the underlying factor of coherence in TPD in the same direction. In CFA the association between the items is purported to be due to their common dependence on a single underlying factor or latent variable (Brown, 2006). As this chapter aims to identify the existence of this unobserved explanatory variable in each of the four countries of interest, it follows that such results should be fairly similar across the comparator nations -unless that cultural specificities are playing a role in this regard. In order to explore this aspect, Table 3.3 displays the means, standard deviations, inter-item correlation and bivariate correlations between these five items in every school system of interest.

Table 3.3 Descriptives and correlation matrices under continuous assumption

	Item	Mean	SD	Item 1	Item 2	Item 3	Item 4
US	**1**	0.79	0.41	1			
n=153	**2**	2.37	0.84	0.28	1		
$\|r_{ij}\|$=**0.31**	**3**	3.72	1.42	**0.34**	**0.30**	1	
	4	3.42	1.17	0.17	**0.34**	**0.47**	1
	5	3.86	1.21	0.21	0.13	**0.40**	**0.49**
UK	**1**	0.65	0.48	1			
n=353	**2**	2.95	0.64	0.21	1		
$\|r_{ij}\|$=**0.30**	**3**	4.14	1.41	0.06	0.24	1	
	4	3.37	1.13	0.09	0.20	**0.62**	1
	5	3.52	1.18	0.09	0.28	**0.60**	**0.66**
JPN	**1**	0.36	0.48	1			
n=190	**2**	1.97	0.77	0.07	1		
$\|r_{ij}\|$=0.17	**3**	2.48	0.86	0.09	0.00	1	
	4	2.65	0.78	0.11	0.07	**0.32**	1
	5	2.61	0.88	0.09	0.18	**0.32**	**0.42**
FIN	**1**	0.47	0.50	1			
n=289	**2**	2.07	0.84	0.03	1		
$\|r_{ij}\|$=0.16	**3**	2.74	1.09	0.21	0.15	1	
	4	3.12	0.83	0.03	0.15	**0.33**	1
	5	2.88	0.98	-0.02	0.16	0.27	0.28

Source: PISA 2012 database

Notes: weighted data; $|r_{ij}|$=average inter-item Pearson correlation (absolute value); SD=standard deviation; bold values indicate correlations over 0.3.

According to the data, there is important variation in the average values of item scores across these four countries. For example, the first item shows that the implementation of a standardised policy for mathematics orientated to quality assurance and school improvement is highly reported in the US and UK (79% and 65%), whereas this occurs in less than a half of schools in Finland (47%) and approximately one third of schools in Japan (36%). A similar pattern of results can be described for the other four items, with head-teachers from the US and UK generally reporting higher rates than Finnish and Japanese leaders, which suggests important contrasts between English speaking countries and the other two comparator nations in terms of the prevalence of the coherence of TPD in schools.

Likewise, it is worth noting that although the absolute values of the average inter-item correlations were not particularly strong in each country, they can be considered fair in the US (.31) and UK (.3), and poor in Japan (.17) and Finland (.16), according to the guidelines suggested by Chan (2003) and Dancey and Reidy (2014). In particular, out of the ten bivariate correlations presented in each matrix, the US sample yielded six with fair values (between .3 and .5) and the UK showed three with strong values (over .6), whereas Japan and Finland showed only three and one fair coefficients, respectively (these are highlighted in bold). Detailed examination revealed that this level of association was mainly observed in the US, UK and Japan among the items related to the in situ supervision of TPD activities carried out by head-teachers (items 3, 4 and 5). In particular, the frequency that US leaders check whether school goals are being accomplished (items 3) and attend TPD events (item 4) also show appreciable links with the presence of a standardised policy for mathematics (item 1) and their willingness to give feedback to teachers that include opportunities for TPD (item 2). The case of Finland was especial because it showed only one non-weak bivariate correlation –between the items 3 and 4 (head-teachers monitor school goals and attend TPD). It is worth mentioning, however, that the correlations among these items and item 5 (opportunities to discuss with teachers about their recent experiences of TPD) were closer to the value that is deemed as a fair relationship (.27 and .28, respectively).

In the light of these results, we are able to refine our initial supposition that in every country the underlying construct of coherence in TPD evenly drives the observed association amongst the five observed indicators. National particularities are evident in

this regard across the selected samples, thus a suitable and flexible analytic strategy is required to pursue the aim of this chapter.

3.2.4. Analytic strategy

I use a combination of factor analysis techniques to examine whether the selected group of items of the head-teacher questionnaire can be deemed across countries as adequate indicators of a unidimensional latent variable defined as the coherence of TPD. Three approaches are executed consecutively for this purpose: Exploratory Factor Analysis (EFA), Confirmatory Factor Analysis (CFA) and Multiple-Group Confirmatory Factor Analysis (MG-CFA). Then, in order to examine whether equivalent measures of this construct are associated to student achievement in mathematics (as an outcome nested within schools), a series of Hierarchical Linear Modelling (HLM) analysis are developed.

3.2.4.1. Exploratory factor analysis

EFA is proposed because it is appropriate to identify the minimum number of latent dimensions that can satisfactorily describe the pattern of correlations among the set of observed indicators. This technique is a useful method for the calibration of psychometric measures as it helps to understand the structure of underlying variables and decide on the items that belong to such constructs (Baglin, 2014; Browne, 2001). In this case, I use the technique to examine whether the same group of suggested items can be accounted for by only one factor that is a contender to represent the key explanatory variable for this study.

The process of conducting an EFA involves mainly three steps: extraction, rotation and interpretation (Baglin, 2014; Beavers *et al.*, 2013; Costello and Osborne, 2005; Kline, 1994). The extraction process refers to determining the number of factors that best explains the correlations among the observed indicators, which can be examined under three criteria. The first criterion is based on eigenvalues, which represents the sum

of the squared factor loadings[26] for a given factor; estimates greater than 1 indicate a relevant proportion of the variance in the observed indicators that is explained by the factors[27]. In some cases, it can be difficult to make a correct decision about the appropriate number of factors where eigenvalues are close to 1, thus scree plots have to be also examined to support the extraction process. Such graph depicts the eigenvalues (vertical axis) versus the number of factors (horizontal axis) in order to visually inspect the point where the main inflexion is created (e.g. "the elbow") to indicate that all factors before this point should be retained. A third technique utilised to decide on the number of factors to be extracted is parallel analysis, which adds in the scree plot the results of examining mean eigenvalues calculated from a large number of random datasets based on the same number of observed indicators and observations. In this case, the intersection of this result with the one obtained in the original scree plot is used to decide on the number of factors to be extracted.

The second step in an EFA –i.e. rotation- can be conducted when more than one factor has been extracted and the relationship between them is established. Depending on considerations of underlying theory, factors can be deemed as correlated (oblique solution) or uncorrelated (orthogonal solution). And according to this assumption different levels of maximisation of the factor loadings will result for the items that best measure their respective factor. In this case, varimax orthogonal rotation under maximum likelihood estimation is used[28]. The final step –i.e. interpretation- relies on the previous processes and focuses on examining the meaning of factor loadings in terms of the strength of coefficients, as in general with any correlation[29]. Although there are no explicit guidelines in this regard, loadings greater than .3 or .4 are generally deemed as evidence of salient association between the item and the corresponding factor because these values indicate approximately 10% to 15% of overlapping variance with the rest of

[26] Factor loadings refers to the standardised partial correlation between each observed variable and a factor, controlling for the contribution of the other factors extracted.

[27] The division of a factor eigenvalue by the number of items included in the analysis yields the exact percentage of common variance among items that is the explained by the factor (e.g. communality).

[28] However, given that the solution of the EFA in this case is expected to be unidimensional, rotation would not proceed because the relationship with other factors would be trivial.

[29] Interpretation might also include the detection of indicators that load importantly in more than one factor (e.g. cross-loading), which would not proceed in this case as only one factor is expected for extraction.

items loading in the same factor (Brown, 2006; Costello and Osborne, 2005; Tabachnick and Fidell, 2001).

This three-step procedure is initially applied to the data collected from each country of interest in order to attempt to calibrate a scale score based on the five item correlations and, consequently, explore whether the same observed indicators could be used to define the key explanatory variable in all these nations. As the items included in this analysis have been measured at the school level, the data is weighted using the corresponding sampling factor (i.e. grade nonresponse adjusted school base weigh, see page 65).

3.2.4.2. Confirmatory factor analysis

Subsequently, a CFA (Jöreskog, 1969) is used to validate in each country the measurement model under analysis –e.g. that just one latent factor explains the variance of all the observed items contained in different versions of the scale-. This technique evaluates the hypothesised structure of the factorial solution by means of examining the variance-covariance matrix of the observed indicators. The aim of a CFA is to estimate all the parameters included in a measurement model that generate a predicted matrix that reproduces as closely as possible the properties of the sample matrix (Brown, 2006). Such parameters can represent either variances (e.g. factor loadings, factors variance and unique variances -also noted as indicator residuals or error), covariances (e.g. between factors and/or between errors) or means (e.g. factor means and/or indicator intercepts), depending on the aims of the analysis. They are estimated through an iterative process in which different functions (e.g. estimators) are applied to the data taking into account the scale of measurement assumed for the indicators –in general, maximum likelihood for continuous and weighted least squares for categorical variables (Muthén and Muthén, 1998-2011). In this chapter, the five items are assumed to be continuous, thus maximum likelihood is employed.

As a result of this procedure, indices of goodness-of-fit are reported to confirm that the predicted parameters reproduce the observed variance-covariance matrix based on the hypothesised measurement model. In this chapter, the expected unidimensional latent structure of the item inter-correlations is assessed using the root mean square error of approximation (RMSEA), the comparative fit index (CFI) and the Tucker-Lewis index

(TLI)[30,31]. The RMSEA evaluates the discrepancy between matrices in the light of the number of degrees of freedom and the sample size; values in this index that are closer to or smaller than .06 (with the upper limit of the confidence interval lower than .08) suggest good model fit, whereas values closer to or greater than 1 indicate the contrary. In turn, the CFI and TLI evaluates the extent to which the hypothesised model fits the data compared to a baseline model where there is no covariance between observed indicators. In both cases, values closer to or greater than .95 are deemed as evidence of good model fit; whereas TLI values smaller than .9 indicates the opposite case (Bartholomew et al., 2008; Brown, 2006; Chen, 2007).

Statistical identification of the measurement model is a crucial condition for the estimation of parameters in a CFA. This aspect describes the capacity to determine a unique set of estimates for each unknown parameter in the model (e.g. freely estimated parameters) on the basis of the known information provided by the sample variance-covariance matrix (e.g. input matrix) (Bartholomew et al., 2008; Brown, 2006). The subtraction of the number of freely estimated model parameters from the number of pieces of information in the input matrix[32] corresponds to the degrees of freedom (*df*) used in the analyses. This number is employed to determine whether the model is under-identified ($df<0$), just-identified ($df=0$) or over-identified ($df>0$). Under-identified models cannot provide estimates of parameters because the input matrix provides insufficient information; just-identified models allows such estimation, but their goodness-of-fit indices cannot be considered because such solutions always yield perfect fit by definition. In contrast, over-identified models produce parameter estimates that can be evaluated on the basis of goodness-of-fit indices, thus CFA aims to measurement models that can fulfil this condition.

[30] The Pearson chi-squared test statistic (χ^2) will be reported, but not necessarily be employed as goodness-of-fit index because it is sensitive to the size of the samples under analysis (Bartholomew *et al.*, 2008).

[31] As in the previous step (e.g. EFA), the data will be weighted using the corresponding school sampling factor (i.e. grade nonresponse adjusted school base weigh, see page 65).

[32] $p(p+1)/2 - t$; where p=number of observed indicators; t=number of freely estimated model parameters.

3.2.4.3. Multiple group – Confirmatory factor analysis

The application of CFA across several groups gives rise to MG analysis (Vandenberg and Lance, 2000; Wu, Li and Zumbo, 2007) whereby the measurement invariance properties of the hypothesised model are formally tested. This is relevant because cross-national analyses in this chapter require that the meaning of the coherence of TPD remains invariant over different school systems, which also enhances the validity of the measurement model that sustain the key explanatory variable. This strategy allows, for instance, to examine whether the suggested items measure the same construct and evidence equivalent associations with it in all the countries under analysis. In this regard, different unidimensional measurement models validated in specific groups of nations is assessed upon the results of the CFA executed in each country. Then a MG-CFA (Brown, 2006; Byrne, 2012; Wu, Li and Zumbo, 2007) is used to evaluate different levels of invariance in each pool of countries using their combined datasets.

Firstly, configural invariance is examined to confirm the equivalence of the factor structure across nations (e.g. that the factor is specified by the same items in all the four countries). In other words, whether head-teachers from different countries employ the same framework to respond the items included in the model (otherwise, lack of configural invariance indicates that unequal constructs were measured across states). This is verified by the examination of the goodness-of-fit indices reported for the simultaneous analysis of the combined dataset when all parameters are freely estimated. If this hypothesis stands, then weak invariance is examined to evaluate whether all the observed variables possess equivalent meanings (e.g. similar factor loadings) for head-teachers from each country. In quantitative terms, this level of invariance implies that per one unit variation in the score of each item, the same unit variation in the score of the factor should be observed across nations. This is evaluated by comparing goodness-of-fit indices of the configural analysis against a model with all factor loadings constrained to be equal. According to Chen (2007), invariance can be inferred if the change in the CFI value is equal to or less than .01, and the change in the RMSEA value is equal to or less than .015.

Finally, in order to support mean-level comparisons across the countries of interest, strong (or scalar) invariance is assessed considering results from previous stages. In this case, not only factor loadings but also the intercepts of the indicators are set to be equal in order to assure that the centres of the factor are scaled identically across

countries. This is an ideal level of cross-national invariance because it allows an even interpretation of the value zero of the factor given that the same loadings from the same items produce the same intercept. In the context of this chapter, strong invariance would mean that head-teachers from different countries employ a similar framework to report the construct (e.g. same items), that each item share the same meaning (e.g. equal loading) and that when a school shows no coherence of TPD, this is informed by the same combination of item scores (e.g. similar intercepts). As in the previous level of invariance, this is also evaluated by mean of goodness-of-fit indices and variations in CFI and RMSEA values[33].

3.2.4.4. Hierarchical linear modelling

In the analysis I consider the coherence of TPD as a school-level variable interpreted as an attribute of students (Rutkowski et al., 2010). Recall that students represent the main unit of analysis in PISA, therefore all findings must be interpreted at the student-level analysis. Nevertheless, schools are selected with probability proportional to their size in the design of PISA, therefore the measurement model related to the explanatory variable might be also interpreted as an attribute of schools. In this case, a HLM approach (O'Connell and McCoach, 2008; Raudenbush and Bryk, 2002; Snijders and Bosker, 1999) becomes suitable to model the hypothesised positive association with student achievement. As a result, the variance of the outcome variable is separated into the between and within schools components, thus the contribution of the coherence of TPD on student achievement is assessed by the specification of a mean-as-outcomes model using background characteristics of students and schools as controls[34] (Raudenbush and Bryk, 2002).

[33] As in the previous steps of the factor analysis strategy, the MG-CFA will use data weighted at the school level with the respective sampling factor (i.e. grade nonresponse adjusted school base weigh, see page 65).

[34] At the student level, gender, immigrant status and socioeconomic status will be used as controls. At the school level, administration (public/private), location, average class size and school size.

The final form of the model to be separately estimated for each country is:

$$Y_{ij} = \gamma_{00} + \gamma_{01}(X_j) + \mu_{0j} + r_{ij}$$

Where:

Y = Student achievement measured as the overall mathematics scale of PISA 2012.

X = Coherence of TPD measured as a standardised score of each measurement model under analysis[35].

γ_{00}= Average intercept across schools.

γ_{01}= Average regression slope across schools.

μ_{0j} = Random intercept effect of unit j, with variance τ_{00}.

r_{ij} = Error term, with variance σ^2.

i = Student i.

j = School j.

In this model the selection probabilities at each stage of sampling need to be considered for parameters' estimation because by only using the overall inclusion weight for each student (w_{ij}), the respective scaling may affect point estimates in the results (Rabe-Hesketh and Skrondal, 2006; Stapleton, 2013). Therefore, both the conditional sampling weight within schools at level-1 (w_{ilj}) and the schools sampling weight (w_j) are required for this two-level model[36]. However, there is not a unique method for standardising weights and no consensus on the best approximation to this problem (Stapleton, 2013). Scaling can proceed by standardising w_{ij} or w_{ilj} to sum to the actual or the effective sample size; although under a third method the sampling weights at the school level can be set to the cluster averages of w_{ij}, which sets accordingly the students' weights to the unit in every school. At this juncture, it is recommended practice to

[35] Factor scores for each measurement model were calculated in Mplus (Muthén and Muthén, 1998-2011) as a result of the CFA procedures implemented in each case, thus values corresponding to the coherence of TPD were assigned to each school. Such factor scores were then standardised within each country to facilitate interpretation, with value zero referring to the average coherence of TPD across schools and each unit as one standard deviation in the original factor score.

[36] Both w_{ij} and w_j are provided in the PISA 2012 dataset (variables W_FSTUWT and W_FSCHWT, respectively; see Table 3.1 in page 65) whereas w_{ilj} can be calculated as: $w_{ilj} = w_{ij} / w_j$.

contrast estimates across all the three methods of scaling –or if you prefer a 'sensitivity analysis' (StataCorp, 2011b)- as well as in relation to results yielded when no scaling or weights are applied –labelled as 'informativeness of weights analysis' (Stapleton, 2013). Therefore, estimates from the HLM in this chapter are computed for these five conditions, namely: overall weight (w_{ij} and no scaling), unweighted data, and scaling methods 1 (actual size), 2 (effective size) and 3[37].

In summary, after the successive application of EFA, CFA and MG-CFA, it is expected that unidimensional measurement models of the key explanatory variable are developed across specific groups of countries. Finally, of any resulting predictor of the coherence of TPD based on these measurement models, their association with student achievement is evaluated through HLM. All the analysis is executed with Mplus v. 7.3 (Muthén and Muthén, 1998-2011) and STATA 12 (StataCorp, 2011a).

In the following section the main results from each stage of the analytic strategy are reported. Firstly, successive EFA are developed to examine the dimensionality of the latent construct of coherence in TPD and suggest specific measurement models –e.g. items and countries-. Secondly, the goodness-of-fit of each of these models are evaluated within each country via CFA in order to validate their use as suitable scales for the assessment of the key explanatory variable. Thirdly, configural, weak and strong invariance of each of these models are examined through MG-CFA in the pooled datasets of the countries included in each measurement model to assess the level of cross-cultural comparability of the scales. Finally, factor scores representing the coherence of TPD are included as key explanatory variables in specific HLM analyses to evaluate whether this construct makes a positive difference to students' learning in mathematics in each of the countries of interest.

[37] Detailed results are presented in appendices J and K.

3.3. Results

Results from the factor analytic evaluation in each country outlined previously are initially reported in this section to detail latent structures of the key explanatory variable that are satisfactory. Following these results, findings from a MG-CFA are reported in order to know whether the resultant measurement models are invariant across the same set of school systems. For those cases in which valid and/or equivalent measurement models represent the key explanatory variable, the association with student achievement is reported through HLM.

3.3.1. Exploratory factor analysis

The underlying dimensionality of the initial scale based on five items was examined in each country using an EFA with orthogonal rotation under maximum likelihood estimation with robust standard errors (MLR). The use of this estimator is specified in Mplus v.7.3 (Muthén and Muthén, 1998-2011) when sampling weights are included in the analysis. Table 3.4 displays the (total) eigenvalues and the corresponding percentage of variance explained by each factor across the four selected nations.

Table 3.4 Eigenvalues from EFA using MLR for the US, UK, Japan and Finland

Country	Factor	Eigenvalues (items 1 to 5) Total	% of variance	cumulative %
US	1	2.28	46%	46%
	2	0.96	19%	65%
	3	0.78	16%	80%
	4	0.55	11%	91%
	5	0.43	9%	100%
UK	1	2.40	48%	48%
	2	1.09	22%	70%
	3	0.76	15%	85%
	4	0.41	8%	93%
	5	0.34	7%	100%
JPN	1	1.78	36%	36%
	2	1.02	20%	56%
	3	0.95	19%	75%
	4	0.68	14%	89%
	5	0.56	11%	100%
FIN	1	1.71	34%	34%
	2	1.06	21%	55%
	3	0.90	18%	73%
	4	0.72	14%	88%
	5	0.62	12%	100%

Source: PISA 2012 database

Notes: weighted data.

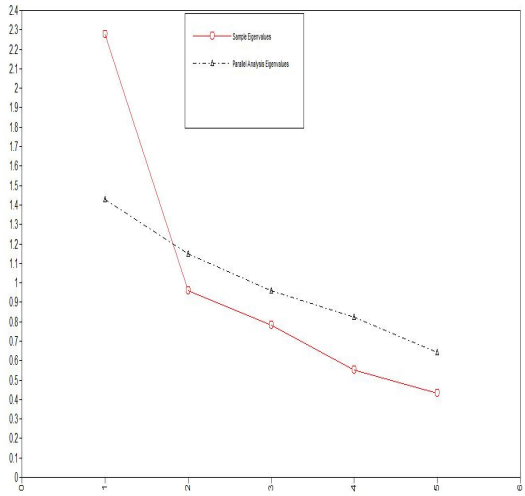
According to these results, it is initially difficult to suggest the same number of factors to be extracted across countries, as either one, two or three latent dimensions could be seemingly advised. For instance, the US sample indicates that one dimension could be satisfactorily extracted because just one eigenvalue is greater than 1, which accounts for by 46% of the common variance of the scale[38]. However, in the other three countries, two-dimensional solutions seem to be adequate, as for example in the UK where the pair of factors with an eigenvalue greater than 1 explains for 70% of the variance. Even the extraction of a third factor might be arguable in Japan and Finland, because an additional eigenvalue is closer to the unit (.95 and .9, respectively) and accounts over 73% of the variance along with the other two previous factors.

[38] As Beavers *et al.* (2013) point out, there is no consensus in the literature on how much variance should be explained by a factor to decide in favour of its extraction. However, these values will be reported in this chapter for purposes of comparison.

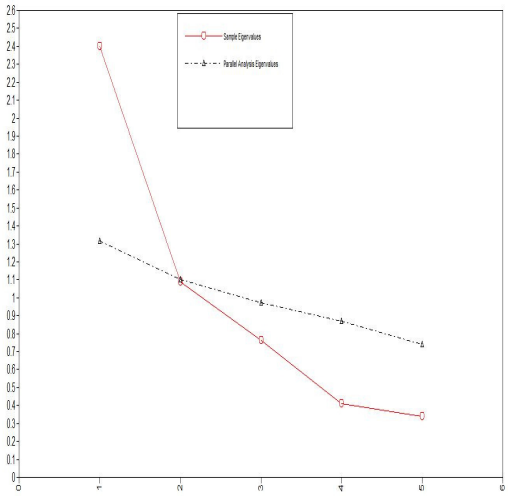
At this juncture, additional criteria are needed to decide on the adequate number of factors to be extracted. Figure 3.1 shows scree plots with parallel analyses produced by the EFA of the initial scale in each country.

Figure 3.1 Initial scree plots (5 items) for factor eigenvalues for EFA using MLR for the US, UK, Japan and Finland

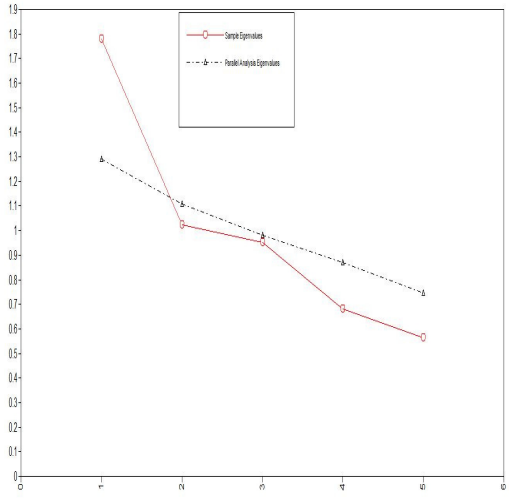
US



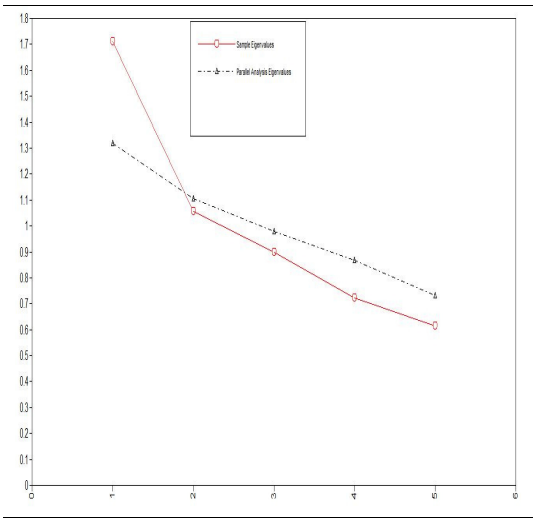
UK



JPN



FIN



Source: PISA 2012 database

Notes: weighted data; horizontal axis indicates the number of factors and vertical axis indicates the value of total eigenvalues; red lines correspond to "sample eigenvalues" and dashed lines correspond to "parallel analysis eigenvalues".

Results indicate that a clear main point of inflexion can be described in each country before the second factor, whereas the lines of parallel analysis intersects the samples at the same point, too. This provides strong argument to retain only one factor across all countries, thus if the five items are included in the scale, the extraction of a single latent dimension seems suitable to the data from every nation.

Considering the extraction of a unique latent dimension as an adequate solution for the initial scale, interpretation of factor loadings proceeds. Recall that interpretation of an EFA relies on the salience of such coefficients and the conceptual contribution of each indicator, thus decisions on the dimensionality of the construct are also constrained by these considerations. In this regard, Table 3.5 provides information on the factor loadings of the five items for all the four countries under analysis, with the percentage of variance explained for by the single factor in the final row.

Table 3.5 Item loadings from EFA with one factor solution using MLR for the US, UK, Japan and Finland

Item	US	UK	JPN	FIN
1	**0.38**	0.12	0.16	0.17
2	**0.42**	**0.31**	0.17	0.27
3	**0.68**	**0.75**	**0.48**	**0.62**
4	**0.72**	**0.81**	**0.63**	**0.54**
5	**0.60**	**0.81**	**0.67**	**0.46**
% of variance	46%	48%	36%	34%

Source: PISA 2012 database

Notes: weighted data.

The coefficients displayed indicate that only in the US all the five items present loadings considered as salient (greater than .3 or .4), that this condition is only met by items 2 to 5 in the UK and by items 3 to 5 in Japan and Finland. Thus a five item scale score for the US can be deemed as an adequate measure of the underlying dimension of the coherence of TPD in schools, as informed by head-teachers from this nation[39].

[39] Furthermore, results from a CFA indicated a good adequacy of the fit of this model (CFI = .97, TLI = .94, RMSEA = .04), which makes of this scale a valid measure of the construct in this country. This scale will be named the "US Measurement Model" (US MM) for the purposes of the HLM analysis.

However, this is not the case of the other three countries, where only the last four or three items reported salient factor loadings under an unidimensional solution.

In this context, a further EFA is conducted once item 1 is removed from the scale, considering that this observed variable yielded the lowest loading across all countries. Table 3.6 shows the corresponding eigenvalues.

Table 3.6 Eigenvalues from EFA using MLR for the US, UK, Japan and Finland

		Eigenvalues (items 2 to 5)		
Country	**Factor**	**Total**	**% of variance**	**cumulative %**
US	1	2.09	52%	52%
	2	0.88	22%	74%
	3	0.57	14%	89%
	4	0.46	12%	100%
UK	1	2.37	59%	59%
	2	0.88	22%	81%
	3	0.42	10%	92%
	4	0.34	8%	100%
JPN	1	1.74	43%	43%
	2	1.01	25%	69%
	3	0.68	17%	86%
	4	0.57	14%	100%
FIN	1	1.68	42%	42%
	2	0.90	23%	65%
	3	0.75	19%	83%
	4	0.67	17%	100%

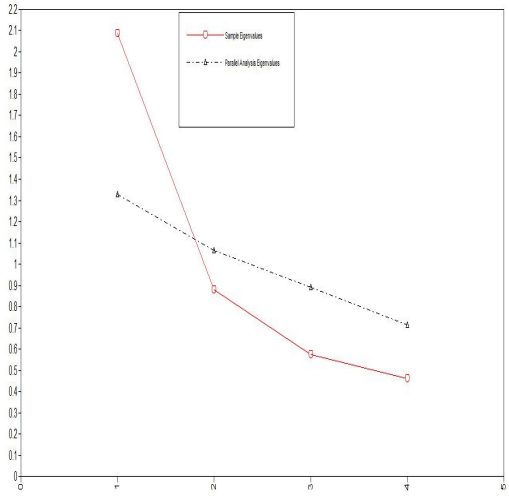
Source: PISA 2012 database

Notes: weighted data.

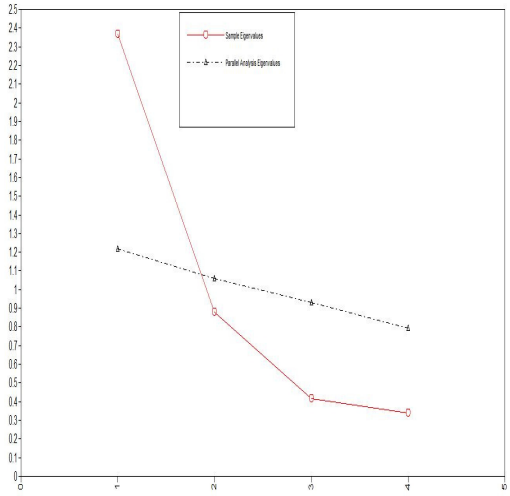
As in the initial analysis, estimates are not either conclusive to advice on the exact number of factors to be retained across countries. For example, results in the US, UK and Finland indicate that one dimension is satisfactory, with only the first eigenvalue greater than 1 and accounting for 52%, 59% and 42% of the variance of the scale, respectively. However, in Japan, a two-dimensional solution might be supported, as two factors with an eigenvalue greater than 1 explains for by 69% of the variance. Figure 3.2 displays the corresponding scree plots with parallel analyses.

Figure 3.2 Scree plots for factor eigenvalues for EFA using items 2 to 5 and MLR for the US, UK, Japan and Finland

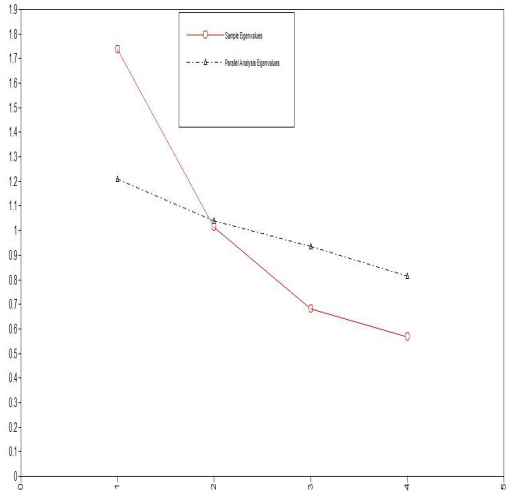
US



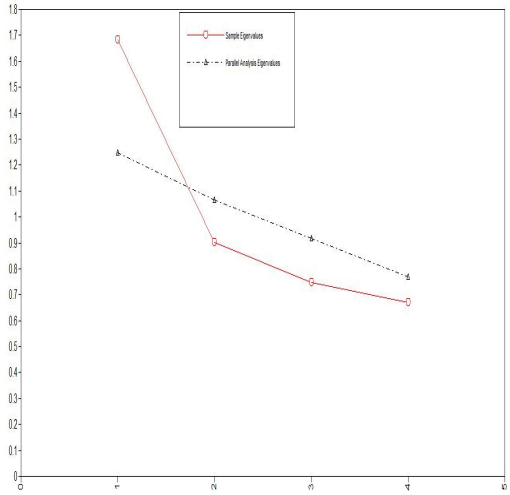
UK



JPN



FIN



Source: PISA 2012 database

Notes: weighted data; horizontal axis indicates the number of factors and vertical axis indicates the value of total eigenvalues; red lines correspond to "sample eigenvalues" and dashed lines correspond to "parallel analysis eigenvalues".

In this case, a clear inflexion supports the extraction of a unique dimension in the US, UK and Finland; however, the case of Japan is less straightforward as it might be difficult to show in which point the elbow creates. Nonetheless, the intersection with the line of parallel analyses provides once again definitive argument to retain only one factor across all countries based on items 2 to 5. Table 3.7 details the factor loadings of the four

items for all the four countries of interest, with the percentage of variance explained for by the factor in the last row.

Table 3.7 Item loadings from EFA with one factor solution using MLR for the US, UK, Japan and Finland

Item	US	UK	JPN	FIN
2	**0.40**	**0.30**	0.17	0.28
3	**0.62**	**0.75**	**0.48**	**0.56**
4	**0.79**	**0.82**	**0.62**	**0.58**
5	**0.61**	**0.80**	**0.68**	**0.49**
% of variance	52%	59%	43%	42%

Source: PISA 2012 database

Notes: weighted data.

As in the initial EFA, factor loadings from this solution indicate that only in the US and UK all these four items report salient values, however this quality is only achieved by items 3 to 5 in the other two countries. As a result, this stage of the EFA suggests the validation of two different unidimensional measurement models through a CFA applied to data from each corresponding nation. The first of these models is labelled as "US and UK Measurement Model" (US&UK MM) and specified in these two countries by items 2 to 5. The second is provisionally named as "Japan and Finland Measurement Model" (JPN&FIN MM) and based on items 3 to 5 for these two nations. If such models fits adequately their national data, then MG-CFA could be applied to each pair of countries.

Finally, in order to examine whether exactly the same structure of an underlying dimension can be extracted from each national sample, a final EFA was conducted by removing item 2 from the scale on the basis of its comparatively lower loading at this stage across the four countries. Eigenvalues corresponding to this solution are provided in Table 3.8.

Table 3.8 Eigenvalues from EFA using MLR for the US, UK, Japan and Finland

		Eigenvalues (items 3 to 5)		
Country	**Factor**	**Total**	**% of variance**	**cumulative %**
US	1	1.91	64%	64%
	2	0.60	20%	84%
	3	0.49	16%	100%
UK	1	2.25	75%	75%
	2	0.42	14%	89%
	3	0.34	11%	100%
JPN	1	1.71	57%	57%
	2	0.71	24%	81%
	3	0.58	19%	100%
FIN	1	1.55	52%	52%
	2	0.76	25%	77%
	3	0.69	23%	100%

Source: PISA 2012 database

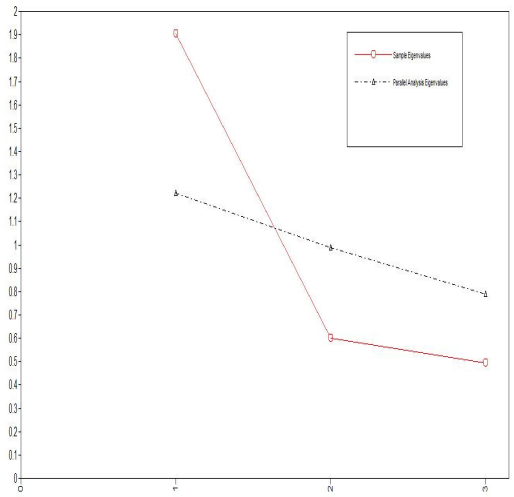
Notes: weighted data.

In this case, the data suggest the extraction of only one factor that account for 64% of the variance of the scale in the US, 75% in the UK, 57% in Japan and 52% in Finland. Scree plots displayed in Figure 3.3 are in line with the decision to create a score based on items 3 to 5 and, unlike previous analyses, all factor loadings yield values over .5, which largely support the saliency of all these three indicators –see Table 3.9 below. Based on these results, it is suggested to validate the aforementioned JPN&FIN MM additionally in the US and UK through a CFA[40] in order to proceed to a MG-CFA in the pooled dataset of all the four states. Therefore, this is then labelled as the "All the 4 countries Measurement Model" (ALL4 MM), as it is specified by items 3 to 5 in all the countries of interest.

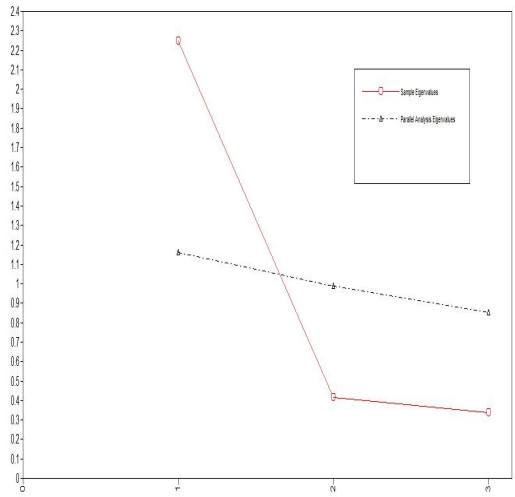
[40] Notwithstanding that this model is just-identified because the number of unknown is equal to the known parameters from the input matrix. Therefore, absence of degrees of freedom shall yield by definition perfect indices of goodness-of-fit (CFI = 1, TLI = 1, RMSEA = 0) in each country.

Figure 3.3 Scree plots for factor eigenvalues for EFA using items 3 to 5 and MLR for the US, UK, Japan and Finland

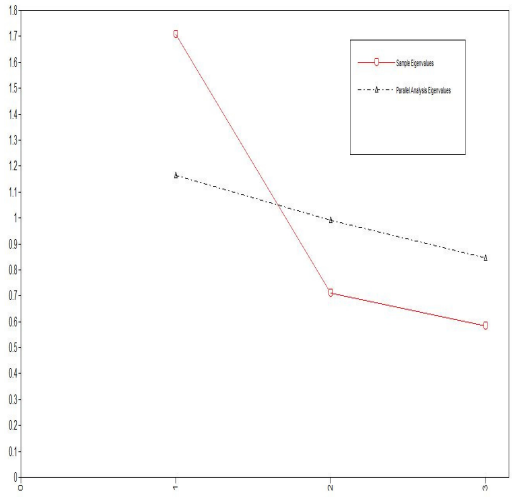
US



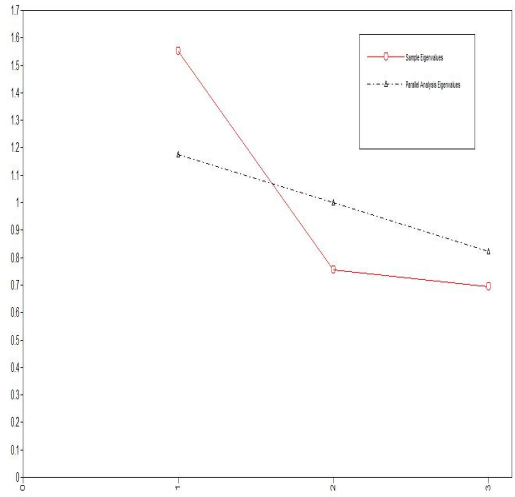
UK



JPN



FIN



Source: PISA 2012 database

Notes: weighted data; horizontal axis indicates the number of factors and vertical axis indicates the value of total eigenvalues; red lines correspond to "sample eigenvalues" and dashed lines correspond to "parallel analysis eigenvalues".

