Identification and estimation of the effects of school quality on achievement of Italian students

Abstract

The aim of this dissertation is to assess the causal effect of school quality on students' acquisition of skills during the academic year. I exploit information at individual and school level obtained by integrating the data from the second wave of the OECD *Programme for International Student Assessment* (PISA 2003) with the administrative data on schools from the archive *Sistema Integrato Segreterie Scolastiche Italiane* (SISSI). The rich dataset allows to overcome the issue of non-random selection of students into schools of different quality. A general statistical procedure is used to allow for latent 'outcome' and 'treatment' variables—variation of skills and quality, respectively. The results confirm the main findings of literature, i.e. school quality, as measured by the usual observable characteristics of schools—like class size, does not matter once we control for individual endowment, family background and peer characteristics. Nevertheless, I provide evidence that schools have multiple dimensions of quality and there is some indication of an impact arising from teachers, whose effectiveness in promoting education seem to be explained by their type of job contract in terms of tenure.

Acknowledgments

I would like to thank my supervisor Erich Battistin for his comments and insights, as well as his patience and moral support. I also wish to acknowledge Enrico Rettore for many other useful suggestions and Daniele Checchi for giving me access to the SISSI data. I am grateful to the members of the research group "Evaluating the impact of public interventions: methods and case studies" of Padova and Giorgio Brunello for constructive discussions. Finally, this dissertation benefits of helpful remarks and comments of participants in seminars at the Department of Statistics of Padova and in the 6th workshop "Brucchi Luchino" in Salerno. I could never have accomplished my thesis without the encouragement of my beloved boyfriend and family, in particular my mother. Of course, I sincerely thank my fellow Ph.D. students, especially Federico and Sofia. I dedicate this work to my dear six nephews Aurora, Margherita, Davide, Andrea, Celeste and Sofia, who have always cheered me up during times of difficulty.

Contents

Chapter 1

Introduction

The economic literature has provided convincing evidence of the positive effects of education, in terms of people's knowledge and competences, on individual earnings and, therefore, on the distribution of income and on the economic growth of a country (see, for example, Hanushek and Woessmann [2007]). There are three groups of educational inputs in the production function for cognitive skills: family background, formal education and peers—i.e. the members of the same group, such as school, neighbourhood, ethnic community and the like (Kramarz, Machin and Ouazad [2006]). It is on these factors that we can play to increase learning and to promote well-being. Henceforth, from the perspective of the policy maker, it is of primary importance to establish the role of schools in determining cognitive skills, because interventions on the educational system are the simplest way to act: has school quality an independent impact on students' performance, after all other educational inputs have been controlled for?

The aim of my dissertation is to answer this question for the Italian educational system. Precisely, the goal is to evaluate the impact of school quality on students' learning during a given school year using data from the second wave of the *Program for International Student Assessment* (PISA 2003). These data provide information on a sample of about 11,000 Italian 15-year-old students from around 500 high secondary schools (30 students per school, on average), tested in their mathematical competences in April 2003.[1] Hence, the variable of interest is students' learning, specifically for mathematics, during the 2002/03 academic year.

Throughout the thesis, by "achievement" I will mean students' skills (competences) in mathematics at some point in time during the academic year. Thus "gain in achievement" will refer to students' variation in the level of skills (what I have called learning, above) over some time period, for instance the whole academic

[1] Competences to be intended as the ability to use scholastic notions in everyday life, instead of pure curricular knowledge (OECD [2004]).

year.

The PISA data, both at the student and at the school level, are obtained from questionnaires completed by students and school heads at the moment of the test. Since the threat of subjective perceptions and voluntary misreporting by school heads, I also use the more reliable information available from the administrative archive *Sistema Integrato Segreterie Scolastiche Italiane* (SISSI) at school level. The resulting database allows me to benefit of a large number of student and school characteristics, necessary to make work my estimation strategy (see Bratti, Checchi and Filippin [2007] for a detailed description of the data).

In order to assess the causal effect of school quality on the variation of student achievement in mathematics during the academic year, I deal with three kinds of problems. First, the assignment process of students into schools of different quality is not at random: for instance, more able students are more likely to attend schools characterized by high quality. Thus the simple comparison of students attending schools with different levels of quality leads to wrong conclusions about the importance of school quality. In such way we estimate a spurious effect of school quality on the gain in student achievement that incorporates some of the effects of other variables (like ability) affecting school choice and learning simultaneously (see, for example, Dearden, Ferri and Meghir [2002]).

Second, school quality corresponds to a concept with a high degree of abstractness and not a clear definition. This concept has not a straightforward concrete counterpart of reference and we are capable to measure it only indirectly through a series of proxies, such as class size, level of expenditure per student, teacher education and experience, and so forth.[2] Therefore, differently from the usual settings in program evaluation, our "treatment" variable is a latent variable, which I will assume continuous and unidimensional in the first instance. To estimate the causal effect of school quality, I will follow the procedure proposed by Black and Smith [2006] for the unidimensional case; then I will extend this strategy to the general case of more than one dimension of school quality. It amounts to the estimation of a factor analysis model by instrumental variables.

The third problem is that the outcome variable, gain in achievement, is a latent object itself. Indeed achievement, like school quality, is a very abstract concept measurable with error by ad hoc proxies, such as test scores. PISA 2003 provides two proxies for student achievement: the score at the PISA test, taken in April 2003, and the teacher's mark of the last school report in January 2003. Clearly this is not the simple case of repeated measures of the same latent factor in two occasions, because the variables considered correspond to different latent dimensions of student achievement. I will develop a method that combines these

[2]See Bollen [1989] for a careful explanation of the process that links a concept to one or more observed variables.

sources of information available from PISA to form a measure of the change in student achievement during the entire academic year, which is a variable of growing interest in the economic literature (see Hanushek, Kain and Rivkin [2005]). The development of a statistical procedure to apply a value-added approach using data from PISA represents a useful contribution of this thesis to the literature.

The results of the analysis seem to confirm the main findings of literature, i.e. school quality, as measured by the usual proxies of quality—like class size, does not matter once we control for individual endowment, family background and peer characteristics (Hanushek [2003]). Anyway, schools have multiple dimensions of quality with potential different weights in students' acquisition of knowledge (Black and Smith [2006]).

If there is an impact of school quality, it arises from teachers[3], who appear to be the most important ingredient of the educational process also from most of the other empirical studies (see the work by Hanushek, Kain and Rivkin [2005] among the most significant contributions). In particular the data exploited in this analysis reveal that the schools with a larger fraction of teachers with a permanent contract are those which succeed, to some extent, in raising students' performance[4], being permanent teachers more motivated and more effective in promoting education.

However, there are some unclear issues to go deep into in future research, for instance to shed further light on the role played by school enrolment. Then it would be interesting to apply my value-added strategy to other educational systems, in addition to the Italian one. Specifically, the international survey of PISA allows to move to a cross-country analysis, where other school quality inputs such as school organization and institutions—recognized to play a key role in the production function of education[5]—can be evaluated as well.

The remainder of the thesis is organized as follows. After a survey of the literature on school quality in Chapter 2, Chapter 3 briefly describes the design of the PISA 2003 study and the data. Chapter 4 begins with placing the investigation in the specific econometric framework of program evaluation, giving a solution to the first problem of nonrandom allocation of students across schools of different quality. It then discusses the other two econometric issues related to achievement and school quality nonobservability, solved both separately and jointly in order to ease the explanation. Here I assume school quality unidimensional; the more general case of more than one dimension of school quality is tackled in Chapter 5,

[3]Also class size is related to teachers, in the sense that class size, or equivalently the student-teacher ratio, is a measure (with reverse sign) of the "teaching time per student", which could not exert any influence once we control for teachers' ability (i.e. able teachers provide good instruction both in small and in large classes).

[4]To the best of my knowledge, this teachers' characteristic (from SISSI) has never been considered in the school quality literature, especially in other countries than Italy.

[5]See Hanushek and Woessmann [2007] for a review.

where I derive a method to estimate non-iteratively the resulting factor analysis model, which is a straightforward extension of the estimation strategy proposed in Chapter 4 for the simple case of unidimensional school quality. Results are given in the last sections of Chapter 4 and Chapter 5, for unidimensional and multidimensional school quality, respectively. Finally, in the Appendix, is described an alternative naive method of estimation showing the robustness of my strategy to solve the issue of different measures, of different dimensions, of student achievement. I also show the uselessness to stratify this strategy across population strata, as suggested by some reports on PISA in Italy.

La letteratura economica ha fornito convincente evidenza sugli effetti positivi dell'istruzione, in termini di conoscenze e competenze degli studenti, sui redditi individuali e, quindi, sulla distribuzione della ricchezza e la crescita economica di un paese (Hanushek e Woessmann [2007]). Ci sono tre gruppi di fattori che determinano l'acquisizione delle competenze: il contesto familiare, la scuola e il "gruppo"—ovvero i coetanei a scuola, nel quartiere, nella comunità etnica, etc. (Kramarz, Machin and Ouazad [2006]). Dunque, dal punto di vista politico, è di fondamentale importanza stabilire il ruolo della scuola nel determinare le competenze, perchè gli interventi sul sistema scolastico sono i più semplici da realizzare: la qualità della scuola ha impatto sul rendimento degli studenti una volta che gli altri fattori sopraddetti sono tenuti sotto controllo?

Lo scopo della mia tesi è rispondere alla domanda in questione per il sistema scolastico italiano. Precisamente, l'obiettivo è valutare l'impatto della qualità della scuola sull'apprendimento degli studenti nel corso dell'anno scolastico usando i dati dalla seconda indagine del *Program for International Student Assessment* (programma internazionale per la valutazione degli studenti)—PISA 2003. Questi dati forniscono informazione su un campione di circa 11.000 studenti quindicenni italiani provenienti da circa 500 scuole superiori (30 studenti per scuola, mediamente), valutati sulle loro competenze in matematica ad Aprile 2003.[6] Quindi la variabile d'interesse è l'apprendimento degli studenti, specificamente in matematica, durante l'anno scolastico 2002/03.

Nel corso della tesi per *achievement* intenderò il livello di competenza degli studenti in matematica in un certo punto nel tempo durante l'anno scolastico. Perciò il termine *gain* in *achievement* si riferirà alla variazione delle competenze degli studenti (quello che ho chiamato apprendimento, sopra) in un certo intervallo di tempo, ad esempio l'anno scolastico.

I dati PISA, sia a livello di singolo studente che a livello di scuola, derivano dai questionari compilati dagli alunni e dai dirigenti scolastici al momento del test. Dato il rischio di percezioni soggettive ed errori volontari di compilazione da parte dei presidi, uso l'informazione più attendibile proveniente dall'archivio amministrativo *Sistema Integrato Segreterie Scolastiche Italiane* (SISSI) a livello di scuola. Dal *dataset* risultante è possibile conoscere un gran numero di caratteristiche degli studenti e delle scuole, necessario per far funzionare la mia strategia di stima (vedi Bratti, Checchi e Filippin [2007] per una descrizione dettagliata dei dati).

Per poter valutare l'effetto causale della qualità scolastica sulla variazione di *achievement* degli studenti in matematica durante l'anno scolastico, devono essere risolti tre problemi. Per prima cosa, il processo di assegnazione degli studenti

[6] Le competenze valutate da PISA sono l'abilità degli studenti ad usare le nozioni scolastiche per la risoluzione dei problemi della vita quotidiana, piuttosto che la conoscenza fine a se stessa delle materie scolastiche.

a scuole di qualità diversa non è casuale: per esempio, è probabile che siano gli studenti più abili a frequentare scuole di qualità superiore. Di conseguenza il semplice confronto di studenti che frequentano scuole con livelli differenti di qualità conduce a conclusioni errate circa l'impatto della qualità scolastica. In questo modo andiamo a stimare un effetto spurio della qualità della scuola sul *gain* in *achievement* degli studenti, che ingloba in parte gli effetti di altre variabili (come l'abilità) che influenzano la scelta della scuola e l'apprendimento nel contempo (si veda, per esempio, Dearden, Ferri and Meghir [2002]).

Vi è poi il problema che la qualità scolastica corrisponde ad un concetto estremamente astratto e privo di una chiara definizione. Questo concetto non ha una concreta controparte di riferimento e siamo in grado di misurarlo solo indirettamente attraverso una serie di *proxies*, come la dimensione della classe, il livello di spesa per studente, il titolo di studio e l'esperienza degli insegnanti, e cosi via.[7] Quindi, a differenza dai contesti usuali di valutazione dei programmi, la nostra variabile "trattamento" è una variabile latente, che assumerò unidimensionale in prima istanza. Per stimare l'effetto causale della qualità scolastica, seguirò la procedura proposta da Black e Smith [2006] per il caso unidimensionale; successivamente estenderò questa strategia al caso generale di molteplici dimensioni di qualità. Essa corrisponde alla stima di un modello di analisi fattoriale attraverso variabili strumentali.

Il terzo problema è dato dal fatto che anche la variabile risultato, *gain* in *achievement*, è latente. Infatti l'*achievement*, come la qualità della scuola, è un concetto confuso e molto astratto, misurabile con errore da *proxies* costruite ad hoc, come il punteggio a tests. PISA 2003 fornisce due *proxies* per l'*achievement* degli studenti: il punteggio al test PISA, condotto ad Aprile 2003, ed il voto dell'insegnante di matematica nell'ultima pagella a Gennaio 2003. Chiaramente questo non è il semplice caso di due misure ripetute dello stesso fattore latente in due occasioni, perchè le variabili considerate corrispondono a dimensioni (aspetti) diversi di *achievement*. Svilupperò un metodo che combina queste informazioni disponibili da PISA per costruire una misura della variazione di *achievement* dello studente nel corso dell'intero anno scolastico, la quale costituisce una variabile di crescente interesse nella letteratura economica (si veda Hanushek, Kain and Rivkin [2005] in proposito). Lo sviluppo di una procedura statistica che permette di poter applicare un approccio valore-aggiunto usando dati PISA rappresenta un utile contributo in letteratura.

I risultati dell'analisi sembrano confermare quelli trovati in letteratura: la qualità della scuola, misurata dalle usuali *proxies* di qualità—come la dimensione della classe, non conta controllando per abilità individuale, contesto familiare e

[7] Si veda Bollen [1989] per una spiegazione accurata del processo che lega un concetto ad una o più variabili osservabili.

caratteristiche del gruppo (Hanushek [2003]). Comunque la scuola ha molteplici dimensioni di qualità con peso diverso sull'acquisizione delle competenze (Black and Smth [2006]).

Se esiste un impatto della qualità della scuola, esso proviene dagli insegnanti[8], che anche la maggior parte degli studi in letteratura ha mostrato essere l'ingrediente fondamentale del processo di apprendimento (l'articolo di Hanushek, Kain e Rivkin [2005] è l'esempio più significativo). In particolare, i dati utilizzati nell'analisi rivelano che le scuole con una frazione maggiore di insegnanti di ruolo sono quelle che hanno successo nell'accrescere il rendimento degli studenti[9], essendo gli insegnanti di ruolo più motivati e più efficaci nel produrre istruzione.

Ad ogni modo, ci sono alcune questioni da approfondire nelle ricerca futura, ad esempio chiarire il ruolo giocato dalla dimensione della scuola (in termini di numero di studenti). Poi sarebbe interessante applicare il mio approccio valore-aggiunto anche ad altri sistemi scolastici, oltre a quello italiano. Specificamente, l'indagine internazionale PISA permette di poter passare ad un'analisi *cross-country*, dove altre *proxies*, come l'organizzazione scolastica e le istituzioni (che sembrano rivestire un ruolo chiave nella funzione di produzione dell'istruzione), potrebbero essere valutate.

Il resto della tesi è organizzato come segue. Dopo un rassegna della letteratura sulla qualità della scuola nel Capitolo 2, il Capitolo 3 descrive brevemente il disegno dell'indagine PISA ed i dati. Il Capitolo 4 inizia col collocare lo studio nello scenario econometrico della valutazione dei programmi, dando una soluzione al primo problema di selezione non casuale degli studenti. Sono poi discusse le altre due questioni relative alla non osservabilità dell'*achievement* e della qualità scolastica, risolte sia separatamente che congiuntamente per facilitare la spiegazione. In questo capitolo assumo una dimensione per la qualità scolastica; il caso più generale di molteplici dimensioni di qualità è affrontato nel Capitolo 5, dove derivo un metodo per stimare non iterativamente il modello risultante di analisi fattoriale, che è una diretta estensione della strategia di stima adottata nel Capitolo 4 per il caso semplice di qualità unidimensionale. I risultati empirici sono riportati nelle sezioni finali dei capitoli 4 e 5, per il caso unidimensionale e multidimensionale, rispettivamente. Infine, nell'appendice, è descritto un metodo alternativo di stima, che mostra la robustezza della mia strategia di risoluzione del secondo problema di misure diverse, di dimensioni diverse, di *achievemenet* degli studenti. Mostro

[8]Anche la dimensione della classe è una *proxy* legata all'insegnamento, nel senso che la dimensione della classe, o equivalentemente il rapporto studenti-insegnanti, è una misura (con segno negativo) del "tempo di insegnamento per studente", il quale potrebbe non esercitare alcuna influenza controllando per abilità degli insegnanti (cioè gli insegnanti abili forniscono buona istruzione tanto nelle classi grandi che nelle classi piccole).

[9]In base alla mia conoscenza, questa caratteristica degli insegnanti (da SISSI) non è mai stata usata in letteratura, in particolare in paesi diversi dall'Italia.

inoltre l'inutilità di applicare questa strategia separatamente in diversi gruppi della popolazione, come suggerito da alcuni rapporti sull'indagine PISA in Italia.

Chapter 2

Literature review

In this chapter I review the empirical evidence available so far on school quality, intended to be the ability of schools to promote knowledge and skills. I explain the existence of two parallel lines of research: one measures school quality by students' performance on standardized tests, whose effects on individual income and economic growth of the country are of interest; the other addresses itself to identify the determinants of students' learning among observed characteristics of schools (such as class size, teachers' experience, teachers' education, and so forth), which now play the role of school quality proxies. In this survey I refer to this second strand of literature, but I present some results also for the first one in the introduction, where further I discuss what observable school characteristics are usually adopted as measures of quality.

The striking evidence on the importance of students' competences for economic development prompts the evaluation of the effective role played by schools on students' acquisition of skills, as compared to other educational inputs like family background and peers. It is also interesting to evaluate the effects of schools on other individual outcomes than pure educational achievement, such as the performance in the labour market. Therefore, the studies of the second group can be divided accordingly to the outcome of interest. The major part of works enters the 'test score literature' or the 'earnings literature', which I treat in Section 2.2: they look at test scores and post-schooling earnings, respectively, and they provide conflicting results at least until the end of the 1990s. The most recent contributions are instead discussed in Section 2.3, where I consider separately the two main aspects of schools investigated by researchers: teachers' quality (Section 2.3.1) and class size (Section 2.3.2). I also summarize the relevant results of the few recent studies addressed to estimate the impact of school inputs on the level of education achieved, in Section 2.3.3. Finally, in Section 2.4, I review the separate and growing strand of literature dealing with college quality.

Another important aspect that distinguishes the large body of studies under examination is the analytical procedure employed, for example fixed effects formulations of the education production model and/or completely non parametric estimation approaches, especially in the last few years (Section 2.3.1). A set of more fortunate studies can exploit experiments; more recent works make use of quasi-experimental strategies, in particular for class size (Section 2.3.2).

2.1 Introduction

The early empirical works on human capital focus on the economic returns to school attainment, the latter being measured by years of schooling. They follow the seminal study of Mincer [1970, 1974], who investigates the relationship between earnings and years of schooling through a simple linear regression of the logarithm of wages on schooling, labour market experience and other individual characteristics (the well known 'Mincer equation')[1]. The overall finding is that an additional year of education is associated with an average increase in earnings of about 10 percent. This result is robust to an imperfect control for ability, because the upward bias generated by the omission of relevant factors approximately cancels out the downward bias produced by measurement errors in reported years of schooling (Griliches [1977]).

Since researchers can easily access data on 'school quantity', only recently the empirical economics literature has turned to consider 'school quality', the real concern of the political debate. As I have said above, there are two parallel lines of research on school quality. Let me consider the one that identifies school quality with people's knowledge and skills, as measured by standardized test scores at a given level of education.

The aim is to establish the role of school quality in determining individual earnings and the economic growth of a country. At the individual level this corresponds to estimate a standard Mincer equation with test performance as additional regressor. The most recent US studies are the ones of Mulligan [1999], Murnane et al. [2000] and Lazear [2003] (see Hanushek and Woessmann [2007] for a detailed description of these works and a complete review, also regarding other countries than the US). They consistently find that an increase of one standard deviation in mathematics performance at the end of high school is associated with an increase of 12 percentage points in subsequent annual earnings.

Hanushek and Zhang [2006] offer an up-to-date across-country comparison of the returns to school quality on individual earnings, using the International Adult

[1] The most recent studies are the ones of Harmon, Oosterbeek and Walker [2003], Psacharopoulos and Patrinos [2004] and Heckman, Lochner and Todd [2006]. See Hanushek and Woessmann [2007] and Card and Krueger [1996] for a review. Card and Krueger also describes other usual approach to estimate the return to years of schooling.

Literacy Survey data: there are substantial advantages in terms of individual income for the countries with higher achievers in the test. They also show the misleading conclusions from considering quantity of schooling alone in the estimation of the Mincer's returns, whose values considerably decrease after adjusting for literacy scores.

Therefore school quality has an indirect effect on earnings through continuation in school. Many works document the strong relationship between school attainment and school quality, measured by grades or scores in achievement tests. Hanushek [2005] provides an overview of this part of literature.

Other interesting results about the impact of school quality on economic growth come from the recent work by Hanushek and Woessmann [2007][2], who estimate a regression of the average annual growth rate of GDP per capita during the period 1960 to 2000 on a measure of school quality, given by a combination of international standardized tests over the 40-year period considered, plus the initial level in GDP per capita, the average number of years of schooling and other control variables. The strongly significant positive effect of school quality on economic growth indicates that one standard deviation increase in test performance translates into 2 percent higher average annual growth rate in GDP per capita. Instead the effect of school quantity turns out to be statistically insignificant, after school quality has been taken into account.

Therefore, having recognized that people's knowledge and skills cause the economic and social well-being of a country, to shed light on the effective role played by schools in the production function of human capital becomes of primary importance. This is the goal of the second line of research which I refer to, where students' test scores are viewed as a product of school quality, rather than school quality itself, now measured by specific school and teacher characteristics (the so-called 'school resources'), such as class (or, equivalently, student-to-teacher ratio), teacher experience and education, and so on. Researchers aim at determining the effect of each input, netted out by the influence of other dimensions of school quality and other educational inputs that enter along with school quality in the production function of human capital. The ultimate purpose is to identify the factor on which policy can more conveniently intervene in order to increase students' performance (and hence future earnings and economic growth, by the previous discussion).

To be clear, the term 'school quality' now refers to attributes of schools that are influenced by educational spending. Class size plus teacher experience and education constitute the 'real resources' available at school for students. Teacher experience and education concur to determine teacher pay, which represents the

[2] They extend the work of Hanushek and Kimko [2000], the first ones to emphasize school quality differences across countries.

major component of expenditures per student with class size. Expenditures per student and teacher salary constitute, on the other hand, the 'financial resources' of schools. These real and financial resources are the most common dimensions of school quality employed in literature, because they (i) are well-measured and readily available from administrative archives and (ii) are easy targets of political manoeuvring. Hanushek [2003] documents the strong increase in expenditures per student over the last forty years in the US, explained by the contemporaneous drop in pupil-to-teacher ratio. What is surprising is that these changes have not been accompanied by an improving in student performance, which has instead declined.

The major part of literature focuses on class size because the favourite intervention of policy-makers is reduction of class size, being the most visible to the voter and the most easy one to implement (Dustman, Rajah and van Soest [2003]). Moreover class size is a discrete well-defined treatment variable suitable for several statistical procedures, differently from other abstract concepts such as teacher quality (Hanushek and Woessmann [2007]). Anyway we will see that the most recent studies of the school quality literature look just at teacher quality.

But the school quality literature can also be divided accordingly to student outcomes as well as school resources. Indeed school quality is likely to affect other outcome variables than test scores, such as educational attainment, labour market participation and wages, family formation and so forth. There are, mainly, two strands of research: the test score literature and the earnings literature, which consider test scores and wages as outcome variable, respectively (Section 2.2). The puzzle is that they represent two opposite lines of research, showing the former insignificant effects and the latter statistically significant effects of the common school resources seen above (at least until the mid-1990s). Quite a few researchers are also interested in the impact of school quality on years of schooling (Section 2.3.3).

2.2 Test score literature vs. earnings literature

The school quality literature ages nearly forty years and stems from the Coleman Report in 1966, which was commissioned in response to the Civil Rights Act of 1964 to assess the equal distribution of educational quality by race, colour, religion and national origin. This was the first systematic survey to gather detailed nationwide data on schools, test scores and students' characteristics. The main finding is that there are differences in test scores among students of different race, but they are attributable to family environment and socioeconomic status rather than to common measures of school quality. Hence school quality does not influence student performance, once the effects of family background and peers have been netted out. The results do not change with many of the studies following the

Coleman Report and looking at students' test scores (see Hanushek [1986], [2003] for a review).

On the other hand, school quality seems to strongly matter for wages (see Card and Krueger [1996] for a review of this strand of literature). The first noteworthy work is the one by Johnson and Stafford [1973][3], who find high returns to investment in expenditures per student (they, also, are the first to emphasize the role of school quality on years of schooling; see Section 2.3.3). But the study of Card and Krueger [1992a] is the milestone reference in papers asserting the positive impact of school quality on students' subsequent labour market success (and school attainment). Differently from the previous works, which estimate simple regression models of earnings on one or more measures of school quality and other control variables (at the district or state level), Card and Krueger [1992a] identify the impact of school quality comparing the earnings of individuals who live in the same state but were born, and hence educated, in different states (with different levels of quality). They find higher returns for individuals born in states with higher quality schools in terms of pupil-to-teacher ratio, term length and relative teacher pay (note that school quality enters the slope relative to years of education in their model).