Table 3.9 Item loadings from EFA with one factor solution using MLR for the US, UK, Japan and Finland

Item	US	UK	JPN	FIN
3	**0.62**	**0.75**	**0.50**	**0.50**
4	**0.76**	**0.83**	**0.65**	**0.59**
5	**0.65**	**0.80**	**0.64**	**0.49**
% of variance	64%	75%	57%	52%

Source: PISA 2012 database

Notes: weighted data.

3.3.2. Confirmatory factor analysis

On the basis of results produced by the EFA, the US&UK MM was specified as a unidimensional measurement model based on a scale that includes items 2 to 5. In total 12 parameters are freely estimated in this model –e.g. 4 factor loadings, 4 unique variances and 4 intercepts-, thus over-identification is fulfilled because only 10 pieces of information are provided by the input matrix (6 bivariate correlations between items and their 4 standardised variances), which results in a positive value of two degrees of freedom. Results from a CFA using MLR showed a good adequacy of the fit of this measurement model in the US (χ^2 = 1.372, p= 0.5, *df*=2; CFI = 1, TLI = 1.07, RMSEA = 0) and UK (χ^2 = 1.383, p= 0.5, *df*=2; CFI = 1, TLI = 1.01, RMSEA = 0) samples. Following recommended practice to report results from this type of analysis (Brown, 2006), Table 3.10 details all the 12 parameter estimates for each country.

Table 3.10 Parameter estimates of US&UK MM (items 2 to 5 in the US and UK)

	Item	Loading	SE		Intercept	SE		Residual	SE	
US	**2**	0.40	0.13	***	2.82	0.30	***	0.84	0.11	***
	3	0.62	0.10	***	2.61	0.24	***	0.62	0.12	***
	4	0.79	0.13	***	2.92	0.25	***	0.38	0.20	*
	5	0.61	0.12	***	3.19	0.32	***	0.63	0.14	***
UK	**2**	0.29	0.10	***	4.50	0.31	***	0.92	0.06	***
	3	0.75	0.05	***	2.98	0.23	***	0.44	0.08	***
	4	0.82	0.05	***	3.01	0.19	***	0.33	0.09	***
	5	0.81	0.04	***	2.99	0.26	***	0.35	0.07	***

Source: PISA 2012 database

Notes: weighted data; *p < .1, **p < .05, ***p < .01; SE=Standard Error.

These data suggest that either in the US or UK all these four items are meaningfully related to the latent factor that will form the key explanatory variable of this chapter. This is noticeable because the salience of loadings already noticed in the previous EFA is here statistically significant, as well as the intercepts and the residuals, which indicates that all these values differ consistently from zero in their respective national target populations. Therefore, these four items can be deemed as a valid scale to assess the coherence of TPD within each of these two school systems, whilst it is suggested the assessment of the invariance of this measurement model through a MG-CFA of their pooled dataset.

Following the EFA results, the "ALL4 MM" was specified as a unidimensional model based on items 3 to 5 across all four countries. As noted, this model is just-identified due to the absence of degrees of freedom in the analysis; therefore, results from a CFA using MLR indicate a perfect fit in each country. Because of this, only estimates of the 9 free parameters are reported in Table 3.11.

Table 3.11 Parameter estimates of ALL4 MM (items 3 to 5 in the US, UK, Japan and Finland)

	Item	Loading	SE		Intercept	SE		Residual	SE	
US	**3**	0.62	0.11	***	2.61	0.24	***	0.62	0.13	***
	4	0.76	0.18	***	2.92	0.25	***	0.43	0.28	
	5	0.65	0.12	***	3.19	0.32	***	0.58	0.16	***
UK	**3**	0.75	0.05	***	2.98	0.23	***	0.45	0.08	***
	4	0.83	0.06	***	3.01	0.19	***	0.31	0.09	***
	5	0.80	0.05	***	2.99	0.26	***	0.36	0.08	***
JPN	**3**	0.50	0.11	***	2.90	0.22	***	0.75	0.10	***
	4	0.65	0.12	***	3.41	0.25	***	0.58	0.16	***
	5	0.64	0.12	***	2.96	0.21	***	0.59	0.15	***
FIN	**3**	0.50	0.12	***	2.50	0.10	***	0.75	0.12	***
	4	0.59	0.15	***	3.71	0.26	***	0.66	0.17	***
	5	0.49	0.12	***	2.93	0.20	***	0.76	0.12	***

Source: PISA 2012 database

Notes: weighted data; *p < .1, **p < .05, ***p < .01; SE=Standard Error.

As in the US&UK MM, all the items and parameters of the ALL4 MM appear as meaningfully related to the latent factor under analysis. Therefore, a scale based on these three items can be also accepted as a valid measuring instrument of the coherence of TPD

in each country, though potential cross-national comparisons are constrained to the assessment of its invariance through a MG-CFA.

3.3.3. Multiple group – Confirmatory factor analysis

Once the three identified measurement models (e.g. US MM, US&UK MM and ALL4 MM) were validated within specific countries through CFA, the invariance of the cross-national instances was assessed through MG-CFA. As already mentioned, this technique aims to determine whether the set of items of each measurement model produces equivalent associations with the underlying construct of the coherence of TPD in all the respective countries under analysis. A sequence of increasingly restrictive analyses with the two corresponding pooled datasets (e.g. the US and UK, and all the four countries) allows to evaluate the three levels of measurement invariance: configural, weak and strong.

These levels respectively describe whether equivalence can be established at the level of the structure (e.g. equal form or the same items used in the model), the meanings (e.g. equal loadings) or the average values of the factors (e.g. equal intercepts). Accordingly, this analysis starts with the examination of the goodness-of-fit of the *configural* level using similar assessment criteria to those utilised in the CFA above; then such results are contrasted against the goodness-of-fit indices yielded at the *weak*, and finally at the *strong* level of invariance. In these two latter stages, variations within the thresholds advised by Chen (2007) (CFI_{change}=<.01; $RMSEA_{change}$=<.015) are as criteria to evaluate whether the hypothesis of invariance successively remains stable.

Table 3.12 displays the results of the tests of measurement invariance applied to the US&UK MM, in which only items 2 to 5 were employed and the combined dataset of the US and UK schools analysed. Model fit coefficients and their changes from one level of invariance to another are detailed for the Chi-squared, CFI, TLI and RMSEA indices. In particular, the last two columns indicate the values used to assess the invariance of the measurement model across these two countries: the change (Δ) in the CFI and RMSEA indices.

Table 3.12 Tests of measurement invariance of US&UK MM (the coherence of TPD as measured by items 2 to 5 in the US and UK)

	χ^2	df	p-value	χ^2 diff	Δdf	RMSEA	(90% CI)	CFI	TLI	ΔRMSEA	ΔCFI
Configural (equal form)	2.87	4	0.58			0.00	(0 - .081)	1.00	1.02		
Weak (equal factor loadings)	5.28	7	0.63	2.42	3	0.00	(0 - .064)	1.00	1.02	0.00	0.00
Strong (equal indicator intercepts)	39.63	10	0.00	34.35	3	0.11	(.074 - .144)	0.83	0.80	0.11	-0.17

Source: PISA 2012 database

Notes: weighted data; *p < .1, **p < .05, ***p < .01; χ^2diff, nested χ^2 difference; RMSEA, root mean square error of approximation; 90% CI, 90% confidence interval for RMSEA; CFI, comparative fit index; TLI, Tucker - Lewis Index.

The values in the first row indicates that the assumption of *configural* invariance is met (CFI = 1, TLI = 1.02, RMSEA = 0), thus the form of the model defined by one factor and the specified four indicators is reported as identical across these two countries. In view of this, it is important to highlight that head-teachers from the US and UK are likely to use a similar conceptual framework to answer the questions included in this model. In particular, they seem keen to consider the coherence of TPD as a key part of their role that involves linking the outcomes from teacher evaluation processes with these activities and implementing multiple strategies for their supervision and control.

Further, the model yielded similar estimates of goodness-of-fit when equality constraints are imposed to the factor loadings, with change indices (CFI_{change}=0; $RMSEA_{change}$=0) within the expected thresholds. This indicates that *weak* invariances is also met, suggesting that items in the US&UK MM evaluate the coherence of TPD in a broadly equivalent manner among both US and UK head-teachers. This implies that leaders in both countries share a similar meaning of the coherence of TPD, in the sense that unobserved variations in this factor affect similarly the scores in the observed items, regardless the nationality of the respondents.

In contrast, when restrictions on factor intercepts were included into the analysis, a substantial impairment of the model fit indices was reported in relation to the previous test of invariance (CFI_{change}=-.17; $RMSEA_{change}$=.11). As a result, the assumption of *strong* invariance is not met by this model, thus the centres of the factor are not scaled identically across both countries. In other words, absence of this level of invariance implies that the hypothetical average value of the coherence of TPD is radically dissimilar between these two nations, thus no comparison of mean performances in this construct could be validly interpreted in the head-teachers from each country.

The results of the tests of measurement invariance applied to the ALL4 MM are presented in Table 3.13, in which items 3 to 5 were used in the combined dataset of all the four countries of interest.

Table 3.13 Tests of measurement invariance of the ALL4 MM (the coherence of TPD as measured by items 3 to 5 in the US, UK, Japan and Finland)

	χ^2	df	p-value	χ^2 diff	Δdf	RMSEA	(90% CI)	CFI	TLI	ΔRMSEA	ΔCFI
Configural (equal form)	0.00	0	0.00			0.00	(.00 - .00)	1.00	1.00		
Weak (equal factor loadings)	1.42	6	0.96	1.42	6	0.00	(000 - .00)	1.00	1.04	0.00	0.00
Strong (equal indicator intercepts)	53.71	12	0.00	52.29	6	0.12	(.09 - .15)	0.81	0.81	0.12	-0.19

Source: PISA 2012 database

Notes: weighted data; *p < .1, **p < .05, ***p < .01; χ^2diff, nested χ^2 difference; RMSEA, root mean square error of approximation; 90% CI, 90% confidence interval for RMSEA; CFI, comparative fit index; TLI, Tucker - Lewis Index.

In this case, *configural* invariance is met by definition because the model is just-identified and yields perfect goodness-of-fit indices (CFI = 1, TLI = 1, RMSEA = 0), thus the form of the model based on one factor and the specified three items is reported as identical across countries. Likewise, when equality constraints were imposed to the factor loadings, the model produced similar estimates of goodness-of-fit (CFI_{change}=0; $RMSEA_{change}$=0). This result suggests that *weak* invariance is also met, thus the three items of the ALL*4* MM elicit a similar meaning among head-teachers from all the four countries. However, when the similarity of factor intercepts was assessed across these samples, the fit of the model was importantly affected regarding the test of *weak* invariance (CFI_{change}=-.19; $RMSEA_{change}$=.12), thus *strong* invariance cannot be supported in this case. On balance, the implications derived from the analysis of the US&UK MM also apply in this model, as only *weak* measurement invariance can be suggested for these four countries.

In sum, the results from the MG-CFA show that both cross-national measurement models satisfactorily reflect the coherence of TPD in the schools of their corresponding sets of countries. The items included in the first of these models (US&UK MM) describe the opportunities for TPD derived from teacher appraisals (item 2), the head-teachers' monitoring of schools goals in TPD (item 3), their presence in such activities (item 4) and their discussion with teachers about their recent experiences in this area (item 5). Only the first of these items is excluded from the second measurement model evaluated (ALL*4* MM). In both cases, the items employed possess the same meaning for head-teachers regardless their nationality, however, the levels of their intercepts cannot by deemed as equivalent across countries.

3.3.4. Hierarchical linear modelling

Given the results of the CFA applied to the US MM (items 1 to 5) and the MG-CFA executed on the US&UK MM (items 2 to 5) and the ALL*4* MM (items 3 to 5), standardised factor scores were generated in these three cases to operationalise the latent construct of the coherence of TPD. HLM was then implemented to statistically analyse within countries the relationship of the key explanatory variable with the achievement of students in the context of a nested data structure (e.g. level-1 students and level-2 schools). A means-as-outcomes model using background characteristics of students and

schools as controls (Raudenbush and Bryk, 2002) was applied in order to examine whether the coherence of TPD makes a difference to the average achievement of students across schools. Consequently, this section presents three specific analyses to evaluate the association of each measurement model of the key explanatory variable with student achievement in PISA 2012[41].

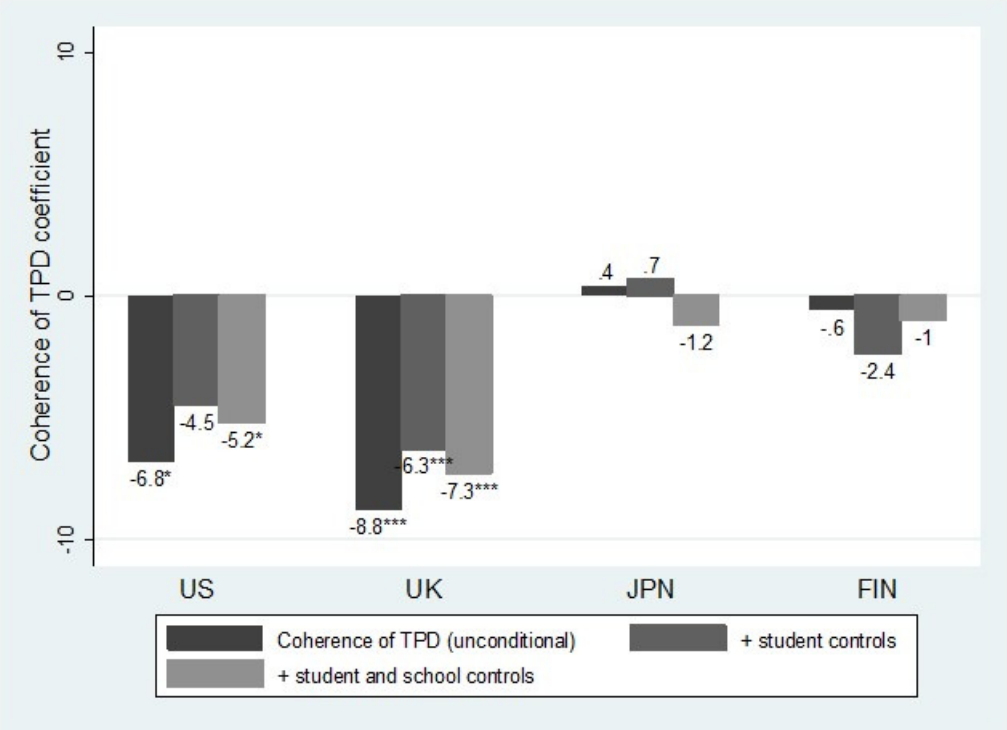
3.3.4.1. HLM models using the ALL4 MM

The first set of HLM models of this analysis uses as level-2 key explanatory variable the standardised factor score comprising three specific items of the coherence of TPD that showed satisfactory metric invariance across all the four countries of interest (e.g. ALL4 MM). These items describe the frequency of actions in which school leaders ensured that TPD were in line with the goals of the school (item 3), the regularity of their presence these events (item 4) and the number of times they shared information on this topic with teachers (item 5).

Figure 3.4 shows the results of the respective mean-as-outcomes model used to explore the coefficient of association with students' achievement in mathematics. For each country, bars represent the magnitude of the regression coefficient in the context of three different HLM models. Whereas the first bar indicates the results of the unconditional model (e.g. bivariate correlation), the second and third bars illustrate how this estimate varied when control variables at the student and school level were introduced.

[41] Following recommended practice to report HLM analysis (Raudenbush and Bryk, 2002), unconstrained (null) models were firstly estimated for each country to provide information on the average achievement of the students (γ_{00}), the variances between ($\boldsymbol{\tau}_{00}$) and within ($\boldsymbol{\sigma}^2$) schools and the corresponding intra-class correlations (ρ) produced by the five methods of estimation employed in the context of the sensitivity analysis (e.g. unweighted data, overall weight, and scaling methods 1, 2 and 3). Results from the ρ coefficients indicated that the proportion of variance of student achievement between schools is more than a half of the total variance in Japan (.54, .55 and .56), whereas this is approximately one third in the UK (.30, .31 and .36), one quarter in the US (.24, .25, .28, .30) and between .12 and .29 in Finland, depending on the method employed. These results allow to hypothesise that a number of school level predictors (e.g. the coherence of TPD) can play a role in the national performance of students in PISA 2012.

Figure 3.4 Variation of the conditional association between student achievement and coherence of TPD across Means-as-Outcomes HLM models for the US, UK, Japan and Finland (ALL4 MM)



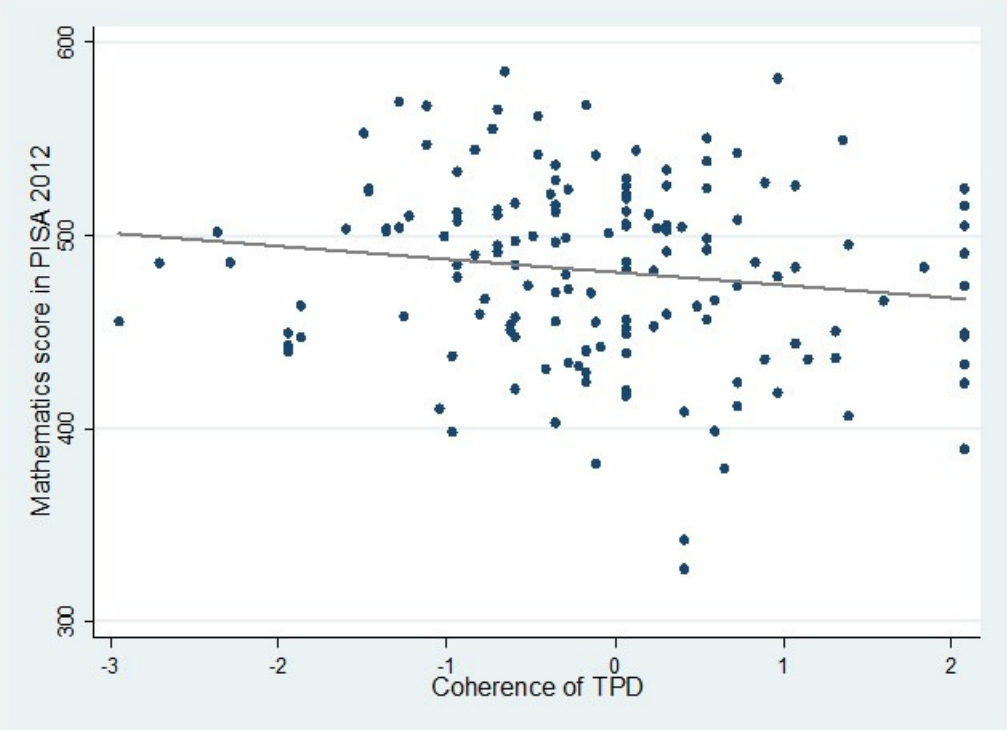
Source: PISA 2012 database

Notes: Outcome variable: Mathematics score in PISA 2012; *p < .1, **p < .05, ***p < .01; Coherence of TPD is measured by a standardised factor score with metric invariance across all the four countries of interest (ALL4 MM).

Results indicated that the contribution of the coherence of TPD to students' outcomes was null or small and in those cases where it was statistically different from zero, it showed a negative direction. The maximum absolute value of the coefficient across all the 12 HLM models was 8.8 points, which only represented about 8% of one standard deviation in the PISA scale across all the OECD countries. In more than a half of the specified models the results indicated that the association of the key explanatory variable with the outcome was likely to be zero, with all the models in Japan and Finland showing this pattern of results. However, both in the US and UK schools, most of the results suggested that the coherence of TPD made evident a significant negative association with the achievement of their students. For instance, the unconditional model indicated that for US and UK students from different schools, one standard deviation improvement in the score of the coherence of TPD was associated with a decrease of 6.8 and 8.8 points in student achievement in mathematics, respectively.

The aforementioned relationships are depicted in Figures 3.5 (US) and 3.6 (UK). The respective scatterplots display in the horizontal axis the score in the coherence of TPD factor and mathematics achievement in the vertical axis. Dots represent schools and the best fit line of the regression is depicted in grey colour.

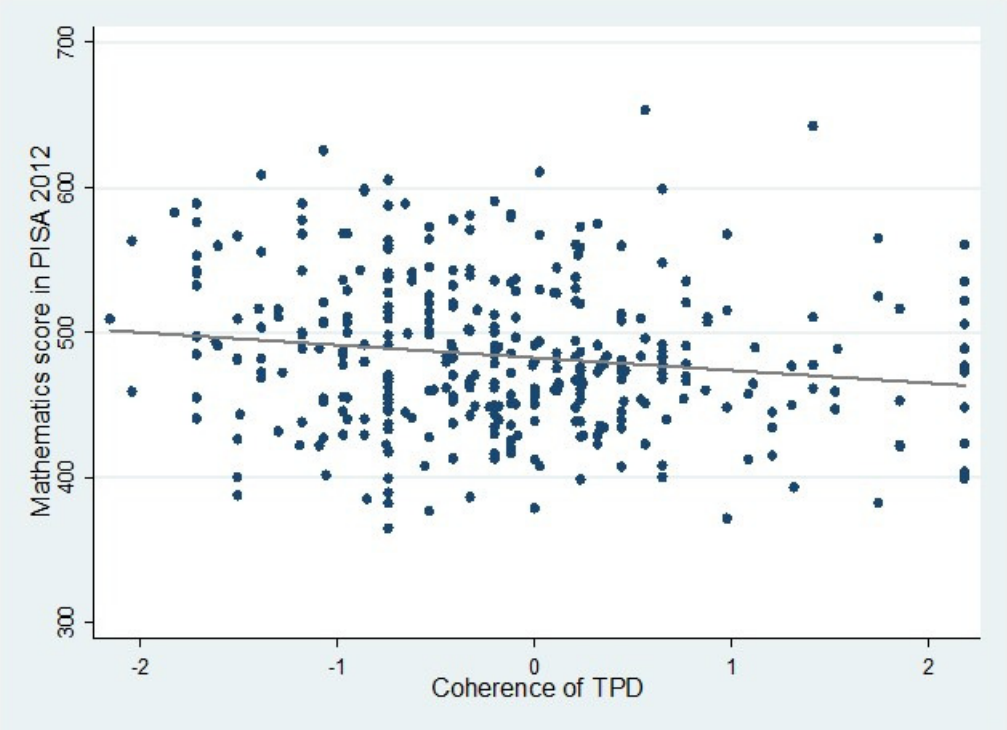
Figure 3.5 Coherence of TPD and student achievement in PISA 2012 in the US



Source: PISA 2012 database

Notes: weighted data; Coherence of TPD is measured by a standardised factor score with metric invariance across all the four countries of interest (ALL4 MM).

Figure 3.6 Coherence of TPD and student achievement in PISA 2012 in the UK



Source: PISA 2012 database
Notes: weighted data; Coherence of TPD is measured by a standardised factor score with metric invariance across all the four countries of interest (ALL4 MM).

The downward slope of the line of best fit in each graph confirmed the slight and negative association between these two variables in each of these nations. Furthermore, such trends were still significant when control variables at the student and school level were included in the model, in which case the absolute values of the coefficients reduced to 5.2 (US) and 7.3 (UK) points[42].

3.3.4.2. HLM models using the US&UK MM

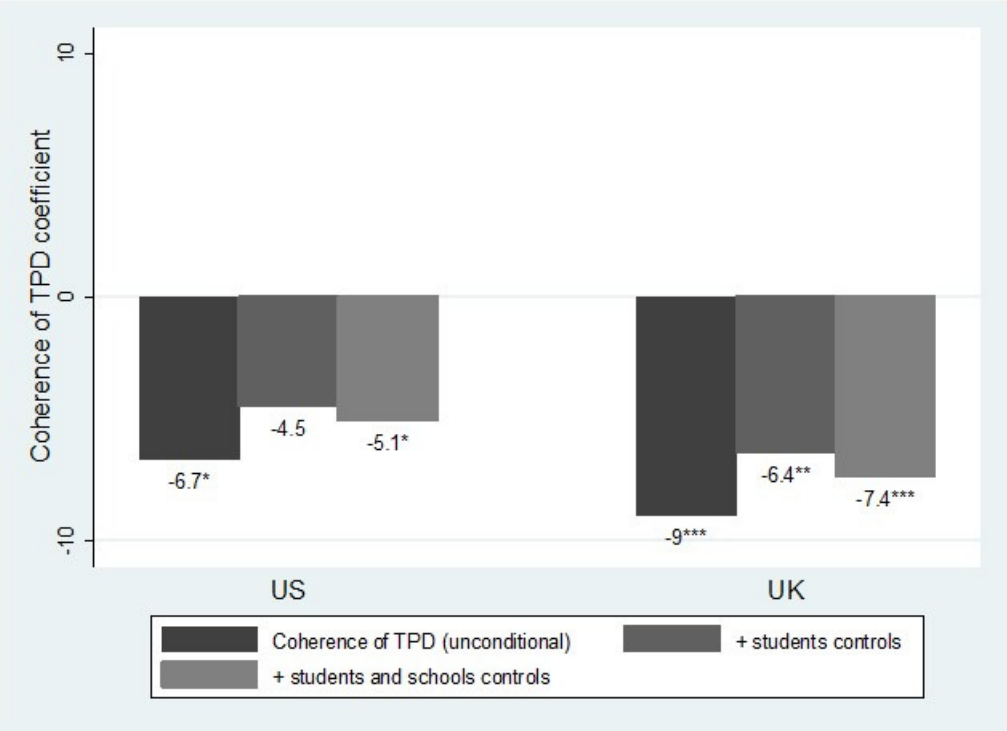
The second group of HLM models uses as level-2 key explanatory variable the standardised factor score based on four specific items of the coherence of TPD that showed satisfactory metric invariance between the US and UK (e.g. US&UK MM). Apart from the items related to the supervision of TPD activities carried out by head-teachers (items 3 to 5), this measurement model also includes the extent to which TPD

[42] Further analyses on the influence of the items removed from the original scale (e.g. "standardised maths policy" and "teachers' appraisals and TPD") are provided in Appendix L.

opportunities were linked to the procedures for teacher evaluation within schools (item 2).

Figure 3.7 presents the coefficients of association with student achievement in the respective mean-as-outcomes models. The first bar in each country indicates the results of the unconditional model (e.g. bivariate correlation) and the second and third bars describe how this estimate changes when student and school level characteristics are controlled.

Figure 3.7 Variation of the conditional association between student achievement and coherence of TPD across Means-as-Outcomes HLM models for the US and UK (US&UK MM)



Source: PISA 2012 database

Notes: Outcome variable: mathematics score in PISA 2012; *p < .1, **p < .05, ***p < .01; Coherence of TPD is measured by a standardised factor score with metric invariance across the US and UK (US&UK MM).

As in the previous analyses with these two countries, results indicated that the size of the relationship was null or small (in general, less than 10 points) and when it was statistically different from zero, it showed a negative association. The unconditional model indicated that for students from different schools, one standard deviation increase

in the coherence of TPD was associated with a decrease of 6.7 (US) and 9 (UK) points approximately in the PISA assessment. This pattern of association held regardless exogenous characteristics of schools and students, in which case the size of the association was -5.1 in the US and -7.4 in the UK[43].

3.3.4.3. HLM models using the US MM

The final HLM models examined are based on the results of the CFA applied to the set of five items that composed the initial scale of coherence in TPD, which was found to be a valid measuring instrument for the US sample (e.g. US MM). In addition to the items included in the previous scales, this measurement model contains data on the presence of a standardised policy for mathematics that include a coherent link between TPD and other teaching resources. Table 3.14 shows the results of the mean-as-outcomes model when the respective standardised factor score was used in this country as a level-2 key explanatory variable under three successive conditions: bivariate association, inclusion of students' control variables and inclusion of school's control variables.

Table 3.14 Means-as-Outcomes HLM models for the US

	Coef	SE		Coef	SE		Coef	SE	
School-level variables									
Coherence of TPD	-6.4	(3.5)	*	-4.5	(3.0)		-5.1	(3.1)	
Administration (public)							-0.9	(11.3)	
Location							-4.4	(4.0)	
Class size							-0.6	(0.6)	
School size							0.0	(0.0)	*
Student-level variables									
Gender (male)				8.7	(2.5)	***	8.5	(2.6)	***
Immigrant				-0.5	(5.7)		-1.0	(5.8)	
SES				23.7	(1.9)	***	23.9	(2.0)	***
Intercept	481			474			491		
Between-school variance	47.1			38.5			38.0		
Within-school variance	77.1			74.1			74.3		

Source: PISA 2012 database

Notes: Outcome variable: Mathematics score in PISA 2012; SE=Standard Error; Weighted data; *p < .1, **p < .05, ***p < .01; Coherence of TPD is measured by a standardised factor score (US MM).

[43] As in the previous analysis, the contribution of the item related to "teachers' appraisals and TPD" that was removed from the original scale is detailed in Appendix L.

As found in the previous analysis, one standard deviation increase in the level of coherence of TPD was associated to 6.4 points less in the average achievement of US schools; however, when individual characteristics of students and their schools were controlled, the association of the key explanatory variable was close to zero.

In summary, the results generally indicated that the coherence of TPD was not likely to be associated to student achievement and in those cases in which its coefficient was statistically different from zero, it showed a small and negative direction. In particular, it is striking that mathematics performance in the US and UK was slightly lower in schools where head-teachers put greater effort in making TPD consistent with school goals (relative to schools that had less coherence of TPD). Likewise, it is worth noting that this aspect was likely to be not associated to the relatively high performance of Japanese and Finnish students in PISA 2012. On balance, these findings differ from the hypothesised positive direction of the association between the coherence of TPD and student achievement[44].

3.4. Discussion and conclusion

The cross-national analyses developed in this chapter have been aimed to provide an empirical inspection of the concept of coherence in TPD and its statistical association with student achievement. In order to examine how the coherence of TPD works across the US, UK, Japan and Finland, three specific questions were raised: do the variables comprising a measuring instrument of this construct operate equivalently across these countries? What is the performance of each nation in this dimension? Does a coherent approach to TPD in schools relate to student achievement? Combining both EFA and CFA together with MG-CFA to provide evidence of measurement invariance across countries, subsequent HLM analyses were adopted to examine the data for an association between student achievement and a latent construct of coherence in head-teacher's TPD reports.

[44] These interpretations of model estimates are generally consistent with the results yielded by the sensitivity analysis of the different methods of scaling (Appendix J) and the informativeness of weights analysis (Appendix K).

One of the main findings of this chapter indicated that only a group of items yielded measurement properties that enabled effective comparisons across all the four countries of interest. These questions requested head-teachers to indicate the steadiness to which they made sure that school goals were achieved (item 3), the regularity that they attended such events (item 4) and discussed with teachers their experiences in this area (item 5). Evidence of configural and metric invariance derived from this chapter suggested that a construct of the coherence of TPD measured through these observed variables is able to evoke a similar meaning among US, British, Japanese and Finnish head-teachers, and, in this sense, operate equivalently for cross-national comparisons.

On the contrary, those measurement models that added other indicators to the construct were not interpreted as part of the coherence of TPD among Japanese and Finnish school leaders. Interestingly, the measurement model that added data on the link between teacher appraisals and TPD (item 2) showed configural and metric invariance only between US and British head-teachers, whereas the initial model that also included information about the implementation of a standardised policy for mathematics (item 1) was only assumed by leaders in the US. In other words, for these two measurement models there is evidence of construct bias that cannot be neglected to compare the coherence of TPD across all the four countries under analysis.

It is probable that head-teachers in Japan and Finland consider that making TPD a coherent practice within schools is sufficiently accomplished by simply monitoring the implementation of these activities themselves (item 3 to 4). Linking TPD with teacher evaluation (items 2) or with a specific school policy (item 1) may be superfluous for them given the high-quality of the teacher's workforce in these two countries. In contrast, for English head-teachers, the coherence of TPD would increase when such opportunities are closely related to the individual performance of the staff (item 2), whereas in the US it would involve and additional effort on standardising such link with the curriculum in mathematics (item 1). The latter case of the English speaking countries may reflect their concern on efficiently managing TPD at the school level given the shortage of high-quality teachers, particularly in mathematics. This scenario demands that head-teachers implement extra measures to actively align TPD with the learning goals of schools.

Given that none of the measurement models yielded satisfactory estimates of strong invariance, it is important to remark that the average level of the coherence of TPD

in each country cannot be equivalently interpreted in a cross-national perspective. An alternative way of representing this aspect is that comparative appraisals of the performance of countries in relation to the coherence of TPD are meaningless if they are based in the national mean estimates of the factor scores. For example, sorting these four countries into a ranking of average performance of the coherence of TPD could not be empirically sustained.

In this context, the question about the contribution of the coherence of TPD to student achievement was cautiously examined in this chapter considering that the key explanatory variable showed three different versions that allowed valid estimates for one, two or all the four countries in each case. What I wish to emphasise here is that regardless the measurement model employed –as well as the method of scaling (Appendix J), the use of sampling weights (Appendix K) and the characteristics of students and schools used as controls- the coherence of TPD seemed to be weakly associated to the achievement of students in mathematics. In particular, the performance of US and British students tended to slightly decrease insofar as the coherence of TPD in schools was enhanced by their head-teachers, whereas this construct was likely to be not associated to the good results of Japan and Finland in the PISA assessment.

Thereby, the beneficial influence of the coherence of TPD on student achievement should be put into question. In this regard, the estimates that I present in this analysis disprove in two ways what studies in the field have been suggesting in relation to the impact of this feature on school outcomes. Firstly, given that coefficients calculated with data from the US revealed consistent inverse associations with the achievement of students, findings conflicted with the expected positive effect that has been continuously reported in this country (Desimone, 2009; Garet *et al.*, 2001; Penuel *et al.*, 2007). My analysis of the PISA 2012 data described that for the US students from different schools, the level of coherence of TPD was faintly detrimental to their performance in this subject –one standard deviation increase in the scale of coherence of TPD led in average to 5.2 points less in the assessment of mathematics across schools. It is possible to speculate in this case that all the efforts implemented by head-teachers may distract the work of teachers with students, because the staff becomes more accountable to their individual progress (item 2) and to the school policy for mathematics (item 1). To put another way, students in the US that attend schools with more coherent TPD would be taught by teachers that spend more time satisfying the expectations of head-

teachers, which would hinder their complete focus on improving their opportunities to learn.

A point that can be made to explain the disagreement with previous research is that the number of observed variables that were available in the school questionnaire of PISA was limited to capture the construct under analysis in this country. Indeed, it is worth noting that previous measures of the coherence of TPD developed in the US (Murray, 2012; Sebastian and Allensworth, 2012) contain a more ample list of components and observed variables that allows a more sensitive assessment of this complex latent factor. Likewise, the fact that teacher learning promoted by the coherence of TPD does not necessarily translate into better student outcomes can be attributed to the potential lack of other relevant dimensions related to the coherence of TPD and student achievement in the HLM models specified (see subsection 1.3). Further research in the US orientated to model this complex path of effective TPD with data about classroom practices, teachers' knowledge and their beliefs about teaching and learning, would be certainly fruitful to elucidate the actual scope of the coherence of TPD in the schools.

However disappointing the findings are, the strength of a comparative design of observational data is that it makes possible to assess this type of arguments with the actual running of different school systems and, thereby, shed light on the way that complex constructs are perceived and associated across varying contexts. In this regard, the findings reported in this chapter for the cases of the UK, Japan and Finland also refuted the expected positive contribution of the coherence of TPD. On one hand, because the HLM models here implemented were able to detect that the specific contribution of the key explanatory variable was both negative and small (US and UK) or definitively led to neutral results (Japan and Finland) by taking advantage of a number of control variables. On the other hand, because the measurement models employed to cross-nationally estimate the coherence of TPD yielded adequate properties that enabled an efficient analysis of the influence of this construct on student achievement.

The point is that using equivalent measures to compare countries, the pattern of results that emerges indicate that in Japan and Finland the coherence of TPD does not make any difference to student outcomes in mathematics, whereas in the UK and US the trend is to some minor extent unfavourable. The matter becomes relevant taking into

account that the prevalence of the observed indicators of the coherence of TPD was comparatively higher in the English speaking countries than in Japan and Finland. The probability of reverse causality cannot be discarded as it may well be the case that instead of a negative impact of the coherence of TPD, it could rather be that head-teachers that more actively make TPD consistent with the goals of their schools are those leading schools with low achiever students. Unfortunately, the cross-sectional design of PISA does not allow to control for by individual variation in scores as longitudinal studies do by including components of prior or posterior achievement of students in order to fully contrast the contribution of the key explanatory variable (Goldstein, 2008).

Taken as a whole, this chapter reveals that the extent to which TPD activities are actively managed to be consistent with the overall goals of schools, in particular with those related to students' learning, is unlikely to make a positive difference on mathematics student achievement, at least as measured in an international large-scale assessment like PISA 2012. In contrast to what common sense and research in the US have reported so far, the achievement of US and British students is observed to slightly decrease insofar as the coherence of TPD in schools improves, whereas this construct shows no relationship with the outcomes of Japanese and Finnish schools.

Considering that a similar pattern of results was reported in the previous chapter (e.g. mathematics content-focused TPD showed a null or slightly negative association with student achievement), the next chapter examines whether the remaining three quality features of TPD (e.g. collective participation, active learning and duration) are associated with the way teachers teach in the classroom.

Chapter 5

Quality features of teacher professional development, teacher learning practices and classroom instruction: a cross-national analysis of TALIS 2013

4.1. Introduction

This chapter seeks to examine whether the remaining three quality features of TPD (*collective participation, active learning* and *duration*) or the practices of teacher learning undertaken in schools are correlated with classroom instructional approaches, using data from the 2013 cycle of the Teaching and Learning International Survey (TALIS). Quality features of TPD refers to a number of distinctive attributes of in-service teacher training that have been identified as empirically associated with positive teaching practices (Caena, 2011; Desimone, 2009), and successfully tested with national samples of teachers in studies implemented in the US and England (Garet *et al.*, 2001; Opfer and Pedder, 2011b). Five measurable dimensions of TPD are typically nowadays examined in the literature[45] –e.g. *coherence, content focus, active learning, extended duration* and

[45] See page 14.

collective participation (Desimone, 2009)- which proceed from specific theories that account for the effectiveness of these experiences[46] (Van Veen, Zwart and Meirink, 2012; Wayne *et al.*, 2008).

In the context of the *theory of teacher change*, TPD is supposed to play a major role in facilitating the autonomous and collaborative development of teachers' practices and beliefs about learning (Meirink *et al.*, 2009b; Putnam and Borko, 2000), thereby organisational and learning activities should be organised in a way that facilitates this process in order to improve classroom practices. The extent to which TPD is implemented with *active learning, extended duration* and *collective participation* is essential because it determines the manner by which teachers improve their practice. In other words, whereas *content focus* and *coherence*[47] describe what is to be learned and under which organisational conditions, *active learning, extended duration* and *collective participation* define the way that the content of the training is presented to the staff (Loucks-Horsley and Matsumoto, 1999). This chapter attempts to clarify whether these three latter quality features of TPD are associated to teachers' practices in the classroom.