The work of Betts [1995] is the first to find an insignificant effect of school quality on earnings, consistently with the test score literature. Using panel data from the National Longitudinal Survey of Youth he first estimate a Mincer's regression by adding a dummy variable for each school in the sample among the regressors. The rejection of the null hypothesis that all dummy coefficients are equal to zero indicates that there are significant differences among schools, but they are not related to the standard measures of school quality. Indeed, in the second step, when the dummies are replaced by the school quality measures (the same employed by Card and Krueger [1992a], now at the level of school), none of them result statistically significant.

To the same conclusion about the irrelevance of school quality on earnings comes Grogger [1996a, b] using data from the High School and Beyond Survey. He shows that teacher-to-student ratio and expenditures per student have not effect on wages: specifically, a 10 percent increase in school spending is associated to only 0.68 percent higher wages. This study plus the previous one by Betts differ from the study by Card and Krueger for (i) the level of aggregation of the school quality data—school level instead of state level—and (ii) the age of the workers—in the last two studies young men who have just entered the labour market. Furthermore, Card and Krueger consider men born in the 1920s, 1930s and 1940s, whereas Betts and Grogger consider a younger cohort of men born in the 1950s and later.

At this point of the school quality history in the US literature, an intense

[3] See Betts [1995] and Card and Krueger [1992a] for other references.

debate has taken place to reconcile the contradictory findings of the test score and earnings literature, in the light of new evidence (cf. Moffit [1996]). Sometimes it has turned into a dispute, which sees, chiefly, Eric Hanushek and Alan Krueger at opposite poles; consider, for instance, the controversy between Hanushek [2003] and Krueger [2003] with respect to the effectiveness of class size reductions (Section 2.3.2).

Let me reconsider the paper by Betts [1995]. He provides a series of plausible reasons for the inconsistent findings between the test score and the earnings literature (some of them can be found again in other researchers' discussions, of course). First of all, we can think of a 'cohort effect' to explain the divergence of the results of Betts [1995] from those of Card and Krueger [1992a] and the other ones in the earnings literature, which consider old cohorts of workers (men born during the first half of the century)[4]. Betts [1995] hypothesizes that there could be a structural shift during the last decades which has led to weaken the initial strong relationship between wages and school quality (see Heckman, Layne-Farrar and Todd [1996], Loeb and Bound [1996] and Hanushek [2002] for similar considerations). Then he questions the validity of test scores as measures of student performance, referring to Griliches and Mason [1972], who find low correlations between test scores and subsequent labour market outcomes (see also Card and Krueger [1992a] and Johnson and Stafford [1973]). In addition test scores could be *a poor indicator of what is learned at school and subsequently rewarded in the labour market* (Card and Krueger [1996]). But the results discussed in Section 2.1 about the strong link between test scores and earnings are enough to dampen these assertions (cf. Hanushek [2002], [2003]). Anyway, we cannot deny that test scores are an imperfect measure of student achievement owing to measurement errors and the difficulty in standardizing tests. Finally Betts asserts that the strategy of Card and Krueger [1992a] leads to substantial measurement errors because of the assumption that a person is educated in the same state where is born. This is the first blame of a long series of attacks on the analytical approach of Card and Krueger [1992a]; see, for example, Heckman, Layne-Farrar and Todd [1996] and Hanushek in all his papers.

Heckman, Layne-Farrar and Todd [1996] replicate the study by Card and Krueger [1992a] extending the analysis to cover additional years. The authors find again significant positive effects of school quality on earnings, but these results are sensitive to the estimation procedure pursued. In particular, two identifying assumptions of the Card and Krueger's strategy are rejected by the data: the hypotheses of random migration and constant returns to school quality across

[4]The data on test scores are more reliable for young cohorts of men born in the second half of the century. On the contrary, there is more stable information on income (when workers have gained experience in the labour market) for men born in the first half of the century.

regions of residence. Migrants are self-selected and their school quality is priced differently across regions of residence. Further, data support a non-linear specification in years of schooling of the log wage equation: the returns to years of schooling vary with their level. Under this new evidence, the estimated effect of school quality becomes weaker, accordingly to the test score literature.

Hanushek, Rivkin and Taylor [1996] also question the results by Card and Krueger [1992a], attempting to reconcile the opposite findings of the test score and achievement literature. Their paper starts with a review of 377 studies on the effects of school quality (teacher-to-pupil ratio and expenditures per student, specifically) in 90 different publications. The studies that show positive impacts of school quality are those which use aggregate school quality data at the state or district level, whatever is the student outcome variable considered. Moreover, the analyses that examine multiple state show, mainly, positive effects. The authors talk of an 'aggregation effect' generated by omitted variables at the state level: state regulations, state teacher certification, state funding formulas, and the like. They develop a formal statistical model demonstrating that the omitted variable bias is largest at the level of the omitted factors. Hence, the contradictory findings of the two strands of literature are due to the fact that they use different levels of aggregation of the school quality data. With the exception of Betts [1995] and Grogger [1996a, b], all studies focusing on earnings aggregate school data at the level of state or district (consider, for example, Card and Krueger [1992a] or Johnson and Stafford [1973]). These studies find strong positive effects because they omit relevant variables at the state level, which biases upward the estimated coefficients of school resources. Clearly, omitted state factors (common to all schools or districts in the same state) will not bias the estimated effects if the analysis is conducted entirely within state. Therefore, we have to believe the results of the test score literature and Betts [1995], which measure school quality resources at the school level. This implies that the standard measures of school quality matter little, whatever the student outcome considered is.

Card and Krueger [1996] defend themselves and the overall earnings literature (excluding Betts [1995] and Grogger [1996a, b], of course) pointing out the risks of working at the school level, rather than at the state level. The 'aggregation effect' could operate in an opposite direction to that claimed by Hanushek et al. [1996]—i.e. there could be downward bias at the school level—because of the considerable measurements errors in the resources of schools reported in administrative archives, which disappear, or at least weaken, taking the average district or state value. But, even worse, there is a problem of endogeneity due to residential location choices, arising from parents' decision, at the disaggregated level (Tiebout [1956]). This problem is still an open issue in the school quality literature, as well as, more in general, the non random matching of both students and teachers with schools (see, for example, Dearden, Ferri and Meghir [2002] and Hanushek, Kain and Rivkin

[2005]).

However, the pattern of results in the review by Hanushek, Rivkin and Taylor [1996] (see also Hanushek [2003]) points toward misspecification (upward) bias instead of errors-in-variables (downward) bias, since the impact of resources seems to be stronger in multi-state samples, regardless of the level of aggregation of the quality data (also at school level with a sample of students in more than one state). Furthermore, by an empirical analysis, the authors demonstrate that the magnitude of the coefficients does not increase aggregating school quality measures at other levels than state.

Then Card and Krueger [1996] points out the young age (20s to early 30s) of the workers considered by Betts [1995] and Grogger [1996a, b] (see also Betts [1996]). We can think of an 'age effect', such that earnings increase with age or, more exactly, with work experience. Consider, for instance, the literature on the effects of school attainment and test scores on wages (Altonji and Pierret [2001] is an example). Nevertheless, this age dependence hypothesis is rejected in the study of Betts [1996], who replicates the analysis of Card and Krueger [1992a] examining also younger workers: surprisingly, the effect of school quality does not grow as the workers gain experience.

As regards, instead, the 'cohort effect' issue—raised by Betts [1995]—Loeb and Bound [1996] deal with it. They have been able to locate a dataset (from the General Social Survey) with information on test scores for individuals born in the first half of the century (like the individuals examined by Card and Krueger [1992a]). The (state level) school quality measures result in producing significantly positive effects on achievement, differently from the remainder of the test score literature. Really, there are also three studies that attain the same results considering achievement of men born in the first half of the century: Orazem [1987], Margo [1990] and Schmidt [1995].

Therefore, the contrast between the earnings and the test score literature may be simply explained by a diminishing impact of school quality (on both achievement and wages) over time. One reason could be that the educational system has changed (Borland and Howsen [1992], Peltzman [1993] and Hoxby [1996]) and now school expenditures are much higher than they were in the past. Hanushek [2002] talks about a diminishing marginal productivity of school quality, which has been confirmed by many studies in developing countries where the level of school resources is still low (see Hanushek and Harbison [1992], for example); consider, also, Johnson and Stafford [1973], who find that the (positive) return to school quality decreases with the level of quality. Another related explanation of the 'cohort effect' is a declining variability in school quality over time (cf. Heckman, Layne-Farrar and Todd [1995], and Loeb and Bound [1996]), due to the convergence toward high levels of quality across schools. The variation in school resources was greater at the beginning of the century, so these quantities were more reliable indicators of school quality. Further, the small variability prevent us from detecting small, but significant, effects of school quality.

Also Card and Krueger [1996] talk about low power and imprecise estimates of small effects, but they refer to the small samples of the studies in the achievement literature. Hanushek, Rivkin and Taylor [2005] deal with the same issue, too (see forward).

Finally, another noteworthy study is the one by Altonji and Dunn [1996], who still find positive effects of school inputs (teacher's pay, expenditures per student and a composite index of school quality measures) on wages using disaggregated data from the National Longitudinal Survey. They solve the problem of unobserved family characteristics by comparing the outcomes of siblings who attend different schools (with different quality) but have the same family background; more exactly they are able to control for unobservable variables common to siblings by adding family fixed effects in the wage model. Note that the estimates remain positive, but are smaller, when family fixed effects are not controlled for, so the omission of relevant family variables generates downward bias. However, contrary to Card and Krueger [1992a], the authors find that school quality has no effect on the return to an additional year of schooling. On the whole this study suggests that an important issue to be taken into account is the inadequate control for family background characteristics.

2.3 Recent results and alternative approaches

There is agreement among researchers that the main reason of discrepancy in findings is the omission of relevant educational variables, such as parental inputs or individual ability, in the estimated educational production functions (see, for example, Card and Krueger [1996] and Hanushek [2002]).

Many studies have included few or no controls for family background and individual endowment, with the result of estimating spurious relationships between achievement (earnings) and school quality, i.e. they have erroneously attribute to school quality higher achievement (earnings) generated by better family factors or higher skills. Furthermore, school quality often refers to just one school year and, in general, only contemporaneous inputs are considered, having researchers access to cross-sectional data. But the acquisition of skills is a cumulative process and all historical and present educational factors influence achievement at some time point.

Fortunately, the works of these last few years can benefit from very rich datasets which allow to overcome the problems of omitted variables bias. In addition, it is common practice to use a value-added approach in the test score literature, where the control for prior achievement through a 'pre-test' score allows to cancel out (time invariant) individual ability and lessen the endogeneity problem due to omitted historical factors.

Despite these better conditions, school resources continue to have insignificant effects on performance in tabulated results. Consider, for instance, the works of Dustman, Rajah and van Soest [1998] and, later, Feinsten and Simmons [1999], for the UK educational system, which use the same rich dataset from the National Child Development Survey. They both show that pupil-to-teacher ratio has no effect on student performance at the first UK national exam, after having controlled for parental background, peers and previous achievement.[5]

The main argument for these findings is that school quality really matters for student performance, but the standard school resources fail to capture school quality, especially for teachers whose fundamental aspects, like motivation and skills, cannot be measured. There is evidence in literature that refined measures of teacher quality have positive effects on students' performance. See the references in Goldhaber and Brewer [1998], who show that some teacher observables characteristics regarding their specific preparation affect student outcomes. They assert that the usual teacher measures are also inadequate for being averaged at the school level: in such way school measures do not take into account the considerable amount of variability within-school and end by not representing any teacher in the school. Goldhaber and Brewer [1998] use data from the National Educational Longitudinal Survey which allows them to link students to a specific class and teacher, and include fixed effects in their statistical model so as to control for unobservables. Accordingly to what I have said above, observable school, teacher and class characteristics explain a little part of the overall variance in student performance.

Therefore, a growing body of research, which makes use of panel data, has turned to evaluate teachers' quality by the systematic differences in test scores among students of different classes with different teachers (Section 2.3.1). Now there is broad consensus that teachers have a central role in determining achievement, but the variation in student performance across teachers is not related to their common observable characteristics, such as level of education. Clearly, it is a dilemma in the political debate the inability to identify specific teacher characteristics on which to intervene to promote student achievement (see Hanushek [2002]).

The teacher quality literature employs fixed effects models to circumvent the problems of both inadequate proxies and endogeneity bias (see Hanushek and Woessmann [2007] for a review). A variety of other more recent approaches have been also implemented to disentangle the causal effect of school quality: they

[5]Anyway, Dustman et al. [1998] show a positive effect of pupil-to-teacher ratio on school attainment; see Section 2.3.3.

essentially try to identify factors which generate exogenous variability in school resources (quasi-experimental strategies). More fortunate studies can even use data from natural or ad hoc experiments. These are common practices mostly to evaluate the impact of class size, which is well suitable for many statistical treatments, being a discrete variable. Further class size constitutes the principal object of interest of the school quality literature for its policy implications (revise Section 2.1). See Hanushek [2003] and Woessman [2005] for a review.

An example of natural experiment is the 'Operation Solomon' in May 1991, during which 15,000 Ethiopian Jewes were brought to Israel and here randomly distributed across the country. Therefore the random assignment of children to schools of different qualities has allowed to evaluate the casual effect of school quality, free of the influence of other disturbing factors such as parents' choices and family background. Gould, Lavy and Pasermann [2003] try to estimate the impact of the quality of the elementary schools across which Ethiopian pupils have been sorted, measured by average test scores a year before the immigration, on their later achievements in high schools. They find strong effects of elementary school quality on high school outcomes such as dropout, repetition and exam performance. Another example of natural experiment is provided by the segregation school system in the South US during the first half of the century (see Card and Krueger [1992b] and [1996] for more details). As regards ad hoc experiments, consider the Tennessee Project STAR (see Section 2.3.2).

2.3.1 Teacher quality literature

We have seen that Betts [1995] finds an effect of school quality on wages by including in the Mincer's regression dummies for schools, whose coefficients turn out to be statistically significant. On the other hand the effect disappears by replacing the dummies with school observable characteristics, which are instead not significant. As we have pointed out above, schools make a difference, but their quality is not adequately measured by the standard proxies used in the empirical literature. This is in particular true for teacher quality, which has been demonstrated to have a key role in the production of student achievement by most current research.

The more direct method to assess the importance of teacher quality is identifying systematic differences in the average performance of students assigned to different teachers (adjusted for all other factors affecting student achievement, of course); the more able teachers are those who obtain higher students' performance. This involves estimating fixed effects models (see Hanushek [2002]), data permitting, which translate into covariance analyses.

Hanushek [1971] is the first to proceed in this way and finds that classroom dummy variables are statistically significant in the regression of students' test scores after conditioning upon initial performance and other factors relevant on achievement gains. These results are confirmed by other studies (cf. Murnane [1975], Murnane and Phillips [1981], Armor et al. [1976] and Hanushek [1992]). Anyway, all these early studies are not able to disentangle the impact of teacher quality from the effects of other classroom factors, because teachers are generally observed in one classroom with cross-sectional data. Only recently, it has been possible to obtain more reliable estimates, thanks to the availability of panel data that allow to observe more classes with the same teacher. So it has been possible to include student fixed effects (by observing students' test score in more than one year) and school fixed effects, exploiting these richer datasets.

We have already considered an example above, with the study of Goldhaber and Brewer [1998]. Another example is the study of Rockoff [2004], who estimates statistical significant differences among teachers in elementary schools: a one standard deviation increase in teacher quality corresponds to a growth of about 0.20 and 0.24 standard deviations in standardized reading and math test scores, respectively. He also finds that teacher experience has positive effects on student test scores. See Hanushek and Woessmann [2007] for references to the most recent studies; Hanushek [2002] and Rockoff [2004] provide a review of the early results.

The most interesting evaluation study is the one by Hanushek, Kain and Rivkin [2005], for both methodology and data availability. They have access to a huge dataset with information on three cohorts of students in adjacent grades—more than 200,000 students in over 3,000 public schools for each cohort—for six consecutive years. It allows them to circumvent the problems of omitted or mismeasured variables and nonrandom matching of students, teachers and schools by the use of complex fixed effects models, with student, school-by-grade and school-by-year fixed effects. Indeed, in the second part of the study, when the author estimate the effect of observable teacher and school characteristics, they are able to account for all the fixed components of the error term, differently from the previous literature. Moreover, unlike other studies, the big sample size permits to obtain precise estimates of even very small effects.

The estimation results show that one standard deviation increase in teacher quality raises the annual gain in achievement by at least 0.11 standard deviations. But the variance in teacher quality declines as students progress through school and teacher quality comes to explain nothing of the variation in student performance in junior high schools. This is confirmed by the findings for the effect of class size, which is statistically significant only for the first grades considered (4th and 5th). It is interesting to note that a policy that increases teacher quality by one standard deviation—and so generates an increase of 0.11 standard deviations in annual variation of student achievement—is equivalent to a policy that reduces class size of 10 students in the 4th grade, 13 or more students in the 5th grade and an implausible larger number in the 6th grade. Finally there is no evidence that better teacher education raises student performance. Teacher experience, instead, seems to matter, but solely for the first two years of experience, accordingly to Rockoff [2004].

Therefore, teacher attributes explain only a small part of the overall variance in teacher quality and this inability to identify specific factors on which to play creates a grave dilemma in the policy debate. For Woessmann and Hanushek [2007] the only way to act in order to improve student performance is giving incentives to all actors involved in the educational process, i.e. all persons involved in education—specifically teachers—must be rewarded/penalized on the basis of students' outcomes.

Another study worth mentioning is the recent work by Kramarz, Machin and Ouazad [2006]. Like Hanushek, Kain and Rivkin [2005] (i) they have access to a rich panel dataset for six cohorts of English pupils in adjacent grades followed from primary to secondary education and (ii) they use fixed effects models without relying on observable characteristics. Their goal is to determine the relative contribution of students, schools and peers in the education production function without using any proxy for peers and school quality (there is a wealth of information on pupils' characteristics, differently from Hanushek, Kain and Rivkin [2005]). The estimation results confirm the Coleman Report's findings, i.e. family background and peers are much more important than school quality for student achievement.

Finally, an interesting recent work focusing on teacher quality is provided by Strayer [2002]. It is very interesting especially for the methodology, different from the other ones seen until now (a similar approach is employed by Dustman, Rajah and van Soest [2003] for the UK context; see Section 2.3.2). Modelling the college choice jointly with the wage determination, Strayer [2002] is able to account for two separate effects of high school quality on earnings. In addition to a direct effect of high school quality on wages, usually considered in the research, there is an indirect effect working through college choice, because the quality of the high school attended by an individual affects his choice of college, which in turn affects his/her future post school earnings. For example, students attending higher quality secondary schools are more likely to attend also higher quality universities, which we expect to give an higher reward in terms of labour market outcomes. Therefore a part of the overall return from college choice comes from the high school choice; see also Behrman, Rosenzweig and Tauban [1996], who emphasize the importance to control for attributes of schools attended before and after the school level analyzed in order to correctly estimate its effect on wages if school quality is correlated across levels (see Section 2.4). Strayer's [2002] results show that the indirect effect of high school quality is very strong: a one standard increase in the fraction of teachers with a graduate degree (in high school) raises the predicted probability to attend a 4-year and 2-year college by 3.3 and 2.4 percentage points, respectively; besides the predicted wages for individuals who attend a 4-year college and a 2-year college are 23 and 9 percent higher than the predicted wages of individuals who do not attend any college, respectively. The author also obtains some evidence of a direct effect of high school quality on earnings, but it is weak (fixed the quality of the college attended, an increase in the fraction of teachers with a graduate degree makes some difference to wages).

2.3.2 Class size literature

A large body of literature has addressed itself to uncover the causal effect of class size on student achievement, for the reasons said before (Section 2.1). Unfortunately, findings are inconsistent, and even the most recent studies do not provide clear indications.

A source of endogeneity bias for class size is the non-random allocation of students in classes of different dimension within schools, due to compensatory funding schemes, such that lower achievers are located in smaller classes where they may be taught better. On the contrary, policy makers may want high-performers in special small classes in order to create a talent elite. Then, parents could choose to live where schools have relatively smaller classes, if particularly interested in sons' education (evidently they also act in many other ways to support sons' learning).

Because of these endogeneity issues, lots of researchers have attempted to find sources of exogenous variation in class size, such as quasi-experimental strategies employing instrumental variables approaches. Angrist and Lavy [1999] have found out that the Maimonides' rule requires that Israeli classes cannot be larger than forty students. Therefore, when a grade enrolment is just above a multiple of forty there is a truly exogenous fall in class size: the observed differences of outcome between students before and after the fall can be attributed to a casual effect of class size. In other words, the class size predicted value from grade enrolment on the basis of the role is a valid instrument for the actual class size (clearly it does not depend on students' performance). Angrist and Lavy's final results indicate a positive effect of class size on student achievement.

Instead Hoxby [2000] finds an insignificant impact of class size exploiting class-size variation caused by natural fluctuations in cohort sizes. There are many other studies exploiting quasi-experimental instances, but they are quite varied as for results and difficult to generalize. See Woessmann [2005] for a detailed review.

Sometimes researchers can benefit from ad hoc experiments where individuals are randomly assigned to different treatments. The only example for class size is the Project STAR in Tennessee in the mid-1980s. A group of kindergarten pupils were randomly assigned to small and large classes. They have been followed until the third grade and tested at the end of each year. The standard reported result is that pupils in smaller classes perform better than pupils in larger classes (Finn and Achilles [1990]; Mosteller [1995]), especially disadvantaged students (Krueger [1999]). But this assertion undervalues some important aspects (see Hanushek

[1999] for a careful explanation). First of all the actual implementation of the experiment leads to cast doubts on the quality of the randomization: (i) 51% of students exit the experiment before the end; (ii) each year lots of students do not take the annual test; (ii) lots of students change the treatment group during the experiment. Then, the substantial advantage observed in the first year of the experiment for students in small classes remain constant in the following treatment years: it is natural to expect that the divergence between the two groups instead increases year by year. Further Hoxby [2000] points out the threat of unusual behaviour of students, teachers and administrators who know to be controlled. Anyway, the estimated positive impact (whatever it is genuine) is relatively small compared to the large reduction in class size and it can be very difficultly generalized to other educational contexts.

We see that another dispute between Hanushek and Krueger has taken place around the evaluation of the Project STAR. In general, it extends to the interpretation of the overall findings of the school quality literature. Both Hanushek [2003] and Krueger [2003] provide a summary of the previous literature, but they come to opposite conclusions, showing that school resources do not matter (the estimates are mainly statistically insignificant) and matter (the estimates are mainly statistically significant and positive), respectively. The principal reason for this discrepancy is the aggregation procedure of results used by the two authors: while Hanushek weights each study through its number of estimates, Krueger derives a unique composite estimate from each study[6]—this lessens the risk of giving an high weight to publications with lots of estimates of low reliability and, conversely, giving a low weight to publications with few estimates but better statistical procedures. Hanushek shows that the publications with single estimates mainly use cross-sectional data, aggregate measures of school quality and multi state samples, with the drawbacks seen before.

Among the most recent studies on the effects of class size reductions, consider Strayer [2002], Woessman [2005], and Dustman, Rajah and van Soest [2003] for the UK school system. As in their previous work, Dustman, Rajah and van Soest [2003] use data from the National Child Development Survey. In general all UK last studies employ this dataset for its wealth of information on the individuals' educational achievements, family background and work histories, that make it unique, also in the US, and so allows to solve the problem of the school quality endogeneity. There is particular concern in the British political debate, centred around class size, about the low proportion of youths who continue to study after

[6]There is another procedure to summarize results of different studies known as meta-analysis (see Hanushek [2002] for more details). It consists in doing formal statistical tests of results, viewed as estimates of a common parameter. An example is provided by Hedges, Laine and Greenwald [1994, 1996] for the school quality literature—he concludes that school quality makes some difference.

the minimum school leaving age of 16. Hence Dustman, Rajah and van Soest evaluate the impact of class size on the decision to continue to study at the end of the compulsory school, finding that it has a statistical significant negative effect, as it is natural to expect. Their procedure consists of two phases and then is estimated the impact of the decision to stay at school on wages. The results suggest that staying at school positively affects wages, which implies that class size has a significant negative impact on earnings. This is in contradiction with the main findings of the previous literature, that estimates directly reduced form wage equations, more likely to give misleading conclusions for the authors. They also argue that school quality influences positively attainment through improved examinations performance and attachment to school of the student.

Instead Woessman [2005] finds no statistically significant effects of class size on achievement for 15 West European countries using data from the Third International Mathematics and Science Study (TIMSS). These data are such that Woessman can implement both the quasi-experimental strategies employed before by Angrist and Lavy [1999] and Hoxby [2000] (which exploit national maximum class size rules and natural fluctuations in cohort sizes[7], respectively; see above).

As regards Strayer [2002], we have seen in the previous section his innovative approach (similar to the 2-steps approach employed by Dustman, Rajah and van Soest [2003]) and the results with respect to teacher quality (the fraction of teachers with a graduate degree). He also considers class size as measure of high school quality. The estimation results suggest the existence of an indirect effect of high school class size on earnings through college choice, which in turn affects the following earnings (see Section 2.3.1 for details), i.e. attending an high school with smaller class sizes increases the probability of attending a 4-year college relative to a 2-year college.

2.3.3 Effect of school quality on student attainment

There are relatively few studies on the impact of school quality on school attainment. Anyway the results are less ambiguous than those for other outcomes and there is substantial agreement that better school quality improves educational attainment. Here I provide a brief overview of some studies.

Card and Krueger [1992a] are among the first ones to handle this issue. We have seen in Section 2.2 the principal features of their work, which uses a new approach to estimate the effect of school quality on earnings based on the comparison of wages of individuals who work in the same region but have attended school in different regions. The simple reduced form estimates of the effect of school quality on years of schooling suggest that better school quality increases school attainment.

[7]The TIMSS survey tests two adjacent grades in each sampled school.

In the study of 1996, Card and Krueger develop a theoretical model to explain the joint determination of educational attainment and earnings with varying levels of school quality. Their result highlight that better school quality leads to more educational attainment by increasing the premium associated with one additional year of schooling.