Teacher learning practices, in turn, summarise the experiences of co-operation undertaken by teachers in their daily interaction in the school setting and in response to the particular needs of the teaching and learning environment (Bakkenes, Vermunt and Wubbels, 2010; Hardy, 2010; McRae *et al.*, 2001; Meirink *et al.*, 2009a). Whilst TPD is generally delivered by external providers to schools, teacher learning practices are naturally developed by teachers themselves through exchanging teaching materials, engaging in discussions about aspects of teaching, observing other colleagues' lessons and providing feedback. Recent research suggests that teachers' participation in these practices is importantly related to how they teach their students in the classroom (de Vries, Jansen and van de Grift, 2013; de Vries, van de Grift and Jansen, 2013; Opfer, Pedder and Lavicza, 2011a), however there are no studies addressing the potential joint influence with TPD on instruction.

The study of the relative influence of TPD and teacher learning practices on school processes is important for a number of reasons. Firstly, because those components

[46] See Appendix A.

[47] I have examined the relationship between these two features and student achievement in the two previous empirical chapters.

of TPD that are necessary to make a difference in the classroom merit further investigation in the context of the actual implementation of teacher learning practices in schools. In cases where teamwork among teachers is effective and seen as part of the ongoing learning process throughout their careers, even the most intense TPD activity may play a superfluous role in order to boost teachers' performance. This is relevant also because the successful implementation of the quality features of TPD can involve further and major efforts for school systems. For example, TPD with *extended duration* might conflict with the actual distribution of teaching loads and the traditional short-term implementation of these activities (Schwile, Dembele and Schubert, 2007), which would demand a change in the structure and culture for learning in schools.

In all these cases, the engagement in teacher learning practices is worth investigation as it may work, compared to the quality features of TPD, as a better predictor of instructional practices. Common sense suggests that teacher learning practices are to some extent independent from the exposure to the quality features of TPD. Indeed, the quality of TPD cannot be formally anticipated by teachers because only after experiencing these activites the features of its delivery can be judged. In other words, teachers who are prone to participate in practices of teacher learning are not necessarily prone to select into doing TPD with more *active learning, collective participation* and *longer duration*. Therefore, as both methods of teacher learning (e.g. teacher learning practices and high quality TPD) seem to be unrelated, it follows that their independent association with instruction can be systematically evaluated. Thereby, the strength of the link between classroom teaching practices, the quality features of TPD and teacher learning practices can be properly compared to orientate policy and practice in this area[48].

[48] This is an advantage of using the quality features of TPD as predictors of teaching practices, compared to using the "type of activity" (e.g. courses, workshops, networks, etc.; see Appendix A, introduction). The type of activity is typically related to the participation in teacher learning practices. To illustrate, in all the 23 countries analysed in the TALIS 2008, teachers that collaborated more with their colleagues were more likely to engage in networks and peer observation (Vieluf *et al.*, 2012) –see also footnote 50, page 112. Teachers can decide on selecting into such TPD activities because they know in advance what they will experience and can choose according to their needs or liking. Therefore, when the association of TPD (operationalised by the type of activity) with instruction is evaluated, the participation in teacher learning practices must be necessarily controlled in order to obtain a precise estimate of its specific contribution.

On the other hand, it is worth noting that the benefits of the quality features of TPD have been so far only studied in the US and English contexts[49], so the question remains in relation to their association with teaching practices in diverse countries. Comparative evidence is critical in this regard because countries often have to choose between funding national TPD programmes or improving school conditions to develop effective teachers' professional communities (Kruse, Louis and Bryk, 1994; Newmann, 1994; Schwile, Dembele and Schubert, 2007). Certainly, an alternative view is needed to determine what type or amount of TPD is sufficient in school systems with different levels of teacher learning practices. By implication, the design of effective in-service training can be refined at the national level through the examination of the complex interdependence between these two methods of teaching improvement. In concrete terms, a cross-national approach to this issue may contribute to the contextualisation of guidelines and standards for quality teacher learning and the respective quality assurance procedures.

At this juncture, the two cycles of the TALIS programme (OECD, 2009a; OECD, 2014d) have facilitated significant progress on understanding how TPD, teacher learning practices and classroom practices are associated in more than twenty nations. For example, several studies derived from both rounds (Hendriks *et al.*, 2010b; OECD, 2013a; OECD, 2014d; OECD, 2015) have remarked that traditional forms of TPD (e.g. workshops, seminars, etc.) are still more prevalent across countries than innovative designs (e.g. teachers' networks, mentoring, etc.). Furthermore, they have shown that in the majority of the countries assessed, such innovative forms of TPD are significantly associated with teacher learning practices[50], as well as with specific instructional

This is not the case of the quality features of TPD. The particular association of TPD (operationaled by these indicators) with instruction can be properly assessed and, what is more important, compared against the size of the specific contribution of teacher learning practices.

[49] Garet *et al.* (2001) reported that one unit increase in the measure of *active learning* was associated with .14 standard deviations increase in the knowledge and skills of teachers in the US, while the measures of *content focus*, *coherence* and background characteristics of teachers were held constant. Likewise, Opfer and Pedder (2011b) found that secondary teachers in the highest performing schools in England took part in TPD activities with longer *duration*, more *active learning* and *collective participation*.

[50] In TALIS 2008, the correlations between the participation in co-operative forms of TPD and measures of teacher learning practices (labelled as "professional learning communities") were positive in all the 24

practices[51] (OECD, 2009a; OECD, 2013a; OECD, 2014d; Vieluf *et al.*, 2012). On the other hand, recent analyses of the 2013 round of the survey (OECD, 2015; Opfer, 2015) have revealed that job-embedded TPD has a stronger association with the self-efficacy of teachers than non-job embedded TPD, which would suggest that innovative forms of TPD are related to the quality of teaching.

However, all of these studies have operationalised TPD in terms of the type of activity attended by teachers (workshops, seminars, teachers' networks, etc.), thus there are no analyses that focus on how such events were implemented. The most recent cycle attempts to redress this gap and investigate aspects of *active learning*, *collective participation* and *extended duration*[52]. This extension opens an interesting opportunity to analyse whether teacher learning practices or the quality features of TPD are related to the way teachers develop their lessons in the classroom. Therefore I examine in this chapter the association between specific classroom practices and each of these two approaches (teacher learning practices and the quality of TPD) across the the US, England, Japan and Finland. The question to be asked here is: does TPD carried out either with greater degrees of *active learning*, *collective participation* or longer *duration* relate to specific classroom teaching practices, when the participation in teacher learning practices is taken into account? This specific question contributes to respond the overarching question of the thesis, which seeks to examine whether differences in teachers' exposure to these three features of TPD might be related with differences in the teaching methods they use with their students[53].

In sum, the chapter aims to examine whether the positive association found in the US and England (Garet *et al.*, 2001; Opfer and Pedder, 2011b) between the quality

countries evaluated; coefficients ranged between .08 (Norway) and .76 (Estonia), approximately (OECD, 2013a).

[51] To illustrate, in approximately half of the countries assessed in TALIS 2013, teachers that participated in individual or collaborative research activities were: (a) 27% (Italy) to 88% (Serbia) more likely to implement project-based learning; and (b) 23% (Poland and Spain) to 98% (Norway) more likely to use ICT in the classroom.

[52] This is certainly curious. The organisers even developed a complex index based on these questions which yielded adequate measurement properties for cross-national analyses (OECD, 2014e). However, neither the individual items nor the index were used for further analyses in any of the two official reports (OECD, 2014a; OECD, 2014d).

[53] See page 23.

features of TPD and educational outcomes is replicated with current data, and whether such results are also evident in Japan and Finland, two top performing countries recognised as with high-quality TPD (Robinson, 2014; Stewart, 2011; Williams, 2013).

It must be noted that participation in TPD is compulsory in all these four countries, however the role of TPD is different due to the characteristics of the teaching profession in each case (see Appendix B). For instance, the identification of the key features of high-quality TPD may be more relevant in the English-speaking countries because their systems focus more in supporting in-service teachers than increasing the requirements to become teacher. Teachers' shortage is a serious concern in these countries, thus policies that effectively enhance the quality of the teacher workforce are critical. Conversely, there is surplus of applicants for teaching positions in Japan and Finland, and the quality of their staff is regarded as comparatively outstanding (Hanushek, Piopiunik and Wiederhold, 2014), whereas their results are systematically high in international large-scale assessments of students' achievement (Mullis *et al.*, 2012a; OECD, 2014b). However, the contribution of high-quality TPD to teaching practices might be hindered by other aspects in these two countries, such as the overloading of teachers' workforce (Japan) or the low attractive of TPD activities (Finland).

Taking into account these contextual conditions, this chapter examines whether teacher learning practices and such attributes of TPD also makes a difference to the same instructional practices, as reported by teachers in these countries. By statistically modelling these relationships in recent international large-scale data produced by the TALIS 2013, this secondary analysis applies an Ordinal Regression Model (ORM) (Agresti, 2002; 2007; Long and Freese, 2006; Winship and Mare, 1984) that takes into account the complex design of the survey. Thereby, the chapter provides statistical evidence to compare the role of these key components of the quality of TPD and teacher learning practices in all the four countries selected.

In the next section I describe the methodological strategy to be adopted to answer the key question above, and the relevant features of the dataset to be analysed, i.e. TALIS 2013. Section 4.3 provides estimates of association with teaching practices, which are followed by a discussion of findings and conclusions in section 4.4.

4.2. Data sources and methodological strategy

4.2.1. Survey design

TALIS is a programme of surveys orientated to monitor every five years the teaching and learning environments from the educational systems of the OECD country members and its partner economies (OECD, 2010; Rutkowski *et al.*, 2013). The main international target population of TALIS are classroom teachers employed in lower secondary education (e.g. ISCED 2, equivalent to "Key Stage 3" in England)[54]. In order to collect representative data from this population, TALIS implements a two-stage stratified sampling procedure: in the first stage, a minimum of 200 schools are selected from each national frame by systematic random sampling with probability proportional to size; in the second stage, a minimum of 20 teachers are randomly selected within each school. In TALIS 2013, teachers and head-teachers from 34 countries took part of the survey (OECD, 2014e).

In order to preserve the desired representation of the target populations, TALIS organisers set minimum standards of participation for the sampled schools (75%) and teachers within them (75%). Countries with response rates below the standard are allowed to improve these numbers by substituting with replacement schools, which correspond to the next school in the national lists of eligible units. These lists have sorted schools according to different characterisitics (e.g. geography, source of financing, size, etc.) in order to warrant a proportional representation (i.e. implicit stratification). In TALIS 2013, the response rates of the countries of interest of this thesis were satisfactorily accomplished after replacement. Only schools in the US were the exception (62%), which unfortunately represents a limitation for the analyses developed with this sample. For this reason, US estimates were excluded from calculations of international averages in the analyses of the official reports[55].

[54] TALIS 2013 also offered countries to analyse data from their primary (ISCED 1) and upper secondary (ISCED 3) schools. Among the countries of interest of this thesis only Finland chose these additional options.

[55] Nonetheless, the organisers considered such proportion as a fair participation, provided that more than a half of schools participated and that 83% of teachers responded within them (OECD, 2014e).

Sampling weights are calculated in each round by the organisers in order to correct for the unit non-response at the school and teacher level, thereby a specific weighting factor (Rust, 2013) informing the probability of selection and adjustments for nonparticipation is assigned to each school and teacher. By applying these inverse probability weights to the variables of interest, estimates are adjusted to be representative of the national target populations (OECD, 2014e). Therefore, weighting factors are employed in every analysis included in this chapter.

Finally, it is worth noting that teachers that have attended at least one TPD activity during the 12 months previous to the survey are the population of interest of this chapter. This is because those who have recently experienced TPD are able to rate the key explanatory variables here assessed –i.e. the quality features of TPD. Thus, the analysis conditions upon the filtered design of the teacher questionnaire which define as "missing by design" (de Leeuw, 2001) those cases that do not comply with this condition[56]. In other words, out of the total of teachers that took part on TALIS 2013, the analysis is based on data collected from participants that undertook some type of TPD in the 12 months prior to the survey.

The disadvantage of this approach is that sample selection is introduced in the model, which may differ by country, given different rates of participation in TPD. In other words, some individual characteristics of teachers that attended TPD are likely to be related to their teaching methods and the quality features of TPD, which could bias the estimates of association between the key explanatory variables and outcomes. Table 4.1 details the number of schools and teachers sampled in each of the countries of interest, followed by the percentage of teachers reporting that they took part in any form of TPD. According to these data, the target population of this chapter corresponds to 95% of the original sample from the US, 92% from England, 83% from Japan and 79% from Finland.

[56] The section related to TPD in the TALIS teacher questionnaire includes the following indication: "If you did not participate in any professional development activities during the last 12 months, please go to Question [x]" (International Project Consortium, 2013, p. 30).

Table 4.1 Samples of schools and teachers in TALIS 2013 and percentages of teachers that attended TPD as reported for the US, England, Japan and Finland

	US	ENG	JPN	FIN
Schools	122	154	192	146
Teachers	1926	2496	3484	2739
Participation in TPD[a]	95%	92%	83%	79%

Source: OECD (2014e) and TALIS 2013 database.
Notes: [(a)] weighted data.

4.2.2. Classroom teaching practices

Teachers' instructional practices were deemed as the second most important theme to be explored in TALIS 2013 by the participant countries out of the twenty policy foci suggested by the OECD (2014e). Data about this aspect were included in the teacher questionnaire of the study (International Project Consortium, 2013), which also collected information on teachers' background, the school where they work, their experiences of TPD and the professional feedback received in the school. Among the questions orientated to collect information about their classroom teaching practices teachers were requested to indicate the frequency that they put into practice particular instructional approaches to promote students' learning. The information requested was restricted to their experience in a particular lesson taught during the week previous to the survey, i.e. "target class" (OECD, 2014e, p. 48).

This chapter uses the same three instances of classroom teaching practices that were specifically analysed in the official report in order to complement the discussion ensued by the authors about their relationship with teacher learning. The measured items are listed as follows:

- Item 1. Students work in small groups to come up with a joint solution to a problem or task.
- Item 2. Students work on projects that require at least one week to complete.
- Item 3. Students use ICT (information and communication technology) for projects or class work[57].

[57] Likert-type scales recoded as: 0= Never or almost never; 1=Occasionally; 2=Frequently; 3=In all or nearly all lessons.

These instructional approaches are conceptualised by the OECD as “active practices” (2014d, p. 154), in the sense that they give a central role to the students in their own process of learning. Further, they are considered key teaching practices that develop crucial skills for future success in higher education and the labour market. On the other hand, each item represent particular aspects of teaching and derive from different theory sources, thus it is not necessarily expected that they measure a common underlying construct related to classroom instruction.

For example, the first statement alludes to the concept of *cooperative learning*, which is described as a teaching method that facilitate the achievement of shared goals within a small group of learners (Johnson and Johnson, 1989; 2009). A substantial body of literature has indicated that cooperative learning strategies outperform competitive and individualistic approaches in different measures of students’ outcomes, such as achievement, problem-solving skills and cognitive development (Johnson and Johnson, 1974; Marzano, Pickering and Pollock, 2001). Therefore, it is perhaps not surprising that the use of cooperative small groups has been extensively implemented in several educational levels and systems throughout the world (Johnson and Johnson, 2009).

The second and third items illustrate the implementation of the *project-based learning* approach, introduced by Kilpatrick (1918) nearly one century ago to highlight the importance of students’ motivation for the development of successful learning activities in the classroom. In this case, the term "project" accounts for instances of class work in which –with the facilitation of the teacher- the contents and processes of learning are suggested, planned, executed and evaluated by the students themselves. Nowadays, the method is usually carried out with the exploration of real-life problems as a mean to engage students in challenging learning activities that require application of new knowledge. Research indicates a positive effect of this strategy on students’ learning, notwithstanding that well-developed projects demand adequate conditions for success, such as time allocation and a collaborative culture for learning in the school (David, 2008). A number of studies have indicated that the use of project-based learning (PBL) in conjunction with information and communication technologies (ICT) seems particularly effective and promising (Blumenfeld *et al.*, 1991; Chang and Lee, 2010; De La Paz and Hernández-Ramos, 2013).

Table 4.2 illustrates the distribution, means, standard deviations, inter-item and bivariate polychoric correlations between these three items in every country of interest.

Table 4.2 Distribution, means, standard deviations and polychoric correlations for outcome variables by country

	Item	Distribution (%)[(a)]				m%	M	SD	Small groups	PBL	$\|r_{ij}\|$
		0	1	2	3						
US	**Small groups**	8	38	42	12	21	1.60	0.80	1		0.29
	PBL	19	44	25	12	21	1.29	0.91	0.23	1	
	ICT	17	37	34	12	21	1.40	0.90	0.24	0.40	
ENG	**Small groups**	4	36	45	15	20	1.70	0.77	1		0.17
	PBL	20	40	26	13	20	1.32	0.94	0.08	1	
	ICT	14	48	29	9	20	1.32	0.82	-0.00	0.48	
JPN	**Small groups**	18	49	26	7	11	1.22	0.81	1		0.17
	PBL	54	32	8	6	11	0.67	0.87	0.09	1	
	ICT	59	31	7	3	11	0.54	0.75	0.23	0.25	
FIN	**Small groups**	9	52	31	8	16	1.39	0.76	1		0.32
	PBL	45	40	9	6	17	0.76	0.85	0.15	1	
	ICT	24	56	17	3	17	0.98	0.73	0.28	0.53	

Source: TALIS 2013 database

Notes: weighted data;
Small groups=Use of small groups; PBL=Use of projects-based learning; ICT=Use of information and communication technology and PBL;
[(a)] Distribution of valid cases; 0= Never or almost never; 1=Occasionally; 2=Frequently; 3=In all or nearly all lessons weighted data;
m%=Percentage of missing cases; M=Mean; SD=Standard deviation; $|r_{ij}|$=Average inter-item correlation (absolute value).

According to these data, although teachers report different frequencies of implementation of these methods of instruction, the majority use them either occasionally or frequently in each country. The only exception in this regard is the use of project-based learning and ICT in Japan and project-based learning in Finland, in which cases the distributions indicate that about a half of the teachers do not employ these methods in their classrooms (54%, 59% and 45%, respectively)[58]. In this context, the percentage of missing data in each variable is in the range of 11% to 21% across all the four countries of interest[59]. Finally, the average inter-item polychoric correlations indicate weak albeit

[58] It is striking that among all the countries evaluated in TALIS 2013, the Finnish and Japanese teachers (along with their Croatian counterparts) reported the lowest levels of use of the three classroom teaching practices here selected (OECD, 2014d).

[59] Having such levels of missing data is certainly not an ideal situation. A likely implication of this scenario is that results might be biased because estimates could systematically differ between teachers that reported their teaching practices and those who did not. Replacing such not reported observations with a set of

positive associations among these three variables for each school system. Only the correlation between project-based learning and ICT yielded coefficients of moderate association in the US (.4), England (.48) and Finland (.53), which is an expected figure given the conceptual framework shared by these two variables.

4.2.3. Key explanatory variables

Data about the "quality features of TPD" *collective participation*, *active learning* and *duration* are also drawn from specific items contained in the teacher questionnaire of the study. Among the questions used to collect information about recent experiences of TPD, the participants were requested to indicate the extent to which some characteristics of this provision were present during these events. What continues are the statements of the question utilised and each item:

Considering the professional development activities you took part in during the last 12 months, to what extent have they included the following?

- Item 1. A group of colleagues from my school or subject group.
- Item 2. Opportunities for active learning methods (not only listening to a lecturer).
- Item 3. An extended time-period (several occasions spread out over several weeks or months)[60].

As mentioned above, these items represent some of the dimensions of TPD that are currently examined in the specialised literature, respectively: collective participation, active learning and extended duration (Caena, 2011; Desimone, 2009). *Collective participation* refers to the interaction of groups of teachers from the same school that is necessary in TPD activities to develop collaborative and meaningful learning amongst professionals. Similarly, TPD programmes based on *active learning* provide opportunities orientated either to observe, design, perform or expose teaching practices, as a manner to engage teachers in inquiry-based learning experiences. On the other hand, it is argued that TPD programmes with longer *duration* are more effective, both with

plausible values –i.e. Multiple Imputation (de Leeuw, 2001; Rubin, 1996; Schafer and Olsen, 1998)- might efficiently address this issue.

[60] Likert-type scales recoded as: 0= Not in any activities; 1=Yes, in some activities; 2= Yes, in most activities; 3=Yes in all the activities.

regard to the overall amount of time that the activity takes and the total amount of hours spent. Studies implemented with national samples of teachers both in the US (Birman *et al.*, 2000; Garet *et al.*, 2001) and England (Opfer and Pedder, 2011b) suggest a positive association between these features and school outcomes. Table 4.3 compares the distributions, means, standard deviations, inter-item and bivariate polychoric correlations between these three items for each country.

Table 4.3 Distribution, means, standard deviations and polychoric correlations for key explanatory variables (collective participation, active learning and duration) by country

	Item	Distribution (%)[(a)]				m%	M	SD	CollPar	ActLea	$\|r_{ij}\|$
		0	1	2	3						
US	**CollPar**	10	36	32	22	2	1.67	0.93	1		0.42
	ActLea	12	44	31	13	3	1.45	0.87	0.49	1	
	ExtDur	38	38	16	8	3	0.94	0.92	0.30	0.46	
ENG	**CollPar**	13	43	28	16	4	1.49	0.91	1		0.42
	ActLea	14	50	28	8	5	1.29	0.80	0.45	1	
	ExtDur	47	34	14	5	5	0.77	0.87	0.34	0.46	
JPN	**CollPar**	26	43	25	6	18	1.08	0.84	1		0.29
	ActLea	17	52	26	5	18	1.19	0.76	0.25	1	
	ExtDur	79	13	5	3	19	0.33	0.72	0.26	0.37	
FIN	**CollPar**	17	41	26	16	4	1.40	0.95	1		0.25
	ActLea	21	47	24	8	5	1.18	0.85	0.26	1	
	ExtDur	63	24	9	4	5	0.54	0.83	0.07	0.43	

Source: TALIS 2013 database

Notes: weighted data;
CollPar=Collective participation; ActLea=Active learning; ExtDur=Extended duration;
[(a)] Distribution of valid cases; 0= Not in any activities; 1=Yes, in some activities; 2= Yes, in most activities; 3=Yes in all the activities;
m=Missing cases; M=Mean; SD=Standard deviation; $|r_{ij}|$=Average inter-item correlation (absolute value).

The reported average ratings suggest that experiences of TPD with *collective participation* and *active learning* are more prevalent than activities with *extended duration*, although the majority of teachers report that these two characteristics were only featured in some of the activities attended over the year. Accordingly, most of the teachers in England (47%), Japan (79%) and Finland (63%) reported having attended TPD activities without *extended duration* (this proportion corresponds to 38% in the US).

The percentage of missing data is negligible across countries, whilst for Japan it is in the range of 18% to 19% across the three variables.

The average inter-item polychoric correlations reveal relatively weak (.25, Finland and .29, Japan) and moderate (.42, England and US) relationships among these three indicators. Detailed examination of the matrices reveals that in England all the items are moderately correlated, whereas in the US this only applies to the two relationships of *active learning* (.49 with *collective participation* and .46 with *extended duration*). Both, in Japan (.37) and Finland (.43), the only moderate associations are found between *extended duration* and *active learning*.

4.2.4. Teacher learning practices as a covariate

Finally, "teacher learning practices" are measured using the "Co-operation among teaching staff scale" developed by the organisers of TALIS 2013 (OECD, 2014e). This is a latent factor composed by two sub-scales, namely: "Exchange and coordination for teaching" and "Professional collaboration". Each of these sub-scales is measured by a group of variables of the teacher questionnaire. The majority of such items were utilised in the previous round of TALIS to analyse relevant characteristics of schools as "professional learning communities" –e.g. co-operation, shared vision, focus on learning, reflective inquiry and de-privatisation of practice (Vieluf *et al.*, 2012). The items in the 2013 questionnaire included each sub-scale under the question "On average, how often do you do the following in this school?" (International Project Consortium, 2013, p. 19) are listed below:

Exchange and coordination for teaching:

- Item 1: Exchange teaching materials with colleagues.
- Item 2: Engage in discussions about the learning development of specific students.
- Item 3: Work with other teachers in my school to ensure common standards in evaluations for assessing student progress.
- Item 4: Attend team conferences.

Professional collaboration:

- Item 5: Teach jointly as a team in the same class.

- Item 6: Observe other teachers' classes and provide feedback.
- Item 7: Engage in joint activities across different classes and age groups (e.g. projects).
- Item 8: Take part in collaborative professional learning[61].

As reported in the technical report of the study (OECD, 2014e), these two sub-scales yielded acceptable reliability estimates –e.g. Cronbach's alpha above .60- in all the four countries of interest of this study, with the only exception of the "Professional collaboration" sub-scale in the Japanese sample that had relatively poor internal consistency (Cronbach's alpha = .50). All in all, this did not affect the adequacy of the fit of the measurement model, which is described with satisfactory metric invariance for cross-national analyses by the TALIS 2013 organisers. Thereby, values corresponding to the engagement in teacher learning practices were assigned to each individual in the dataset, which I standardised within each country to have a mean of zero and a standard deviation of one for ease of interpretation[62]. The percentage of missing data in this variable across the four countries under evaluation is less than 4% (England).

It is worth noting that teacher learning practices are suggested as a covariate in this chapter provided that the engagement of teachers in such activities is not importantly related with their exposure to different levels of the quality features of TPD. Regarding this latter aspect, Table 4.4 displays the polychoric correlations between each feature and the covariate in all the four countries of interest.

Table 4.4 Polychoric correlations between teacher learning practices and the quality features of TPD (collective participation, active learning and extended duration) by country

Key explanatory variable	US	ENG	JPN	FIN
Collective participation	0.22	0.10	0.14	0.12
Active learning	0.22	0.22	0.14	0.10
Extended duration	0.19	0.20	0.11	0.16

Source: TALIS 2013 database

Notes: Outcome variable: teacher learning practices (covariate); weighted data.

[61] Likert-type scales coded as: 1= Never; 2=Once a year or less; 3= 2-4 times a year; 4= 5-10 times a year; 5=1-3 times a month; 6= Once a week or more.

[62] I also standardised this variable across countries using the whole pooled dataset. The corresponding estimates did not particularly differed from the within countries standardisation.

Although all the coefficients indicated a positive associations between each pair of variables, their small magnitude (between .1 and .22) showed that differences in the participation in teacher learning practices are not necessarily related with the quality features of TPD. Consequently, the independent contribution of the key explanatory variables (collective participation, active learning and extended duration) and the covariate (teacher learning practices) to the use of different teaching methods in the classroom is specifically estimated and compared.

Considering the characteristics of the variables presented above, it is clear that their self-reported nature might become an important source of measurement error that could hinder the validity of the results. With this point in mind it is worth inquiring, for example, whether survey methods can accurately provide information about teaching methods implemented in the classroom. Previous studies (Burstein *et al.*, 1995; Mayer, 1999) have shown that aspects such as the quality of the engagement of teachers with education reforms or the exact amount of time allocated to each method is difficult to be reported with enough precision by the teachers themselves. However, if these topics are not part of the focus of the investigation, then survey data can provide an accurate indication of the prevalence and combination of the general teaching approaches implemented in the classroom with adequate levels of consistency and generalisability.

For example, Mayer (1999) compared the answers of 124 secondary teachers to questions regarding their instructional practices and measures derived from classroom observations undertaken by independent researchers. The results indicated a strong and positive correlation (Pearson's r = .85) between these two sources of data, which suggested that self-reported practices adequately reflect those reported by external observers. Likewise, the review of specialised literature developed by Desimone (2009) featured several studies in favour of the convergence among data collected through observations, interviews and surveys when the latter focuses on factual information (instead of evaluative) about teacher learning. Thereby, as long as the information requested about the experiences of TPD is descriptive (about facts) and not based on personal judgements about facts, self-reported data generally elicit similar evidence to that gathered by means of the other two research methods. Although TALIS data cannot be contrasted with alternative sources of information, the literature surveyed here would appear to suggest that the variables included in this chapter are reasonably valid.

4.2.5. Analytic strategy

The three outcome variables of this study are operationalised as Likert-scaled items based on four ordered levels that represent frequencies of implementation of classroom teaching practices over a one year period –e.g. "Never or almost never", "Occasionally", "Frequently" or "In all or nearly all lessons". Provided that such levels represent a meaningful order of categories, I adopt an Ordinal Regression Model (ORM) (Agresti, 2002; 2007; Long and Freese, 2006; Winship and Mare, 1984) to analyse for each country of interest the relationships between each outcome, the key explanatory variables and the covariate.

Under an ORM analysis, the cumulative probability across the sequence of levels of the outcome variable is used to reflect the order of the categories of implementation. In this chapter, the cumulative probability means the probability of that teachers from a particular country indicate one specific level of implementation of a teaching method along with the preceding levels in the order. For example, the cumulative probability that a teacher in a given country will report a "Frequent" implementation of a teaching method is equivalent to the sum of the probabilities that such instructional practice is used "Never or almost never", "Occasionally" and "Frequently".

Based on this measure, the odds of reporting a particular level of implementation are calculated as the ratio between the cumulative probability of such level and the probability of the rest of levels that indicate more frequency of the outcome variable. To take the previous example, the odds for the level "Frequently" corresponds to the cumulative probability of having answered "Frequently", "Occasionally" or "Never or almost never" divided by the probability of the level "In all or nearly all lessons".

The natural logarithm of such odds –e.g. labelled as the logit function (Agresti, 2007)- can be modelled through a linear expression based on a number of intercepts and regression coefficients. Specific intercepts are estimated for each cumulative distribution of levels, so each parameter represents the logit function of a particular level of the outcome variable when all the predictors take the value zero. To illustrate, each model developed in this chapter estimates three intercepts because the outcome variables have

four levels each, so three cut points between cumulative distributions of levels are calculated[63].

In turn, the regression coefficients describe the extent to which one unit increase in the respective predictor changes the log-odds of reporting a higher level of implementation of the teaching method under analysis, while the rest of the variables in the model are held constant. For instance, the regression coefficient of *collective participation* indicates the degree to which one unit increase in such key explanatory variable changes the log-odds of implementing a particular teaching method with more frequency in the classroom (given that the covariate and the other key explanatory variables are held constant).

The estimated regression coefficients obtained in ORM describe the relationship between one specific level versus all higher levels of the outcome variable are assumed to be the same as those that describe the relationship between the next lowest level and all higher categories –and so forth. This is referred as the proportional odds assumption or the parallel regression assumption (Agresti, 2007; Long and Freese, 2006). In other words, the size and direction of each regression coefficient are supposed to be constant regardless of which cut point is taken as reference. Thus, for example, the degree to which one unit increase in the *collective participation* predictor changes the log-odds of implementing a teaching method "Never or almost never" versus "Occasionally" is similar to the degree of changing the log-odds of implementing it "Frequently", "Occasionally" or "Never or almost never" versus "In all or nearly all lessons". It is worth mentioning that such premise can become a strong assumption because the regression coefficients across the levels of the outcome variable do not always produce similar values.

ORM is used to estimate the logit function of each classroom teaching practice given the linear combination of the quality features of TPD (key explanatory variables) and teacher learning practices (covariate). The analysis proceeds by examining how the relationships between each predictor and the outcome variables change once different conditions are included in the distinct modelling steps. Model 1 includes a number of background characteristics of the sample as control variables –i.e. gender, teaching

[63] As the cumulative probability equals one in the highest level, there are no odds associated to this category of response.

experience and the completion of initial teacher training. I chose to use these variables because it is probable that teachers with different characteristics might experience different levels of the quality features of TPD. Model 2 restricts the ORM to a school fixed-effects model (Clarke *et al.*, 2010; Snijders, 2005) that utilises only variation within schools across teachers in order to remove any across-school differences in unobserved variables that could bias the results. This model relies on the intuition that the association between teaching practices and the quality features of TPD, as well as their link with teacher learning practices, can be observed with sufficient accuracy among teachers from the same school. In this case, the between-schools variance in the use of teaching methods can be dispensed for the ORM analysis, in favour of focusing the model only on the differences in these outcomes that are observed within-schools.

Finally, Model 3 introduces teachers' attitudes towards teaching and learning in order to separate out their contribution to outcomes from the contribution of the covariate. Teacher learning practices are closely related to this factor, to the extent that they are regularly studied as a single concept in the literature, i.e. Teachers' Orientation to Learning[64] (Opfer and Pedder, 2011a; Opfer, Pedder and Lavicza, 2011a; Remillard and Bryans, 2004). Holding this factor constant in the context of the school fixed-effects model will allow to better understand the specific contribution of teacher learning practices on the way that teachers teach in the classroom. For this purpose, I use the index of constructivist beliefs developed by the TALIS 2013 organisers (OECD, 2014e), which I have standardised within each national sample to have a mean of zero and a standard deviation of one.

In sum the inclusion of control variables and school fixed-effects is used to diminish the potential bias on the estimates of the association between the key explanatory variables and classroom teaching practices. Odds-ratios obtained from ORM modelling are reported to ease the interpretation of parameters for the quality features of TPD and the covariate[65].

[64] See Appendix A.

[65] Odds-ratios are the exponential of the corresponding regression coefficients; they share a similar interpretation, but in relation to the odds (not the log-odds) of observing the outcome variable with higher frequency. Thus, for example, they refer to the extent to which one unit increase in the predictor changes the odds of implementing the outcome variable "Never or almost never" versus "Occasionally", or "Frequently", "Occasionally" or "Never or almost never" versus "In all or nearly all lessons".

The final form of the model to be separately estimated for each country is:

$$\ln\left(\frac{Pr(y_i \leq j)}{1 - \left(Pr(y_i \leq j)\right)}\right) = \alpha_j + \beta_1 CollPar_{in} + \beta_2 ActLea_{in} + \beta_3 ExtDur_{in} + \beta_4 TLP_{in} + \beta_5 Block_{in} + \gamma_2 S_2 + \ldots + \gamma_n S_n + \beta_6 Att_{in} \quad \forall k$$

Where:

i = Teacher i;

j = Level j of the outcome variable;

n = School n;

k = Country k;

y = Classroom teaching practice (ordinal outcome variable);

CollPar = Collective participation (key explanatory variable);

ActLea = Active learning (key explanatory variable);

ExtDur = Extended duration (key explanatory variable);

TLP = Teacher learning practices (covariate);

Block = A set of three variables about teacher background characteristics;

S = School (dummy-variable);

Att = Index of constructivist beliefs.

Furthermore, the cross country variation between each estimate of the quality features of TPD in Model 3 is formally evaluated using an Independent Samples t-test analysis[66]. This procedure allows to evaluate whether the relationship between each key explanatory variable and the outcomes is different between each pair of countries. For

[66] As at this point the individual hypothesis stating potential differences between pairs of countries will be tested in multiple occasions, the 95% level of statistical confidence of each comparison might no longer represent the error rate of the set of comparisons among countries as a whole. Nevertheless, as the analysis involves only a small number of simultaneous planned hypotheses, I considered this not being a substantial issue and I preferred do not execute a multiple hypothesis testing approach based on Bonferroni correction (Shaffer, 1995).

this purpose, teacher learning practices and the index of constructivist beliefs is standardised across the pooled dataset of countries to have a mean of zero and a standard deviation of one[67].

The following section sets out to report the results of the ORM analysis for each of the three classroom teaching practices of interest. Firstly, odd-ratios are compared to answer the research question of this chapter. Secondly, results from the cross-national analysis of Model 3 parameters are presented. All the analytic strategy is executed using Stata 12 (StataCorp, 2011a).

4.3. Results

The results from the analytic strategy outlined above in section 4.2.5 are divided into four sub sections. The first three sub sections examine the relationship between each classroom teaching practice and the set of predictors. The final subsection addresses the cross-country analysis of the parameters yielded by the final Model 3.

4.3.1. Classroom teaching practice 1: Students work in small groups to come up with a joint solution to a problem or task.

This sub section develops the results for the first outcome variable of this chapter, e.g. the extent to which "students work in small groups to come up with a joint solution to a problem or task". As shown in Table 4.2, more than 82% (Japan) of teachers put into practice this method in their classrooms at least occasionally (92% US, 94% Finland and 96% England); however, a frequent or total use was more observed in the English speaking countries (54% US and 60% England) than in Japan (32%) and Finland (39%).

Table 4.5 shows the results of the three ORM models fitted with data from each of the four countries of interest. In each case, the first three rows present the odds-ratio,

[67] Item mean substitution (Eekhout, 2014; Hawthorne and Elliott, 2005) is used to maximise the amount of information available in al the ORM models. This procedure substitutes the missing values for a teacher on one variable with the weighted mean value of all other teachers that reported valid values for that variable. Dummy variables were created for each predictor to indicate those cases in which the missing values were replaced. These indicators were used in the corresponding ORM models.

standard errors and statistical significance of the key explanatory variables (using conventions *p < .1. **p < .05. ***p < .01), whereas in the following row the same estimates are reported for the covariate. Columns indicate the three models suggested in the analytic strategy, which are detailed in the three bottom rows with check symbols (✓).

Table 4.5 Ordinal Regression Models for the classroom teaching practice "Students work in small groups to come up with a joint solution to a problem or task" in the US, England, Japan and Finland

	Model 1			Model 2			Model 3			Model 1			Model 2			Model 3		
	OR	**(SE)**		**OR**	**(SE)**		**OR**	**(SE)**		**OR**	**(SE)**		**OR**	**(SE)**		**OR**	**(SE)**	
					US									**ENG**				
CollPar	0.99	(0.08)		0.92	(0.07)		0.91	(0.07)		0.89	(0.08)		0.87	(0.10)		0.87	(0.10)	
ActLea	1.02	(0.08)		0.99	(0.09)		1.00	(0.09)		1.07	(0.08)		1.08	(0.09)		1.06	(0.08)	
ExtDur	1.12	(0.09)		1.12	(0.11)		1.13	(0.11)		1.20	(0.08)	***	1.21	(0.10)	**	1.19	(0.10)	**
TLP	1.41	(0.11)	***	1.42	(0.15)	***	1.37	(0.14)	***	1.42	(0.07)	***	1.35	(0.07)	***	1.33	(0.08)	***
N	1416			1416			1416			1823			1823			1823		
					JPN									**FIN**				
CollPar	1.15	(0.06)	***	1.08	(0.07)					1.02	(0.06)		1.06	(0.07)		1.05	(0.07)	
ActLea	1.19	(0.08)	**	1.20	(0.08)	***				1.16	(0.08)	**	1.10	(0.08)		1.09	(0.08)	
ExtDur	1.01	(0.06)		1.03	(0.08)					1.04	(0.08)		1.06	(0.09)		1.06	(0.09)	
TLP	1.58	(0.07)	***	1.63	(0.09)	***				1.30	(0.06)	***	1.32	(0.08)	***	1.31	(0.08)	***
N	2551			2551						1790			1790			1790		
Teachers' characteristics	✓			✓			✓			✓			✓			✓		
School fixed-effects				✓			✓						✓			✓		
Teachers' attitudes							✓									✓		

Source: TALIS 2013 database

Notes: Outcome variable: Students work in small groups to come up with a joint solution to a problem or task; *p < .1. **p < .05. ***p < .01; OR=odds-ratios; SE=standard error; CollPar=Collective participation; ActLea=Active learning; ExtDur=Extended duration; TLP=Teacher learning practices; N=number of observations; weighted data; Model 3 did not converge in Japan.

According to these data, the amount of error in the odds-ratios reported for the US precluded a rejection of the null hypothesis in all the key explanatory variables and across all of the three models. Similarly, for example, the results of the features *collective participation* in England and Finland, *extended duration* in Japan and Finland, and *active learning* in England suggested that these measures were not likely to be not related to the implementation of the teaching method examined.