Both the studies by Dustman, Rajah and van Soest [1998, 2003], who consider the decision to take further education at the end of compulsory school, confirm the findings of Card and Krueger [1992a]. Another contribution from the UK is provided by Dearden, Ferri and Meghir [2002], who employ, like the other UK studies, the National Child Development Survey data. They consider the effect of pupil-to-teacher ratio at the end of both primary and secondary school, and the effect of type of secondary school (selective and non-selective). The outcome variables of interest, in addition to school attainment, are wages at ages 23 and 33. Only the secondary pupil-to-teacher ratio seems to have effect, but only on wages at age 33 of women, especially of low ability (in terms of prior achievement). This confirms the presence of an 'age effect' (revise Section 2.2), such that school quality influence begins to act when workers have gained some experience in the labour market. Instead the fact that lower ability women are particularly advantaged supports the findings of Lazear [2001], who shows that pupil-to-teacher ratio has a stronger effect on lower ability pupils not disruptive (young females in general are more mature than males of the same age). As regards school attainment, which we are interested in, it is influenced by the type of school for both males and females. School type affects wages, instead, solely for men at age 33. An interesting aspect of this study is the use of propensity score matching in this last case, to evaluate the effect of school type on male wages at age 33; this is the only example in the school quality literature. Matching results show that attending selective schools has a stronger effect in the group of students who actually do not attend selective schools ('treatment effect on the non-treated' in the language of the matching literature).

Another study looking at the effect of school quality on attainment is provided by Long [2005a], who emphasizes the existence of multiple dimensions of school quality, differently affecting school attainment. This issue has also been pointed out, for the first time, by Loeb and Bound [1996], who are in particular concerned for the correlation among school inputs. Using the National Education Longitudinal Study data, Long carries out a systematic study examining a large number of outcome variables (41) and school quality dimensions (72), these last ones measured at several grade levels as in the study of Dearden, Ferri and Meghir [2002]. In fact the same school input may have a different impact on different outcomes. It turns out that school resources have a stronger effect on attainment, test scores and family formation, than on labour market outcomes, being the workers still young (age effect). There is a group of school inputs that consistently bear positive effects, for example schools that encourage discipline and college attendance, private schools and smaller schools.

Finally, Brunello and Checchi [2003] attempt to evaluate the impact of school quality on attainment for a sample of Italians born between 1941 and 1970, using data from the Survey on the Income and Wealth of Italian Households. They also are interested in the evaluation of the effect of school quality on the returns to education and use a much more sophisticate version of the 2-steps model employed by Card and Krueger [1992a], by taking into account (i) the effects of endogenous migration (see Heckman, Layne-Farrar and Todd [1996] and Section 2.2), (ii) the fact that educational attainment is the result of an individual choice (they include predicted years of schooling in the wage equation, following Strayer [2002]; see Section 2.3.1) and (iii) the interaction between school quality and family background in the production function of human capital. Note that school quality is measured by pupil-to-teacher ratio and it is an average over all levels (kindergarten, primary school, low secondary school, etc.). The estimation results suggest that a lower pupil-to-teacher ratio is associated with higher educational attainment and bigger returns to education, particularly for individuals with worse family background, measured by parental education (indeed the effects of school quality on returns to education turn out to be negative for individuals of cohort and regions with higher parental education). Then the author demonstrate that school-quality and family background are technical substitutes in the production function of human capital.

2.4 College quality literature

The literature on college quality counts a lower number of contributions than the corresponding literature for schools, but results are more clear and concordant in indicating that the college an individual attends matters for future labour market performance and other outcomes.

Loury and Garman [1995] provide a survey of the early studies, which mainly (i) fail to control for other educational inputs affecting the outcome in addition to college quality and (ii) are difficult to generalize by considering highly unrepresentative samples of the population of college attendees. Note that these early works and almost all the later ones typically use the terms quality and selectivity interchangeably, indeed they measure school quality by the average Scholastic Aptitude Test (SAT) score of the incoming freshmen, that is the test employed in entrance examinations by all four-year colleges in the US[8]. Moreover most studies consider wages as outcome variable. Only recently the attention has been addressed to a wider array of quality attributes and outcomes, and now researchers agree on viewing college quality as a multi-faceted entity (cf. Long [2005b] and Black and

[8]The corresponding test for the UK is the A level score (see Chevalier and Conlon [2003]).

Smith [2006]). See Brand and Halaby [2003] and Black and Smith [2006] for an overview of the most recent contributions.

The standard approach in the literature to evaluate the impact of college quality is the estimation of a simple linear regression model where the logarithm of wages is regressed on the chosen college quality attribute and other control variables, expected to be all the variables related to wages that affect college choice, and hence college quality ('selection on observables assumption')—essentially controls for individual ability, family background and high school quality, in addition to labour market experience and years of schooling.

Consider, for instance, Daniel, Black and Smith [1997], who assert to be the first ones to use a dataset such as to allow (i) to rely on the selection on observables assumption and (ii) to carry out an evaluation on a representative sample of male college attendees. In fact Daniel, Black and Smith [1997] use a rich dataset from the National Longitudinal Survey of Youth, which also permits to observe a recent cohort of male students, unlikely the previous studies that examine old cohorts of men. Hence there is the possibility to assess the changes in the impact of college quality over time, likewise the school quality literature. This last one has demonstrated a diminishing effect of school quality over the last decades (see Section 2.2); on the contrary Behrman, Rosenzweig and Tauban [1996] argue that the returns to college quality could have been increasing. Daniel, Black and Smith [1997] find that several dimensions of college quality (tuition, expenditures per student, faculty to student ratio, rejection rate, average SAT score and faculty with Ph.D.) have a strong positive impact on wages, but while their effects are consistently significant and positive when the measures are inserted one at a time in the wage equation, some of them become insignificant and even negative when the measures are inserted jointly. Since these indicators are highly correlated, the authors decide to combine them into a unique index via factor analysis. Another result is that the payoff to attending an higher quality college (an 'elite' college, in the terminology of this literature) is larger for blacks than for whites, according to the previous findings of Loury and Garman [1995].

To my knowledge, there are few studies that attempt to adjust for selection on unobserved variables in evaluating the impact of college quality. Behrman, Rosenzweig and Tauban [1996] are probably the first ones. Having available a dataset with information on a sample of female twins, they are able to control for family and individual fixed effects by comparing twins who have attended different colleges, but have the same family and individual endowments. Further they control for observed attributes of home environment and prior schools; in fact they stress the importance of considering the quality of all levels of schooling (primary school, secondary school, college) attended, since school quality is likely to be correlated across levels (revise Section 2.3.1). They find that Ph.D. granting colleges, private universities with highly paid senior faculty and smaller enrolments reward positively students in terms of future earnings, but contrary to Altonji and Dunn [1996] (who, instead, consider siblings and high school quality; revise Section 2.2) these positive effects increase when family and individual fixed effects are not controlled for. Indeed they also find that individuals with higher endowments attend school longer and attend better colleges, with the result of upward bias in the standard estimates (which do not adequately control for human capital endowments) and exacerbation of starting inequalities (in human capital endowments).

Surely Dale and Krueger [2002] provide the most original contribution in this direction by the fact of having access to a unique dataset with information on the colleges which students apply to, and hence are accepted or rejected by, at the moment of college choice. They are able to control for unobservable variables that determine the selection process on the part of both (i) students and (ii) institutions by grouping students (i) who have applied to colleges with similar qualities—students with similar aspirations—and (ii) have been accepted and rejected by two sets of similar colleges (in terms of quality)—students with similar capacity. Therefore the authors compare the wages of students in the same group and with similar other observable characteristics. Clearly the students in the same group are equivalent with respect to the unobservables 'aspiration' and 'capacity', upon which depends the college selection process on the part of students (who choose colleges) and institutions (which admit some students), respectively. The findings are very interesting, because students who attend more selective colleges (colleges with an higher average SAT score) do not earn more than students who attend a less selective college but are in the same group. Moreover, and relatedly, the selectivity of the college a student has been rejected by has a stronger effect than the selectivity of the college actually attended. Anyway the average tuition cost of the college actually attended (rather than the average SAT score) seems to matter for wages.

Another study which attempts to take into account the selection on observables is provided by Brewer, Eide and Ehrenberg [1999], who exactly control for the fact that students choose the college they attend on the basis of the net cost they bear. First Brewer, Eide and Ehrenberg model the choice of the type of college, given by the combination of quality and control (six categories: public/private elite/middle/bottom selective), on the basis of the net cost (tuition minus financial aid) and other control variables (individual ability, family background, etc.), so as to obtain for each student an estimate of the probability to attend a specific type of college (six 'selectivity terms' for each student). Then separately for each college type the authors estimate the standard wage equation adding the selectivity term of the category considered among the regressors and they use the resulting estimated coefficients to calculate average wage differentials—the mean difference (in the whole sample) of predicted wage values of different categories (for example private elite colleges vs. public bottom colleges). Unlike previous studies, the authors also use longitudinal data, which allow them to observe how change the returns to college quality type across time for a given cohort and between different cohorts at some time point of the labour market experience. The estimation results suggest that there is a strong return to attending an elite private college instead of a bottom public college and this premium is larger for students who attend college more recently, consistently with the findings of Behrman, Rosenzweig and Tauban [1996].

Black and Smith [2004] also use an alternative approach to the standard Ordinary Least Squares (OLS) regression, but their concern arises from the assumption of linear conditioning on observables, rather than selection on observables in itself. In fact the linearity assumption could mask the failure of the "common support" condition, i.e. in correspondence of certain values of the conditioning variables there are not attendees of both high quality (elite) and low quality (non-elite) colleges. For example, if more able students attend only elite colleges, their counterfactual outcome, in the case of attendance of a non-elite college, is not non-parametrically identified, because it cannot be approximated by the observed outcome of equally able students attending actually non-elite colleges. We have, instead, to rely on a linear projection, given by the predicted value for the outcome variable from its linear relationship with ability estimated for students who attend non-elite colleges in correspondence of low values of ability (supposing, for simplicity, to consider only ability, in addition to college quality, among the regressors). Furthermore, leaving aside the problem of the common support, there is no theoretical reason for the linear dependence of the outcome variable on certain variables such as ability, considering wages (cf. Tobias [2003]).

Therefore, Black and Smith [2004] apply propensity score matching to estimate the impact of college quality on earnings, using data from the National Longitudinal Survey of Youth, as in their previous work seen above (these data are rich enough to make work the selection on observables assumption). This method consists in comparing the wages of elite and non-elite college attendees with similar values of the propensity score, which is the probability to attend an elite college conditionally on a set of observable variables, estimated through a flexible parametric logit model (the use of the propensity score is equivalent to consider all variables collectively, but allows to reduce the dimensionality problem). The first result the authors obtain is that the distribution of students into colleges of different quality (in terms of the average SAT score) is related to their ability (measured through the scores on the Armed Services Vocational Aptitude Battery), in that they find more able students who attend colleges of higher quality, but this sorting based on ability is disproportional since there are much more able students attending non-elite colleges than non-able students attending elite colleges. Hence the common support condition holds only weakly and there are few students who attend non-elite colleges where the propensity score is high, with the result to have very imprecise matching estimates, with very large standard errors. Anyway, the point matching estimates of the effect of college quality is quite similar to those obtained by OLS, especially for men, and overall they indicate that college quality matters.

In reality the first study to apply matching to estimate the effect of college quality is provided by Brand [2000], using data from the Wisconsin Longitudinal Study. The continuation is in following joint work with Halaby in 2003, which is also unique for looking at the effect of college quality (i) across all career (until age 53 approximately) (ii) not only on wages but also on employment and occupational status, and job satisfaction, as well as educational attainment. Differently from Balck and Smith [2004], Brand and Halaby find that there is no causal effect of college quality (measured by the Barron's categories[9]) on any outcome, estimating by OLS. Instead the matching estimates suggest that there is some benefit only early in the career, contrary to the findings of the literature related to school quality (see the 'age effect' in Section 2.2; see also Warren, Hauser and Sheridan [2002] for similar conclusions and a justification). Anyway, these insignificant results refer to the 'average treatment effect on the treated', i.e. the impact of college quality for the group of elite college attendees; considering the effect for a random sample of college attendees the results seem to indicate, on the contrary, a positive impact.

Both matching and the approach of Dale and Krueger [2002] are employed again by Long [2005b] on the National Education Longitudinal Study data. His goal, in fact, is to see how vary the estimates of the returns to college quality with alternative estimation strategies. In addition to matching and the Dale and Krueger method, he uses OLS and the Instrumental Variables (IV) approach, where the average quality of the colleges within a certain radius of the student's residence location is used as instrument for the endogenous college quality[10], following Card [1995], who uses instead college closeness as instrument for educational attainment in the estimation of the return to an additional year of schooling. This work is also interesting for considering several college quality dimensions and outcome variables besides selectivity and labour market outcomes, respectively. See, for example, Behrman, Rosenzweig and Tauban [1996] and Monks [2000], who consider colleges which offer Ph.D. degrees, or Bowen and Bok [1998], who considers the incidence of divorces and separations as outcome variable. Long, hence, emphasizes the importance of viewing college quality as a multidimensional variable with many other attributes than selectivity, with their own effect on each out-

[9] This categorical indicator combines information on several measures of college selectivity (admission test scores, percentage of applicants who are accepted, high school grade point average, etc.).

[10] The author expects that the average quality of the nearby colleges (i) is correlated with the quality of the college actually attended but (ii) it is not directly related to his/her earnings. Clearly this last condition could be difficult to hold.

come (see Long [2005a] for similar considerations with school quality; cf. Section 2.3.3). He obtains that all methods consistently estimate a positive effect of all college quality inputs on educational attainment. The results are instead mixed for the effects on earnings, probably because the young workers considered have just entered the labour market and they will be rewarded only successively (this conclusion is opposite to that of Brand and Halaby [2003] and Warren, Hauser and Sheridan [2002]; see above).

Another recent study looking at various college quality dimensions and outcome variables is provided by Black, Daniel and Smith [2005], who use data from the National Longitudinal Survey of Youth, as in their previous studies in 2004 and 1997, seen above. This dataset is so rich in individual information that the selection on observables assumption is likely to be satisfied. It contains, in fact, detailed information on family background, home environment and high school characteristics, as well as a measure of individual ability extracted from the Armed Services Vocational Aptitude Battery test, administered at the beginning of the college[11]. Moreover the detailed information on the individual employment histories allows to evaluate how change the labour market effects over time. Therefore the authors estimate a simple regression model, including also an interaction term between college quality and years of schooling in order to assess a 'dose-response effect' of college quality, rather than a simple 'signaling effect'. The estimation results suggest that there is an indirect effect of college quality through years of schooling, which does not interact with college quality, i.e. college quality is only a signal of individual skills and does not improve, actually, students' human capital. Then the effects of college quality on wages are weaker for women than for men, but they are stable over time for both sexes. There is also some evidence of non-linearity in the effects of college quality, i.e. there is a larger premium at high levels of quality[12]. Finally, it turns out that college quality affects other outcome variables than wages: it increases the probability of both completing college and achieving Ph.D.; it also increases spousal earnings (especially for women) and reduces the number of children (especially for men).

But the most original (in my opinion) contribution by Black and Smith is in 2006, when they review the previous college quality literature on the basis of two alternative ways of viewing college quality: as (i) a set of interconnected college inputs (dimensions; see, for example, Monks [2000] and Long [2005b] among my references) or (ii) a latent variable measured with error by a certain number of proxies, which is the implicit assumption of most of the literature. Furthermore almost all previous studies insert only one dimension in the log wage equation so as

[11] Clearly it is not affected by college learning.

[12] The authors have combined their measures of college quality into a single index and they have inserted in the regression model three dummies indicating between which quartiles it is located.

to obtain biased estimates of the parameter of interest with both interpretations. In the first approach the estimated parameter incorporates the effects of all the other dimensions not included in the log wage equation, with the result of upward bias if the college inputs are positively correlated (misspecification of the model due to the omission of relevant variables). Instead, in the second traditional approach the measurement error of the proxy used as indicator of college quality leads to a downward bias in the estimated value of the parameter of interest (this is the first standard result we encounter in the literature on measurement errors, when one explicative variable is badly measured in a regression model). In their paper Black and Smith take the second approach and show how to reduce, and even eliminate, the attenuation bias induced by the measurement error using more than one proxy. For instance, it is possible to form a quality index that combine the multiple quality measures via factor analysis (see also Black, Daniel and Smith [2005], Long [2005b] and Daniel, Black and Smith [1997] among my references), whose measurement error is lower than the corresponding one by considering the proxies individually. A second possibility is to employ the IV technique and using another proxy (or all the other proxies) as instrument for the measure inserted in the log wage equation. This strategy is equivalent to a factor analysis where wages (the outcome variable) become an additional indicator for college quality, but this is preferable because allows, without loss of generality, to normalize to one the variance of the latent quality variable and correct the additional bias arising from the fact that the proxies are defined on different scales. Therefore the authors apply these alternative methods to the National Longitudinal Survey of Youth data, using five inputs as measure of college quality: the faculty to student ratio, the rejection rate among appliers, the freshman retention rate, the mean faculty salary and the usual average SAT score of the incoming class. By considering five proxies, the covariance structure of the factor analysis model becomes over-identified, thus they estimate the model by the Generalized Method of Moment. They obtain strong evidence of a positive impact of college quality on wages, which indicates that one standard deviation increase in college quality is associated with an increase of 4.3 percentage points in wages. Anyway the authors point out that this result is preliminary, because the assumption of only one dimension for college quality is very strong, and much more work is required.

Finally, a recent original contribution is provided by Hoffmann and Oreopoulos [2006], who are the first ones to investigate the effects of teacher quality on student achievement in colleges[13] (this literature in instead wide for primary and secondary schools; cf. Section 2.3.1). They have access to a large rich dataset rel-

[13]Really also Borjas [2000], Bettiniger and Long [2004, 2005], and Ehrenberg and Zhang [2005] are interested in the relationship between teacher quality and college attendees' achievement, but they examine small samples and do not control adequately for students' background.

ative to a Canadian university, that allows to match students with their teachers. Specifically the authors consider the entering class, whose teachers for each course change year by year owing to scheduling problems or, simply, replacements. Furthermore the authors can control more successfully for the non-random selection on the part of students in the first year, because the names of teachers are unknown when students choose the course to attend, often in the summer period. Therefore they estimate the variability in a certain student outcome due to different teachers (teacher quality) by a fixed effects model, controlling for several observable individual characteristics, such as high school grade, within a specific type of course. They also replicate the analysis using a simple regression model, where they insert a variable relative to students' evaluation of teachers as measure of teacher quality, plus the course, the year and other control variables. There is some evidence of variation in student outcomes by instructors, within a given course: for instance, an increase of two standard deviations in teacher quality translates into a 1.5 percent decrease in the rate of dropping out the course. There is, instead, no evidence that the measure of teacher quality provided by students has a positive effect. Hence the main findings of the school quality literature (Hanushek, in the first place) are confirmed for colleges: teachers are very important in determining student achievement, but the differences among students in achievement are not related to standard measures of teacher quality (like students' evaluation, in this case).

As regards Italy, Brunello and Cappellari [2005] attempt to investigate the Italian university system, using data from the "Indagine statistica sull'inserimento professionale dei laureati" survey, which provides a wealth of information on a sample of individuals graduated in 1998, interviewed three years after graduation. The college quality information comes from another dataset, made available by the National Statistical Office for the academic year 1996/97. Note that the Italian university system is very little differentiated, compared with the US system, and the tuition fees are relatively low; furthermore few universities adopt entrance examinations, which are not comparable with each other. Therefore the authors consider the university location and control (public/private) plus other college observable characteristics, such as the student to teacher ratio, as measures of college quality. The outcome variables they consider are employment status and earnings three years after graduation. The empirical strategy proceeds in two steps and is similar to that used by Card and Krueger [1992a] for the US school system, because are the individuals who work in a region different from the region of the college they have graduated from that allow to identify the effects of the clusters 'province-per-faculty' on earnings (or employment) in the first stage. In the second stage these parameters are regressed on the set of faculty and province dummies plus a vector of observable university characteristics. The authors also use an average measure of the labour market performance at the province-per-faculty level as outcome variable in the second step, given by the logarithm of the average employment ratio multiplied by the average earnings in a given province-per-faculty cluster, which they call ‘employment weighted earnings’. On the whole, the estimation results suggest that the college an individual has graduated from matters for early labour market outcomes. More precisely, there is a strong effect of the location of the university (North-South divide) and the type of college, in fact attending a private university yields an increases of 18 percentage points in the employment weighted earnings relative to attending a public college. Besides, the student to teacher ratio is important; indeed a decrease of 10 percentage points in the student to teacher ratio translates into a 2.4 percent increase in employment weighted earnings.

Chapter 3

Data

The main source of data exploited throughout this thesis is the second wave of the *Program for International Student Assessment* (PISA 2003), whose main features are briefly described in Section 3.1 (the interested reader is referred to OECD [2005a] for more technical details). These data provide information on a sample of 11,565 Italian 15-year-old students from 382 high secondary schools (30 students per school, on average), tested in their mathematical competences in April 2003. The questionnaires completed by students and school heads at the moment of the test allow me to rely on a wealth of information both at the student and at the school level; see Section 3.2.1. In particular, the detailed information on student characteristics, family background and home environment is such that the conditioning on observable variables alone help solve the problem of non-random selection into schools of different qualities (cf. Section 4.1). As regards schools, I have access to several proxies for school quality. Anyway, since the threat of subjective perceptions and voluntary misreporting by school principals, especially with respect to equipment and personnel (cf. Checchi [2004]), I prefer to employ the more reliable data available from the administrative archive *Sistema Integrato Segreterie Scolastiche Italiane* (SISSI) at the school level; see Section 3.2.2. Descriptive statistics are reported in Section 3.3.

3.1 The design of the PISA 2003 survey

What PISA is

PISA is a 3-yearly international standardized survey promoted by the Organization for Economic Co-operation and Development (OECD) to assess 15-year-old students' knowledge and skills. It was carried out in all 30 OECD countries plus other 11 partner countries.[1] The age of 15 was chosen because in most tested countries students are near the end of compulsory schooling, and an objective of PISA is to compare the 'yield' of different compulsory educational systems. In this regard, the regularity of the evaluation enables the participating countries to monitor their changes in students' outcomes over time.

What PISA assess

The fields covered by the assessment are mathematics, reading and science, but it focuses on only one of them in each cycle. Specifically, the domain of interest in the second wave of PISA is mathematics, to which is devoted the 70 percent of the testing time. Unlikely the other cross-country comparative tests, PISA was designed to evaluate students' ability to use what they have learned at school in the resolution of every-day problems ('know-how'), rather than an end in itself acquisition of curricular subjects ('know-that'), variable from country to country (see OECD [2004] for more details about the broader concept of knowledge measured by PISA).

Who PISA is addressed to

As stated above, 15-year-old students, attending an educational institution in grades 7 or higher within the country, constitute the 'international' target population for each state. Hence, citizen students who attend a school abroad are excluded from the population and foreign students who are enrolled in a school located in the country are instead not. Precisely, in Italy, as in the major part of the participating countries, the period of assessment was April 2003 and only students aged from 15 years and 3 months (completed) to 16 years and 2 months (completed) enter the sample, therefore the reference population can be defined as students born during 1987, whatever is the grade attended. The choice of an age-based sample is another difference with the other international surveys, which use, by contrast, grade-based samples. This is mainly related to the purposes of the program, aimed at evaluating the 'yield' of the whole compulsory educational system.
The 'international' target population can differ from the 'national' target population, since some schools and students may be a priori excluded. In the first case, at the level of school, this can be due to practical or organizational reasons, for

[1] The OECD countries are: Australia, Austria, Belgium, Canada, Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Japan, Korea, Luxembourg, Mexico, Netherlands, New Zealand, Norway, Poland, Portugal, Slovak Republic, Spain, Sweden, Switzerland, Turkey, United Kingdom, United States. The partner countries are: Brazil, Hong Kong-China, Indonesia, Liechtenstein, Latvia, Macao-China, Russian Federation, Thailand, Tunisia, Uruguay, Serbia.

instance the inaccessible location of some schools or the difficulty to administer the assessment in other schools. In the second case, at the within-school level, the exclusion can involve disabled students or foreign students with a poor knowledge of the national language.