On the contrary, the odds-ratios reported by England for the *extended duration* score, as well as by Japan for the measure of *active learning*, would suggest that a one unit increase in such features of TPD would result in approximately 20% increase in the odds of implementing the outcome variable with a higher frequency, while the other predictors in all the three models tested were held constant. As the odds-ratios across the three models were particularly similar, it is likely that the relationship was not importantly biased by omitted variables at the school-level. Therefore, these results indicated that in England and Japan such features of TPD were positively associated to the implementation of cooperative learning.

It is worth noting that the odds-ratios observed in Model 1 by the features *collective participation* in Japan and *active learning* in Finland were not likely to reject the null hypothesis once school fixed-effects models were implemented. Although in both cases Model 1 estimates indicated that a one unit increase in these predictors would approximately increase the odds of using cooperative learning by 1.15:1, such results of this order of magnitude were no longer reproduced in the successive models executed. Consequently, this would suggest that the odds-ratios of the first model on these two quality features of TPD may have been importantly biased by confounders at the school level in the respective countries.

At this juncture, where only two quality features of TPD seemed to be directly associated to the outcome variable in just two countries, it is worth highlighting that the participation in teacher learning practices was positively associated with the implementation of cooperative learning in all the national samples and across all the statistical models planned. In this regard, for example, if a teacher from the US were to increase his or her teacher learning practices score by one standard deviation, his or her odds of implementing such instructional practice with more frequency would be expected to increase by 37% (Model 3). Likewise, whereas such expectation would improve approximately by one third in England (e.g. 33%, Model 3) and Finland (31%, Model 3),

this proportion is practically twice for Japanese teachers (63%, Model 2). These results are particularly relevant for Finland and Japan, the two countries with the lowest prevalence of this teaching practice in the classroom among the four nations evaluated. I wish to emphasise here that these results rejected the null hypothesis across the three models implemented.

In short, the participation in teacher learning practices was consistently associated in all the four countries with the use of cooperative learning. In contrast, only some quality features of TPD yielded significant results (*extended duration* and *active learning*), in some countries (England and Japan) and with a smaller magnitude, when compared to the association between teacher learning practices and the outcome. In particular, the chance of observing that students work in small groups to come up with a joint solution to a problem or task was apparently not associated to the quality features of TPD in the US. However, it was directly related to the participation of English teachers in TPD activities with longer *duration*, and Japanese teachers in events with more *active learning*.

4.3.2. Classroom teaching practice 2: Students work on projects that require at least one week to complete.

Results for the second outcome variable of this chapter, e.g. the extent to which "students work on projects that require at least one week to complete" (e.g. project-based learning) are reported in this sub section. As reported in Table 4.2, more than 19% (US) of teachers never or almost never implemented this instructional practice over the year with their target classes (about the same as 20% for England, yet higher in Finland, 45% and Japan, 54%). However, a frequent or total use was more likely to be observed in the English speaking countries (37% US and 39% England) than in Japan (14%) and Finland (15%). Table 4.6 reports the corresponding results of the ORM models for the US, England, Japan and Finland.

Table 4.6 Ordinal Regression Models for the classroom teaching practice "Students work on projects that require at least one week to complete" in the US, England, Japan and Finland

	Model 1			Model 2			Model 3			Model 1			Model 2			Model 3		
	OR	**(SE)**		**OR**	**(SE)**		**OR**	**(SE)**		**OR**	**(SE)**		**OR**	**(SE)**		**OR**	**(SE)**	
						US									**ENG**			
CollPar	1.08	(0.07)		1.03	(0.08)		1.02	(0.07)		0.97	(0.05)		0.96	(0.06)		0.96	(0.06)	
ActLea	0.91	(0.07)		0.99	(0.09)		0.99	(0.09)		1.08	(0.09)		1.11	(0.10)		1.10	(0.10)	
ExtDur	1.11	(0.08)		1.04	(0.10)		1.04	(0.10)		0.98	(0.06)		0.96	(0.07)		0.95	(0.07)	
TLP	0.96	(0.06)		1.01	(0.08)		0.96	(0.08)		1.11	(0.05)	**	1.11	(0.05)	**	1.09	(0.05)	*
N	1416			1416			1416			1817			1817			1817		
						JPN									**FIN**			
CollPar	1.09	(0.07)		1.14	(0.08)	*				0.88	(0.05)	**	0.86	(0.06)	**	0.85	(0.06)	**
ActLea	1.15	(0.08)	**	1.17	(0.09)	**				1.14	(0.07)	**	1.14	(0.08)	*	1.12	(0.08)	*
ExtDur	1.16	(0.07)	**	1.11	(0.07)					1.18	(0.07)	**	1.15	(0.09)	*	1.15	(0.09)	*
TLP	1.04	(0.05)		1.04	(0.06)					1.17	(0.06)	***	1.12	(0.07)	*	1.10	(0.07)	
N	2547			2547						1781			1781			1781		
Teachers' characteristics	✓			✓			✓			✓			✓			✓		
School fixed-effects				✓			✓						✓			✓		
Teachers' attitudes							✓									✓		

Source: TALIS 2013 database

Notes: Outcome variable: Students work on projects that require at least one week to complete; *p < .1. **p < .05. ***p < .01; OR=odds-ratios; SE=standard error; CollPar=Collective participation; ActLea=Active learning; ExtDur=Extended duration; TLP=Teacher learning practices; N=number of observations; weighted data; Model 3 did not converge in Japan.

Results in the US indicated no evidence of association between any predictor and the teaching method analysed, which contrasts with the results of the previous outcome variable that showed at least a positive association with the engagement in teacher learning practices. In this case, neither the quality features of TPD nor the covariate appeared to be related to the implementation of project-based learning. All of the estimates in the US formally failed to reject the null hypothesis. The same was true for the odds-ratio of the key explanatory variables in England and the covariate in Japan. None of these estimates reached statistical significance, which indicated that sampling variation could not be discarded in the context of the conditions attached to each model.

However, the parameters on teacher learning practices in England suggested that differences in such variable were directly related to different levels of implementation of project-based learning. To illustrate, if an English teacher was to increase his or her teacher learning practices score by one unit, his or her odds of using with more frequency this method in the classroom would be likely to increase by 11%. Given that estimates from the first two models for this variable were strikingly similar in magnitude, it can be argued that the parameter estimate for the covariate could not have been biased by school-level confounders. The fact that the estimate decreased to 9% in Model 3 would alternatively suggest that such association may have been partially explained by the beliefs of teachers about teaching and learning in the English sample.

Likewise, the odds-ratios reported by Japan on *active learning* indicated that holding constant the characteristics of teachers used as controls, a one unit improvement in such feature was significantly associated with a 15% increase in the odds of implementing the outcome variable with higher frequency. Two percentage points were added to this value in Model 2, which may indicate that the association on *active learning* was slightly stronger for teachers working in the same schools.

Finland represents a singular case in which almost all the predictors included in the model seemed to be associated, either negatively or positively, to the teaching method evaluated. To illustrate, the odds-ratios of the three quality features of TPD rejected the null hypothesis consistently across all of the models implemented. On one hand, a one unit increase in the *collective participation* score was associated with 12% decrease in the odds of using the instructional practice with more frequency in the classroom, while the characteristics of teachers were held constant. Given that this parameter did not vary wildly across the three ORM models, it can be argued that it was not importantly biased

by omitted variables at the school-level or by selection of teachers into certain types of schools.

On the other hand, both *active learning* and *extended duration* reported odds-ratios that reflected a direct association with the use of project-based learning in the Finnish schools. Considering results in Model 3, a one unit increase in *active learning* and *extended duration* would result in 12% and 15% increase in the odds of developing such instructional method, respectively. It must be noted that such positive direction of the relationship between these features of TPD and the outcome consistently failed to reject the null hypothesis across all the three statistical models executed for the Finnish sample.

However, some parameters did appear to be significant yet not consistently across each model. In Japan, for example, the estimates on *extended duration* were significantly related with the outcome variable in Model 1, but when only variation within schools was exploited (Model 2), this result did not longer stand. A similar case was described by the participation in teacher learning practices in England, in which case the magnitude of the parameter tended to decrease across the successive models executed (from 1.17:1 in Model 1 to 1.10:1 in Model 3). In this instance, the association between teacher learning practices and project-based learning seemed to be accounted for by the attitudes of teachers towards teaching and learning.

To sum up, the quality features of TPD seemed to be not related to the implementation of the teaching method examined in the two English speaking countries, the US and England, whereas all of them produced consistent associations in Finland, either in a negative direction – when considering *collective participation*- or in a positive direction –when considering *active learning* and *extended duration.* Consistent positive relationships across models were also found for *active learning* in Japan and the engagement in teacher learning practices in England.

Finally, I consider the results for the third item of classroom teaching practice – e.g. “students use ICT (information and communication technology) for projects or class work”.

4.3.3. Classroom teaching practice 3: Students use ICT (information and communication technology) for projects or class work

The results corresponding to the third outcome variable of this chapter are developed in this sub section. As illustrated in Table 4.2, whereas 59% of Japanese teachers never or almost never implemented such method over the year in their target classes, more than 76% of teachers in Finland, 83% in the US and 86% in England used it at least occasionally. Like in the previous two classroom teaching practices analysed, a frequent or total use of this method was more likely to be observed in the English speaking countries (56% US and 38% England) than in Japan (10%) and Finland (20%). Table 4.7 details the estimates of the ORM models executed in each country of interest.

Table 4.7 Ordinal Regression Models for the classroom teaching practice "Students use ICT (information and communication technology) for projects or class work" in the US, England, Japan and Finland

	Model 1			Model 2			Model 3			Model 1			Model 2			Model 3		
	OR	**(SE)**		**OR**	**(SE)**		**OR**	**(SE)**		**OR**	**(SE)**		**OR**	**(SE)**		**OR**	**(SE)**	
					US									**ENG**				
CollPar	1.06	(0.08)		1.02	(0.08)		1.00	(0.07)		1.04	(0.06)		1.04	(0.07)		1.04	(0.07)	
ActLea	0.97	(0.08)		0.88	(0.09)		0.87	(0.09)		1.12	(0.09)		1.11	(0.11)		1.11	(0.11)	
ExtDur	1.10	(0.09)		1.17	(0.11)	*	1.17	(0.11)	*	0.95	(0.07)		0.95	(0.08)		0.95	(0.08)	
TLP	1.09	(0.09)		1.15	(0.11)		1.11	(0.10)		1.09	(0.06)	*	1.09	(0.07)		1.09	(0.07)	
N	1414			1414			1414			1821			1821			1821		
					JPN									**FIN**				
CollPar	1.11	(0.07)	*	1.09	(0.07)					0.85	(0.04)	***	0.85	(0.05)	***	0.84	(0.05)	***
ActLea	1.15	(0.07)	**	1.21	(0.09)	**				1.15	(0.09)	*	1.14	(0.10)		1.12	(0.10)	
ExtDur	1.08	(0.07)		1.11	(0.07)					1.03	(0.06)		1.05	(0.07)		1.04	(0.07)	
TLP	1.20	(0.06)	***	1.19	(0.07)	***				1.34	(0.08)	***	1.34	(0.09)	***	1.31	(0.08)	***
N	2545			2545						1782			1782			1782		
Teachers' characteristics	✓			✓			✓			✓			✓			✓		
School fixed-effects				✓			✓						✓			✓		
Teachers' attitudes							✓									✓		

Source: TALIS 2013 database

Notes: Outcome variable: Students use ICT (information and communication technology) for projects or class work; *p < .1. **p < .05. ***p < .01; OR=odds-ratios; SE=standard error; CollPar=Collective participation; ActLea=Active learning; ExtDur=Extended duration; TLP=Teacher learning practices; N=number of observations; weighted data; Model 3 did not converge in Japan.

As in the previous sub section, the results indicated that the quality features of TPD and the participation in teacher learning practices were poorly related to the use of ICT in the English speaking countries. Only the estimates on *extended duration* in the two school fixed-effects models in the US and the odds-ratios on the covariate in Model 1 in England seemed to reject the null hypothesis with a 90% level of confidence. In the first case, holding constant the characteristics of teachers, their beliefs about teaching and learning, and omitted school characteristics, a one unit increase in the *extended duration* score would result in 17% increase in the odds of using this strategy with students. In contrast, variations in the same feature of TPD seemed to be not associated to the outcome variable in Japan and Finland (in any of the three models executed the null hypothesis was rejected).

Interestingly, the attendance of Finnish teachers for TPD who report more *collective participation* appear to decrease the odds of implementing the teaching method under examination. A similar contradiction was also observed in the previous sub section. In this case, a one unit increase in such key explanatory variable would result in a reduction of 16% in the likelihood of using ICT in the classroom, regardless the attributes of teachers and other potential confounders at the school-level (Model 3). However, teachers from this country that participated more in teacher learning practices were more likely to use this instructional practice. In concrete terms, if a teacher in Finland were to increase his or her engagement in the covariate by one unit, his or her odds of using more frequently this strategy would be likely to increase by a third (31%, Model 3).

A direct association between the participation in teacher learning practices and the teaching method evaluated was also found in Japan, in which case the odds of using this practice in the classroom were 19% greater per each unit increase in the covariate. In this country, teachers that attended TPD with greater degrees of *active learning* were more probable to use this method, too. A one unit increase in this quality feature of TPD would result in 21% increase in the odds of using it with Japanese students. In both cases, the odds-ratios remained statistically significant across the models imposed.

In summary, the opportunity of observing that students use ICT (information and communication technology) for projects or class work seemed to be not linked to the quality features of TPD for the US and England samples. On the contrary, this teaching method was found to be positively associated to the engagement of Japanese and Finnish teachers in teacher learning practices. For Japanese teachers, their participation in TPD

activities with greater degrees of *active learning* also reported a direct relationship with the outcome. In this context, it is worth highlighting that Finnish teachers that reported more TPD activities that included *collective participation* were less probable to implement ICT.

Finally, in order to assess whether estimates of association in Model 3 were statistically different across the countries of interest, Independent Samples t-tests were carried out to contrast country differences.

4.3.4. Cross-national comparison of parameters

Table 4.8 displays the p-values yielded by each test organised by quality features of TPD (*collective participation, active learning* and *extended duration*), classroom teaching practices and pairs of countries compared. Values in bold indicate statistically significant differences between the corresponding parameters at the 90% level of statistical confidence.

Table 4.8 P-levels based on t-tests for Independent Samples under Model 3 for each country pairing on each item of teacher practice

p-values		Small groups		PBL		ICT	
		ENG	**FIN**	**ENG**	**FIN**	**ENG**	**FIN**
Collective participation	**US**	0.861	0.146	0.789	0.141	0.538	0.176
	ENG		0.220		0.262		**0.031**
Active learning	**US**	0.964	0.978	0.927	0.883	0.210	0.213
	ENG		0.425		0.402		0.364
Extended duration	**US**	0.192	0.178	0.650	0.667	**0.084**	**0.087**
	ENG		**0.025**		0.444		0.523

Source: TALIS 2013 database

Notes: weighted data; Small groups=Use of small groups; PBL=Use of projects-based learning; ICT=Use of information and communication technology and PBL Model 3 did not converge in Japan.

In general, results suggested that the magnitude of such estimates did not differ significantly across countries, therefore the majority of the odds-ratios reported in the previous sub sections can be considered as equivalent proportions when they are compared from one country to another. In fact, all the odds-ratios relative to the

implementation of PBL seemed to be similar across the countries of interest. Only four cases produced significant estimates.

4.4. Discussion and conclusion

The comparative analysis developed in this chapter sought to examine whether the quality features of TPD or the participation in teacher learning practices were statistically associated with the implementation of a number of classroom teaching practices. An ORM strategy was developed to examine variations in the relationships between each predictor and the outcome variables once specific conditions were imposed to successive statistical models. The odds-ratios of the most restrictive model were then formally contrasted across countries using t-tests for Independent Samples in order to assess whether the corresponding parameters were consistently different at the macro level.

One of the main findings of this chapter indicated that in the US and England the quality features of TPD were rarely associated to the implementation of the classroom teaching practices analysed. Contrary to expectations, only the relationships between the *duration* of TPD and two outcome variables –e.g. the use of ICT in the US and cooperative learning in England- seemed to reject the null hypothesis. My analyses showed that a one unit increase in this quality feature was likely associated with 17% and 19% increase in the odds of using such methods, respectively. Such results are relevant considering that an important proportion of teachers in these countries reported having participated of TPD without this characteristic (38% in the US and 47% in England)[68].

However, apart from this specific result, no other key explanatory variable seemed to relate to the way that teachers teach in these countries. The findings are not consistent with Garet *et al.* (2001) who found a direct association between *active learning* and the knowledge and skills of teachers in the US, and with Opfer and Pedder (2011b), who reported a positive association between the three quality features of TPD and school achievement in England. In this sense, it is striking that the quality features of TPD

[68] In particular, this result could extend the findings of the official report of TALIS 2013 for England, in which case not only TPD activities based on individual or collaborative research would result in increases on the odds of implementing ICT, but also TPD with *extended duration* (OECD, 2014d).

examined seemed not to be associated to the implementation of any of the teaching methods selected for the analyses.

It is possible to argue that the quality features of TPD examined in this chapter can be insufficient to influence the practices of teachers given the characteristics of the profession in the US and England. For example, engaging in TPD with greater or lower levels of *active learning, collective participation* and/or *duration* could be definitely irrelevant in systems where teachers' attrition is high. Under such circumstances, teachers are less likely to put efforts on improving their practices, thus the participation in TPD is interpreted just as an administrative requisite which not necessarily leads to changes in the way they teach. This may also reflect the specific positive association found for teachers that reported TPD with longer *duration* in relation to the use of ICT (US) and cooperative learning (England) in the classroom. In this instances, teachers engaged in this type of TPD are probably those who plan to remain in the profession. For such teachers, the participation in a longer term TPD activity is coherent with their willingness to improve their instructional practices.

All in all, the research question posed in this chapter[69] can be satisfactorily answered in reference to the implementation of the three classroom teaching practices as reported by Finnish and Japanese teachers. For example, results from Japan suggested that greater levels of *active learning* were consistently associated with more frequent implementation of cooperative learning, project-based learning and ICT. To be more precise, a one unit increase in this feature would result in about 20% increase in the odds of implementing these instructional methods. Such result is interesting considering that more than a half of the teachers in this country (52%) informed that this feature was present only in some of the activities of TPD attended over the year (17% indicated that their TPD had no *active learning*). It also adds insights to the results of Japan reported in the official report of TALIS 2013 (OECD, 2014d), which described that no type of TPD activity was likely to increase the odds of implementing project-based learning or ICT in this country. My analysis showed that at least TPD with *active learning* would do so.

[69] Does TPD carried out either with greater degrees of active learning, collective participation or longer duration relate to specific classroom teaching practices when the participation in teacher learning practices is taken into account?

These findings are relevant for the design of high-quality TPD in contexts with important levels of overloading. As commented in Appendix B, teachers in Japan spend in average 53 hours per week in work related tasks, which translates in approximately a 10 hours daily schedule. In this context, one may expect that no attribute of TPD would motivate teachers to put additional efforts on changing the way they teach. Indeed, TPD is likely to be experienced just as a bureaucratic requirement because it is also compulsory to maintain employment. However, the analyses included in this chapter showed that Japanese teachers who engaged more in TPD with *active learning* were more likely to use active teaching practices in the classroom. TPD delivered with such feature seems to raise the high potential of Japanese teachers and motivate them to improve even more their instruction.

For Finnish teachers, in turn, improvements in *active learning* and in the *duration* of TPD were positively related with the use of project-based learning. My analyses showed that the odds of implementing this method of instruction would have increased by 20% per each unit increase in these variables. Such result also adds insights to the results of Finland documented in the official report of TALIS 2013 (OECD, 2014d), as not only observation visits to other schools would increase the odds of using project-based learning, but TPD activities with greater degrees of *active learning* and longer *duration.* In this context, it is important to point out that these features have also low prevalence in this country, with 68% of teachers having experienced none or only some activities of TPD with *active learning* and 63% without *extended duration*. Given the small level of implementation of some of these instructional practices in both countries –e.g. practically half of the teachers never used project-based learning or ICT with their target classes-, these results may shed light on how to enhance this aspect.

On the other hand, the results for *collective participation* in Finland were unexpected. As reported above, a one unit increase in this feature was found to be associated to approximately 15% decrease in the odds of implementing either project-based learning or ICT. In other words, at least in this country and in relation to these particular classroom teaching practices, the *collective participation* in TPD would have not been positively associated with instructional change. This aspect contrasts with recent reviews of the literature (Caena, 2011; Desimone, 2009) which have pointed that this attribute is one of the features of TPD that are critical to improve the quality of teaching.

In short, TPD delivered in Finland with more *active learning* and longer *duration* increases the likelihood of using project-based learning, whereas *collective participation* reduces the chances of this teaching method and its alternative version through ICT. It must be acknowledged that project-based learning is difficult to implement because teachers have to guide and support the work of students over a longer period of time (e.g. more than one lesson). However, Finnish teachers are likely to be well prepared for such task, as they are selected through very competitive processes (see Appendix B) and their skills are outstanding compared to colleagues from other parts of the world (Hanushek, Piopiunik and Wiederhold, 2014). In this context, TPD with *active learning* and longer *duration* seems suitable to raise the strong potential of Finnish teachers as it provides a learning environment which is coherent with the implementation of project-based learning. On the contrary, such strong potential seems to reduce when teachers from the same school attend the same TPD activity, which is normally provided outside the school premises. In this case, concentrating efforts in a demanding instructional method (i.e. projects-based learning) is likely to be hindered by the fact that teachers have to focus their own learning in the TPD event shared with colleagues.

Taking all these findings together, it seems that the quality features of TPD here evaluated were not related to outcomes in the countries where such association was expected (US and England), whereas they were in Japan and Finland, but only with some teaching methods and not always following a positive direction. At this juncture, the results of the participation in teacher learning practices are interesting. They show, for instance, that the engagement of teachers in this kind of activities were directly associated with the implementation of cooperative learning in all the four countries of interest. In England and Finland this variable is also positively associated with project-based learning, as it is with the use of ICT in Japan and Finland. In all these cases, the size of the coefficients are comparatively greater than those produced by the quality features of TPD, which would suggest that teacher learning practices are more closely associated to instructional change.

The relevance of this chapter lies in the implications that the enactment of the quality features of TPD has for the enhancement of classroom teaching practices at the national level. Policy developers and programme designers of in-service teacher training could use these findings to consistently adjust teacher learning activities with outputs in order to support their teachers and improve the quality of teaching. Nonetheless, it must

be warned that further research focused on the causal analysis of this topic is necessary (see subsection 1.3), as well as more analyses of cross-national large-scale data. In other words, to ascertain that specific quality features of TPD are more appropriate to enhance the implementation of certain teaching methods in specific national contexts, repeated measures designs using teachers' data from different nations might be a step forward in this area.

Nonetheless, this chapter revealed that the extent to which TPD activities were implemented with *collective participation, active learning* and *extended duration* in the US and England was unlikely to make a difference on the odds of using teaching methods based on cooperative learning, project-based learning or ICT. Unlike the evidence reported in the specialised literature to-date, the instructional approaches of US and English teachers seemed to be not associated to such quality features of TPD, whereas in contrast many of them were linked to such practices in Japan and Finland.

Chapter 6

Conclusions

Revisiting the research and policy context

The aspects that determine how TPD is implemented in order to improve educational outcomes are increasingly important for school systems where this type of activities are compulsory for the teaching profession. The countries analysed in this thesis (the US, England, Japan and Finland) represent four country-specific instances, with participation rates that demonstrate that almost all of their teachers engage in TPD activities every year. Data taken from TIMSS 2011 (Chapter 2) and TALIS 2013 (Chapter 4) showed that nowadays 8 out of 10 Japanese and Finnish teachers engage in TPD, whereas 9 do so in the two English-speaking countries. At this juncture, the qualities of TPD that can effectively enhance teachers' and students' learning become an urgent matter of study.

As mentioned in page 14, research has highlighted five key indicators in this regard, namely: *content focus*, *coherence*, *active learning*, *collective participation* and *duration* (Caena, 2011; Desimone, 2009). Studies conducted with national probability samples of teachers in the US (Garet *et al.*, 2001) and England (Opfer and Pedder, 2011b) have provided empirical evidence that support the favourable association of these features with teaching practices and student achievement. By the same token, experimental research in the US has started to use these dimensions as appropriate descriptors of the quality of TPD (Greenleaf et al., 2011; Heller et al., 2012; Penuel, Gallagher and Moorthy, 2011; Walker et al., 2012). However, most of this research has

failed to provide an accurate account of the contribution of these variables under the normal conditions of schools. On the one hand, they have assumed that all of these features are equally important to improve variables at the teacher and student level, which contradicts the set of theories underlying the influence of each dimension on educational outcomes. In addition, no study has assessed the generalisability of their contribution in teaching and learning environments which are outside of the US and England. Considering that the application of TPD with these characteristics can involve significant costs for school systems, it is crucial to determine whether they are related -and to what extent- to differences in teaching practices or student achievement across multiple countries.

Overarching research aims

The aim of this thesis was to contribute to the research on effective TPD by comparing the potential of three theory-based relationships between its quality features and national educational outcomes, as observed in the US, England, Japan and Finland. Individual pieces of research sought to elucidate the following issues:

1) Does mathematics *content-focused* TPD relate to student achievement in this subject (controlling for the effect of characteristics of students and teachers)?

2) Does a *coherent* approach to TPD in schools relate to student achievement (controlling for the effect of characteristics of students and schools)?

3) Does TPD carried out either with greater degrees of *active learning*, *collective participation* or longer *duration* relate to specific classroom teaching practices, when the participation in teacher learning practices is taken into account?

The overarching question was:

Are there differences in teachers' exposure to the quality features of TPD that might be associated with differences in national educational outcomes at the student and teacher level?

Key findings

In general terms, the evidence generated in this thesis demonstrates that variations in educational outcomes can be linked to changes in the quality of TPD, as measured by the five features outlined above. However, the direction of such relationships seems to be conditioned by the level at which the outcome is measured (teachers or students). For example, all the positive associations reported in the empirical chapters were found at the teacher level. In Chapter 4, I show that TPD which is delivered with greater degrees of *active learning* increases in Japan the likelihood of using all the teaching practices evaluated, whereas it seems to be positively related to the use of project-based learning in the Finnish classrooms. *Collective participation* is also positively associated with this teaching method in Japan, whereas TPD implemented with longer *duration* increases the chances of using ICT in the US, cooperative learning in England and project-based learning in Finland. On the contrary, only inverse associations were found at the student level. In Chapter 3, I revealed that the achievement in mathematics slightly decreases in the US and UK insofar as head-teachers strengthen the *coherence* of TPD. Likewise, a negative association was also reported in Chapter 2 for English and Japanese students in relation to the engagement of their teachers in mathematics *content-focused* TPD.

On the one hand, these results reinforce the view that changes in instructional practices can be promoted by the way in which the content of TPD is organised and presented to teachers (Desimone *et al.*, 2002; Garet *et al.*, 2001; Ingvarson, Meiers and Beavis, 2005; Loucks-Horsley and Matsumoto, 1999). As predicted, TPD that provide on several occasions (*duration*) more opportunities to plan, observe or perform teaching practices (*active learning*) among teachers from the same school (*collective participation*) increases the chances of using active teaching methods, such as cooperative learning, project-based learning and ICT. What is more, the data suggest that these associations can be observed independently from the attitudes of teachers towards teaching and learning, as well as from their engagement in teacher learning practices within schools, which are two variables also strongly associated with the way they teach (de Vries, Jansen and van de Grift, 2013; Opfer, Pedder and Lavicza, 2011a).

On the other hand, the results obtained at the student level do not support the hypotheses that learning achievement in mathematics can be enhanced by engaging in TPD which is either *focused* specifically on subject-matter content (Blank and de las

Alas, 2009; Kennedy, 1998) or managed *coherently* with the goals of schools in this area (Newmann *et al.*, 2001a; Newmann *et al.*, 2001b). In relation to the feature *content focus*, the findings rather coincide with those of Telese (2012) in the sense that students' scores tend to decrease when their teachers attend this kind of TPD. In the case of the *coherence* of TPD, findings also indicate negative associations with mathematics' scores. To be more precise, schools with more *coherent* TPD do not perform better in PISA in these four countries, relative to schools with less *coherence*.

There is insufficient evidence for the generalisability of the relationships between the quality features of TPD and educational outcomes across the four school systems analysed. Certainly, none of the suggested dimensions of the quality of TPD rejected the null hypotheses of each empirical chapter in more than a half of the countries of interest and there was no clear pattern about the countries where the associations were more regularly observed. Likewise, it was striking to find either negative or null associations in the US and England, considering that these two countries provided the empirical rationale for this investigation. Contrary to expectations, *active learning*, *collective participation* and *duration* were rarely related to educational outcomes in these two countries, whereas the *coherence* of TPD yielded a negative estimate of association in the US.

Limitations

There are obvious limitations to the findings of this thesis that can be extended to all its three empirical chapters. Firstly, this study does not claim to be able to support the causality underlying the reported statistical associations between the quality features of TPD and national educational outcomes. Given the observational nature of the datasets utilised, teachers had inequivalent probabilities to be exposed to different levels of the quality of TPD, as measured by the suggested five features. Therefore, extraneous variables might affect both the key explanatory variables and the outcomes, an aspect which clearly hinder the potential causation of these links.

Secondly, all the results here presented are only applicable to students, teachers and head-teachers at lower secondary education from the four school systems selected (the US, England, Japan and Finland). Opfer and Pedder (2011b) have reported differential patterns of association between the quality features of TPD and school

achievement for primary and secondary teachers in England. This aspect could be also expected for teachers from other countries given the particular organisation of teaching in each of these levels.

Implications for research

The results presented in this thesis provide a number of new avenues for research. A number of specific implications are given below in relation to the key findings and limitations:

- Considering the association between the quality features of TPD and classroom practices (Chapter 4), further research is needed to understand the mechanisms through which teachers' exposure to TPD with *active learning, collective participation* and longer *durantion* translates into using the particular instructional methods examined. One strategy would be to explain whether this occurs simply because teachers enact teaching practices deemed as models to follow or as a result of a process of critical reflection on their own way of teaching (Clarke and Hollingsworth, 2002; Dewey, 1933; Guskey, 1986).
- Regarding discrepancies with previous research in relation to the contribution of mathematics content-focused TPD to student achievement (Chapter 2), this thesis recognises that methodological differences between experimental and observational designs might have contributed to inconsistencies in results (Wayne *et al.*, 2008). Accordingly, it is worth comparing further results from follow-up studies of TIMSS and NAEP versus results of randomised controlled trials that evaluate mathematics knowledge only via standardised assessments.
- In relation to the study of the *coherence* of TPD as an unobservable (latent) indicator of the quality of TPD (Chapter 3), further research is clearly necessary to extend the use of head-teachers' perceptions in international large-scale surveys as a valid measure of this factor. This thesis is the first attempt to do so.
- The limited extent of generalisability of the features examined in this thesis lends weight to the argument that the links between the quality of TPD and educational outcomes is likely to be country-specific or affected by contextual variables at the macro level (CERI, 1998; Hardy, 2012). Therefore, future research should model the contribution of organisational and cultural characteristics that may moderate the

pathways between teachers' exposure to high-quality TPD and outcome measures at the teacher or student level.

- Taking into account the intrinsic limitations of observational data (see subsection 1.3) and the issues of feasibility in experimental designs conduced in multiple countries (UNICEF, 2010), the use of longitudinal designs may become a step forward in this area. In this case, prior levels of the outcome variables and other relevant confounders can be used to fully compare the effect which is attributable to teachers' exposure to the quality features of TPD (Goldstein, 2008).
- Finally, including more countries into the analysis may be important in re-examining the cross-national generalisability of the association between the quality of TPD and educational outcomes. One question is to what extent the conclusions of this study may be applicable to the rest of participant countries of TIMSS 2011, PISA 2012 and TALIS 2013. Substantial knowledge on the policies and structures that define the role of TPD in each nation would be necessary in order to guide empirical analyses and provide meaningful interpretations of findings at the macro level.

Consequences for policy

The evidence presented in this thesis is based on a thorough cross-national inquiry of the relationship between the quality features of TPD and national educational outcomes at the teacher and student level. Results are stronger in relation to the way teachers teach in the classroom and whilst there are a number of interesting links with student achievement it is difficult to make robust conclusions that hold in each country. This in turn points to the need to know more about the feasibility and merit of applying universal criteria to monitor and evaluate the quality of TPD on the world stage. In particular, the type of cultural readiness of head-teachers and teachers which is favourable to sustain the learning promoted by TPD may have wider implications for the debate about global policies in this area.

Nevertheless, the findings presented here may be of practical importance to the design of national strategies for TPD in the US, England, Japan and Finland. For example, they may deter policymakers in England from focusing on *content-focused* TPD as a mean to increase students' scores in mathematics. Moreover, for municipalities in Finland interested in enhancing the use of ICT and project-based learning in the

classrooms, it would be unwise at present to offer TPD mainly based on *collective participation*. All in all, whilst a number of new avenues for research are suggested, these results provide a starting point for further examination of the key qualities of TPD that truly sustain the quality of education.

Appendices

Appendix A: Theoretical approaches to the influence of teacher professional development on educational outcomes: a literature review

Introduction

Evidence from educational research suggests that TPD is one of the most important policies which can be identified at the school and teacher level to improve student learning. The meta-meta-analysis[70] developed by Hattie (2008) confirmed this relevance by reporting that TPD was ranked 19th out of 138 conditions of successful teaching and learning in schools. Research with a focus on the influence of TPD on educational outcomes has approached these encouraging results from several perspectives. Nonetheless, whilst an important development has been made in recent years about the influence of the quality features of TPD, the most common and prominent slant has considered that some types of TPD activities are more effective than others.

For instance, the meta-analysis developed by Wade (1985) found that activities such as class observation, microteaching, video and audio feedback were more successful than coaching, modelling and production of instructional materials. Likewise, in a more recent review of the literature, Schwile, Dembele and Schubert (2007) claimed that problems associated to the ineffectiveness of TPD were due to the mainstream implementation of workshops and seminars. On the basis of findings of a group of studies in the area (Feiman-Nemser, 2001; Lieberman and Miller, 1991; NCRTL, 1993), the authors asserted that not all types of TPD activities were equally effective and that isolated activities delivered by external providers lacked of the necessary support to sustain changes in teaching practices over time. It is worth noting, however, that these investigations were not empirically based and their conclusions rather derived from personal reflections of researchers when implementing and organising TPD programmes.

In contrast, recent research has remarked that the type of activities of TPD is a raw indicator of its quality and can become delusive to analyse its influence on educational outcomes. In this line of argumentation, Timperley et al. (2007) synthesised

[70] The study developed by Hattie (2008) is a (meta) analysis of over 800 meta-analyses on educational research.

nearly one hundred of empirical investigations and concluded that the type of TPD activities were not radically different when comparing between successful and unsuccessful instances. Furthermore, there is a myriad of types of TPD activities reported in the literature, whereas it is difficult to classify them as they share essential characteristics that overlap. As an illustration, Villegas-Reimers (2003) reported at least 22 types of activities in her extensive review of the literature, whereas in the first round of TALIS (OECD, 2009a), the organisers evaluated a set of 9. All in all, the criteria used to define, select and organise discrete features for each case were usually too vague to be specifically identified in practice, which represents a limitation that could certainly hinder the evaluation of their influence on educational outcomes.

Despite these constraints, accounts of improvements in school outcomes due to the implementation of different categories of TPD has given room to a "new paradigm" in this field (Schwile, Dembele and Schubert, 2007; Villegas-Reimers, 2003), in terms of studies focusing on general characteristics of innovative types of activities. In this regard, a bulk of literature reviews about the effectiveness of TPD (CERI, 1998; Craig, Kraft and du Plessis, 1998; Guskey, 1994; Loucks-Horsley and Matsumoto, 1999; Timperley *et al.*, 2007; Villegas-Reimers, 2003; Wilson and Berne, 1999) have highlighted that instead of provision, some key aspects of these types of TPD programmes would contribute to improvements in the performance of students and the skills of their teachers.

In general, these studies were orientated to policy purposes and presented lists of considerations for the implementation of what seemed to work well in practice. These lists were regularly inspired on the characteristics of communities of learning, a particular type of TPD activity in which informality, collaboration, school organisational support and focus on teaching practice are dominant themes. However, taking all these reviews together, only few of these ideas were grounded on empirical evidence and they rather referred to interpretations developed by the authors regarding the execution of unconventional types of TPD activities.

Despite such limitations, the development of such ideas has been at the same time beneficial because they have emphasised that the quality of TPD activities is an important aspect to be considered among the predictors of better teaching and learning outcomes. Indeed, they have nurtured the understanding that instead of only considering the access

to opportunities for TPD, the key features of these activities are worthy of further exploration in terms of their influence on educational outcomes. In this regard, recent literature has remarked that TPD that is *focused on content* knowledge, delivered *coherently*, and with greater degrees of *active learning*, *collective participation* and longer *duration*, is consistently associated with better teaching practices and student achievement (Caena, 2011; Desimone, 2009).

The investigations that nurture the critical features approach to the effectiveness of TPD can be classified according to whether the main outcome of interest is at the teacher or student level (Supovitz, 2001; Wayne et al., 2008), thus each of the quality features of TPD also correspond to specific literatures in the field. For instance, the first of these dimensions (e.g. TPD activities focused on subject matter knowledge) has a particular relevance, as a result of meta-analyses that have consistently remarked its effect on student achievement (Blank and de las Alas, 2009; Kennedy, 1998; Salinas, 2010; Scher and O'Reilly, 2009).

A clear distinction between *content focus* and the rest of features is also remarked by Wayne et al. (2008) when the authors conceptualise the types of treatments or interventions that experimental studies in the field are able to assess. At least two theories inform these treatments: *theory of instruction* and *theory of teacher change* (Van Veen, Zwart and Meirink, 2012; Wayne et al., 2008). The former refers to studies in which the main educational outcome of interest is student achievement and the main explanatory feature of TPD is its *focus*. The *theory of teacher change*, in turn, gathers investigations mainly concerned with variations on teaching practices due to the influence of any of the quality features of TPD, as well as about the assumed mechanisms by which these features would affect this outcome.

Theory also indicates that TPD activities should be analysed in the light of the systems in which they are implemented (Ganser, 2000; Hoban, 2002; Villegas-Reimers, 2003), as the quality features of TPD are particularly sensitive to characteristics of the nature of the context of delivery (Desimone, 2009; Wayne et al., 2008). In the most recent literature review in the field, Van Veen, Zwart and Meirink (2012) have remarked that in addition to the *theory of instruction* and the *theory of teacher change*, a *theory of context* is necessary in order to understand the assumed conditions that support the influence of

TPD on educational outcomes. A more detailed description of these three theories will be developed in the following three sub-sections.