Sample selection

As regards the sampling procedure, a two-stage stratified sample of the population of interest is investigated. The first stage units are schools with 15-year-old students (students with an age between 15 years and 3 months, and 16 years and 2 months, hereafter, for Italy). They are sampled with a probability proportional to their size, to be intended as a function of the estimated number of 15-year-old students enrolled in the school at the moment of the test. The overall school sampling frame (the list of schools entitled to be sampled) is divided into a series of strata and the sampling takes place separately in each of them, also divided into other implicit strata (see further). At the second stage, 35 students within each sampled school are selected with equal probability; all the 15-year-old students enrolled in the school enter the sample if their number is lower than 35. Exactly, for each school the measure of size (MOS, in the terminology of PISA) is equal to the maximum between the 15-year-olds enrolment (ENR) and the prefixed target cluster size of 35. Consequently, there is a simple random sampling of small schools (schools with ENR$\leq$35, i.e. MOS=35).
The explicit stratification produces one separate list of schools for each combination of stratification variables, where different sampling schemes may be adopted. As a general rule, the number of schools to be drawn in every single stratum is such that the proportion of eligible students—not schools—in the sample is equal to the corresponding proportion in the population. However, a country may decide to implement a disproportionate allocation of schools over the strata and extract, for example, the same number of schools in each stratum regardless its relative size (in terms of ENR), in order to ensure the representativeness of all groups of the population and obtain sufficiently reliable estimates in every one of them, thanks to an adequate sample size in small clusters as well (consider, for instance, small but economically influential provinces or regions). This disproportional allocation is also the case of Italy, whose explicit stratification variables are school track (lyceum, technical school and professional school) and geographical location, exactly five macro-areas (North-West, North-East, Centre, South and South-Islands[2]) and four *regioni* (Lombardia, Piemonte, Toscana and Veneto) plus

[2]North-West comprises Liguria, Lombardia, Piemonte and Valle d'Aosta; Nort-East comprises Emilia Romagna, Friuli Venezia Giulia, Veneto and Trentino Alto Adige; Centre comprises Lazio, Marche, Toscana and Umbria; South comprises Abruzzo, Campania, Molise and Puglia; finally South-Islands comprises Basilicata, Calabria, Sardegna and Sicilia. This composition of the

two *province autonome* (Bolzano and Trento), which take part in the program with representative samples of their own territory.
Instead, the implicit stratification simply sorts schools within each explicit stratum according to the values of another set of variables, thus the allocation of schools across implicit strata is proportional. Within the cells resulting from the double stratification, schools are also ordered by ENR. The reason is to assign two substitute schools to each school in the sample frame, which have to be replaced in the case of no participation to the program. Precisely, the replacement schools are the ones immediately preceding and following the reference school in the list: the schools in the same implicit strata with ENR-1 and ENR+1, respectively. This choice ensures that the substitute schools are similar to the reference school in terms of size and implicit stratification variables. These last ones, as well as the explicit stratification variables, must be highly correlated with students' performance in the test in order to improve the reliability of the survey estimates. We have only one implicit stratification variable for Italy, that is school type, i.e. being the school public or private.
With reference to the selection of schools within each explicit stratum, it follows a specific systematic procedure, which allows to minimize the difference between the sum of school weights (in the first stage) and the actual number of schools in the population for the stratum considered, in combination with the probability proportional to size school sampling scheme. This last one prevents the unfortunate cases of only small or large (in terms of MOS) schools sampled, so as to have equal final student weights whose sum exactly estimates student population size. I omit a description of the systematic selection procedure, to which I refer in OECD [2005a]; see OECD [2005b] for the underlying idea.
Nevertheless, final student weights can be highly variable, as a consequence of the complex sample design. One explanation is the over-sampling or under-sampling of some strata of the population for the reasons seen before. In addition, there is the fact that PISA administers use an estimate of the actual number of 15-year-old students at the time of the text when they construct MOS and sample schools, in the first stage. In the most fortunate cases they have access to the information relative to the previous academic year, concerning quantities such as the ratio between the total enrolment and the number of grades in the school or the enrolment in the modal grade for 15-year-olds. As a result, this approximation leads to an inconsistency between the a priori and the final student weights. Finally, school and within-school adjustments for school and student non-response are another source of variability in final student weights. Regrettably, Italy is one of the countries characterized by the most variable sampling weighs, with Canada and United Kingdom (see OECD [2005b]). For instance, the unweighted mean

macro-areas is the same one used in other international and national surveys.

of the mathematical score is about 7 percentage points higher than the weighted mean for Italy.
Therefore, the PISA manual (OECD [2005b]) recommends to not use the standard methods of inference, adopted under the assumption of simple random sampling, with the PISA data and weight observations to achieve unbiased estimates of any quantity of interest. Of the same mind are Bratti, Checchi and Filippin [2007], with regard to Italy, who also follow the procedure provided by the manual to estimate standard errors. Indeed the sampling design is so complex that does not exist, or it is too difficult to obtain, an explicit formula for the sampling variance of even the most simple statistics, such as mean, and it is necessary to resort to replication methods. The manual suggests the employ the Fay's variant to the Balanced Repeated Replication method. Roughly speaking, it consists in generating alternative samples, with different sampling weights (provided by the PISA 2003 dataset for each replication), of the original students' sample and comparing the (weighted) estimates of the parameter of interest in each replication with the original one in order to attain an unbiased estimate of the sampling variance of the statistic in question.
However, in my following analysis, I will not use sampling weights, because there is no need to use them when the researcher takes a structural approach and the stratification which occurs in the sampling procedure is not on the outcome variable—students' achievement in our case (see Cameron and Trivedi [2005] for a convincing justification). Bratti, Checchi and Filippin [2007] employ the sampling weights just because their *paper has descriptive aim, since it estimates statistical association without giving any causal interpretation.*

The PISA score

Another issue to deal with are the 'plausible values'. Indeed, the PISA score for the specific domain assessed is not simply the share of correct answers, but a set of five numbers, called plausible values. These are five random draws from an ability distribution estimated for students with similar item response patterns and backgrounds by the modern techniques of Item Response Theory (IRT)[3]. See, for

[3]At the basis of this approach there is the Rash model, which uses the number of correct answers and the difficulty of the items administered to a particular student to estimate his or her ability, defined on the same scale of the item difficulty. In short, the first step is to assign a level of difficulty on a continuum scale to each item of the test. This is simply obtained considering the distribution of successful students in each item, so as to get a relative scale, and fixing two anchor points in this scale (usually centring the distribution on zero), so as to move to an absolute scale. Then a logistic function of student ability and item difficulty allows to calculate the probability to be successful in a specific item for a specific level of ability. The probability to observe a specific item pattern is the product of the success probability of each single item, assuming independence of the answers to different questions. By replicating this computation

example, Hambleton and Swaminathan [1989] for a general treatment of this subject; I refer to Adams, Wilson and Wang [1997] for the model used to scale the PISA data and to OECD [2005a] for its application to PISA and the computation of the plausible values. Further, each plausible value has been standardized in a subsequent step so that its sample mean and standard deviation, across all countries, equal 500 and 100, respectively.
The matter with the plausible values is that the PISA manuals (OECD [2005a,b]) recommend to not use their mean in statistical analyses, and to use, instead, the mean of the five estimates of the parameter of interest obtained with each plausible value. Consequently, two components constitutes the sampling variance of a parameter estimator: the mean of the five sampling variances and the variance of the five estimated parameters for each plausible value (see, also, Bratti, Checchi and Filippin [2007]). Anyway, almost all researchers analysing PISA simply compute the mean of the plausible values and so I do. Indeed, my aim is to measure the knowledge and skills of every single student and I have to take the average value to minimize the measurement error associated with each individual's estimate.

Method of assessment

As regards the assessment, it consists of a mixture of multiple-choice items and free answer questions, organized in groups referring to a passage describing a real-life situation. The whole set of items lasts about seven hours, but different combinations of them are administered across students, who are engaged for about two hours with the test. The modern IRT techniques allow to obtain fully comparable student's ability estimates, even if students are assessed with different subsets of items—it is only sufficient that there are students assessed with common items. Also the fact that only a subset of students is evaluated in the minor domains is not a problem for the assignment of plausible values in the specific not tested domain to those students without questions on it (see OECD [2005a]). Anyway, I am interested in mathematical literacy, and all different combinations of items have questions on the major domain.

Finally, students have to answer a background questionnaire providing information about themselves, their home and their school. At the same time, school principals have to complete a questionnaire about the school. I describe this information in the next section, followed by some descriptive statistics for the variables actually exploited in the analysis in Section 3.3.

for each level of ability, it is possible to obtain the distribution of the probability to observe the specific item response across all levels of ability. This distribution turns out to be centred on a specific ability value, which also is the most likely ability value, that is chosen as the estimate of the ability of any students with the specific response pattern observed.

3.2 Data

3.2.1 PISA

Individual information

It stems from the student questionnaire, which mainly provides information on home environment (family structure, economic and cultural family background, parents' involvement in child's education), school environment (students' opinions about the school and teachers), and specific students' characteristics, such as their motivation, attitudes to learning and self-confidence, specifically in relation to mathematics. Students have also to fill in other two questionnaires regarding their educational career and acquaintance with Information and Communication Technology. Several indices at the student level have been obtained from the questionnaires combining the answers to different items, through the IRT techniques. The interested reader can refer to the OECD PISA manual [2005b] for the complete list of variables, the questionnaires and an explanation on how indices are derived. Here I only describe the variables which I use in the empirical analysis. Table 3.2 displays some summary statistics.

Mathematical achievement

PISA score	Mean of the five plausible values in mathematics
Teacher's mark	Teacher's appraisal in mathematics of the last school report in January 2003

Demographic and family background

Female	Dummy equal to 1 if student is female; 0 otherwise
Age	Student's age
Single parent	Dummy equal to 1 if student lives with only one parent; 0 otherwise
Parents' occupation	Index of the highest occupational level of parents: higher values of this index indicates higher levels of occupation
Parents' education	Highest level of educational attainment of parents in years of schooling

Home possessions	Index of home possessions: positive values of this index indicate higher levels of home possessions
Computer facilities at home	Index of computer facilities at home: positive values of this index indicate higher levels of computer facilities at home
No. of books at home	Number of books at home
Dialect at home	Dummy equal to 1 if student speaks dialect at home; 0 otherwise
Educational career	
Kindergarten attendance	Dummy equal to 1 if student attended kindergarten for *more than one* year; 0 otherwise
Grade repetition	Dummy equal to 1 if student failed *exactly one time* during his/her scholastic career; 0 otherwise. This variable is equivalent to a dummy indicating if student started primary school one year in advance by focusing on students attending grade 10
Loss of months	Loss of two or more consecutive months during primary/low secondary school
Change of school	Change of school during primary/low secondary school
University degree expected	Dummy equal to 1 if student expects to complete a university degree; 0 otherwise
Never latecomer	Dummy equal to 1 if student never arrives late at school; 0 otherwise
Strategies of learning	
Memorisation learning	Index of memorisation/rehearsal learning strategies for mathematics: positive values of this index indicate preferences for this learning strategy

Elaboration learning	Index of elaboration learning strategies for mathematics: positive values of this index indicate preferences for this learning strategy
Competitive learning	Index of competitive learning for mathematics: positive values of this index indicate preferences for this learning strategy
Cooperative learning	Index of cooperative learning for mathematics: positive values of this index indicate preferences for this learning strategy

School information

The questionnaire completed by school heads provides information about the type of school and its composition, the quality of human and material resources, and decision-making processes. I refer again to the OECD [2005b] manual for the description of the whole available information. I use only a small subset of variables related to school quality, because they are highly unreliable stemming from the inaccurate and misleading statements of school principals. Table 3.2 reports the corresponding descriptive statistics. Also from the school questionnaire have been derived indices describing specific aspects of schools, involving responses to multiple questions. Some variables (class size and peer's characteristics) are the school average of individual variables.

School quality

School and class size

Total enrolment	Total number of students
Teacher-student ratio	Number of teachers per student; the number of part-time teachers contributes 0.5 to the total number of teachers
Maths teacher-student ratio	Number of maths teachers per student; the number of part-time teachers contributes 0.5 to the total number of teachers
Class size	School mean of the number of students during maths lessons stated by each student

Teachers' quality	
Prop. of full-time teachers	Proportion of full-time teachers
Prop. of certified teachers	Proportion of fully certified teachers; the number of part-time teachers contributes 0.5 to the total number of teachers
Prop. of maths teachers with a major in maths	Proportion of maths teachers with a major in mathematics; the number of part-time teachers contributes 0.5 to the total number of teachers
Computer resources	
Prop. of pcs connected to www	Proportion of computers connected to Internet
Prop. of pcs connected to a lan	Proportion of computers connected to a local network
Pc-student ratio	Number of computers per student
Pc use	Frequency of computer use at school
School selectivity	
Prop. of non-repeating students	Proportion of students who are not repeating the year (they could come from other schools)
Prop. of students who have changed school	Proportion of students who have changed high secondary school (if high, it corresponds to a 'remedial' school)
School climate	
Student-teacher relation	School mean of the index of good student-teacher relation: positive values of this index indicate students' perception of good student-teacher relations
Teachers' support	School mean of the index of teachers' support in mathematics lessons: positive values of this index indicate students' perception of higher levels of teachers' support

Disciplinary climate	School mean of the index of good disciplinary climate in maths lessons: positive values of this index indicate students' perception of a positive disciplinary climate

Additional school information

School location

Small town	Dummy equal to 1 if school is located in a small town (fewer than 15,000 people); 0 otherwise
City	Dummy equal to 1 if school is located in a city (over 100,000 people); 0 otherwise

Peer's characteristics

Mean parents' occupation	School mean of the index of the highest occupational level of parents: higher values of this index indicates higher levels of occupation
Mean parents' education	School mean of the highest level of educational attainment of parents in years of schooling
Mean no. of books at home	School mean of the number of books at home

3.2.2 SISSI

Since the subjective perceptions of school heads could be affected by reporting errors, the Italian Ministry of Public Education (MIUR) has merged the PISA dataset with the more reliable information coming from the administrative archive SISSI at *school level*. SISSI is a software distributed to all state schools in Italy by MIUR in the 1998/99 academic year. Its applications cover all operations of school administration and gathers information on students, personnel, book-keeping, stocktaking, text-books and library management. The additional data attached by MIUR regards teachers, the type of school and its composition. They also allow me to collocate the school at the level of *provincia*, where *province* constitute a subdivision of the Italian territory at an intermediate level between *comuni* (lowest level) and *regioni* (highest level). Bratti, Checchi and Filippin [2007], who use the same dataset exploited in this analysis, provide a discussion about the extra information at our disposal. Bear in mind that, both for PISA and for SISSI, the information on school quality refers to the 2002/03 academic year. Summary statistics of the MIUR variables are reported in Table 3.2.