Theory of instruction

Research aimed to analyse the effectiveness of the focus of TPD on student learning is originally related to the distinction between content, pedagogy and curriculum that is addressed within the theory of teacher knowledge. In this context, Shulman (1986) described how in the US the traditional expectation of content expertise in teachers confronted the emerging policies of standards focused on the enacting of research based teaching competencies. Such policies introduced an analytical dividing line between the realms of subject field content (e.g. mathematics, science, etc.) and pedagogical practices (e.g. instruction, assessment, etc.). According to the author, this sharp separation did not fit with the actual work of teachers, thus in order to solve this limitation he proposed the well-known three categories of teacher knowledge: (a) subject matter content (b) curricular content, and (c) pedagogical content knowledge (PCK). Whereas the first domain contained the academic knowledge related to the discipline to be taught, and the second one addresses the aims, programmes and strategies that norm teaching, PCK referred to "the ways of representing and formulating the subject that make it comprehensible to others" (Shulman, 1986, p. 9). To be more precise, PCK would connect content, curriculum and pedagogy in the context of classroom.

Fennema and Franke (1992) conducted an extensive critical review of literature questioning different categories of teacher knowledge that would help in the improvement of mathematics education and on what students learn. Their model of teachers' knowledge developing in context reproduces some of the categories introduced by Shulman (1986) and postulates "context specific knowledge" as a synthesis between "math knowledge", "pedagogical knowledge" and the "knowledge of learners' cognition in mathematics" (Fennema and Franke, 1992). Using this framework to analyse research in the field, the authors concluded that teacher knowledge focused on mathematics content and on how students learn is related to better classroom practices and student learning in mathematics. In this sense, the discussion turned to confirm the pre-eminence of content knowledge in the field of mathematics education, highlighting in addition the general participation of contextual categories such as students' characteristics.

In the field of mathematics and sciences education, other reviews focused on the effectiveness of TPD came to confirm this conclusion. Loucks-Horsley and Matsumoto (1999) executed an exhaustive revision of the theoretic and empirical literature using a model of influences on the relationship between TPD and student learning. The framework is broad in order to stress the link between teacher learning and student learning, including a complex set of concepts and categories that allows a deeper understanding of the phenomenon. Within the domain of TPD, the authors placed content as one of the four features of the quality of TPD, along with processes, strategies/structures and contexts. Content in this model is a composite of three subcategories: subject matter, learners and learning, and teaching methods; and is posed as a contextualised knowledge. Narrative systematic reviews in the field also gave prominence to content focus as a key factor of effective TPD. Wilson and Berne (1999) synthesised empirical research on teacher learning questioning the kind of professional knowledge that teachers acquire across the multiple opportunities of TPD. As in Loucks-Horsley and Matsumoto (1999), the categories of teacher knowledge that were considered to be worthy of development in TPD activities were subject matter, students and learning, and teaching.

Only in recent years has a set of meta-analyses given support to the pre-eminence of content instead of other foci of TPD activities in order to explain the influence on student achievement. The study developed by Kennedy (1998) is the first synthesis in the topic that considers effect size (*d*) measures as a means to analyse studies on effective TPD in mathematics and science education. By analysing different pathways of assumed influence between TPD and student learning, the author concluded that activities focused on content and on how students learn made a positive difference in student achievement.

Evidence on the impact of content on students' mathematics achievement has been confirmed in further meta-analyses. For instance, Scher and O'Reilly (2009) explicitly compared the impact of TPD programmes focused either on content or pedagogy and reported a positive effect size in favour of the former category (*d*=.38). Criticising the narrow breadth of these two categories in order to summarise research in the field, Salinas (2010) additionally adopted Sowder (2007) model of multiple foci to analyse the effect of substantive content focus of TPD in students' mathematics achievement. By examining 15 empirical studies, the analysis yielded a positive impact of TPD that is mainly focused on PCK (*d*=.57). Blank and de las Alas (2009) reported

similar positive results for TPD focused on content knowledge. By synthesising 16 empirical studies, the effect sizes on student mathematics achievement was .21 for pre-post measures and .13 for only post measures.

However, and despite these encouraging results, Yoon et al. (2007) commented that the limited number of studies in the field accomplishing minimum standards for the inclusion in meta-analysis constrained interpretations on the effect of features such as *content focus*. Furthermore, the wide diversity of foci categories employed by Salinas (2010) and the significant heterogeneity among effect sizes reported by Blank and de las Alas (2009), introduce caveats about the consistency of content focus's effect.

Theory of teacher change

In parallel to the *theory of instruction*, a number of investigations in the field of TPD have focused on the influence on teaching practices. More complex models have also been developed in this area (Clarke and Hollingsworth, 2002; Hoban, 2002) which endeavour to explain nexuses between features of TPD and/or specific attributes of teachers. In this sense, the theory introduces a foremost distinction between the phenomenon of teacher learning and the opportunities designed to trigger it. In other words, research about cognitive processes of learning involved when teachers participate in TPD has started to be investigated separately from the characteristics of these activities (Franke et al., 2001; Putnam and Borko, 2000).

Loucks-Horsley and Matsumoto (1999) used this distinction as a strategy to organise research on TPD in mathematics and science. By treating research about teacher learning as separate from research interested on the quality of TPD activities, specific features of each phenomenon were accounted. As a result, whereas quality features were relevant in the domain of TPD, processes such as knowledge, skills and beliefs about disciplinary content, student's cognition, pedagogy and leadership were highlighted as relevant aspects of research developed in the domain of teacher learning.

Research on teacher learning has acknowledged the 'situatedness' feature of the process (Putnam and Borko, 2000) by postulating that the phenomenon is led by construction and participation, rather than by a passive acquisition of contents. As a construction, teacher learning is described as an active operation by which teachers

autonomously elaborate new and existing knowledge; as a participation process, it is defined as taking place in specific social contexts that provide support and meaning (Meirink et al., 2009b). Furthermore, this line of argument has remarked that for learning to take place, at least a change in teacher beliefs about teaching and learning and/or teaching practices should occur (Bakkenes, Vermunt and Wubbels, 2010; Meirink et al., 2010; Meirink et al., 2009b; Vermunt and Endedijk, 2011; Voogt et al., 2011).

In both cases, teacher learning becomes a relevant concept as it would make permanent the influence of TPD over teachers' performance. In general, these beliefs could be orientated to prior understandings about subject matter or learning processes of students, whilst practices generally relate to classroom behaviours of teachers. Nevertheless, some authors have argued that processes aimed to change teacher beliefs are not easy to achieve (Ng, 2010) and, in addition, it is not consistently clear the sequence of change between beliefs and practices. Albeit most of the models about teacher learning posit that changes in beliefs occur first, others point a converse relationship in which changes in practices would trigger changes in beliefs (Guskey, 1986; 1994; 2002).

In recent years, teacher learning theory has been approached using the conceptual framework of Teachers' Orientation to Learning (Opfer and Pedder, 2011a; Opfer, Pedder and Lavicza, 2011a; Remillard and Bryans, 2004). Teachers' Orientation to Learning refers to patterns of interdependence between teacher beliefs about teaching and learning, and teacher learning practices that are naturally undertaken in the school setting. In contrast to TPD programmes (e.g. activities provided by external agents), teachers undertake individual and collective school-based activities in order to improve mastery in their role, such as knowledge updating, reflection about practice, and collaboration among peers (Kwakman, 2003; Vermunt and Endedijk, 2011). These activities are linked to the degree of conviction teachers develop about what is vital for the teaching and learning process in the classroom, i.e. teacher beliefs (Block and Hazelip, 1995). Recent research has provided empirical evidence on the association between the participation in teacher learning practices and teacher beliefs, either when these activities are individually undertaken by teachers (Bakkenes, Vermunt and Wubbels, 2010), or when they involve a natural collaboration with peers (Meirink et al., 2010) or both (de Vries, van de Grift and Jansen, 2013; Meirink et al., 2009a).

Interestingly, attempts to analyse the relationship between some quality features of TPD and Teachers' Orientation to Learning have been developed using cross-national evidence, though variations in outcomes has been accounted for by the types of TPD activities implemented. The study developed by Vieluf et al. (2012) in the context of the first round of TALIS (OECD, 2009a), provided indirect interpretations about the contribution of features such as *collective participation* and *duration*. According to this investigation, in order to predict teaching practices in the classroom, the amount of days attending TPD activities was in general less relevant than the type of TPD activity.

However, this finding would only apply for more cooperative types of TPD activities (e.g. networks and peer observation), which in turn were found to be associated to cooperative teacher learning activities (e.g. collaboration among staff), in the 23 countries under analysis. In contrast, relationships between TPD activities in which *collective participation* would be less emphasised (e.g. courses and workshops) and cooperative teacher learning practices, were found only in about a half of the countries. This aspect would suggest that among the quality features of TPD, *collective participation* would be linked to teacher learning practices such as collaboration among teachers.

Nevertheless, this conclusion is not completely accurate from the perspective of the quality features of TPD, because the authors argued that some types of TPD activities would better represent *collective participation* than others (more typical), instead of assessing this feature directly. This is somehow problematic because current research indicates that variations are not actually explained by engaging in different types of TPD activities, but to variations in core features that inform the quality of its delivery. Thus, in this case, it is necessary to obtain evidence about the actual prevalence of *collective participation* in each of these TPD activities, as it cannot be assumed that this variation is due to categories of TPD activities such as networks, peer observation, courses and workshops.

Some of these results naturally encourage further cross-national exploration of the association between features such as *duration* and *collective participation* -as quality features of TPD- and collaborative teacher learning practices. On the other hand, knowledge about the links between the rest of the features of the quality of TPD (*content focus, coherence* and *active learning*) and teacher learning practices has still not been

explored from a comparative perspective. Taking into account that the aforementioned study also found in practically all of the countries under analysis a strong association between teacher beliefs and teaching practices, it would be worth questioning how the quality features of TPD and Teachers' Orientation to Learning are associated, in order to explain or more fully understand teaching practices across national contexts.

Knowledge about the link between the quality features of TPD and Teachers' Orientation to Learning is currently of great value as empirical literature has recently described consistent patterns of association between Teachers' Orientation to Learning and teaching practices. For instance, de Vries, Jansen and van de Grift (2013) reported that the broader the participation in teacher learning practices, beliefs and teaching practices become both more student-orientated. This study also found no evidence of association between these processes and greater subject-matter orientated teaching practices. Likewise, Opfer, Pedder and Lavicza (2011a) reported that Teachers' Orientation to Learning was associated with changes in classroom practices, as reported by teachers. These studies suggested that Teachers' Orientation to Learning, in terms of patterns of association between individual beliefs about teaching and learning, and the strategies undertaken by teachers themselves to improve their work, contribute to explain the type of practices at the classroom level. Nonetheless, and despite these advances in the field, there are no studies addressing the complex pathway between the quality features of TPD, Teachers' Orientation to Learning and teaching practices.

Opfer and Pedder (2011b) have developed the closest effort in this area by analysing large-scale survey information from teachers in England. Drawing data from a nationally representative sample of teachers from primary and secondary education, the authors examined the links between school achievement and Teachers' Orientation to Learning, as well as with three components of the quality of TPD activities: *duration*, *active learning* and *collective participation*. As a result, evidence of moderate association between Teachers' Orientation to Learning and educational outcomes was confirmed, whereas relevant differences in the quality of TPD experienced by teachers were related to different levels of school achievement. In particular, secondary teachers from the middling and lowest bands of achievement reported having undertaken TPD activities with substantial less quality than their colleagues in schools with higher performance. Such analysis have been recently replicated using data from England in the recent round of TALIS (Micklewright et al., 2014), although yielding somehow the opposite results.

In this case, and using similar measures of school achievement and a summary index of the quality of TPD comprising *active learning, collective participation* and *duration*, lower secondary teachers from the lowest band of school achievement reported significantly better quality in the TPD undertaken.

These results suggested that the key components of the quality of TPD activities experienced by secondary teachers in this country are associated to the level of achievement of the school, though the direction of such link is not clear yet. In this regard, more knowledge is needed to understand how teacher level variables (and among them, Teachers' Orientation to Learning) might play a role as intervening the influence of the quality of TPD on educational outcomes. Unfortunately, the studies discussed above were not aimed to examine differences in educational outcomes due to variations in the quality of TPD activities, nor in the context of other relevant variables related to processes of teacher change (e.g. Teachers' Orientation to Learning). Instead, the authors only preferred to analyse variations in school achievement due to the exposure to some of the quality features of TPD (*duration*, *active learning* and *collective participation*).

Theory of context

Even though the participation of system conditions on the complex pathway between TPD and educational outcomes has been remarked in the literature (Hoban, 2002; Opfer and Pedder, 2011a) few studies has attempted to empirically analyse intervening variables at the school and national level (Hendriks *et al.*, 2010a). Attention to context indeed mirrors one of the quality features of TPD (*coherence*), in terms of the necessary alignment between TPD programmes, teacher individual characteristics, and educational policy goals. Nonetheless, such alignment, instead of being standard for every site, would be unique for every educational system where TPD is implemented, as each context entails distinctive features of their teachers and schools that would constrain or boost the influence of the quality of TPD (Opfer and Pedder, 2011a; Van Veen, Zwart and Meirink, 2012). In this sense, for instance, it is possible to hypothesize that some quality features of TPD might be more relevant than others in certain contexts. Therefore, it becomes necessary to assess under what organisational conditions these dimensions are associated to variations in educational outcomes.

In contrast to the *theories of instruction* and *teacher change*, there is limited literature addressing to what extent higher level variables of educational systems affect the impact of TPD and there are no studies addressing how the quality features of TPD affect educational outcomes in different contexts (Van Veen, Zwart and Meirink, 2012). Research has so far considered School Organisational Conditions as determinants of the influence of TPD activities on teaching practices and student learning (Hendriks *et al.*, 2010a). To be more precise, School Organisational Conditions are regarded as intervening in the relationship between Teachers' Orientations to Learning and teaching practices, thus the connection with TPD activities is rather preceded by the complex patterns of beliefs and learning practices that teachers develop in the school. Studies that feature School Organisational Conditions have tended to document the role of the leadership of head-teachers (James and McCormick, 2009; Supovitz, Sirinides and May, 2010), as well as the kind of culture and structure (Opfer and Pedder, 2011a) that the school community develops in terms of shared beliefs, norms and practices about teaching and learning.

With respect to studies on the role of *leadership*, research carried out in the US and in different European countries have shown that this attribute is frequently linked to the level of participation in teacher learning practices (Geijsel et al., 2009; Gumus, Bulut and Bellibas, 2013; James and McCormick, 2009; King, 2011; Rajala et al., 2008; Runhaar, Sanders and Yang, 2010; Supovitz, Sirinides and May, 2010). In general, these investigations highlight key characteristics that define how head-teachers steer the improvement of teaching practices in the classroom by promoting either individual or collaborative teacher learning practices. Although changes in teacher beliefs about teaching and learning are reported in a minority of studies (Geijsel *et al.*, 2009; James and McCormick, 2009; Runhaar, Sanders and Yang, 2010), results indicates that Teachers' Orientation to Learning are affected as a whole by this attribute of head-teachers. Theoretically, this finding would introduce a competing hypothesis against the contribution of the quality features of TPD to teaching practices. However, as these studies generally assume teacher change as a function of internally school-based processes (leadership, teacher learning practices, and teacher beliefs), the dispute has not been yet resolved.

Regarding *culture and structure*, School Organisational Conditions are considered in terms of the collective beliefs, norms and practices related to teaching and

learning that are prevalent in the school. For instance, Opfer and Pedder (2011a) commented that one of the most relevant School Organisational Conditions found in the literature were school-level beliefs about teaching and learning, a feature that Opfer, Pedder and Lavicza (2011b) and Opfer and Pedder (2011b) also operationalised as the school orientation to learning. In this case, teachers described the extent to which different actions orientated to support the improvement of teaching and learning in the classroom were valued in their schools.

Interestingly, these factors were reported to be associated in England to the average levels of student achievement in national standardised measures, with high achiever schools yielding more positive school orientation to learning (Opfer and Pedder, 2011b). Unfortunately, such analyses did not assess the expected contribution of teaching practices to student achievement, as this outcome measure was available only at the school level for the study. It is plausible to hypothesize that the school orientation to learning could moderate the influence that teaching practices has on student learning, via complex pathways of influence where Teachers' Orientation to Learning and even the quality features of TPD could play a role. An argument of this sort challenges the current understanding of the participation of school-level factors that provide support and sustainability to changes at the teacher level and, particularly, what role could the quality features of TPD play in order to trigger this transformation.

Discussion

The *theory of context* advances the argument that the influence of TPD on educational outcomes can only be understood in an organisational/social setting. For instance, most of the findings contained within the *theory of instruction* proceed from the practice and implementation of Randomised Controlled Trials, which immediately control for the influence of potential confounding variables due to random allocation between treatment and control. Typically, these samples are small and in no case they represent information about the whole educational system where the programmes were implemented. Sample sizes usually involve the participation of less than twenty-five teachers per study and they are regularly drawn from specific regions within the US, thus the external validity of results is limited.

On the other hand, there are no studies within the *theory of teacher change* addressing how Teachers' Orientation to Learning and teaching practices work in multiple school settings. Even though some studies investigated data collected from different schools, they only focused on the association between Teachers' Orientations to Learning and teaching practices at the teacher level, thus no attention was given to the contribution of contextual characteristics of these schools. Albeit Opfer and Pedder (2011b) and Opfer, Pedder and Lavicza (2011b) in parallel and using the same data examined factors at the school level that were associated with teaching practices, these analyses were undertaken separately from the contribution of Teachers' Orientation to Learning. In other words, as in the case of the research about the quality features of TPD, we don't know yet how the link between Teachers' Orientation to Learning and teaching practices might vary according to features of context (Hendriks *et al.*, 2010a).

In general terms, both theories lack of complementary analyses inspired in the *theory of context* as no empirical evidence reports the intervention of variables at the school and/or national level on the influence of TPD on educational outcomes. Indeed, the *theory of context* itself still considers as important only variables at the school level (e.g. School Organisational Conditions), whereas recent cross-national case studies developed by Hardy et al. (2010), Hardy and Rönnerman (2011) and Hardy (2012) have described policy influences at the system level. By critically analysing policy documents related to TPD and practices in specific schools, the pressure introduced by neoliberal and managerial logics is examined by these authors in terms of the actual learning experiences developed by teachers from some Anglo and Nordic countries. Such developments might guide systematic comparisons across countries in order to empirically explore how characteristics of their national school systems and the organisation of teachers' work underlie the complex pathway between the quality of TPD and educational outcomes.

Finally, research presented here in relation to the *theories of instruction, change,* and *context* suggests the need to examine the association between the quality features of TPD and educational outcomes using a cross-national approach. Essentially, it is relevant to verify whether the positive results reported in the US and England from using national probability samples (Birman et al., 2000; Desimone et al., 2002; Garet et al., 2001; Opfer and Pedder, 2011b; Pedder and Opfer, 2011) can be replicated with current data from international large-scale assessments.

Appendix B: The teaching profession in the US, England, Japan and Finland

This appendix puts forward a description of the teaching profession in the US, England, Japan and Finland as a mean to explore relevant characteristics of these countries in this area. What we are mainly concerned with here is demonstrating some patterns about teachers' work in order to highlight contrasts among these countries. Hypotheses and predictions are beyond the scope of this section; otherwise, the following results are presented to inform the reader about relevant differences and commonalities in the national contexts where TPD develops.

From regular census information we are able to calculate the number of students to be enrolled in the system and, accordingly, the number of teachers required to fulfil that demand. In broad outline, such information is useful to contrast the size of the systems under comparison, which is relevant when thinking on national strategies of TPD and the barriers to fidelity in their implementation. It is certainly true that in larger school systems, national reforms steered from the national level may need proportionally more efforts and time to be implemented throughout the whole country.

Table 6.1 provides data on teachers demand at secondary education in the four selected countries from 2007-2011. For each nation, total numbers of students and teachers are presented alongside the respective annual Pupil Teacher Ratio (PTR) – except in the UK, where values are based on data available from the years 2007 and 2008.

Table 6.1 Teachers demand in the US, Japan, UK and Finland at secondary education. 2007-2011

	US			UK			Japan			Finland		
	Student enrolment[(1)]	**Teachers** [(2)]	**PTR** [(4)]	**Student enrolment**[(1)]	**Teachers** [(2)]	**PTR** [(4)]	**Student enrolment**[(1)]	**Teachers** [(2)]	**PTR** [(4)]	**Student enrolment**[(1)]	**Teachers** [(2)]	**PTR** [(4)]
2007	24,731,028	1,698,103	15	5,306,369	378,882	14	7,427,059	607,663	12	432,607	44,170	10
2008	24,692,888	1,717,576	14	5,356,450	375,385	14	7,355,678	607,062	12	431,233	42,991	10
2009	24,524,564	1,756,753	14	5,429,636	384,065[(3)]		7,299,966	609,966	12	428,332	43,319	10
2010	24,192,786	1,758,269	14	5,538,230	391,747[(3)]		7,296,330	613,851	12	426,710	43,076	10
2011	24,214,304	1,671,040	14	5,000,332	353,698[(3)]		7,284,867	617,642	12	422,872	44,493	10

Sources: EdStats/World Bank (2014)

Notes: [(1)] Enrolment in total secondary. Public and private. All programmes. Total; [(2)] Teaching staff in total secondary. Public and private. Full and part-time. All programmes. Total; [(3)] Fitted values holding PTR constant; [(4)] Pupil-teacher ratio

The data suggests relevant differences in the size of school systems under analysis, and it is clear that the US system is very much bigger than the rest of comparator countries. Considering average numbers of student enrolment and teaching staff, secondary education in this country is approximately three times greater than in Japan, five times greater than in the whole UK and represents 57 times the size of the Finnish school system at this level. Such result must be considered when comparing the influence of TPD activities on national educational outcomes across the four countries selected in this study, as the US system is expected to spend proportional efforts in implementing TPD nationwide.

In general, these sizes tend to remain stable over time, however it is worth mentioning that a gentle negative growth rate is observed across countries –especially in the English-speaking countries- between 2007 and 2011, which suggests that the demand of teachers at secondary education is expected to slightly decline in the following years. To take the most striking examples, over twenty seven and twenty five thousand teachers were no longer required by 2011 at secondary education in the US and the UK, respectively, comparing with 2007 data. In such cases, less vacancies for newly qualified teachers might become an opportunity to enhance standards to enter the teaching profession and upgrade in-service staff via TPD activities. However, such opportunity might be conditioned to potential patterns of teacher shortage that have been documented in these systems.

In this regard, it must be noted that both the US and England follow similar high rates of teacher attrition (Hutchings, 2011), whereas only in recent years policy instruments to measure and control this problem have been developed at the state level (DfE, 2013). In the US, approximately 10% of teachers leave every year the profession since 1988 (Keigher and Cross, 2010) and practically a third of teachers and half of the teachers in urban communities resign during their first five years of career (Gregorian, 2001). According to Suell and Piotrowski (2007), these rates of attrition are even higher in fields such as mathematics and sciences, where 20% of teachers leaves annually the career in this country. In England, teacher attrition in secondary education is approximately 11% (Passy and Golden, 2010) and the number of trainee teachers recruited in mathematics and sciences programmes has decreased in the last years. This is more evident in higher education institutions, where a third of vacancies for initial teacher training in mathematics remained unfilled in 2013 (Gardner, 2013). It has been

recently argued that other professions demanding similar domains of knowledge would become more appealing for applicants due to the recovering of England from the international finance crisis (Howson and Waterman, 2013; Roberts, 2013).

Unlike these two countries, in Japan and Finland there is actually surplus of applicants to teaching positions in the school system. In Japan, the proportion of just graduated teachers is larger than the planned teaching positions to be hired, thus competition is high and not all students enrolled in teacher certification programmes actually aim to eventually work in the teaching profession (Japanese Ministry of Education, 2003). In Finland only 10% of applicants are accepted into teacher certification programmes, thus the process is also competitive as in Japan (Sahlberg, 2011). In addition, teacher training programmes are highly attractive in Finland in comparison with the other Nordic countries, where the rates of enrolment in these programmes have drop systematically in the last decade (Nordic Council of Ministers, 2009).

Table 6.2 provides information on characteristics of initial teacher education programmes and requirements to enter the teaching profession in the public system at lower secondary education across the four selected cases of this thesis.

Table 6.2 Requirements for initial teacher training and for teaching in public institutions at lower secondary education in the US, England, Japan and Finland

	US	ENG	JPN	FIN
Competitive examination required to enter pre-service teacher training	No	No	No	Yes
Existence of alternative teacher certification	Yes	Yes	No	No
Duration of teacher-training programme in years	4	3, 4	2, 4, 6	5
Teaching practicum required as part of pre-service training	Yes	No	Yes	Yes
Credential or license, in addition to the education diploma, required to become fully qualified	Yes	No	Yes	No
ISCED type of final qualification(1)	5A	5A	5A+5B, 5A, 5A	5A
Credential or license, in addition to the education diploma, required to start teaching	Yes	Yes	Yes	No
Teaching practicum required to obtain credential/ licence	Yes	Yes	Yes	No
Competitive examination required to enter the teaching profession	Yes	No	Yes	No
Teaching practicum required after being recruited, as an induction/probation period	No	No	No	No
Compulsory requirement for continuing education to maintain employment	Yes	Yes	Yes	Yes

Sources: European Commission/EACEA/Eurydice (2013); Ingersoll (2007); OECD (2012a)

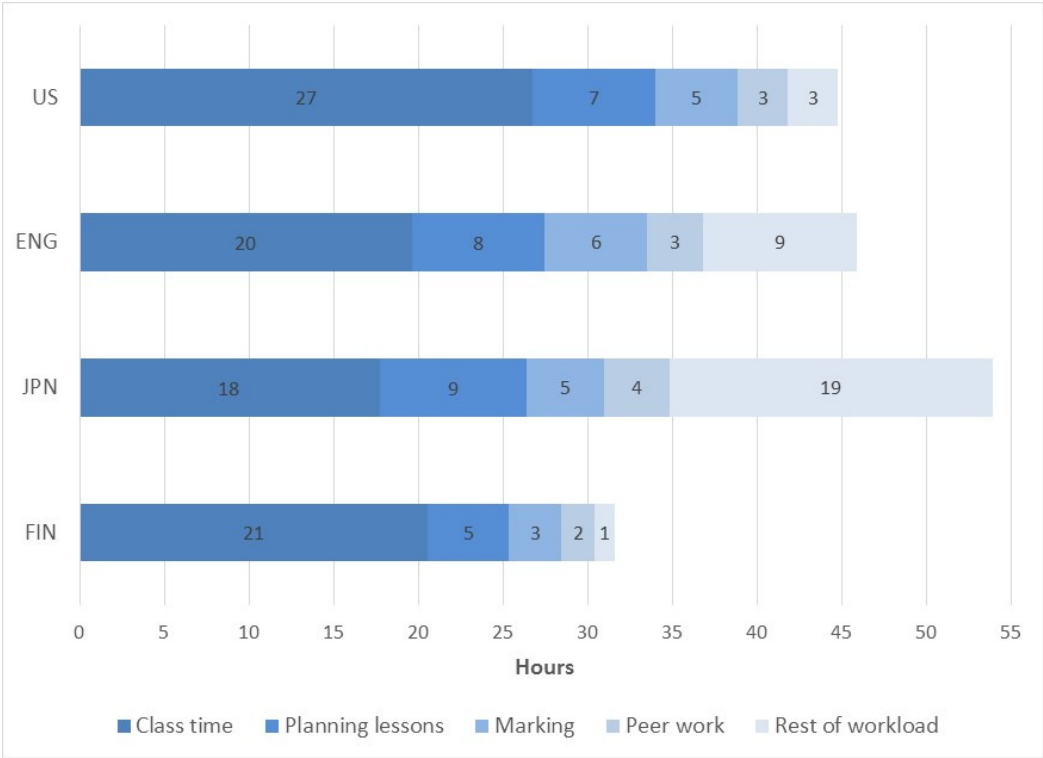
Notes: (1) ISCED 5 refers to qualification in tertiary education. Types A and B share the same level of competence, however type A programmes are more academic and type B are more occupationally orientated.

Detailed examination of policy options for teachers' qualification reveals that requirements to enter initial teacher training are less restrictive in the US and England. As an illustration of such level of control to enter teacher education, compare the presence of a competitive assessment administered to applicants in Finland against the widespread existence of alternative teacher certification programmes in the US and England. Although there are concerns in the literature about the low quality of such pathways to become teacher (Musset, 2010; Suell and Piotrowski, 2007), in the four countries under analysis, graduates obtain an equivalent qualification, which in the International Standard Classification of Education (ISCED) corresponds to a 5A level.

This is a tertiary education level of competence with more emphasis on academic knowledge than occupational skills –nonetheless, both orientations can be found in Japan- which trainees can attain through programmes lasting between three (England) and six (Japan) years. In general, this qualification involves a period of practical training in schools –except in England-, and in the US and Japan it is linked to obtaining a certification granted by the national level. It is important to point out that teacher qualification via completing teacher education programmes does not necessarily allows to work in public schools in all these countries. Only in Finland requirements are met when finishing this phase of the career, whereas in the other three countries graduates still need to complete a practicum period to attain an additional certification granted by the state and in the US and Japan also be examined competitively. Finally, it should be emphasised that in all the four countries selected, participation in TPD is compulsory for teachers in order to maintain employment, therefore it is expected a high proportion of staff undertaking such activities.

In relation to the actual utilisation of teachers' work, Figure 6.1 shows the total weekly workload divided into teaching hours, hours spent in activities related to teaching (e.g. planning lessons, marking and working with peers) and hours spent in tasks not directly related to teaching (e.g. administrative work, management, meetings with parents, etc.), as reported in TALIS 2013.

Figure 6.1 Weekly hours spent on teaching, planning, marking, working with other teachers and undertaking other tasks in the school not directly related to teaching in the US, England, Japan and Finland



Sources: OECD (2014d, table 6.12)

According to these data, teachers in the US, England and Japan spend greater amount of total working hours; this measure is equivalent in the two English-speaking countries (approximately 45 hours) and even higher in Japan (53), whereas Finnish teachers work very much less (32). To put it in a more concrete manner, whereas US and English teachers work in average nine hours per day, Japanese ones do more than ten and Finnish only over six –this represents approximately three-quarters of a normal working day.

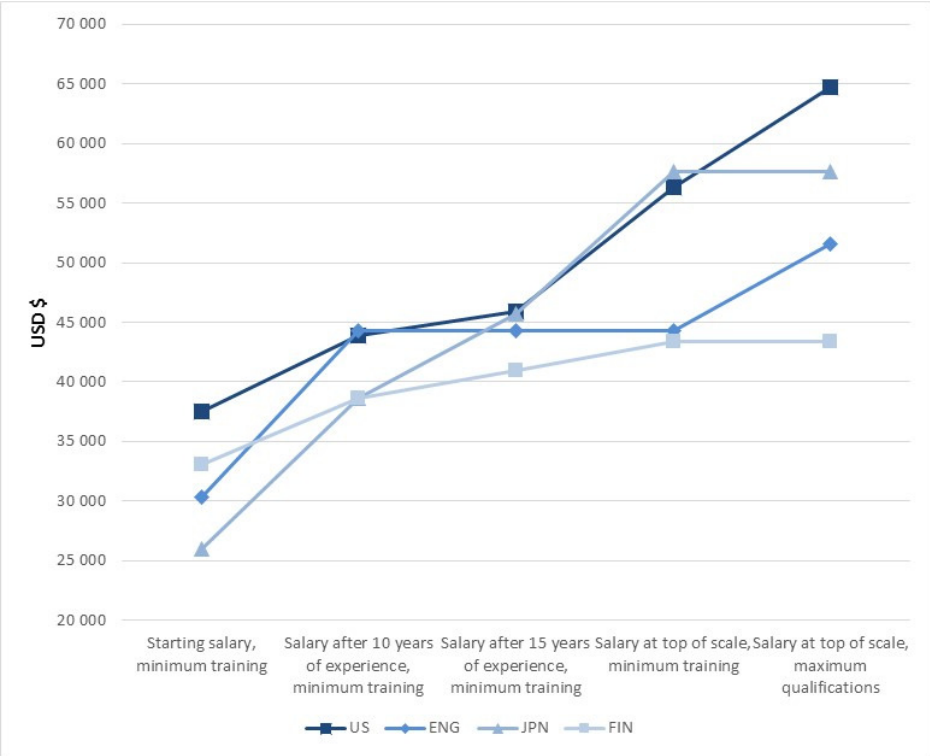
On the other hand, teachers in England, Japan and Finland spend similar time teaching in the classroom (20, 18 and 21 hours per week, respectively); a number that is only surpassed by US teachers, whose spend 27 hours of instruction in an average week. In other words, whereas English, Japanese and Finnish teachers spend approximately four hours of class time per day, students in the US are practically taught one hour extra. Finally, the time spent in activities directly related to teaching are proportionally less in Finland comparing against the three other countries. Considering the hours spent in

preparing lessons, marking and working with peers, Finnish teachers spend only 10 hours, whereas Japanese, English and the US professionals do 18, 17 and 15 hours, respectively.

In this context, it is worth underlining the high cross-country variation in the hours spent in school tasks not directly related to teaching. Whereas teachers in Finland and the US report between one and three hours in this type of tasks, in England and Japan this represents nine and nineteen hours of the total weekly schedule, respectively. This is in particular relevant for the effective implementation of TPD, as it is in this zone of time that such activities can become realizable. Put another way, there is practically no weekly time available for TPD activities in Finland and the US, whereas schedules of English and Japanese teachers yield a large scope to be adjusted for this purpose.

Data about the structure of the teachers' career adds relevant information to assess the attractiveness of the profession across countries. Teachers' career trajectories regarding years of experience and level of training is displayed in Figure 6.2 for the four countries under analysis. Five milestones are described from the initial –e.g. earning the starting salary and having minimum training- to the terminal –e.g. earning the top of the scale having maximum qualifications- stages of the profession (earnings are expressed in USD currency).

Figure 6.2 Teachers' career structure at lower secondary education in the US, England, Japan and Finland



Sources: OECD (2013b)

According to the trends observed in the data, the pool of selected countries can be sorted according to the potential attractiveness of their teachers' career in the following order. The country with the most motivating career is the US, considering that the starting and ending salaries are comparatively higher than the rest of comparator systems, and that between these two milestones there is a steady upward trend. Then Japan is in second place the most attractive teacher career given its own upward trend, even though its starting salary is the lowest among the four countries under analysis. In contrast, less appealing careers are observed in England and Finland, systems in which the data reveal work trajectories that tend to stabilise and flatten out opportunities for promotion over time, respectively. Table 6.3 provides additional data comparing teachers' compensation.

Table 6.3 Comparison of lower secondary education teachers' salaries over time and regarding levels of salary in similar professions in the US, England, Japan and Finland

	US	ENG	JPN	FIN
Ratio of salary to earnings for full-time, full-year workers with tertiary education aged 25 to 64	0.67	0.92	m	0.98
Ratio of salary at top of scale to starting salary	1.50	1.46	2.21	1.31
Number of Years from starting to top salary	m	12	34	20

Sources: OECD (2013b)

Notes: m=missing value

The information reveals important differences across countries in the aforementioned levels of attractiveness of their national careers' structure. The case of the US illustrates how a steep working trajectory might become insufficient when earnings represent in average only two-thirds (0.67) of earnings in similar professions. Likewise, even though the US (1.50), England (1.46) and Finland (1.31) describe substantial increases in salary from the start to the end of the career, only in the latter two countries teachers' earnings are much more equivalent to those in similar professions (0.92 and 0.98, respectively). In these cases, the aforementioned flat trends of the careers in these two school system might be less relevant insofar as average earnings of teachers are more aligned with those in equivalent labour market areas. On the other hand, one of the peculiarities of the Japanese career structure is that the salary at the top of the scale is slowly achieved (34 years) comparing with England (12 years) and Finland (20 years). This aspect implies a longer-term perspective to increase earnings in this country, which yields as result that at the end of their career teachers earn 2.21 times the initial salary, a ratio comparatively higher than in England and Finland.

Taking into account that TPD activities are compulsory for teachers to maintain employment in all the countries under evaluation, it becomes more relevant to interpret the context provided by each national school system in terms of the structure of the career in the national labour market. For instance, in countries like England and Finland, where the trajectory of the career tends to reach a plateau, such activities might be perceived as a less relevant incentive for promotion and more like a bureaucratic requirement. On the contrary, either in the US or Japan, TPD could be seen as an effective opportunity to raise performance and gain recognition leading to upgrades along the career. In particular, this should be more accentuated among US teachers, regarding the fact that the upward trend

of the occupation leads to a greater compensation towards the end of the career when more training has been undertaken.

Concluding this appendix, it may be argued that specific patterns of demand of teachers across the four selected countries might be reasonably linked to the stringency of requirements to become teacher and the utilisation of teachers' workload. On the one hand, the US and England yield a consistent pattern of high attrition, whereas the conditions to enter the teaching profession are less restrictive. Teachers work similar amount of hours in both countries, even though in England more time is spent in activities not directly related to teaching –a time that is destined to teaching in the US. On the other hand, Japan and Finland describe a surplus of applicants to the teaching profession, so requirements are set to a higher standard. Nonetheless, each country yields specific patterns of utilisation, with Finnish teachers working fewer hours in total –from which almost all are devoted to teaching and related tasks-, whereas Japanese staff works proportionally one third more, mainly developing tasks that are not directly linked to teaching.

Given such circumstances, it is important to emphasise that in all these countries TPD activities are obligatory for teachers to keep working in the public system; however, this basic requirement might be differently perceived in the light of dissimilar patterns of career structure. For instance, whilst the US career yields the most ascending trajectory, it is comparatively less attractive than in Finland and England regarding earnings in similar occupations –the salaries in these countries are practically equivalent-. Likewise, as the opportunities for upgrading in these two countries tend to stabilise over time, the average span to achieve the top salary of the national scale is very much shorter than, for example, in Japan. In this sense, the perception of English and Finnish teachers about the role of TPD activities in their own careers should be not as pivotal as it might be expected from professionals in the US and Japan.

Appendix C: Policy background

In recent decades a specific policy debate about TPD has been promoted in developed countries due to the implementation of lifelong learning policies (Day, 1999; OECD, 1998). In contrast to other sectors of these economies, the teaching profession has been recognised as paradigmatic in order to accomplish the aspiration of universal and permanent learning opportunities of individuals within societies. High quality teachers are necessary to ensure adequate access to knowledge through national school systems and, accordingly, teachers need regular opportunities to develop their competencies in the ever-changing context of globalisation. A number of policy documents and meetings of supranational institutions such as the Organisation for Economic Co-operation and Development (OECD), the European Union and the International Summits on the Teaching Profession have highlighted this challenge for policy analysis at the national level.

The International Summits on the Teaching Profession

Since 2011 the OECD jointly with the Asia Society has been organising annual International Summits on the Teaching Profession, in which delegates from Asia, Europe, the US and Canada meet to discuss policy issues on national school reforms and share best practices in this field. In the five versions of the summit (Stewart, 2011; 2012; 2013; 2014; 2015), TPD has been posed as a key issue for improving teacher quality, as well as one of the main mechanism for their support and retention.

According to the analysis of the delegates, pre-service teacher education is not sufficient to prepare teachers for all the challenges they will be expected to meet during their careers, therefore in-service training is considered a necessary strategy for the short and medium goals of each national school system. As such, TPD can address different objectives, such as diminishing the effect of knowledge obsolesce for the teacher; supporting skill's development according to innovative teaching approaches; supporting teachers in the application of changes made to curricula; and supporting less competent teachers to become accomplished professionals.

From the point of view of countries taking part in these summits, the current challenges of TPD are the quality of the provision and the relevance for the purposes of national reforms, thus it is important the role of policy makers in this area. Furthermore, it seems adequate to concentrate efforts on more effective forms of TPD and link them to the improvement of teaching and opportunities of career promotion. Particularly, it has been claimed during the last two versions of the event that as most of the TPD is perceived as not useful by teachers themselves, more opportunities for teacher learning 'on the job' and school-based professional collaboration should be promoted globally in the short term.

Among the pool of countries involved in the debate, Japan, Singapore and Finland have been proposed as examples of good national systems of TPD (Stewart, 2011), even though their contexts describe distinctive characteristics to consider when it comes to evaluate the effectiveness of these activities. However, regardless their unique national characteristics, specific processes at the school level, such as head-teachers leadership and procedures of feedback and appraisal, have been identified in these meetings as key actions to support appropriate TPD in different nations. All in all, knowledge and experience across all the levels of the system –students, teachers, administrators and policy makers- were agreed to be the main elements to consider for improving the quality of education globally.

The European Union

The European Union has been concerned about TPD especially since the Lisbon European Council identified that teachers' mobility and the attraction of high quality applicants were key aspects for the development of the region (European Parliament, 2000). The work programme derived from this agreement posed the development of the teaching profession as a priority also in terms of providing adequate conditions for supporting teachers' learning throughout their careers. The strategy has stimulated the elaboration of a common European framework on teachers' competences, the support for policy enactment at the national level and the monitoring of objectives related to enhancing the provision of TPD. In this context, TPD is deemed as the main mechanism for the development of teachers' skills, knowledge and attitudes (European Commission, 2012a), as a specific domain of policy advice for national school systems (ETUCE, 2008;

European Commission, 2005b) and as one of the indicators that informs the success of the so-called Lisbon strategy (European Commission, 2009).

Since the strategy was launched, policy advice for member states has been regularly delivered through the implementation of peer learning activities in which national delegates meet to discuss issues of common interest (European Commission, 2005b; European Commission, 2012a). Even though the ideas analysed in these meetings do not represent formal agreements of the organisation, the activity becomes a relevant source of information for the enactment of initiatives at the national level, and it feeds decision making at the European Commission and the Council of the European Union.

In 2005, ten member states and a group of experts met in Ireland to examine issues related to TPD in the context of the implementation of national educational reforms (European Commission, 2005b). Among the participants there was general agreement about the importance of different aspects of TPD, such as facilitating consistency with initial teacher education, empowering teachers to be responsible of their own TPD, and supporting appropriate school strategies to assure the implementation of lifelong learning strategies for teachers. The challenge of sustaining TPD in schools was considered decisive to improve the learning experiences of students in the classroom, even though it was argued that this process could be also stimulated through alternative resources. In short, delegates agreed that each national system had to consider its context characteristics to find the adequate balance in this regard (European Commission, 2005b).

In further discussions, the Council of the European Union has formally agreed that in order to attract and retain high quality staff in the school system, teachers require access to high quality TPD, which is evidence-based and aligned with their professional needs (Council of the EU, 2009). In this sense, the European Union has invited the member states to promote the universal participation of teaching staff in TPD activities. Likewise, the European Commission has been invited to enhance cooperation in this area and to document the development of frameworks on teachers' competences, on the basis of common principles for the region (European Commission, 2005a).

In summary, the participation of all European teachers in TPD is deemed as the main strategy for the short and mid-term improvement of the skills, knowledge and attitudes of teachers that ensure the highest standards in their performance (European

Commission, 2012a). It is worth noting that in a recent publication (2012b), the European Commission has remarked that in the light of the effects of the economic crisis on some national systems, a drastic change will be required in the way that TPD is delivered. In this context, features of high quality TPD informed by research and the preference for school-based teacher learning experiences, such as feedback and appraisal, were acknowledged as critical for the future of TPD supply and the purposes of the Lisbon strategy.

The Organisation for Economic Co-operation and Development

The OECD has persistently posed the need of a lifelong learning approach for the teaching profession considering that pre-service education is not sufficient to deal with the challenges of the process of globalisation (CERI, 1998; Coolahan, 2002; Musset, 2010; OECD, 2005). In 1998, the OECD's Centre for Educational Research and Innovation (CERI) dedicated an entire chapter to describe how educational reforms should recognise the participation of teachers for the purposes of lifelong learning policy. Moreover, the OECD commented that the continual updating of teachers' knowledge and skills had to be accomplished to face global demands such as the introduction of new technologies of information and communication into education.

In further publications, the OECD has also claimed the need of a more holistic approach to TPD, considering the spatial and temporal dimensions of the concept of lifelong learning (Coolahan, 2002; Persson, 2005). The temporal dimension recognises that teacher learning develops over time, thus teacher education has to be considered as a three phases process –e.g. Initial, Induction and Continuing-, an aspect that has been adopted in recent years by different national systems (Conway *et al.*, 2009; Department of Education of Northern Ireland, 2010; Walker *et al.*, 2011). On the other hand, the spatial dimension of lifelong learning underlines that natural learning environments could be equally, or even more, effective to promote teacher learning than mainstream TPD activities. In this regard, a set of several informal types of TPD activities have gained status in contrast to more formal practices such as courses, workshops and seminars. As long as the continuum of TPD gives room to more informal approaches to TPD, schools are seen as learning organisations and teachers as inquirers (OECD, 2005).

A distinctive feature of the discourse of the OECD in relation to the delivery of TPD is the permanent critical standpoint in respect of the developments of research and practice in this field as a mean to introduce guidelines for policy enactment at the national level. To illustrate, the CERI (1998) remarked that at the end of the twentieth century, research was not being successful to explain improvements in teaching quality due to the participation in particular types of TPD. In this sense, the OECD encouraged research and policy to enhance the connection between TPD activities and students' achievement, by aiming TPD activities towards pedagogy as well as content knowledge. It is worth noting that in the same publication the OECD commented that TPD would be determined by national characteristics, such as the country tradition in teacher education, the type of institutions and how they steer the system, the status of the teaching profession and the attitudes of teachers towards curriculum.

In its influential report "Teachers Matter. Attracting, Developing and Retaining Effective Teachers", the OECD (2005) discussed the attributes of effective teaching that should serve as learning objectives for TPD programmes. Even though the impact of developing teaching skills and subject matter knowledge was supported by research, a slight criticism was argued in terms of the presumably small explanatory power of this type of attributes on student learning. According to the OECD, research supporting these findings was questionable as it was simply based on the analysis of association of variables, which not necessarily accounted for causal explanations of the effect on students' learning. Consequently, the issue revealed that few meaningful evaluations were available to guide TPD programmes towards the expected impact in the classrooms.

The OECD documents addressing topics related to TPD also underlined flaws in the implementation of such activities. In general, this criticism regarded that TPD was usually delivered with low intensity, given that the most of these activities were still implemented through workshops, courses and seminars (Musset, 2010; OECD, 2005). This type of implementation would only suppose a sufficient time to reactivate knowledge and skills that were purportedly acquired to a high standard during the pre-service phase of the career. In other words, insofar as the efforts of professionalization were concentrated in that early phase, TPD would require a small dose (or 'top-up') of the same format to elicit changes in teaching practices.

However, the OECD's criticism reveals that this approach makes of TPD an activity undertaken without a following support in the classroom and, consequently, a fragmented experience throughout the career. Hence, and in contrast to initial teacher education curricula, TPD lacks of the necessary coherence and structure to face the challenges of teachers' work. As a result, teachers end up participating in different TPD activities throughout their professional career, but only a few of them would be either logically related. What is more critical, the typical implementation of TPD would be unrelated to teaching practices (Musset, 2010; OECD, 2005), thus its learning activities would have nothing to do with the actual classroom needs. This is particularly worrying, given that no TPD programme could influence student achievement if it is not aimed to improve teacher performance.

It is worth mentioning that the OECD documents in the field of TPD also alert about the barriers for the evaluation of such activities (Musset, 2010; OECD, 2005). It is argued that as long as the structure of programmes and the expertise of trainers are not used to gather and analyse relevant information about features of the TPD delivered, the chance to give accounts of outcomes is limited. Likewise, TPD practices are multiple and not easy to classify, thus any evaluation task, especially those deductively oriented, would always face the problem of a valid identification of discrete activities. As Musset (2010, p. 26) claimed, "it is also difficult to analyze precisely the different types of continuing training since it includes many different activities, with also many different purposes, and with many different forms".

Given the difficulties of evaluating TPD and the recognised importance of teacher policies for school and student outcomes, the OECD launched in 2008 the Teaching and Learning International Survey (TALIS), as a tool to monitor, study and inform global and national policy purposes in this area. In this context, the European Union and the OECD has agreed a formal cooperation on the collection of data for monitoring the indicators of the Lisbon strategy related to TPD (European Commission, 2009). Therefore, it is expected that national and supranational decisions on issues concerning the teaching profession will be informed in the coming years by the results of this research programme[71].

[71] A more detailed description of the role of TPD within the framework of TALIS is presented in Appendix D.

In summary, the key policy sources summarised above demonstrate that the current debate on TPD poses that initial teacher education is insufficient to support school staff with the ever changing context in which education and learning are situated. Therefore, national school systems are encouraged to offer permanent opportunities of learning for their teachers on the basis that as greater this provision, more positive the influence on educational outcomes. However, the quality of this supply is also considered an important issue. In short, it seems necessary to assess the improvement of national school systems in the light of the trade-offs between the universal provision of TPD and the quality of its implementation. The documents produced in this area seem to suggest the need of inquiring which dimensions make of TPD an effective mechanism to improve national educational outcomes. Likewise, the contextual characteristics of each national school system are regularly underlined as relevant factors to support the success of such activities. Hence, from the policy standpoint, it seems also worth to assess this influence under different national settings, so that a fair evaluation of the contribution of TPD can be revealed at the macro level.

Appendix D: The Teaching and Learning International Survey (TALIS)

The framework of TALIS is based on the concept of "effective teaching and learning conditions" (Rutkowski et al., 2013, p. 16), defined as those educational practices contributing to the effective learning of students in specific environments. From this concept, TALIS embraces a model for the contextualisation of the conditions of teaching and learning that follows an "input/process/output" approach to visualise the four levels of national school systems: students, teacher/classrooms, schools and countries/systems. In particular, the programme focuses on measurable and malleable factors (e.g. potentially controlled by policy and practitioners) that can describe international benchmarks for the national policy enactment. Such variables are expected to produce complex patterns of interrelationships in order to explain educational outcomes, thus, for instance, the same factor could work as input and output at the same time, or some inputs can be correlated with each other. In concrete terms, TALIS covers a wide number of dimensions of the work of head-teachers and teachers, and is aimed to understand how these aspects relate with levels of job satisfaction and feelings of self-efficacy.

It is worth highlighting that the organisers of TALIS admit some limitations regarding the use of this type of indicators for the causal analysis of school system variables, which is due to the nature of the multilevel cross-sectional design implemented. In addition, it is acknowledged the potential bias that could be introduced by relying in data collected from self-reported questionnaires. Nonetheless, the main value of the survey would reside in the fact that the indicators included in the questionnaires were deemed as relevant for the educational goals at the national level, because they were prioritised by the participating countries themselves. Therefore, insofar as the integrity and clarity of these variables is warranted over time, TALIS could become a valuable resource for the cross-cultural testing of hypothesis using accumulative and large-scale data. As such, TALIS would become a powerful instrument for the longer term monitoring of trends within and across national educational systems.

In this context, it is striking that themes related to TPD were highly rated for their inclusion in the 2013 cycle by the countries taking part in the programme. Measures

related to the topic can be found at different levels and positions of the framework implemented in TALIS. For instance, at the teacher/classroom level, TPD was deemed as an input factor that is expected to influence at least teachers' practices, their feelings of self-efficacy, and rates of retention and job satisfaction. The extent to which teachers perceive that TPD fulfils their learning needs to a high standard, as well as the level of support received from the school environment, were considered as intervening variables on the effectiveness of TPD.

In this sense, the quality of TPD plays a key role for the TALIS programme in improving teaching and learning, insofar as such activities promote collaboration, active learning, continuity, and the differentiation of contents according to teachers' needs. Consistently, policy questions related to TPD in TALIS were concerned with (1) the type of TPD activity and frequency, (2) the perceived impact, and (3) the associations with teaching profiles, school climate, self-efficacy and job satisfaction. In short, TPD was deemed as an important malleable input factor for the improvement of national school outcomes.

However, themes related to TPD can be also found in the framework of TALIS as a process at the teacher/classroom level, in which case the opportunity of teachers to learn from their own practices in the school setting is considered a critical source for teacher learning. Such processes are explicitly described as school-based teachers' professional practices that promote collaboration and cooperation among the staff (e.g. teacher learning practices, see Chapter 4). Examples of this type of practices are "the exchange of instructional materials, developing curricula, meeting to discuss student progress, and collective learning activities" (Rutkowski et al., 2013, p. 36). The implementation of these processes by teachers would be inhibited by structural deficits of schools, such as the shortage of material resources and exhausting working schedules.

Furthermore, it is worth noting that such opportunities were also mentioned as school and country level processes, so specific policies and courses of action were also highlighted in TALIS as necessary elements to sustain the impact of TPD over time. In particular, inadequate management practices and cultural aspects of the system would difficult teachers' professional practices. However, insofar as these factors are controlled by national school systems, teachers' professional practices are expected to enhance teacher reflection, classroom instruction and student learning.

Appendix E. List of variables (Chapter 2)

Key explanatory variable:

- **Mathematics content-focused TPD (TPDContent)**

  TIMSS 2011 variable name: BTBM29A

"In the past two years, have you participated in professional development in any of the following? Check one circle for each line. a) Mathematics content"

Recoded values: Yes=1; No=0

Control variables:

Block 1, Student background variables:

- **Student gender**

  TIMSS 2011 variable name: BSBG01

  "Are you a girl or a boy?"

  Recoded values: 1=Boy; 0=Girl

- **Books (Number of books in the home)**

  TIMSS 2011 variable name: BSBG04

  "About how many books are there in your home?"

  Recoded values:

  1 = None or very few (0–10 books);

  2 = Enough to fill one shelf (11–25 books);

  3 = Enough to fill one bookcase (26–100 books);

  4 = Enough to fill two bookcases (101–200 books);

  5 = Enough to fill three or more bookcases (more than 200).

- **Parental education (Parents' Highest Education Level)**

  TIMSS 2011 variable name: BSDGEDUP

"What is the highest level of education completed by your mother/father?"

Recoded values:

1 = Some Primary, Lower, Secondary, or No School;

2 = Lower Secondary;

3 = Upper Secondary;

4 = Post-Secondary but Not University;

5 = University.

Block 2, Teacher background variables:

- **Teacher gender**

TIMSS 2011 variable name: BTBG02

"Are you female or male?"

Recoded values: Male = 1; Female = 0.

- **Teaching experience (in years)**

TIMSS 2011 variable name: BTBG01

"By the end of this school year, how many years will you have been teaching altogether?"

- **Math majored (Teacher majored in mathematics)**

TIMSS 2011 variable name: BTBG05A

"During your post-secondary education, what was your major or main area(s) of study? Check one circle for each line. a) Mathematics"

Recoded values: Yes=1; No=0

Block 3, Teacher organisational variables:

- **Teaching hours**

TIMSS 2011 variable name: BTBG08C

"In your current school, how severe is each problem? Check one circle for each line. c) Teachers have too many teaching hours"

Recoded values:

1 = Serious problem;

2 = Moderate problem;

3 = Minor problem;

4 = Not a problem.

- **Teacher shortage**

  TIMSS 2011 variable name: BCBG15A

"How difficult was it to fill eighth-grade teaching vacancies for this school year for the following subjects? Check one circle for each line. a) Mathematics"

Recoded values:

1 = Very difficult;

2 = Somewhat difficult;

3 = Easy to fill vacancies;

4 = Were no vacancies in this subject.

- **Teacher satisfaction**

  TIMSS 2011 variable name: BTBG11B

"How much do you agree with the following statements? Check one circle for each line. b) I am satisfied with being a teacher at this school"

Recoded values:

1 = Disagree a lot;

2 = Disagree a little;

3 = Agree a little;

4 = Agree a lot.

Appendix F. Multiple correlation matrices of predictors by country

US

	TPD Content	Student gender	Books	Parental education	Teacher gender	Teaching experience	Math majored	Teaching hours	Teacher shortage	Teacher satisfaction
TPD Content	1.0									
Student gender	0.0	1.0								
Books	-0.1	0.1	1.0							
Parental education	0.0	0.0	0.4	1.0						
Teacher gender	0.3	0.0	0.0	0.0	1.0					
Teaching experience	-0.1	0.0	0.1	0.1	0.1	1.0				
Math majored	0.1	0.0	0.0	0.0	0.0	0.0	1.0			
Teaching hours	0.2	0.0	0.0	0.0	0.3	-0.1	0.0	1.0		
Teacher shortage	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.0	1.0	
Teacher satisfaction	0.2	0.0	0.1	0.1	0.1	0.2	0.0	0.4	0.1	1.0

England

	TPD Content	Student gender	Books	Parental education	Teacher gender	Teaching experience	Math majored	Teaching hours	Teacher shortage	Teacher satisfaction
TPD Content	1.0									
Student gender	0.0	1.0								
Books	-0.1	0.1	1.0							
Parental education	-0.1	0.0	0.3	1.0						
Teacher gender	-0.1	0.1	0.0	0.0	1.0					
Teaching experience	-0.1	0.0	0.0	0.0	-0.1	1.0				
Math majored	0.1	0.0	0.0	0.0	0.2	-0.1	1.0			
Teaching hours	0.3	0.0	0.0	0.0	0.1	-0.1	0.3	1.0		
Teacher shortage	0.1	-0.1	0.0	0.0	0.1	0.0	-0.1	0.0	1.0	
Teacher satisfaction	0.2	0.0	0.1	0.0	0.2	-0.2	0.1	0.5	0.1	1.0

Japan

	TPD Content	Student gender	Books	Parental education	Teacher gender	Teaching experience	Math majored	Teaching hours	Teacher shortage	Teacher satisfaction
TPD Content	1.0									
Student gender	0.0	1.0								
Books	0.1	0.0	1.0							
Parental education	0.0	0.0	0.3	1.0						
Teacher gender	0.0	0.1	0.0	0.0	1.0					
Teaching experience	-0.2	0.0	0.0	0.0	-0.2	1.0				
Math majored	0.4	0.0	0.1	0.1	0.1	0.0	1.0			
Teaching hours	0.2	0.0	0.0	0.0	0.0	-0.3	0.0	1.0		
Teacher shortage	0.0	0.0	0.1	0.0	-0.2	0.1	0.3	0.0	1.0	
Teacher satisfaction	0.1	0.0	0.0	0.0	-0.1	-0.1	0.0	0.2	-0.2	1.(

Finland

	TPD Content	Student gender	Books	Parental education	Teacher gender	Teaching experience	Math majored	Teaching hours	Teacher shortage	Teacher satisfaction
TPD Content	1.0									
Student gender	0.0	1.0								
Books	0.1	0.2	1.0							
Parental education	0.0	0.0	0.3	1.0						
Teacher gender	0.2	0.0	0.0	0.0	1.0					
Teaching experience	0.3	0.0	0.0	0.0	0.0	1.0				
Math majored	0.2	0.0	0.1	0.0	0.1	0.3	1.0			
Teaching hours	-0.1	0.0	0.0	0.0	0.1	-0.1	0.0	1.0		
Teacher shortage	0.0	0.0	0.0	0.0	0.0	0.1	0.2	0.0	1.0	
Teacher satisfaction	0.0	0.0	0.1	0.0	0.2	0.0	0.1	0.2	0.0	1.(

Source: TIMSS 2011 database

Appendix G. OLS national models

UNITED STATES

	MODEL 0		MODEL 1			MODEL 2		
	estimate	stderr	estimate	stderr		estimate	stderr	
TPD Content	-10.82	8.22	-4.75	6.54		-4.65	6.67	
Student gender			-4.88	1.65	***	-4.58	1.64	**
Books			18.50	0.94	***	18.40	0.94	***
Parental education			11.60	1.17	***	11.58	1.17	***
Teacher gender			3.06	5.53		3.77	5.55	
Teaching experience			0.53	0.23	**	0.57	0.23	**
Math majored			2.98	5.04		2.16	5.10	
Teaching hours						0.97	2.94	
Teacher shortage						0.10	3.13	
Teacher satisfaction						-2.41	3.42	
R-squared	0.01		0.20			0.20		
N	10477		10477			10477		

Source: TIMSS 2011 database

Notes: Outcome variable: Overall mean score Grade 8 Mathematics TIMSS 2011; *p < .1. **p < .05. ***p < .01.

ENGLAND

	MODEL 0			MODEL 1			MODEL 2		
	estimate	stderr		estimate	stderr		estimate	stderr	
TPD Content	-24.18	12.29	*	-14.80	9.65		-16.61	10.01	*
Student gender				-3.65	3.80		-3.48	3.62	
Books				23.76	2.18	***	23.08	2.12	***
Parental education				14.86	2.44	***	14.69	2.38	***
Teacher gender				0.62	9.67		-2.25	9.72	
Teaching experience				-0.10	0.48		-0.04	0.48	
Math majored				12.31	10.64		12.12	10.97	
Teaching hours							-3.08	6.93	
Teacher shortage							1.79	5.41	
Teacher satisfaction							10.75	5.91	*
R-squared	0.02			0.25			0.26		
N	4030			4030			4030		

Source: TIMSS 2011 database

Notes: Outcome variable: Overall mean score Grade 8 Mathematics TIMSS 2011; *p < .1. **p < .05. ***p < .01.

JAPAN

	MODEL 0		MODEL 1			MODEL 2		
	estimate	stderr	estimate	stderr		estimate	stderr	
TPD Content	-7.88	6.60	-9.18	5.83		-11.01	4.87	**
Student gender			-5.13	2.78	*	-5.44	2.56	**
Books			14.46	1.17	***	14.15	1.17	***
Parental education			23.20	1.77	***	22.79	1.66	***
Teacher gender			-7.60	6.16		-5.45	5.83	
Teaching experience			0.03	0.22		0.22	0.22	
Math majored			12.97	6.06	**	13.43	5.45	**
Teaching hours						7.72	2.55	***
Teacher shortage						3.06	3.05	
Teacher satisfaction						8.63	3.08	**
R-squared	0.00		0.18			0.20		
N	4593		4593			4593		

Source: TIMSS 2011 database

Notes: Outcome variable: Overall mean score Grade 8 Mathematics TIMSS 2011; *p < .1. **p < .05. ***p < .01.

FINLAND

	MODEL 0		MODEL 1			MODEL 2		
	estimate	stderr	estimate	stderr		estimate	stderr	
TPD Content	5.56	6.98	-2.65	6.54		-2.94	6.50	
Student gender			-1.52	2.03		-1.48	2.01	
Books			12.69	1.03	***	12.55	1.04	***
Parental education			13.36	1.22	***	13.28	1.21	***
Teacher gender			-1.61	3.73		-2.15	3.84	
Teaching experience			0.15	0.16		0.12	0.16	
Math majored			18.01	5.25	***	17.31	5.34	***
Teaching hours						-1.23	2.53	
Teacher shortage						1.00	2.65	
Teacher satisfaction						2.96	2.59	
R-squared	0.00		0.14			0.15		
N	4286		4286			4286		

Source: TIMSS 2011 database

Notes: Outcome variable: Overall mean score Grade 8 Mathematics TIMSS 2011; *p < .1. **p < .05. ***p < .01.

Appendix H. OLS national models using mathematics pedagogy-focused TPD as key explanatory variable

UNITED STATES

	MODEL 0		MODEL 1			MODEL 2		
	estimate	stderr	estimate	stderr		estimate	stderr	
TPD Pedagogy	-6.54	7.98	-0.52	6.18		0.58	6.29	
Student gender			-4.91	1.66	***	-4.63	1.64	***
Books			18.54	0.95	***	18.48	0.95	***
Parental education			11.72	1.18	***	11.68	1.19	***
Teacher gender			2.67	5.53		3.36	5.49	
Teaching experience			0.53	0.23	**	0.58	0.23	**
Math majored			0.47	3.13		1.83	5.11	
Teaching hours						0.98	2.97	
Teacher shortage						0.12	3.11	
Teacher satisfaction						-2.78	3.38	
R-squared	0.00		0.20			0.20		
N	10477		10477			10477		

Source: TIMSS 2011 database

Notes: Outcome variable: Overall mean score Grade 8 Mathematics TIMSS 2011; *p < .1. **p < .05. ***p < .01.

ENGLAND

	MODEL 0		MODEL 1			MODEL 2		
	estimate	stderr	estimate	stderr		estimate	stderr	
TPD Pedagogy	-21.15	14.97	-11.41	11.96		-12.27	12.38	
Student gender			-3.77	3.73		-3.65	3.63	
Books			23.89	2.08	***	23.22	2.06	***
Parental education			15.32	2.32	***	14.86	2.35	***
Teacher gender			3.02	9.68		-0.44	9.79	
Teaching experience			-0.14	0.44		-0.07	0.45	
Math majored			1.41	5.54		11.66	10.78	
Teaching hours						-3.91	6.85	
Teacher shortage						1.51	5.45	
Teacher satisfaction						10.06	5.99	*
R-squared	0.01		0.24			0.26		
N	4030		4030			4030		

Source: TIMSS 2011 database

Notes: Outcome variable: Overall mean score Grade 8 Mathematics TIMSS 2011; *p < .1. **p < .05. ***p < .01.

JAPAN

	MODEL 0		MODEL 1			MODEL 2		
	estimate	stderr	estimate	stderr		estimate	stderr	
TPD Pedagogy	-5.08	6.18	-3.49	5.11		-4.60	4.71	
Student gender			-5.09	2.82	*	-5.53	2.59	**
Books			14.45	1.17	***	14.04	1.17	***
Parental education			23.62	1.80	***	23.01	1.62	***
Teacher gender			-5.11	5.95		-4.23	5.79	
Teaching experience			0.10	0.21		0.29	0.22	
Math majored			3.11	3.56		10.97	5.38	**
Teaching hours						7.39	2.57	***
Teacher shortage						3.16	3.20	
Teacher satisfaction						8.58	3.01	**
R-squared	0.00		0.18			0.19		
N	4593		4593			4593		

Source: TIMSS 2011 database

Notes: Outcome variable: Overall mean score Grade 8 Mathematics TIMSS 2011; *p < .1. **p < .05. ***p < .01.

FINLAND

	MODEL 0		MODEL 1			MODEL 2		
	estimate	stderr	estimate	stderr		estimate	stderr	
TPD Pedagogy	3.08	4.90	3.69	4.41		2.05	4.40	
Student gender			-1.45	2.02		-1.46	2.01	
Books			12.98	1.04	***	12.51	1.04	***
Parental education			13.58	1.26	***	13.28	1.21	***
Teacher gender			-2.17	3.70		-2.41	3.80	
Teaching experience			0.30	0.16	*	0.11	0.16	
Math majored			1.20	2.50		16.89	5.29	***
Teaching hours						-1.37	2.55	
Teacher shortage						0.94	2.65	
Teacher satisfaction						2.91	2.57	
R-squared	0.00		0.13			0.15		
N	4286		4286			4286		

Source: TIMSS 2011 database

Notes: Outcome variable: Overall mean score Grade 8 Mathematics TIMSS 2011; *p < .1. **p < .05. ***p < .01.

Appendix I. OLS national models using mathematics curriculum-focused TPD as key explanatory variable

UNITED STATES

	MODEL 0		MODEL 1			MODEL 2		
	estimate	stderr	estimate	stderr		estimate	stderr	
TPD Curriculum	6.58	8.02	6.29	6.60		6.48	6.62	
Student gender			-4.91	1.66	***	-4.61	1.65	**
Books			18.55	0.95	***	18.47	0.95	***
Parental education			11.70	1.19	***	11.67	1.19	***
Teacher gender			2.85	5.52		3.58	5.50	
Teaching experience			0.53	0.23	**	0.58	0.23	**
Math majored			0.33	3.14		1.45	5.13	
Teaching hours						1.08	2.98	
Teacher shortage						-0.03	3.12	
Teacher satisfaction						-2.97	3.44	
R-squared	0.00		0.20			0.20		
N	10477		10477			10477		

Source: TIMSS 2011 database

Notes: Outcome variable: Overall mean score Grade 8 Mathematics TIMSS 2011; *p < .1. **p < .05. ***p < .01.

ENGLAND

	MODEL 0		MODEL 1			MODEL 2		
	estimate	stderr	estimate	stderr		estimate	stderr	
TPD Curriculum	-11.87	12.56	-5.54	9.60		-5.45	9.27	
Student gender			-3.99	3.67		-3.87	3.60	
Books			24.05	2.14	***	23.40	2.11	***
Parental education			15.55	2.29	***	15.11	2.29	***
Teacher gender			2.23	9.57		-1.30	9.66	
Teaching experience			-0.04	0.46		0.05	0.48	
Math majored			1.39	5.70		11.79	10.89	
Teaching hours						-3.81	6.66	
Teacher shortage						1.45	5.64	
Teacher satisfaction						9.90	6.06	
R-squared	0.01		0.24			0.25		
N	4030		4030			4030		

Source: TIMSS 2011 database

Notes: Outcome variable: Overall mean score Grade 8 Mathematics TIMSS 2011; *p < .1. **p < .05. ***p < .01.

JAPAN

	MODEL 0		MODEL 1			MODEL 2		
	estimate	stderr	estimate	stderr		estimate	stderr	
TPD Curriculum	-0.31	7.10	-0.07	4.96		-0.63	4.59	
Student gender			-5.16	2.79	*	-5.61	2.57	**
Books			14.33	1.17	***	13.92	1.16	***
Parental education			23.66	1.80	***	23.06	1.63	***
Teacher gender			-4.71	5.87		-3.92	5.72	
Teaching experience			0.14	0.21		0.34	0.21	
Math majored			3.09	3.64		10.32	5.47	*
Teaching hours						7.53	2.59	***
Teacher shortage						3.21	3.33	
Teacher satisfaction						8.29	3.06	**
R-squared	0.00		0.18			0.20		
N	4593		4593			4593		

Source: TIMSS 2011 database

Notes: Outcome variable: Overall mean score Grade 8 Mathematics TIMSS 2011; *p < .1. **p < .05. ***p < .01.

FINLAND

	MODEL 0		MODEL 1			MODEL 2		
	estimate	stderr	estimate	stderr		estimate	stderr	
TPD Curriculum	8.81	5.90	7.15	5.41		4.09	5.06	
Student gender			-1.53	2.01		-1.51	2.00	
Books			12.97	1.03	***	12.50	1.03	***
Parental education			13.54	1.26	***	13.26	1.21	***
Teacher gender			-1.82	3.78		-2.30	3.86	
Teaching experience			0.28	0.16	*	0.10	0.16	
Math majored			1.50	2.50		16.85	5.29	***
Teaching hours						-1.16	2.52	
Teacher shortage						1.11	2.65	
Teacher satisfaction						2.83	2.59	
R-squared	0.00		0.13			0.15		
N	4286		4286			4286		

Source: TIMSS 2011 database

Notes: Outcome variable: Overall mean score Grade 8 Mathematics TIMSS 2011; *p < .1. **p < .05. ***p < .01.

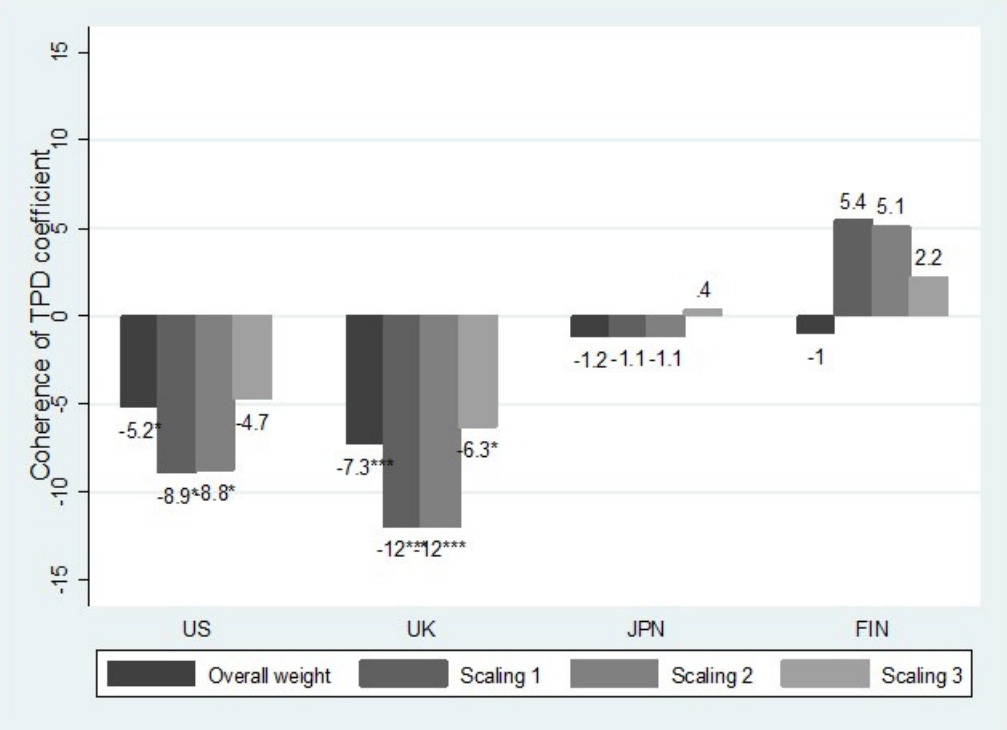
Appendix J. Sensitivity analysis of different methods of scaling

The main findings of the HLM analyses developed in Chapter 3 indicated that the *coherence* of TPD is poorly associated to student achievement, yielding coefficients of small size and negative association. In particular, results suggested that mathematics performance in the US and UK tended to decrease insofar as the *coherence* of TPD in schools was enhanced by head-teachers, whereas this aspect was likely to be not related to the outstanding results of Japan and Finland in the PISA 2012 assessment. The following HLM analyses aims to explore whether similar interpretations of results can be argued when estimates are modelled using the three methods of scaling of the sampling weights discussed in the specialised literature (Rabe-Hesketh and Skrondal, 2006; Stapleton, 2013).

Analysis across the US, UK, Japan and Finland (*coherence* of TPD as measured by the ALL4 MM)

Figure 6.3 compares the regression coefficient of the *coherence* of TPD (controlled for student and school characteristics) of the HLM models that use the overall sampling weight and the three methods of scaling. For each country, the bars represent the size, direction and statistical significance produced when estimates are calculated under each of these conditions.

Figure 6.3 Conditional association between student achievement and coherence of TPD (ALL4 MM) for the US, UK, Japan and Finland using weighted data and different methods of scaling



Source: PISA 2012 database

Notes: Outcome variable: mathematics score in PISA 2012; *p < .1, **p < .05, ***p < .01; coherence of TPD is measured by a standardised factor score with metric invariance across all the four countries of interest (ALL4 MM).

The result that emerges from this exercise is that either the direction or the significance of the estimates do not differ from the initial interpretation based on the model that employs the overall sampling weight. To be more precise, regardless the method of scaling utilised in the HLM, the *coherence* of TPD seems to be negatively associated to student outcomes in the US and UK schools, whereas in Japan and Finland this is not related at all.

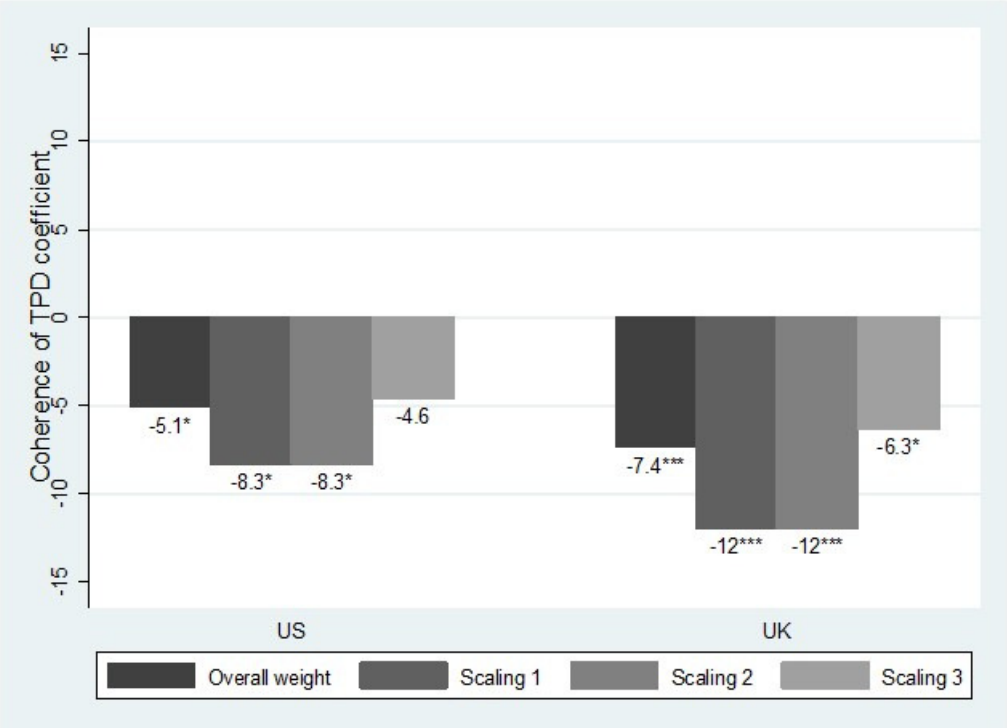
However, the magnitude of the coefficients is affected by the inclusion of the methods of scaling 1 and 2 in the two English-speaking countries, with absolute values that practically double the size of the estimates based on using the overall weight (and the method of scaling 3 in the UK). For instance, for US students from different schools, one standard deviation improvement in the *coherence* of TPD leads to 5.2 points less in the PISA assessment when the model is fitted with the overall weight, whereas this value increases to 8.9 and 8.8 using the methods of scaling 1 and 2, respectively. In the UK,

the estimate transforms from -7.3 (overall weight) or -6.3 (method of scaling 3) points to -12 (methods of scaling 1 and 2).

Analysis across the US and UK (*coherence* of TPD as measured by the US&UK MM)

Figure 6.4 replicates the previous analysis with data from the US and UK and using as key explanatory variable the factor score of the *coherence* of TPD that showed satisfactory metric invariance between these two countries (e.g. US&UK MM).

Figure 6.4 Conditional association between student achievement and coherence of TPD (US&UK MM) for the US and UK using weighted data and different methods of scaling



Source: PISA 2012 database
Notes: Outcome variable: mathematics score in PISA 2012; *p < .1, **p < .05, ***p < .01; coherence of TPD is measured by a standardised factor score with metric invariance across the US and UK (US&UK MM).

Similar results are found in this case, suggesting that the *coherence* of TPD is inversely related to student achievement, whereas estimates from the methods of scaling

1 and 2 are approximately twice the value of the results yielded when the overall weight and the method of scaling 3 are employed. To illustrate, for British students from different schools, one standard deviation increase in the *coherence* of TPD is associated to 7.4 (overall weight) or 6.3 (method of scaling 3) points less in mathematics, whereas this value increases to -12 points using the methods of scaling 1 and 2. In turn, the estimate in the US shifts from -5.1 (overall weight) to -8.3 (methods of scaling 1 and 2) points.

HLM models using the US MM

Finally, the same analysis is performed with data from the US and the scale of *coherence* in TPD that was found to be a valid measuring instrument only for this country (e.g. US MM). Table 6.4 compares the regression coefficient of the key explanatory variable in the four conditions under analysis and additionally details the estimates of the control variables included in the models at the student and school level.

Table 6.4 Means-as-Outcomes HLM models for the US using weighted data and different methods of scaling

	Overall weight			Scaling 1			Scaling 2			Scaling 3		
	Coef	**SE**		**Coef**	**SE**		**Coef**	**SE**		**Coef**	**SE**	
School-level variables												
Coherence of TPD	-5.1	(3.1)		-7.9	(4.8)		-7.9	(4.8)		-4.6	(3.3)	
Administration (public)	-0.9	(11.3)		-17.2	(15.8)		-17.2	(15.8)		2.3	(11.6)	
Location	-4.4	(4.0)		-6.3	(6.0)		-6.3	(6.0)		-0.8	(4.4)	
Class size	-0.6	(0.6)		0.6	(1.0)		0.6	(1.0)		-0.8	(0.7)	
School size	0.0	(0.0)	*	0.0	(0.0)		0.0	(0.0)		0.0	(0.0)	*
Student-level variables												
Gender (male)	8.5	(2.6)	***	7.4	(3.6)	**	7.5	(3.6)	**	8.3	(2.5)	***
Immigrant	-1.0	(5.8)		4.8	(8.3)		4.8	(8.3)		-1.5	(5.6)	
SES	23.9	(2.0)	***	27.0	(2.1)	***	27.0	(2.1)	***	26.4	(1.8)	***
Intercept	491			473			473			484		
Between-school variance	38.0			34.9			34.8			34.0		
Within-school variance	74.3			72.3			72.3			75.6		

Source: PISA 2012 database

Notes: Outcome variable: mathematics score in PISA 2012; SE=Standard Error; *p < .1, **p < .05, ***p < .01; coherence of TPD is measured by a standardised factor score (US MM).

It can be clearly seen that the interpretation of results do not differ across these four methods of estimation, because no significant association is found for the *coherence* of TPD in the US. In other words, in this country the contribution of this version of the key explanatory variable to student achievement seems to be zero, regardless the method of scaling employed, as well as the individual characteristics of students and their schools.

In conclusion, results from this sensitivity analysis of the three methods of scaling of the sampling weights are consistent with the findings reported in the main analysis of Chapter 3. The coefficients yielded under each of these conditions agree in terms of the direction and statistical significance of the contribution of the *coherence* of TPD for each measurement model employed and country analysed. To be more precise, regardless the method used for the HLM –and the characteristics of students and schools included as controls-, mathematics achievement seems to be inversely related to the key explanatory variable in the US and UK, whereas in Japan and Finland there is no evidence of association. In this context, the most striking finding is that the absolute value of the regression coefficients calculated with the methods of scaling 1 and 2 tended to be noticeably greater than those obtained with the overall inclusion weight and the method of scaling 3.

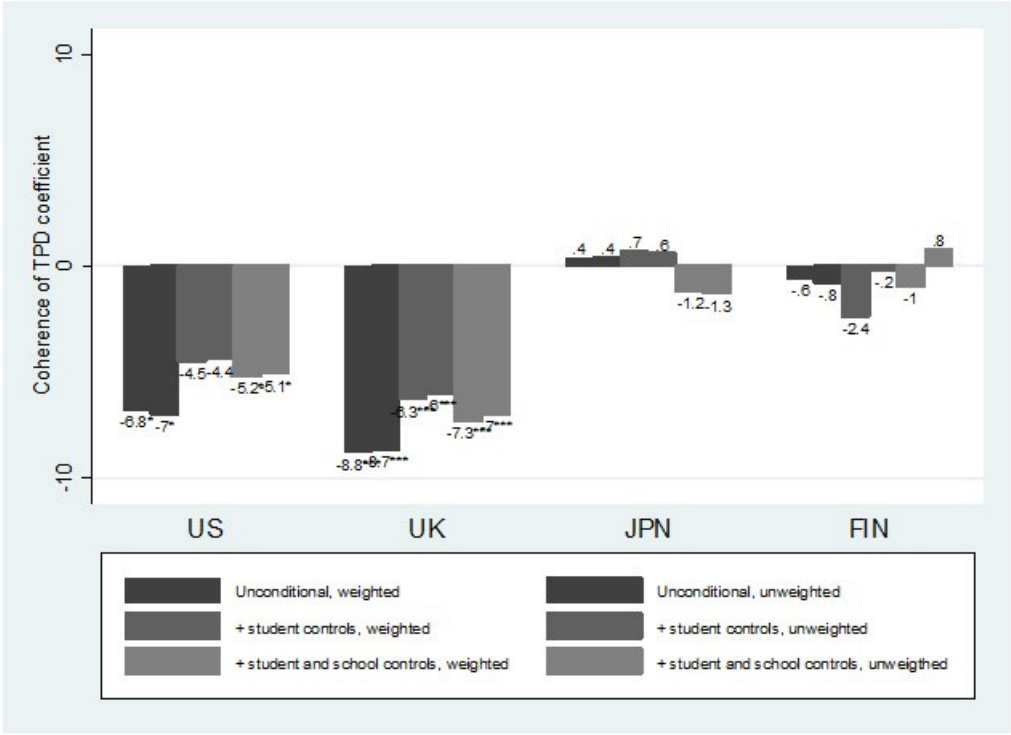
Appendix K. Informativeness of weights analysis

In addition to the sensitivity analysis presented in Appendix J, it is suggested to evaluate the potential effect of the survey design by contrasting the interpretation of the estimates reported in the main HLM analysis of this chapter with results obtained without using sampling weights. Following recommended practice in this regard (Anderson, Kim and Keller, 2013; Kim, Anderson and Keller, 2013; Stapleton, 2013), the next HLM analyses examine whether similar findings emerge when estimates are analysed using unweighted data.

Analysis across the US, UK, Japan and Finland (*coherence* of TPD as measured by the ALL4 MM)

Figure 6.5 compares all the regression coefficients of the *coherence* of TPD already reported in the main analysis of Chapter 3 against their corresponding estimates obtained without using sampling weights. For each country and HLM model, adjoining bars with the same grey colour symbolise the size, direction and statistical significance produced when estimates are calculated under each of these two conditions.

Figure 6.5 Variation of the conditional association between student achievement and coherence of TPD across Means-as-Outcomes HLM models for the US, UK, Japan and Finland (ALL4 MM) using weighted and unweighted data



Source: PISA 2012 database

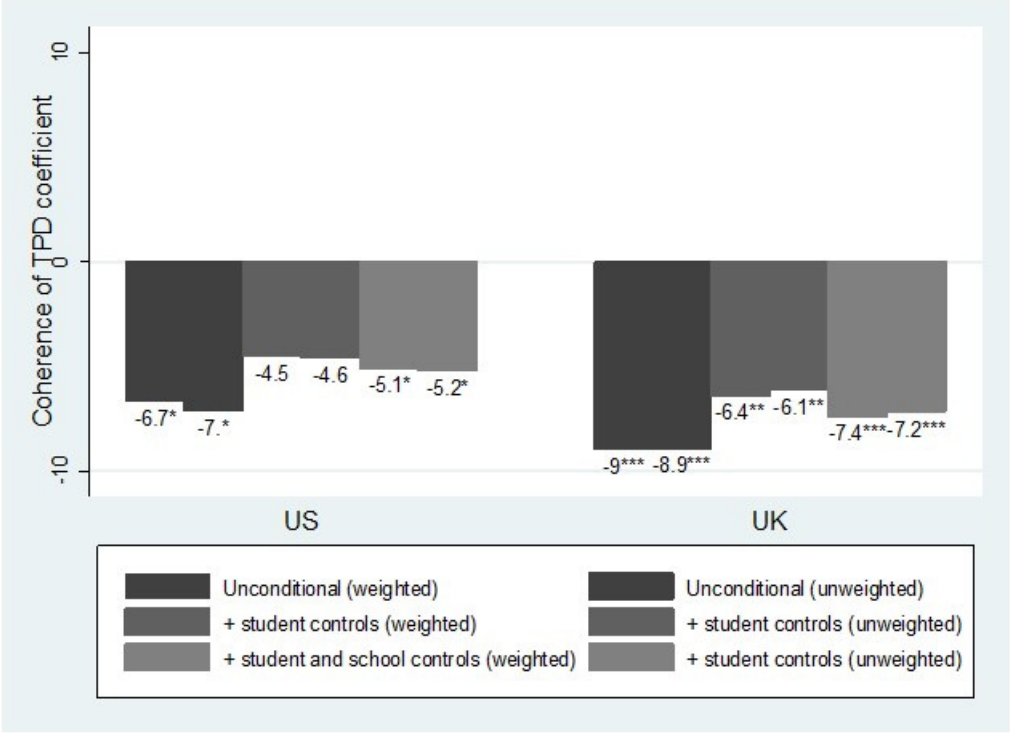
Notes: Outcome variable: mathematics score in PISA 2012; *p < .1, **p < .05, ***p < .01; coherence of TPD is measured by a standardised factor score with metric invariance across all the four countries of interest (ALL4 MM).

The results suggest that there is a null effect of fitting the models with or without the overall inclusion weights, given that general findings across countries do not vary and the difference in the absolute values of significant estimates is trivial (less than 0.3 points in the PISA scale). In this sense, regardless sampling weights are used in the analyses, the *coherence* of TPD seems to be inversely associated to student outcomes in the US (unconditional model and model controlled by student and school variables) and UK schools, whereas in Japan and Finland the association is likely to be zero.

Analysis across the US and UK (*coherence* of TPD as measured by the US&UK MM)

Figure 6.6 reproduces the previous analysis with data from the US and UK and using as key explanatory variable the factor score that showed satisfactory metric invariance between these two countries (e.g. US&UK MM).

Figure 6.6 Variation of the conditional association between student achievement and coherence of TPD across Means-as-Outcomes HLM models for the US and UK (US&UK MM) using weighted and unweighted data



Source: PISA 2012 database

Notes: Outcome variable: mathematics score in PISA 2012; *p < .1, **p < .05, ***p < .01; coherence of TPD is measured by a standardised factor score with metric invariance across the US and UK (US&UK MM).

Similar results are found in this case, as results do not differ whether overall inclusion weights are used or not, and the dissimilarities between the corresponding significant point estimates are also smaller than 0.3 points. The *coherence* of TPD is inversely related to student achievement in these two countries regardless the analysis is undertaken using the sampling weights. To illustrate, for British students from different schools, one standard deviation improvement in the level of *coherence* of TPD leads to

7.4 (weighted) or 7.2 (unweighted) points less in the PISA assessment, controlling for student and school characteristics.

HLM models using the US MM

Lastly, a similar analysis is developed with data from the US and using as key explanatory variable the factor score of *coherence* in TPD that was found to be a valid measuring instrument only for this country (e.g. US MM). Table 6.5 compares the regression coefficients of each HLM model fitted with and without sampling weights.

Table 6.5 Means-as-Outcomes HLM models for the US using weighted and unweighted data

	Weighted			Unweighted			Weighted			Unweighted			Weighted			Unweighted		
	Coef	**SE**		**Coef**	**SE**		**Coef**	**SE**		**Coef**	**SE**		**Coef**	**SE**		**Coef**	**SE**	
School-level variables																		
Coherence of TPD	-6.4	(3.5)	*	-6.9	(3.5)	*	-4.5	(3.0)		-4.6	(2.9)		-5.1	(3.1)		-5.2	(2.9)	*
Administration (public)													-0.9	(11.3)		4.7	(11.6)	
Location													-4.4	(4.0)		-3.0	(3.4)	
Class size													-0.6	(0.6)		-0.6	(0.6)	
School size													0.0	(0.0)	*	0.0	(0.0)	
Student-level variables																		
Gender (male)							8.7	(2.5)	***	10.4	(2.6)	***	8.5	(2.6)	***	10.3	(2.7)	***
Immigrant							-0.5	(5.7)		-1.3	(4.8)		-1.0	(5.8)		-1.9	(4.9)	
SES							23.7	(1.9)	***	26.2	(1.4)	***	23.9	(2.0)	***	26.7	(1.4)	***
Intercept	481			482			474			473			491			485		
Between-school variance	47.1			42.8			38.5			33.4			38.0			33.1		
Within-school variance	77.1			78.2			74.1			75.1			74.3			75.2		

Source: PISA 2012 database

Notes: Outcome variable: mathematics score in PISA 2012; SE=Standard Error; *p < .1, **p < .05, ***p < .01; coherence of TPD is measured by a standardised factor score (US MM).

In general, estimates of the key explanatory variable hold the same characteristics across these two conditions in terms of size, direction and statistical significance. The only exception is the significance of the coefficient in the model that controlled for student and school characteristics. In this case, the unweighted estimate yielded a consistent coefficient (-5.2 points, $p<.1$), whereas the weighted analysis produced a value without statistical significance (-5.1 points), which is due to the smaller standard error yielded by the unweighted analysis.

To sum up, results from this informativeness of sampling weights analysis are consistent with the main findings of Chapter 3. Almost in all cases, the coefficients fitted with or without overall inclusion weights yielded similar magnitudes, directions and statistical significances in relation to the contribution of the *coherence* of TPD for each measurement model employed and country analysed. In other words, regardless the HLM analysis is performed with sampling weights, the achievement of British and US students seems to be negatively related to the degree of *coherence* of TPD in their schools, whereas in Japan and Finland no association is found.

Appendix L. HLM analyses of the influence of items removed from the original scale in conjunction with factors of coherence of TPD across countries of interest

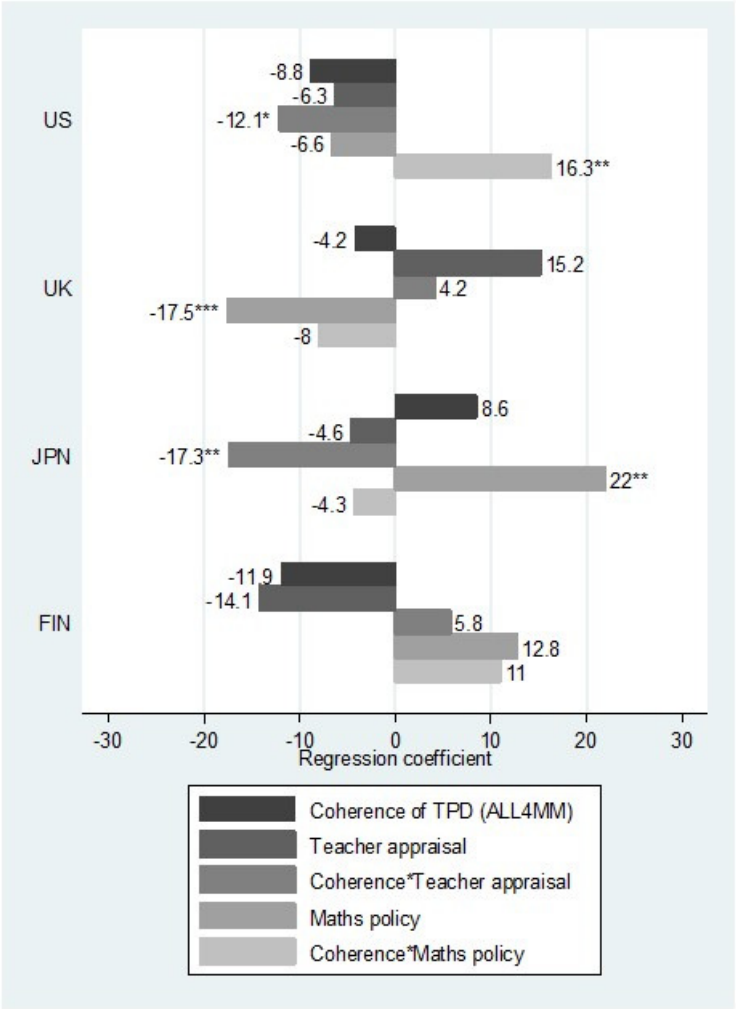
The following analyses introduce models that were specified to take advantage of all the data available, including the two items that were removed from the original scale of *coherence* of TPD. This approach estimates the regression coefficients of the key explanatory variable when such items are simultaneously included in the model as separate dummies[72] and along with a term representing their interaction with the *coherence* of TPD. The purpose here is to develop a more detailed examination of the specific contribution to student achievement of the implementation of standardised policies for mathematics and the extent to which teacher appraisals are linked to TPD, once students' and schools' characteristics are controlled.

Analysis across the US, UK, Japan and Finland (*coherence* of TPD as measured by the ALL4 MM)

The first HLM model developed in this appendix utilises as level-2 key explanatory variable the standardised factor score based on the three items of the *coherence* of TPD that showed satisfactory metric invariance across all the four countries of interest (e.g. ALL4 MM). Figure 6.7 displays the partial contribution of the key explanatory variable, the items removed from the original instrument and their respective interaction terms. For each country, the first bar represents the size of the association of the *coherence* of TPD with student achievement, whereas the second and fourth bars illustrate the partial association of the items related to teacher appraisals and the implementation of standardised policies for mathematics in schools. The third and fifth bars indicate the corresponding interactions with the key explanatory variable.

[72] The original categories of the item related to teacher appraisals ("extent to which appraisals of and/or feedback to teachers have directly led to opportunities for TPD") were recoded as 0=No opportunities for TPD and 1=At least small opportunities for TPD.

Figure 6.7 Means-as-Outcomes HLM models of the coherence of TPD, teacher appraisals and TPD, and standardised policies for Mathematics for the US, UK, Japan and Finland



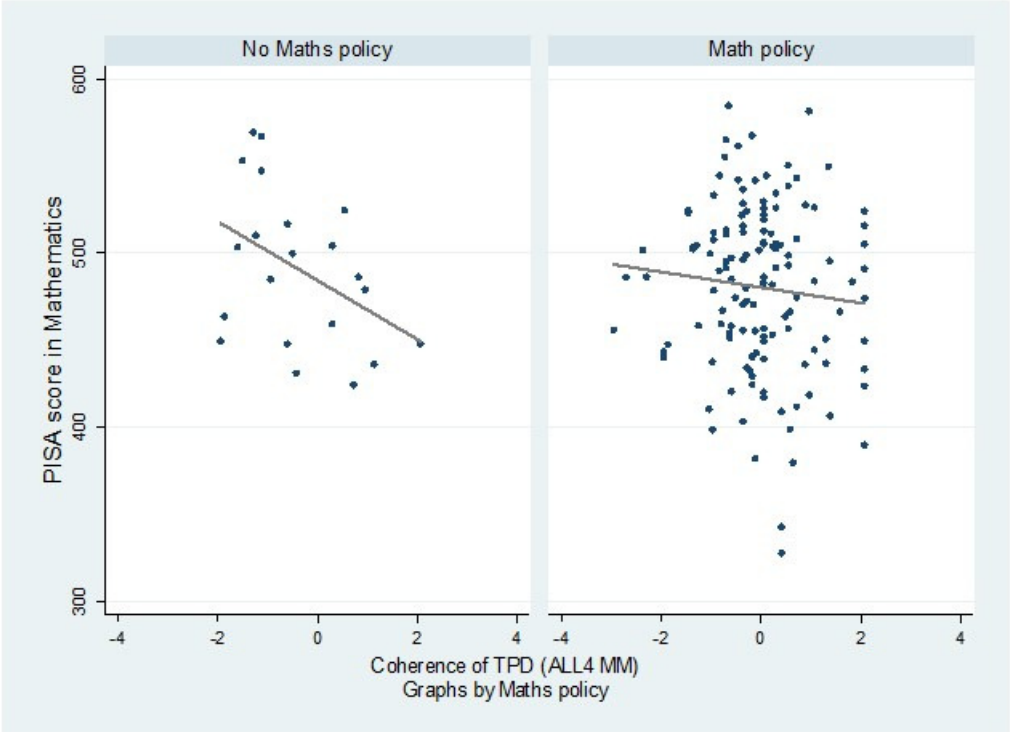
Source: PISA 2012 database

Notes: Outcome variable: mathematics score in PISA 2012; *p < .1, **p < .05, ***p < .01; coherence of TPD is measured by a standardised factor score with metric invariance across all the four countries of interest (ALL4 MM).

In general, results indicated that the variables included in the model were poorly associated to the performance of students in mathematics (absolute values of estimates were smaller than 22% of one standard deviation in the PISA assessment across all OECD countries). However, it is worth noting that they worked differently in each country, as in some cases they showed significant interactions with the key explanatory variable or denoted specific differences in school outcomes. For instance, in the US the relationship between the *coherence* of TPD and student achievement was different for schools whether they implemented standardised policies for mathematics or not,

signalled by the statistical significance of the interaction term (16.3 points) included in the model. To illustrate such interaction, Figure 6.8 depicts the relationship between the key explanatory variable and student achievement separated by the implementation of such plans.

Figure 6.8 Coherence of TPD and student achievement in PISA 2012 in the US, by implementation of a standardised policy for Mathematics in schools



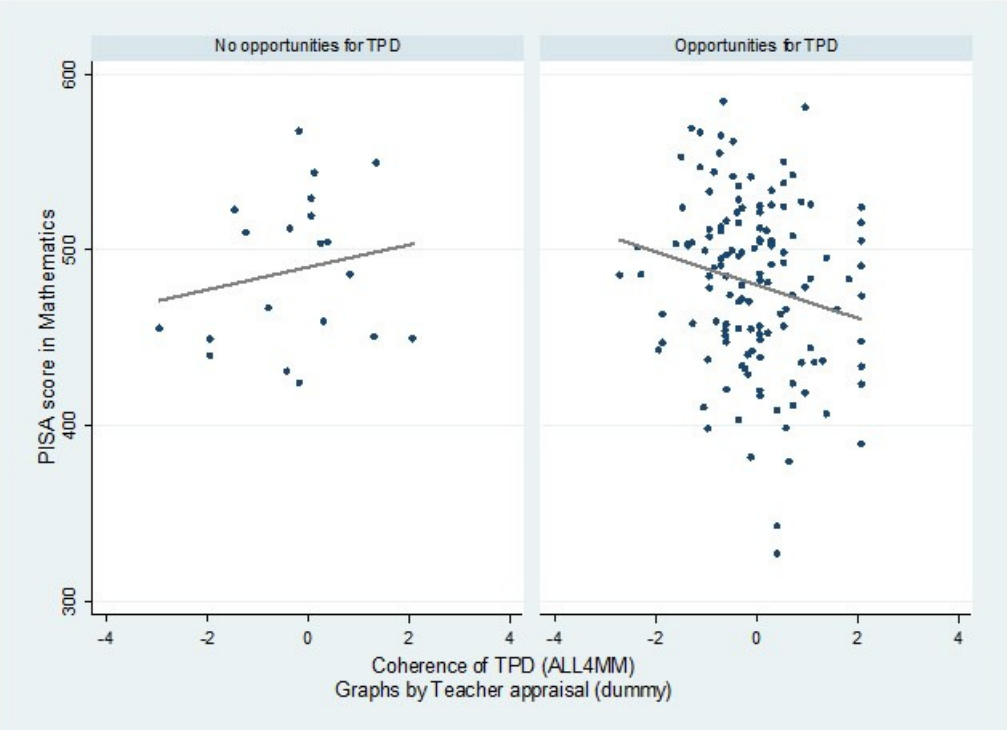
Source: PISA 2012 database
Notes: weighted data; coherence of TPD is measured by a standardised factor score with metric invariance across the US, UK, Japan and Finland (ALL4 MM).

The graphs show that for schools with no implementation of standardised policies for mathematics the relationship was negative –e.g. characterised by a steeper downward slope of the best fit line-, whereas for schools that put into practice such schemes the inverse influence of the *coherence* of TPD was practically overridden. These results suggested that the previously reported negative association between the *coherence* of TPD and student achievement in this country could be attenuated by the enactment of such policies in schools.

For British schools, the influence of implementing a standardised policy for mathematics was significant and associated to -17.5 points in the PISA scale –see Figure 6.7-, regardless the level of *coherence* of TPD in schools and the rest of variables included in the model. On the contrary, the implementation of such schemes seemed to be importantly beneficial for Japanese schools, as results indicated that for Japanese students from different schools this variable was associated to 22 points more in the PISA assessment. In both countries, these estimates held even when the level of *coherence* of TPD, as well as the other school and student characteristics, were considered in the model, and it represented approximately one fifth of a standard deviation in the PISA scale across OECD countries.

On the other hand, both in the US and Japan the analysis yielded a negative interaction between the *coherence* of TPD and the extent to which teacher appraisals led to opportunities for TPD. To be more precise, the direction of the association between the key explanatory variable and student achievement was different whether head-teachers linked their appraisals with opportunities for TPD or not. Figure 6.9 and 6.10 illustrate this aspect by plotting this relationship for each of the two levels of the dummy variable labelled as “teacher appraisals”.

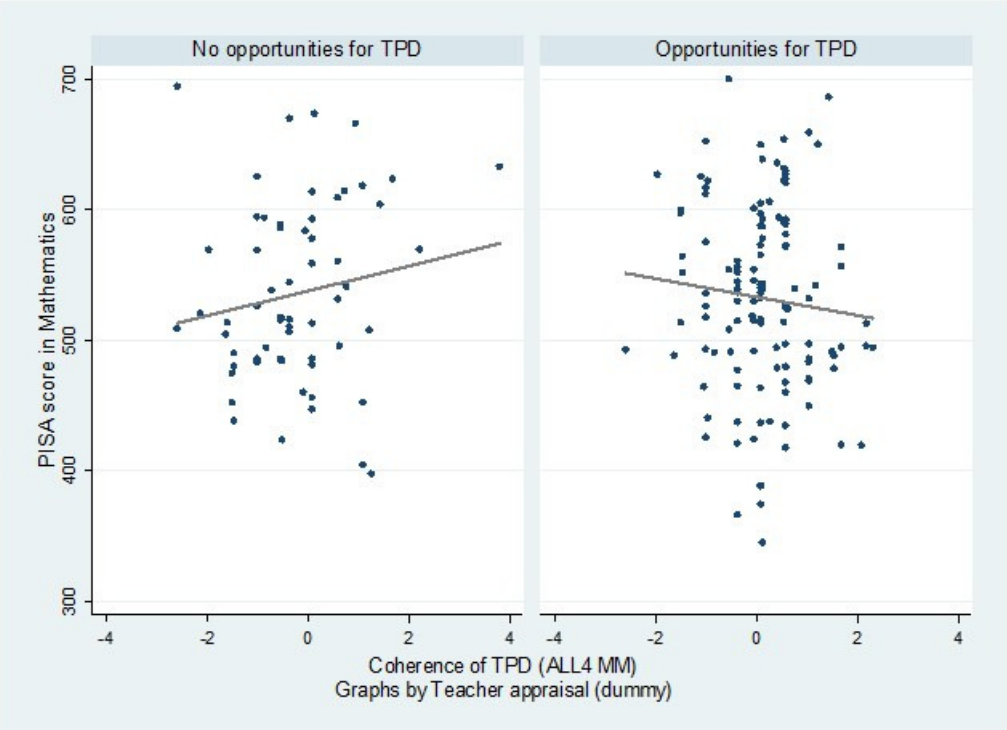
Figure 6.9 Coherence of TPD and student achievement in PISA 2012 in the US, by disposition of head-teachers to link teacher appraisals and opportunities for TPD



Source: PISA 2012 database

Notes: weighted data; coherence of TPD is measured by a standardised factor score with metric invariance across the US, UK, Japan and Finland (ALL4 MM).

Figure 6.10 Coherence of TPD and student achievement in PISA 2012 in Japan, by disposition of head-teachers to link teacher appraisals and opportunities for TPD



Source: PISA 2012 database

Notes: weighted data; coherence of TPD is measured by a standardised factor score with metric invariance across the US, UK, Japan and Finland (ALL4 MM).

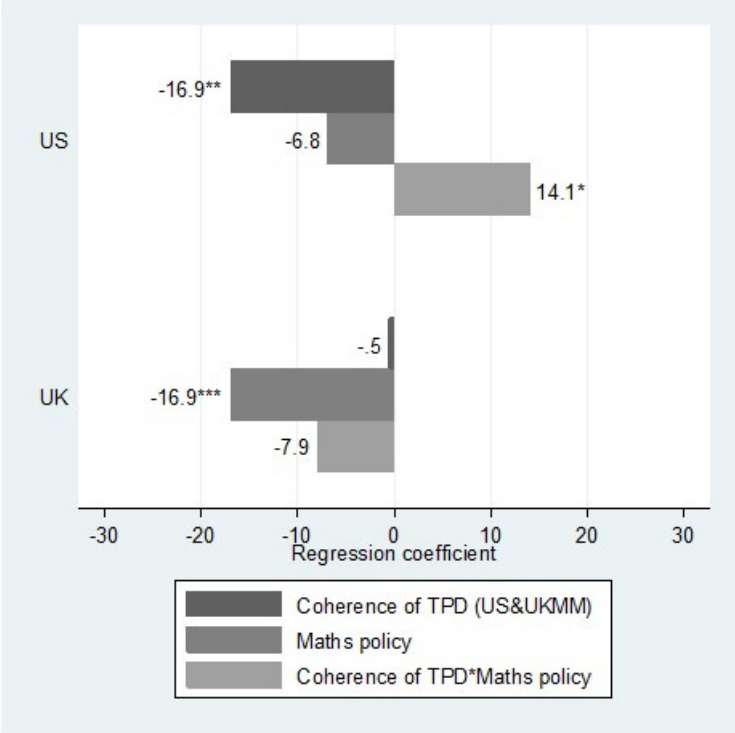
In both cases, the graphs indicated a positive association of the *coherence* of TPD and student achievement in schools where head-teachers did not link their appraisals with opportunities for TPD. On the contrary, the association seemed to be negative when such opportunities were contingent to the evaluation of teachers' performance undertaken by school leaders. This is an interesting finding because it would suggest that the *coherence* of TPD can be beneficial to student outcomes in the US and Japan insofar as the opportunities for TPD are independent from the appraisal carried out by head-teachers.

Finally, it is worth highlighting that in Finland none of the variables suggested in this appendix to measure the *coherence* of TPD were associated to student achievement in any of the HLM models specified in the analysis.

Analysis across the US and UK (*coherence* of TPD as measured by the US&UK MM)

The second HLM model examined in this appendix utilises as level-2 key explanatory variable the standardised factor score based on the four items of the *coherence* of TPD that showed satisfactory metric invariance across the US and UK (e.g. US&UK MM). Figure 6.11 shows the conditional association with student achievement of the *coherence* of TPD, the implementation of a standardised policy for mathematics and the interaction between these two predictors, represented by the corresponding bars for each country.

Figure 6.11 Means-as-Outcomes HLM models of the coherence of TPD, and standardised policies for mathematics for the US and UK



Source: PISA 2012 database

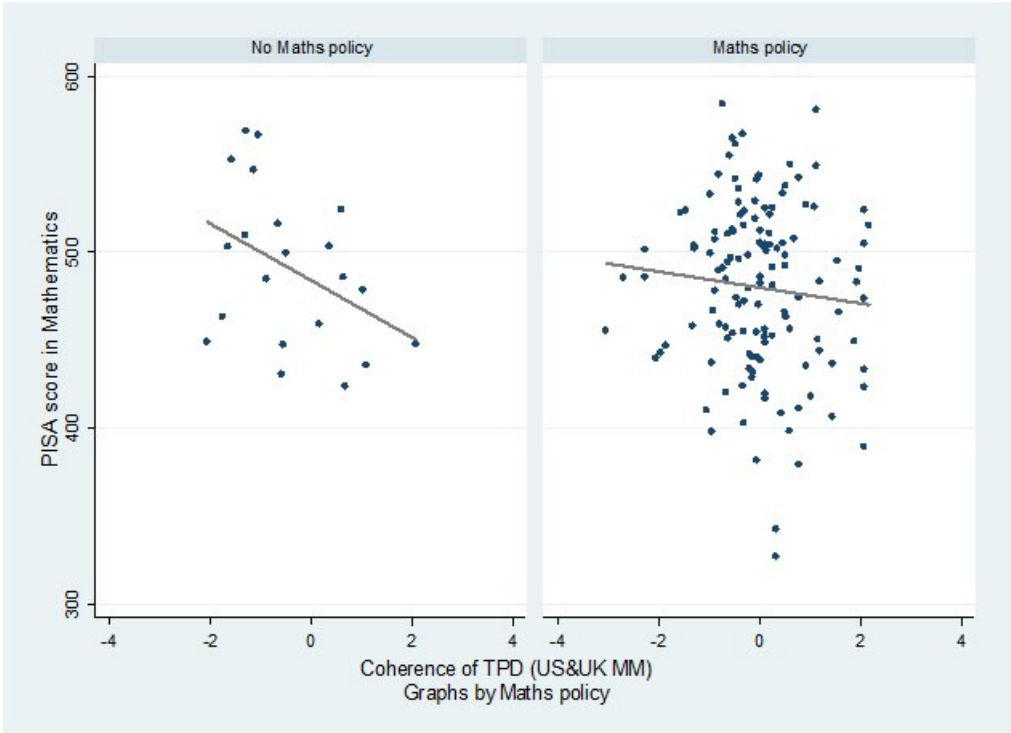
Notes: Outcome variable: mathematics score in PISA 2012; *p < .1, **p < .05, ***p < .01; coherence of TPD is measured by a standardised factor score with metric invariance across the US and UK (US&UK MM).

As in the preceding analysis, results indicated that this group of variables were weakly related to the achievement of students in mathematics, with absolute values of the coefficients smaller than 17% of one standard deviation in the PISA scale across

OECD countries. Nevertheless, each country presented a different pattern of results. In the US, the contribution of the *coherence* of TPD was significantly negative and the implementation of standardised policies for mathematics in the UK revealed a significant negative association, too. To be more precise, for US students from different schools, one standard deviation increase in the *coherence* of TPD was associated to a decline of approximately 17 points in the PISA score, whereas the same drop was observed in British schools that implemented the aforementioned plans for mathematics, regardless the rest of variables included in the model.

On the other hand, results indicated that for US schools there was a positive interaction (14.1 points) between the *coherence* of TPD and the dummy variable under evaluation. In this regard, Figure 6.12 presents the relationship between the key explanatory variable and student achievement separated by the implementation of standardised policies for mathematics in schools.

Figure 6.12 Coherence of TPD and student achievement in PISA 2012 in the US, by implementation of a standardised policy for mathematics in schools



Source: PISA 2012 database

Notes: weighted data; coherence of TPD is measured by a standardised factor score with metric invariance across the US and UK (US&UK MM).

The scatterplots confirms that for schools with no application of such schemes the relationship was more negative than for schools that implemented them, illustrated by a steeper downward slope of the best fit line. These results suggested that the negative relationship between the *coherence* of TPD and student achievement in this country might be weakened by the enactment of standardised policies for mathematics.

In summary, results from this appendix confirms findings remarked in the main analysis of Chapter 3 in terms of a null or small contribution of the *coherence* of TPD to student achievement for the countries under evaluation. In addition, the analysis here presented indicates that in countries such as the UK and Japan, individual variables theoretically related to the *coherence* of TPD (e.g. "Maths policy" and "teacher appraisal") make a difference to the average achievement of students. Further, that in the US and Japan there are relevant interactions between these variables and the *coherence* of TPD that introduce particular patterns of association between the key explanatory variable and the achievement of students for different types of schools.

Bibliography

Agresti, A. (2002). *Categorical data analysis*. New Jersey: John Wiley & Sons, Inc.

Agresti, A. (2007). *An introduction to categorical data analysis*. New Jersey: Wiley-Interscience.

Anderson, C. J., Kim, J.-S. and Keller, B. (2013). 'Multilevel Modeling of Categorical Response Variables'. *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis,* 481.

Asparouhov, T. and Muthén, B. (2009). 'Exploratory Structural Equation Modeling'. *Structural Equation Modeling: A Multidisciplinary Journal,* 16 (3), 397-438.

Baglin, J. (2014). 'Improving Your Exploratory Factor Analysis for Ordinal Data: A Demonstration Using FACTOR'. *Practical Assessment, Research & Evaluation,* 19 (5), 2.

Baker, D. P. and Letendre, G. K. (2005). 'The universal math teacher? International beliefs, national work roles, and local practice'. In D. P. Baker and G. K. Letendre (Eds), *National differences, global similarities* (pp. 104-116). Stanford, California: Stanford University Press.

Baker, D. P., Letendre, G. K., Astiz, M. F. and Wiseman, A. (2005). 'Slouching toward a global ideology. The devolution revolution in education governance'. In D. P. Baker and G. K. Letendre (Eds), *National differences, global similarities* (pp. 134-149). Stanford, California: Stanford University Press.

Bakkenes, I., Vermunt, J. D. and Wubbels, T. (2010). 'Teacher learning in the context of educational innovation: Learning activities and learning outcomes of experienced teachers'. *Learning and Instruction,* 20 (6)*,* 533-548.

Bando, R. (2013). 'Guidelines for Impact Evaluation in Education Using Experimental Design'. *IDB Technical Note (Office of Strategic Planning and Development Effectiveness).*

Barber, M. and Mourshed, M. (2007). *How the World's Best-Performing School Systems Come Out on Top*. London: McKinsey & Co.

Bartholomew, D. J., Steele, F., Galbraith, J. and MoustakI, I. (2008). 'Confirmatory Factor Analysis and Structural Equation Models', *Analysis of Multivariate Social Science Data* (Second edition ed.). London: Chapman and Hall/CRC

Beavers, A. S., Lounsbury, J. W., Richards, J. K., W. Huck, S., Skolits, G. J. and Esquivel, S. L. (2013). 'Practical Considerations for Using Exploratory Factor Analysis in Educational Research'. *Practical Assessment, Research & Evaluation,* 18 (6)*,* 1-13.

Birman, B., Desimone, L., Porter, A. and Garet, M. (2000). 'Designing Professional Development That Works'. *Educational Leadership,* 57 (8)*,* 28-33.

Blank, R. K. and de las Alas, N. (2009). *Effects of Teacher Professional Development on Gains in Student Achievement. How Meta Analysis Provides Scientific Evidence Useful to Education Leaders*. Washington, D. C.: Council of Chief State School Officers.

Block, J. H. and Hazelip, K. (1995). 'Teachers' beliefs and belief systems'. In L. W. Anderson (Ed.), *International encyclopedia of teaching and teacher education* (2nd ed., pp. 25-28). Kidlington, Oxford, UK: Elsevier Science Ltd.

Blumenfeld, P. C., Soloway, E., Marx, R. W., Krajcik, J. S., Guzdial, M. and Palincsar, A. (1991). 'Motivating Project-Based Learning: Sustaining the Doing, Supporting the Learning'. *Educational Psychologist,* 26 (3-4)*,* 369-398.

Borko, H. (2004). 'Professional Development and Teacher Learning: Mapping the Terrain'. *Educational Researcher,* 33 (8)*,* 3-15.

Borko, H., Elliott, R. and Uchiyama, K. (2002). 'Professional development: a key to Kentucky's educational reform effort'. *Teaching and Teacher Education,* 18 (8)*,* 969-987.

Brown, T. A. (2006). *Confirmatory Factor Analysis for Applied Research*. New York: The Guilford Press.

Browne, M. W. (2001). 'An Overview of Analytic Rotation in Exploratory Factor Analysis'. *Multivariate Behavioral Research,* 36 (1)*,* 111-150.

Burstein, L., McDonnell, L. M., Van Winkle, J., Ormseth, T., Mirocha, J. and Guiton, G. (1995). 'Validating National Curriculum Indicators'.

Byrne, B. (2012). *Structural equation modeling with Mplus: basic concepts, applications, and programming*. London: Taylor & Francis.

Caena, F. (2011). *Literature review. Quality in Teachers' continuing professional development*: European Commission. Directorate-General for Education and Culture.

Campbell, D. T. and Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Boston, MA, US: Houghton, Mifflin and Company.

CERI (1998). *Staying ahead. In-service training and teacher professional development*. Paris: Centre for Educational Research and Innovation. OECD.

Chan, Y. (2003). 'Biostatistics 104: correlational analysis'. *Singapore Med J,* 44 (12), 614-9.

Chang, L.-C. and Lee, G. C. (2010). 'A team-teaching model for practicing project-based learning in high school: Collaboration between computer and subject teachers'. *Computers & Education,* 55 (3), 961-969.

Chen, F. F. (2007). 'Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance'. *Structural Equation Modeling: A Multidisciplinary Journal,* 14 (3), 464-504.

Clarke, D. and Hollingsworth, H. (2002). 'Elaborating a model of teacher professional growth'. *Teaching and Teacher Education,* 18 (8), 947-967.

Clarke, P., Crawford, C., Steele, F. and Vignoles, A. F. (2010). 'The choice between fixed and random effects models: some considerations for educational research'. *IZA Discussion Paper* (5287).

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*: Psychology Press.

Conway, P. F., Murphy, R., Rath, A. and Hall, K. (2009). *Learning to teach and its implications for the continuum of teacher education: A nine-country cross-national study. Report commissioned by the Teaching Council*. Cork, Ireland: University College Cork.

Coolahan, J. (2002). *Teacher Education and the Teaching Career in an Era of Lifelong Learning* (OECD Education Working Papers No. 2). Paris: Organisation for Economic Co-operation and Development.

Costello, A. B. and Osborne, J. W. (2005). 'Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most From Your Analysis'. *Practical Assessment, Research & Evaluation,* 10 (7)*,* 1-9.

Council of the EU (2009). *Council conclusions on the professional development of teachers and school leaders*. Brussels: The Council of the European Union.

Craig, H., Kraft, R. and du Plessis, J. (1998). *Teacher development. Making and impact*. Washington, D. C.: World Bank. Human Development Network. Effective Schools and Teachers.

Dancey, C. and Reidy, J. (2014). 'Correlational analysis: Pearson's r', *Statistics Without Maths for Psychology*. London: Pearson.

Darling-Hammond, L. and Ball, D. L. (1998). *Teaching for high standards: What policymakers need to know and be able to do*. Philadelphia: Consortium for Policy Research in Education and the National Commission on Teaching & America's Future.

Darling-Hammond, L., Chung Wei, R., Andree, A., Richardson, N. and Orphanos, S. (2009). *Professional learning in the learning profession: A status report on teacher development in the United States and abroad*. Dallas, TX: National Staff Development Council.

David, J. L. (2008). 'Project-Based Learning'. *Educational Leadership,* 65 (5)*,* 80-82.

Day, C. (1999). *Developing teachers. The challenges of lifelong learning*. London: Routledge Falmer.

De La Paz, S. and Hernández-Ramos, P. (2013). 'Technology-Enhanced Project-Based Learning: Effects on Historical Thinking'. *Journal of Special Education Technology,* 28 (4).

de Leeuw, E. (2001). 'Reducing Missing Data in Surveys: An Overview of Methods'. *Quality and Quantity,* 35 (2)*,* 147-160.

de Vries, S., Jansen, E. P. W. A. and van de Grift, W. J. C. M. (2013). 'Profiling teachers' continuing professional development and the relation with their beliefs about learning and teaching'. *Teaching and Teacher Education,* 33 (0), 78-89.

de Vries, S., van de Grift, W. J. and Jansen, E. P. (2013). 'Teachers' Beliefs and Continuing Professional Development'. *Journal of Educational Administration,* 51 (2), 213.

DeMonte, J. (2013). *High-Quality Professional Development for Teachers. Supporting Teacher Training to Improve Student Learning*. Washington D. C.: Center for American Progress.

Department of Education of Northern Ireland. (2010). Teacher Education Partnership Handbook.

Department of Education. United States of America. (2010). A Blueprint for Reform. The Reauthorization of the Elementary and Secondary Education Act.

Department of Education. United States of America. (2011). Great teachers and great leaders.

Desa, D. (2014). *Evaluating Measurement Invariance of TALIS 2013 Complex Scales*: OECD Publishing. Available [Online] at: http://dx.doi.org/10.1787/5jz2kbbvlb7k-en.

Desimone, L. M. (2009). 'Improving Impact Studies of Teachers' Professional Development: Toward Better Conceptualizations and Measures'. *Educational Researcher,* 38 (3), 181-199.

Desimone, L. M., Porter, A. C., Garet, M. S., Yoon, K. S. and Birman, B. F. (2002). 'Effects of Professional Development on Teachers' Instruction: Results from a Three-year Longitudinal Study'. *Educational Evaluation and Policy Analysis,* 24 (2), 81-112.

Dewey, J. (1933). *How we think: a restatement of the relation of reflective thinking to the educative process*. New York: D.C.Heath.

DfE. (2013). *Teacher supply model: a technical description. DFE-00278-2013*. London: Department for Education, England and Wales.

Eekhout, I. (2014). *Don't Miss Out!: Incomplete data can contain valuable information*. Amsterdam: EMGO+ Institute for Health and Care Research, Department of Epidemiology and Biostatistics, VU University Medical Center.

Elmore, R. F. and Burney, D. (1997). *Investing in teacher learning: Staff development and instructional improvement in Community School District #2, New York City*. New York: National Commission on Teaching and America's Future.

Embretson, S. E. and Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Psychology Press.

ETUCE (2008). *Teacher Education in Europe. An ETUCE Policy Paper*. Brussels: European Trade Union Committee for Education.

European Commission (2005a). *Common European Principles for Teacher Competences and Qualifications*. Burssels: European Commission.

European Commission (2005b). *CPD for teachers and trainers. Report of a Peer Learning Activity held in Dublin, 26 - 29 September 2005*. Burssels: Education and Training 2010 programme. Cluster 'Teachers and Trainers'.

European Commission (2009). *Progress towards the Lisbon objectives in education and training. Indicators and becnhmarks 2009*. Burssels: European Commission.

European Commission (2012a). *Supporting teacher competence development for better learning outcomes*. Brussels: European Commission.

European Commission (2012b). *Supporting the Teaching Professions for Better Learning Outcomes. Communication from the Commission Rethinking Education:*

Investing in skills for better socio-economic outcomes. Strasbourg: European Commission.

European Commission/EACEA/Eurydice (2013). *Key Data on Teachers and School Leaders in Europe. 2013 Edition. Eurydice Report*. Luxembourg: Publications Office of the European Union.

European Parliament (2000). *Lisbon European Council. 23 and 24 of March, 2000. Presidency Conclusions*. Lisbon: European Parliament.

Evans, M., J., K., J., S. and Treiman, D. (2010). 'Family scholarly culture and educational success: books and schooling in 27 nations'. *Research in Social Stratification and Mobility,* 28 (2)*,* 171-197.

Feiman-Nemser, S. (2001). 'From preparation to practice: Designing a continuum to strengthen and sustain teaching'. *The Teachers College Record,* 103 (6)*,* 1013-1055.

Fennema, E. and Franke, M. L. (1992). 'Teachers' knowledge and its impact'. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics* (pp. 147-164). New York, NY, England: Macmillan Publishing Co, Inc.

Firestone, W. A., Mangin, M. M., Martinez, M. C. and Polovsky, T. (2005). 'Leading Coherent Professional Development: A Comparison of Three Districts'. *Educational Administration Quarterly,* 41 (3)*,* 413-448.

Franke, M. L., Carpenter, T. P., Levi, L. and Fennema, E. (2001). 'Capturing Teachers' Generative Change: A Follow-Up Study of Professional Development in Mathematics'. *American Educational Research Journal,* 38 (3)*,* 653-689.

Fuhrman, S. H. (1993). *Designing Coherent Education Policy: Improving the System*. New Brunswick, NJ: Consortium for Policy Research in Education.

Ganser, T. (2000). 'An Ambitious Vision of Professional Development for Teachers'. *NASSP Bulletin,* 84 (618)*,* 6-12.

Gardner, R. (August, 12th, 2013). 'Exclusive: UK faces desperate shortage of science and maths teachers'. *The Independent*.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F. and Yoon, K. S. (2001). 'What Makes Professional Development Effective? Results From a National Sample of Teachers'. *American Educational Research Journal,* 38 (4)*,* 915-945.

Geijsel, F. P., Sleegers, P. J. C., Stoel, R. D. and Kruger, M. L. (2009). 'The Effect of Teacher Psychological and School Organizational and Leadership Factors on Teachers' Professional Learning in Dutch Schools'. *The Elementary school journal,* 109 (4)*,* 406.

Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B. and Vermeersch, C. M. (2011). *Impact evaluation in practice*: World Bank Publications.

Gilleece, L. (2015). 'Parental involvement and pupil reading achievement in Ireland: Findings from PIRLS2011'. *International Journal of Educational Research,* 73*,* 23-36.

Glass, G. V. (1976). 'Primary, Secondary, and Meta-Analysis of Research'. *Educational Researcher,* 5 (10)*,* 3-8.

Goldstein, H. (2008). 'Comment peut-on utiliser les etudes comparatives internationale pour doter les politiques educatives d'informations fiables? '. *Revue Francaise de Pedagogie,* 164 (Juillet-Septembre)*,* 69-76.

Greenleaf, C. L., Litman, C., Hanson, T. L., Rosen, R., Boscardin, C. K., Herman, J., Schneider, S. A., Madden, S. and Jones, B. (2011). 'Integrating Literacy and Science in Biology'. *American Educational Research Journal,* 48 (3)*,* 647-717.

Gregorian, V. (July, 6th, 2001). 'How to train - and retain - teachers'. *New York Times*.

Grilli, L., Pennoni, F., Rampichini, C. and Romeo, I. (2014). *Exploiting TIMSS and PIRLS combined data: multivariate multilevel modelling of student achievement*. Paper presented at the VI European Congress of Methodology. Utrecht, Netherlands.

Gumus, S., Bulut, O. and Bellibas, M. S. (2013). 'The Relationship between Principal Leadership and Teacher Collaboration in Turkish Primary Schools: A Multilevel Analysis'. *Education Research and Perspectives,* 40 (1), 1-29.

Guskey, T. R. (1986). 'Staff Development and the Process of Teacher Change'. *Educational Researcher,* 15 (5), 5-12.

Guskey, T. R. (1994). *Professional Development in Education: In Search of the Optimal Mix*. Paper presented at the Annual Meeting of the American Educational Educational Research Association. New Orleans, LA.

Guskey, T. R. (2002). 'Professional Development and Teacher Change'. *Teachers and Teaching,* 8 (3), 381-391.

Hanushek, E. A., Piopiunik, M. and Wiederhold, S. (2014). 'The Value of Smarter Teachers: International Evidence on Teacher Cognitive Skills and Student Performance'. *National Bureau of Economic Research Working Paper Series,* No. 20727.

Hardy, I. (2010). 'Critiquing Teacher Professional Development: Teacher Learning within the Field of Teachers' Work'. *Critical Studies in Education,* 51 (1), 71-84.

Hardy, I. (2012). *The Politics of Teacher Professional Development*. London: Routledge. Available [Online] at: http://dx.doi.org/10.4324/9780203110386.

Hardy, I. and Rönnerman, K. (2011). 'The value and valuing of continuing professional development: current dilemmas, future directions and the case for action research'. *Cambridge Journal of Education,* 41 (4), 461-472.

Hardy, I., Rönnerman, K., Moksnes Furu, E., Salo, P. and Forsman, L. (2010). 'Professional development policy and politics across international contexts: from mutuality to measurability?'. *Pedagogy, Culture & Society,* 18 (1)*,* 81-92.

Hattie, J. (2008). *Visible learning: a synthesis of over 800 meta-analyses relating to achievement.* Oxon: Routledge Ltd.

Hawthorne, G. and Elliott, P. (2005). 'Imputing Cross-Sectional Missing Data: Comparison of Common Techniques'. *Australian and New Zealand Journal of Psychiatry,* 39 (7)*,* 583-590.

Heller, J. I., Daehler, K. R., Wong, N., Shinohara, M. and Miratrix, L. W. (2012). 'Differential effects of three professional development models on teacher knowledge and student achievement in elementary science'. *Journal of Research in Science Teaching,* 49 (3)*,* 333-362.

Hendriks, M., Luyten, H., Scheerens, J., Sleegers, P. and Steen, R. (2010a). Enhancing educational effectiveness through teachers' professional development. In J. Scheerens (ed), *Teachers' professional development: Europe in international comparison. An analysis of teachers' professional development based on the OECD's Teaching and Learning International Survey (TALIS)*. Belgium: European Union.

Hendriks, M., Luyten, H., Scheerens, J., Sleegers, P. and Steen, R. (2010b). Teachers' professional development – a snapshot from talis of lower secondary education. In J. Scheerens (ed), *Teachers' professional development: Europe in international comparison. An analysis of teachers' professional development based on the OECD's Teaching and Learning International Survey (TALIS)*. Belgium: European Union.

Hoban, G. (2002). *Teacher learning for educational change. A systems thinking approach.* Buckingham: Open University Press.

Hochberg, E. D. and Desimone, L. M. (2010). 'Professional Development in the Accountability Context: Building Capacity to Achieve Standards'. *Educational Psychologist,* 45 (2), 89-106.

Howell, D., C. (2007). *Statistical methods for psychology*. Belmont, CA: Thomson Wadworth.

Howson, J. and Waterman, C. (2013). *The future of teacher education in England: developing a strategy*. Amersham: The IRIS Press.

Hutchings, M. (2011). *What impact does the wider economic situation have on teachers' career decisions? A literature review*. London: Institute for Policy Studies in Education, London Metropolitan University.

Hutchison, D. (2008). 'On the conceptualisation of measurement error'. *Oxford Review of Education,* 34 (4), 443-460.

IEA. (2012). 'TIMSS 2011 Encyclopedia. Education Policy and Curriculum in Mathematics and Science. Volumen 2: L–Z and Benchmarking Participants'. In I. V. S. Mullis, M. O. Martin, C. A. Minnich, G. M. Stanco, A. Arora, V. A. S. Centurino and C. E. Castle (Eds). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College, and International Association for the Evaluation of Educational Achievement.

Ingersoll, R. (2007). 'A comparative study of teacher preparation and qualifications in six nations'. *GSE Publications,* 145.

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

International Project Consortium. (2013). Principal and Teacher Questionnaire. OECD Teaching and Learning International Survey (TALIS). Main Study Version. English, UK Spelling. In International Association for the Evaluation of Educational Achievement (IEA) The Netherlands, IEA Data Processing and

Research Center (IEA DPC) Germany and Statistics Canada (eds): Organisation for Economic Co-operation and Development.

James, M. and McCormick, R. (2009). 'Teachers learning how to learn'. *Teaching and Teacher Education,* 25 (7)*,* 973-982.

Japanese Ministry of Education, C., Sports, Science and Techonolgy (MEXT) (Elementary and Secondary Education Bureau), (2003). *Attracting, developing and retaining effective teachers. OECD activity (Analytical Review). Japanese country background report*: Organization of Economic Cooperation and Development (OECD).

Jarvis, P. (2007). 'Globalisation, lifelong learning and the learning society. Sociological perspectives.', *Lifelong learning and the learning society* (Vol. 2). Oxon: Routledge.

Jerrim, J. (2011). *England's "plummeting" PISA test scores between 2000 and 2009: Is the performance of our secondary school pupils really in relative decline?* DoQSS Working Paper No. 11-09: Department of Quantitative Social Science, Institute of Education, University of London.

Johnson, D. W. and Johnson, R. T. (1974). 'Instructional Goal Structure: Cooperative, Competitive, or Individualistic'. *Review of Educational Research,* 44 (2)*,* 213-240.

Johnson, D. W. and Johnson, R. T. (1989). *Cooperation and competition: Theory and research*. Edina, MN, US: Interaction Book Company.

Johnson, D. W. and Johnson, R. T. (2009). 'An Educational Psychology Success Story: Social Interdependence Theory and Cooperative Learning'. *Educational Researcher,* 38 (5)*,* 365-379.

Jöreskog, K. G. (1969). 'A general approach to confirmatory maximum likelihood factor analysis'. *Psychometrika,* 34 (2)*,* 183-202.

Kaplan, D. (2016). 'Causal inference with large-scale assessments in education from a Bayesian perspective: a review and synthesis'. *Large-scale Assessments in Education,* 4 (1)*,* 7.

Keigher, A. and Cross, F. (2010). *Teacher Attrition and Mobility: Results From the 2008–09 Teacher Follow-up Survey*. Washington, DC: U.S. Department of Education.

Kennedy, M. (1998). Form and Substance in Inservice Teacher Education, *Research Monograph No. 13*. Madison, WI: National Institute for Science Education. University of Wisconsin-Madison.

Kilpatrick, W. (1918). 'The project method'. *The Teachers College Record,* 19 (4)*,* 319-335.

Kim, J.-S., Anderson, C. J. and Keller, B. (2013). 'Multilevel Analysis of Assessment Data'. In L. Rutkowski, M. Von Davier and D. Rutkowski (Eds), *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis* (Vol. 18, pp. 389). London: Chapman & Hall/CRC.

King, F. (2011). 'The role of leadership in developing and sustaining teachers' professional learning'. *Management in Education,* 25 (4)*,* 149-155.

Kline, P. (1994). *A easy guide to factor analysis*. London: Routledge.

Kruse, S., Louis, K. S. and Bryk, A. (1994). 'Building professional community in schools'. *Issues in restructuring schools,* 6 (3)*,* 67-71.

Kwakman, K. (2003). 'Factors affecting teachers' participation in professional learning activities'. *Teaching and Teacher Education,* 19 (2)*,* 149-170.

Lewis, C. (2009). 'What is the nature of knowledge development in lesson study?'. *Educational Action Research,* 17 (1)*,* 95-110.

Lieberman, A. and Miller, L. (1991). *Staff development for education in the '90s*. New York: Teachers College Press, Columbia University.

Little, J. W. (1993). 'Teachers' Professional Development in a Climate of Educational Reform'. *Educational Evaluation and Policy Analysis,* 15 (2)*,* 129-151.

Long, J. S. and Freese, J. (2006). *Regression models for categorical dependent variables using Stata*. Texas: Stata Corporation. College Station.

Loucks-Horsley, S. and Matsumoto, C. (1999). 'Research on Professional Development for Teachers of Mathematics and Science: The State of the Scene'. *School Science and Mathematics,* 99 (5)*,* 258-271.

Macdonald, K. (2014). 'PV: Stata module to perform estimation with plausible values'. [Online]. Available at: https://ideas.repec.org/c/boc/bocode/s456951.html.

Marzano, R. J., Pickering, D. and Pollock, J. E. (2001). *Classroom instruction that works: Research-based strategies for increasing student achievement*. Alexandria, VA: Ascd.

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

McRae, D., Ainsworth, G., Groves, R., Rowland, M. and Zbar, V. (2001). *PD 2000 Australia: A national mapping of school teacher professional development. A report for the Commonwealth Department of Education, Training and Youth Affairs (DETYA)*. Canberra: Commonwealth of Australia.

Meirink, J. A., Imants, J., Meijer, P. C. and Verloop, N. (2010). 'Teacher learning and collaboration in innovative teams'. *Cambridge Journal of Education,* 40 (2)*,* 161-181.

Meirink, J. A., Meijer, P. C., Verloop, N. and Bergen, T. C. M. (2009a). 'How do teachers learn in the workplace? An examination of teacher learning activities'. *European Journal of Teacher Education,* 32 (3)*,* 209-224.

Meirink, J. A., Meijer, P. C., Verloop, N. and Bergen, T. C. M. (2009b). 'Understanding teacher learning in secondary education: The relations of teacher activities to changed beliefs about teaching and learning'. *Teaching and Teacher Education,* 25 (1)*,* 89-100.

Micklewright, J., Jerrim, J., Vignoles, A., Jenkins, A., Allen, R., Ilie, S., Bellarbre, E., Barrera, F. and Hein, C. (2014). *Teachers in England's secondary schools: evidence from TALIS 2013. Research report*. London: Department for Education, Institute of Education, University of London.

Mirazchiyski, P. (2013). *Providing school-level reports from international large-scale assessments: methodological considerations, limitations, and possible solutions*. Hamburg: International association for the evaluation of educational achievement (IEA).

Mullis, I. V. S., Martin, M. O., Foy, P. and Arora, A. (2012a). 'International Achievement in Mathematics', *TIMSS 2011 International Results in Mathematics* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Foy, P. and Arora, A. (2012b). 'Performance at the TIMSS 2011 International Benchmarks', *TIMSS 2011 International Results in Mathematics* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Foy, P. and Arora, A. (2012c). 'Population Coverage and Sample Participation Rates', *TIMSS 2011 International Results in Mathematics* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Foy, P. and Arora, A. (2012d). 'Teacher preparation', *TIMSS 2011 International Results in Mathematics* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Murray, J. M. (2012). 'Development and Psychometric Evaluation of the Independent School Teacher Development Inventory'. *The Journal of Experimental Education,* 80 (3)*,* 219-245.

Musset, P. (2010). *Initial Teacher Education and Continuing Training Policies in a Comparative Perspective: Current Practices in OECD Countries and a Literature Review on Potential Effects* (OECD Education Working Papers No. 48). Paris: Organisation for Economic Co-operation and Development.

Muthén, L. K. and Muthén, B. O. (1998-2011). Mplus User's Guide. Los Angeles, CA: Muthén & Muthén.

NCRTL (1993). *Findings on Learning to Teach*. East Lansing, MI: National Center for Research on Teacher Learning, Michigan State University.

Newmann, F. M. (1994). 'School-Wide Professional Community'. *Issues in restructuring schools,* (6)*,* 2-3.

Newmann, F. M., King, M. B. and Youngs, P. (2000). *Professional development that addresses school capacity: Lessons from urban elementary schools*. Paper presented at the Annual Meeting of the American Educational Research Association. New Orleans.

Newmann, F. M., Smith, B., Allensworth, E. and Bryk, A. S. (2001a). *Improving Chicago's Schools. School Instructional Program Coherence: Benefits and Challenges*. Chicago: Consortium on Chicago School Research.

Newmann, F. M., Smith, B., Allensworth, E. and Bryk, A. S. (2001b). 'Instructional Program Coherence: What It Is and Why It Should Guide School Improvement Policy'. *Educational Evaluation and Policy Analysis,* 23 (4)*,* 297-321.

Ng, C. H. (2010). 'Do career goals promote continuous learning among practicing teachers?'. *Teachers and Teaching,* 16 (4)*,* 397-422.

Nordic Council of Ministers. (2009). *Comparative study of Nordic teacher-training programmes*. Copenhaguen: Norden.

O'Connell, A. A. and McCoach, D. B. (2008). *Multilevel Modeling of Educational Data*. Charlotte, NC: INFORMATION AGE PUBLISHING, INC.

O'Day, J. A. and Smith, M. S. (1993). 'Systemic Reform and Educational Opportunity'. In S. H. Fuhrman (Ed.), *Designing Coherent Education Policy: Improving the System* (pp. 250-312). San Francisco: Jossey-Bass Publishers.

OECD. (1998). *Education Policy Analysis 1998*. Paris: OECD Publishing. Available [Online] at: http://dx.doi.org/10.1787/epa-1998-en.

OECD. (2005). *Teachers Matter. Attracting, Developing and Retaining Effective Teachers*. Paris: OECD Publishing. Available [Online] at: http://dx.doi.org/10.1787/9789264018044-en.

OECD (2009a). *Creating Effective Teaching and Learning Environments. First result from TALIS*. Paris: OECD Publishing.

OECD. (2009b). *PISA Data Analysis Manual: SAS*. Paris: OECD Publishing.

OECD (2010). *TALIS 2008 Technical Report*. Paris: Organisation for Economic Co-operation and Development.

OECD. (2012a). *Indicator D5 Who are the teachers?*: OECD Publishing. Available [Online] at: http://dx.doi.org/10.1787/eag-2012-33-en.

OECD. (2012b). *PISA 2009 Technical Report*: OECD Publishing. Available [Online] at: http://dx.doi.org/10.1787/9789264167872-en.

OECD. (2013a). *Fostering Learning Communities Among Teachers*: OECD Publishing. Available [Online] at: http://dx.doi.org/10.1787/5k4220vpxbmn-en.

OECD. (2013b). *Indicator D3 How much are teachers paid ?*: OECD Publishing. Available [Online] at: http://dx.doi.org/10.1787/eag-2013-27-en.

OECD. (2014a). *New Insights from TALIS 2013: Teaching and Learning in Primary and Upper Secondary Education*. Paris: OECD Publishing.

OECD. (2014b). *PISA 2012 Results: What Students Know and Can Do - Student Performance in Mathematics, Reading and Science*. (Vol. Volume I, Revised edition, February 2014). Paris: PISA, OECD Publishing.

OECD. (2014c). *PISA 2012 Technical Report*. Paris: OECD Publishing.

OECD. (2014d). *TALIS 2013 Results: An International Perspective on Teaching and Learning*. Paris: OECD Publishing.

OECD. (2014e). *TALIS 2013 Technical Report*. Paris: OECD Publishing.

OECD. (2015). *Embedding Professional Development in Schools for Teacher Success*. Paris: OECD Publishing. Available [Online] at: http://dx.doi.org/10.1787/5js4rv7s7snt-en.

Ofsted. (2006). The logical chain: continuing professional development in effective schools. London: Office for Standards in Education, Children's Services and Skills.

Opfer, V. D. (2015). *Understanding types of professional development reported by teachers in TALIS 2013*. Paper presented at the TALIS Conference. Learning from each other. Ministry of Education, Culture and Science of the Netherlands. Amsterdam.

Opfer, V. D. and Pedder, D. (2011a). 'Conceptualizing Teacher Professional Learning'. *Review of Educational Research,* 81 (3), 376-407.

Opfer, V. D. and Pedder, D. (2011b). 'The lost promise of teacher professional development in England'. *European Journal of Teacher Education,* 34 (1), 3-24.

Opfer, V. D., Pedder, D. G. and Lavicza, Z. (2011a). 'The role of teachers' orientation to learning in professional development and change: A national study of teachers in England'. *Teaching and Teacher Education,* 27 (2), 443-453.

Opfer, V. D., Pedder, D. J. and Lavicza, Z. (2011b). 'The influence of school orientation to learning on teachers' professional learning change'. *School Effectiveness and School Improvement,* 22 (2), 193-214.

Passy, R. and Golden, S. (2010). *Teacher resignation and recruitment survey*. Slough: National Foundation for Educational Research.

Pedder, D. and Opfer, V. D. (2011). 'Are we realising the full potential of teachers' professional learning in schools in England? Policy issues and recommendations from a national study'. *Professional Development in Education,* 37 (5), 741-758.

Penuel, W. R., Fishman, B. J., Yamaguchi, R. and Gallagher, L. P. (2007). 'What Makes Professional Development Effective? Strategies That Foster Curriculum Implementation'. *American Educational Research Journal,* 44 (4), 921-958.

Penuel, W. R., Gallagher, L. P. and Moorthy, S. (2011). 'Preparing Teachers to Design Sequences of Instruction in Earth Systems Science'. *American Educational Research Journal,* 48 (4), 996-1025.

Persson, M. (2005). 'Continuing professional development and networking in Europe'. In A. Alexandrou (Ed.), *The continuing professional development of educators: emerging European issues*. Oxford: Symposium Books.

PISA Consortium. (2010). Translation and adaptation guidelines for PISA 2012. Paris: Organisation for Economic Co-operation and Development.

PISA Consortium. (2011). School questionnaire for PISA 2012. Main survey. Paris: Organisation for Economic Co-operation and Development.

Pokropek, A. (2016). 'Introduction to instrumental variables and their application to large-scale assessment data'. *Large-scale Assessments in Education,* 4 (1), 4.

Putnam, R. T. and Borko, H. (2000). 'What Do New Views of Knowledge and Thinking Have to Say about Research on Teacher Learning?'. *Educational Researcher,* 29 (1), 4-15.

Rabe-Hesketh, S. and Skrondal, A. (2006). 'Multilevel modelling of complex survey data'. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 169 (4), 805-827.

Rajala, R., Flores, M. A., Tornberg, A. and Veiga, S. A. M. (2008, 2008). *The role of school leadership in learning at work and professional development in three European countries*. Paper presented at the Annual meeting of the Australian Association for Research in Education. Fremantle, Australia.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical linear models. Applications and data analysis methods*. (Second. ed.). London: Sage Publications, Inc.

Reale, M. (2006). Heteroscedasticity, *Econometrics*. Christchurch, New Zealand Department of Mathematics and Statistics. College of Engineering. University of Canterbury.

Remillard, J. T. and Bryans, M. B. (2004). 'Teachers' Orientations toward Mathematics Curriculum Materials: Implications for Teacher Learning'. *Journal for Research in Mathematics Education,* 35 (5), 352-388.

Roberts, N. (2013). *Initial training for school teachers in England*. London: House of Commons Library.

Robinson, L. (2014). *UK "lacks coherent plan for teacher research and development", finds report*. [Online]. Available at: http://www.thersa.org/about-us/media/press-releases/uk-lacks-coherent-plan-for-teacher-research-and-development,-finds-report. [Last accessed September 11th, 2014].

Rubin, D. B. (1996). 'Multiple Imputation after 18+ Years'. *Journal of the American Statistical Association,* 91 (434), 473-489.

Runhaar, P., Sanders, K. and Yang, H. (2010). 'Stimulating teachers' reflection and feedback asking: An interplay of self-efficacy, learning goal orientation, and transformational leadership'. *Teaching and Teacher Education,* 26 (5), 1154-1161.

Rust, K. (2013). 'Sampling, Weighting, and Variance Estimation in International Large-Scale Assessments'. In L. Rutkowski, M. Von Davier and D. Rutkowski (Eds), *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis* (Vol. 6, pp. 117). London: Chapman & Hall/CRC.

Rutkowski, D., Rutkowski, L., Bélanger, J., Knoll, S., Weatherby, K. and Prusinski, E. (2013). *Teaching and Learning International Survey TALIS 2013. Conceptual framework*. Paris: Organisation for Economic Co-operation and Development.

Rutkowski, L. (2016). 'Introduction to special issue on quasi-causal methods'. *Large-scale Assessments in Education,* 4 (1)*,* 8.

Rutkowski, L., Gonzalez, E., Joncas, M. and von Davier, M. (2010). 'International Large-Scale Assessment Data Issues in Secondary Analysis and Reporting'. *Educational Researcher,* 39 (2)*,* 142.

Sahlberg, P. (2011). 'The professional educator. Lesson from Finland'. *American Educator,* (Summer).

Salinas, A. (2010). *Investing in Our Teachers: What Focus of Professional Development Leads to the Highest Student Gains in Mathematics Achievement?* University of Miami.

Schafer, J. L. and Olsen, M. K. (1998). 'Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective'. *Multivariate Behavioral Research,* 33 (4)*,* 545-571.

Scher, L. and O'Reilly, F. (2009). 'Professional Development for K–12 Math and Science Teachers: What Do We Really Know?'. *Journal of Research on Educational Effectiveness,* 2 (3)*,* 209-249.

Schwile, J., Dembele, M. and Schubert, J. (2007). *Global perspectives on teacher learning: improving policy and practice*. Paris: UNESCO, International Institute for Educational Planning.

Sebastian, J. and Allensworth, E. (2012). 'The Influence of Principal Leadership on Classroom Instruction and Student Learning: A Study of Mediated Pathways to Learning'. *Educational Administration Quarterly,* 48 (4), 626-663.

Shaffer, J. P. (1995). 'Multiple hypothesis testing'. *Annual Review of Psychology,* 46 (Health & Medical Complete), 561 - 584.

Shulman, L. S. (1986). 'Those Who Understand: Knowledge Growth in Teaching'. *Educational Researcher,* 15 (2), 4-14.

Snijders, T. and Bosker, R. (1999). *Multilevel analysis: an introduction to basic and advanced multilevel modeling*. London: SAGE.

Snijders, T. A. B. (2005). 'Fixed and Random Effects'. In B. S. Everitt and D. C. Howell (Eds), *Encyclopedia of Statistics in Behavioral Science* (Vol. 2, pp. 664-665). Chicester: Wiley.

Sowder, T. S. (2007). 'The mathematical education and development of teachers'. In F. K. Lester Jr. (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 157-223). Charlotte, NC: National Council of Teachers of Mathematics.

Stapleton, L. M. (2013). 'Incorporating Sampling Weights into Single-and Multilevel Analyses'. In L. Rutkowski, M. Von Davier and D. rutkowski (Eds), *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis* (Vol. 17, pp. 363). London: Chapman & Hall/CRC.

StataCorp. (2011a). *Stata: Release 12. Statistical Software*. College Station, TX: StataCorp LP.

StataCorp. (2011b). 'xtmixed — Multilevel mixed-effects linear regression. Survey data.', *Stata. Longitudinal-data/panel-data reference manual. Release 12* (pp. 342-354). College Station, TX: Stata Press.

Stewart, V. (2011). *Improving teacher quality around the world: the International Summit on the Teaching Profession*: Asia Society Partnership for Global Learning.

Stewart, V. (2012). *Teaching and leadership for the twenty-first century. The 2012 International Summit on the Teaching Profession*: Asia Society Partnership for Global Learning.

Stewart, V. (2013). *Teacher quality: the 2013 International Summit on the Teaching Profession*: Asia Society Partnership for Global Learning.

Stewart, V. (2014). *Excellence, equity, and inclusiveness. High quality teaching for all: the 2015 International Summit on the Teaching Profession*: Asia Society Partnership for Global Learning.

Stewart, V. (2015). *Implementing highly effective teacher policy and practice: the 2015 International Summit on the Teaching Profession*: Asia Society Partnership for Global Learning.

Suell, J. L. and Piotrowski, C. (2007). 'Alternative teacher education programs: A review of the literature and outcome studies'. *Journal of Instructional Psychology,* 34 (1)*,* 54-58.

Supovitz, J., Sirinides, P. and May, H. (2010). 'How Principals and Peers Influence Teaching and Learning'. *Educational Administration Quarterly,* 46 (1)*,* 31-56.

Supovitz, J. A. (2001). 'Translating teaching practice into improved student achievement'. In S. Fuhrman (Ed.), *From the capitol to the classroom: standards-based reform in the states. One hundreth yearbook of the National Society for the Study of Education* (Vol. 2). Chicago, Illinois: The University of Chicago Press.

Tabachnick, B. G. and Fidell, L. S. (2001). 'Using multivariate statistics'. *Boston Massachusetts Allyn and Bacon*.

Telese, J. A. (2012). 'Middle School Mathematics Teachers' Professional Development and Student Achievement'. *The Journal of Educational Research,* 105 (2)*,* 102-111.

Timperley, H., Wilson, A., Barrar, H. and Fung, I. (2007). *Teacher professional learning and development: best evidence synthesis iteration*. Wellington: Ministry of Education of New Zealand.

TIMSS & PIRLS International Study Center, B. C. (2011). *TIMSS 2011 Teacher Questionnaire Mathematics - Grade 8*. Chestnut Hill, MA: International Association for the Evaluation of Educational Achievement.

UNICEF. (2010). 'Getting ready for school: a child to child approach programme evaluation for year one.[Online]: Retrieved on 2 November 2016 at'. *URL: http://www. unicef. org/education/files/UNICEF_CtC_Year_One_Impact_Evaluation. pdf*.

Van Veen, K., Zwart, R. and Meirink, J. (2012). 'What makes teacher professional development effective? A literature review'. In M. Kooy and K. Van Veen (Eds), *Teacher learning that matters. International perspectives* (pp. 3-21). London: Taylor & Francis.

Vandenberg, R. J. and Lance, C. E. (2000). 'A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research'. *Organizational Research Methods,* 3 (1), 4-70.

Vermunt, J. D. and Endedijk, M. D. (2011). 'Patterns in teacher learning in different phases of the professional career'. *Learning and Individual Differences,* 21 (3), 294-302.

Vieluf, S., Kaplan, D., Klieme, E. and Bayer, S. (2012). *Teaching Practices and Pedagogical Innovation. Evidence from TALIS*. Paris: OECD Publishing, Centre for Educational Research and Innovation.

Villegas-Reimers, E. (2003). *Teacher professional development: an international review of the literature*. Paris: UNESCO, International Institute for Educational Planning.

Voogt, J., Westbroek, H., Handelzalts, A., Walraven, A., McKenney, S., Pieters, J. and de Vries, B. (2011). 'Teacher learning in collaborative curriculum design'. *Teaching and Teacher Education,* 27 (8)*,* 1235-1244.

Wade, R. K. (1985). 'What Makes a Difference in Inservice Teacher Education? A Meta-Analysis of Research'. *Educational Leadership,* 42 (4)*,* 48-54.

Walker, A., Recker, M., Ye, L., Robertshaw, M., Sellers, L. and Leary, H. (2012). 'Comparing technology-related teacher professional development designs: a multilevel study of teacher and student impacts'. *Educational Technology Research and Development,* 60 (3)*,* 421-444.

Walker, M., Jeffes, J., Hart, R., Lord, P. and Kinder, K. (2011). Making the links between teachers' professional standards, induction, performance management and continuing professional development RR075. London: Department for Education.

Wayne, A. J., Yoon, K. S., Zhu, P., Cronen, S. and Garet, M. S. (2008). 'Experimenting With Teacher Professional Development: Motives and Methods'. *Educational Researcher,* 37 (8)*,* 469-479.

Williams, M. (October, 8th, 2013). 'Professional development: what can Brits learn from schools abroad?'. *The Guardian*.

Wilson, S. M. and Berne, J. (1999). 'Teacher Learning and the Acquisition of Professional Knowledge: An Examination of Research on Contemporary Professional Development'. *Review of Research in Education,* 24*,* 173-209.

Winship, C. and Mare, R. D. (1984). 'Regression models with ordinal variables'. *American Sociological Review,* 512-525.

Winship, C. and Morgan, S. L. (1999). 'The Estimation of Causal Effects from Observational Data'. *Annual Review of Sociology,* 25*,* 659-706.

Wooldrigde, J. M. (2003). *Introductory econometrics: a modern approach, 2e*. Ohio: Thomson South Western.

Wu, A. D., Li, Z. and Zumbo, B. D. (2007). 'Decoding the Meaning of Factorial Invariance and Updating the Practice of Multi-group Confirmatory Factor Analysis: A Demonstration With TIMSS Data'. *Practical Assessment, Research & Evaluation,* 12 (3)*,* 1-26.

Yoon, K. S., Duncan, T., Lee, S. W.-Y., Scarloss, B. and Shapley, K. (2007). *Reviewing the evidence on how teacher professional development affects student achievement*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest.

Youngs, P. and King, M. B. (2002). 'Principal Leadership for Professional Development to Build School Capacity'. *Educational Administration Quarterly,* 38 (5)*,* 643-670.

Zhang, X. (2011). 'On Interpreters' Intercultural Awareness'. *World Journal of English Language,* 1 (1)*,* p47.