	SCHOOL QUALITY
School and class size	
Total enrolment	Total number of students
Teacher-student ratio	Number of teachers per student
Class size	Total enrolment divided by total number of classes
Class size grade 10	Enrolment divided by number of classes for grade 10 (I will focus on grade 10 attendees in the empirical analysis)
School selectivity	
Prop. of successful freshmen	Proportion of students who do not repeat grade 9 (the first class of high secondary schools in Italy)
Freshman retention rate	Proportion of students who do not drop out of grade 9 (the first class of high secondary schools in Italy)
Teachers' quality	
Prop. of permanent teachers	Proportion of teachers with permanent contract
Mean permanent teachers' seniority	Average seniority of teachers with permanent contract
Mean teachers' age	Average age of teachers

Additional school information

Type of school

Classic lyceum	Dummy variable equal to 1 if school is a classic lyceum[4]; 0 otherwise
Scientific lyceum	Dummy variable equal to 1 if school is a scientific lyceum; 0 otherwise
Technical school	Dummy variable equal to 1 if school is a technical school; 0 otherwise
Vocational school	Dummy variable equal to 1 if school is a vocational school; 0 otherwise
Private	Dummy equal to 1 if school is private; 0 otherwise

Geographical macroarea

North-east	Dummy equal to 1 if school is located in the northeast of the country; 0 otherwise. Northeast corresponds to regions (*regioni*) Emilia-Romagna, Friuli-Venezia Giulia, Veneto and Trentino-Alto Adige.
North-west	Dummy equal to 1 if school is located in the northwest of the country; 0 otherwise. Northwest corresponds to regions (*regioni*) Liguria, Lombardia, Piemonte and Valle d'Aosta.
Centre	Dummy equal to 1 if school is located in the centre of the country; 0 otherwise. Centre corresponds to regions (*regioni*) Lazio, Marche, Toscana and Umbria.
South and Islands	Dummy equal to 1 if school is located in the south of the country or in islands; 0 otherwise. South and islands correspond to regions (*regioni*) Abruzzo, Calabria, Campania, Puglia, Sardegna and Sicilia.

[4]There are mainly three types of high secondary schools in Italy: *licei*, technical schools and vocational schools. *Licei* prepare students for university and divide into classic *lieci* and scientific *licei*, which provide education focused on humanities and sciences, respectively. Technical schools give more specialized education relating to technical subjects, but their students are also entitled to enrol in universities. While both *licei* and technical schools last five years, vocational schools last only three years. These last ones are job oriented educational institutions, which prepare students for specific professions. They are usually recognized to be lower ability schools; however, also students from vocational schools can apply to colleges if they attend two integrative years.

As said before, I have information on the province (*provincia*) which the school belongs to. Hence, province dummies will also be included among control variables in the subsequent analysis. There are sampled schools in 81 of the 103 overall provinces.[5] See Table 3.2 for the complete list of provinces and the distribution of students across them. Note that the sampling procedure leads to a disproportional distribution of students across geographical regions: for example, only a small percentage of students, equal to 16.51 points, attends school in South and Islands, counter to a share of 36.33 percentage points for North-east.

3.3 Summary statistics

Since, for Italy, the target population of the survey, conducted in April 2003, are students born in 1987, sampled students should be attending grade 10—that is the second class of high secondary schools in Italy—at the moment of the test, if they have never failed and they have started primary school at the compulsory starting age of six. Therefore, hereafter, I consider only the 9,562 regular students enrolled in grade 10, who correspond to the 82.15 percent of the whole sample; see Table 3.1. Clearly, there could be students who started primary school one year in advance and had to repeat exactly one grade (0.5%) among regular students.

Table 3.1 - Distribution of 15-year-old students across grades

Grade	Freq.	Percent
7	11	0.09
8	63	0.54
9	1,775	15.25
10	9,562	82.15
11	228	1.96
Total	11,639	100.00

In Table 3.2 I report some descriptive statistics for the variables described in the previous section. They refer to the sample of interest of grade 10 attendees.

[5]No school belongs to Ascoli Piceno, Avellino, Benevento, Biella, Campobasso, Catanzaro, Chieti, Enna, Forlì, Isernia, Macerata, Matera, Pescara, Piacenza, Potenza, Prato, Reggio Emilia, Rieti, Rimini, Siracusa, Teramo and Terni.

Table 3.2 - Summary statistics

Variable	**Mean**	**Std. Dev.**	**Min.**	**Max.**	**Obs.**
		Individual information – PISA			
PISA score	507.855	85.745	133.931	771.179	9562
Teacher's mark	6.19	1.457	1	10	9391
Female	0.544	0.498	0	1	9562
Age	15.714	0.284	15.25	16.25	9562
Single parent	0.137	0.344	0	1	9447
Parents' occupation	48.259	16.308	16	90	9403
Parents' education	12.829	3.427	0	17	9512
Home possessions	0.067	0.899	-3.787	1.94	9548
Computer facilities at home	0.005	0.888	-1.676	1.051	9534
No. of books at home	184.698	216.172	5	750	9445
Dialect at home	0.227	0.419	0	1	9017
Kindergarten attendance	0.886	0.318	0	1	9525
Grade repetition	0.005	0.071	0	1	9380
Loss of months	0.088	0.283	0	1	9482
Change of school	0.111	0.314	0	1	9435
University degree expected	0.517	0.5	0	1	9554
Never latecomer	0.622	0.485	0	1	9499
Memorisation learning	-0.026	0.851	-3.483	3.292	9518
Elaboration learning	-0.088	0.949	-3.262	3.263	9517
Competitive learning	-0.08	0.958	-2.844	2.45	9504
Cooperative learning	0.082	0.968	-3.134	2.741	9504
		School quality information – PISA			
Total enrolment	627.131	336.235	100	1681	9173
Teacher-student ratio	0.119	0.034	0.028	0.267	9039
Maths teacher-student ratio	0.013	0.004	0.004	0.029	8658
Class size	19.97	3.099	12.182	27.118	9409
Prop. of full time teachers	13.329	15.969	0.383	120	8299
Prop. of certified teachers	0.915	0.135	0.16	1	9186
Prop. of maths teachers with a major in maths	0.745	0.238	0.167	1	8849
Prop. of pcs connected to www	0.733	0.314	0.012	1	9300
Prop. of pcs connected to a lan	0.526	0.41	0	1	9328
Pc-student ratio	0.155	0.095	0.027	0.518	9124
Pc use	13.02	6.269	0.2	27.273	9393
Prop. of non-repeating students	0.919	0.061	0.700	1	9046
Prop. of students who have changed school	0.04	0.045	0	0.227	9472
Student-teacher relation	-0.344	0.321	-1.042	0.486	9390
Teachers' support	-0.264	0.401	-1.343	0.799	9390
Disciplinary climate	-0.067	0.457	-1.139	1.093	9394

Continued on next page...

... Table 3.2 continued

Variable	Mean	Std. Dev.	Min.	Max.	Obs.
School quality information – SISSI					
Total enrolment	646.494	356.013	59	1705	9301
Teacher-student ratio	0.1	0.035	0.007	0.183	7835
Class size	21.047	2.405	13.2	27	9284
Class size grade 10	21.336	2.843	9.5	27	9229
Prop. of successful freshmen	0.836	0.108	0.474	1	7842
Freshman retention rate	0.961	0.05	0.645	1	7838
Prop. of permanent teachers	0.779	0.28	0	1	7967
Mean permanent teachers' seniority	8.872	2.475	1	13.867	7131
Mean teachers' age	46.718	3.7	35.75	53.2	7823
Additional school information – PISA					
Small town	0.184	0.388	0	1	9534
City	0.269	0.444	0	1	9534
Mean parents' occupation	48.103	7.776	29.176	70.469	9562
Mean parents' education	12.856	1.467	8.793	16.406	9562
Mean no. of books at home	184.176	92.897	26.304	537.5	9562
Additional school information – SISSI					
Classic lyceum	0.098	0.298	0	1	9432
Scientific lyceum	0.216	0.412	0	1	9432
Technical school	0.371	0.483	0	1	9432
Vocational school	0.176	0.381	0	1	9432
Private	0.042	0.201	0	1	9532
North-east	0.363	0.481	0	1	9562
North-west	0.289	0.453	0	1	9562
Centre	0.182	0.386	0	1	9562
South and Islands	0.165	0.371	0	1	9562
Province dummies					
Agrigento	0.005	0.07	0	1	9562
Alessandria	0.017	0.128	0	1	9562
Ancona	0.006	0.077	0	1	9562
Aosta	0.003	0.052	0	1	9562
L'Aquila	0.002	0.04	0	1	9562
Arezzo	0.012	0.108	0	1	9562
Asti	0.011	0.106	0	1	9562
Bari	0.02	0.141	0	1	9562
Bergamo	0.014	0.116	0	1	9562
Belluno	0.008	0.091	0	1	9562
Bologna	0.005	0.071	0	1	9562
Brindisi	0.006	0.077	0	1	9562

Continued on next page...

... Table 3.2 continued

Variable	**Mean**	**Std. Dev.**	**Min.**	**Max.**	**Obs.**
Brescia	0.018	0.133	0	1	9562
Bolzano	0.103	0.304	0	1	9562
Cagliari	0.002	0.048	0	1	9562
Caserta	0.01	0.102	0	1	9562
Caltanisetta	0.005	0.068	0	1	9562
Cuneo	0.02	0.141	0	1	9562
Como	0.004	0.06	0	1	9562
Cremona	0.004	0.06	0	1	9562
Cosenza	0.015	0.121	0	1	9562
Catania	0.002	0.047	0	1	9562
Ferrara	0.003	0.051	0	1	9562
Foggia	0.008	0.091	0	1	9562
Firenze	0.033	0.18	0	1	9562
Frosinone	0.008	0.092	0	1	9562
Genova	0.008	0.09	0	1	9562
Gorizia	0.002	0.049	0	1	9562
Grosseto	0.009	0.097	0	1	9562
Imperia	0.002	0.047	0	1	9562
Crotone	0.002	0.045	0	1	9562
Lecco	0.007	0.082	0	1	9562
Lecce	0.014	0.119	0	1	9562
Livorno	0.017	0.131	0	1	9562
Lodi	0.006	0.078	0	1	9562
Latina	0.006	0.074	0	1	9562
Lucca	0.008	0.091	0	1	9562
Messina	0.006	0.076	0	1	9562
Milano	0.047	0.212	0	1	9562
Mantova	0.006	0.079	0	1	9562
Modena	0.003	0.052	0	1	9562
Massa	0.012	0.111	0	1	9562
Napoli	0.022	0.147	0	1	9562
Novara	0.019	0.138	0	1	9562
Nuoro	0.003	0.058	0	1	9562
Oristano	0.003	0.054	0	1	9562
Palermo	0.01	0.1	0	1	9562
Padova	0.023	0.149	0	1	9562
Perugia	0.006	0.077	0	1	9562
Pisa	0.019	0.137	0	1	9562
Pordenone	0.003	0.058	0	1	9562
Parma	0.003	0.058	0	1	9562
Pistoia	0.01	0.1	0	1	9562
Pesaro-Urbino	0.005	0.072	0	1	9562
Pavia	0.005	0.069	0	1	9562
Ravenna	0.002	0.042	0	1	9562
Reggio Calabria	0.002	0.042	0	1	9562
Ragusa	0.003	0.056	0	1	9562

Continued on next page...

... Table 3.2 continued

Variable	**Mean**	**Std. Dev.**	**Min.**	**Max.**	**Obs.**
Roma	0.021	0.144	0	1	9562
Rovigo	0.014	0.118	0	1	9562
Salerno	0.006	0.079	0	1	9562
Siena	0.006	0.076	0	1	9562
Sondrio	0.012	0.11	0	1	9562
La Spezia	0.002	0.04	0	1	9562
Sassari	0.01	0.099	0	1	9562
Savona	0.004	0.067	0	1	9562
Taranto	0.001	0.037	0	1	9562
Trento	0.093	0.291	0	1	9562
Torino	0.058	0.233	0	1	9562
Trapani	0.003	0.059	0	1	9562
Trieste	0.003	0.051	0	1	9562
Treviso	0.016	0.125	0	1	9562
Udine	0.004	0.067	0	1	9562
Varese	0.017	0.128	0	1	9562
Verbania	0.006	0.079	0	1	9562
Vercelli	0.003	0.053	0	1	9562
Venezia	0.041	0.198	0	1	9562
Vicenza	0.021	0.142	0	1	9562
Verona	0.014	0.118	0	1	9562
Viterbo	0.002	0.046	0	1	9562
Vibo Valentia	0.003	0.053	0	1	9562

Chapter 4

Unidimensional school quality

This chapter is organized as follows. The next section places the investigation in the specific econometric framework of program evaluation, giving a solution to the first problem of nonrandom allocation of students across schools of different quality. Section 4.2 discusses the other two econometric issues related to achievement and school quality nonobservability, solved separately and jointly in Section 4.3. Finally a solution to the time-window problem is suggested in Section 4.4, followed by the empirical findings in Section 4.5.

4.1 The basic framework

The economic problem I deal with in this dissertation fits nicely the usual set-up of program evaluation. Let $Y_0^\star$ and $Y_t^\star$ denote student achievement in mathematics in September (0) and the generic month (t) of the 2002/03 academic year, respectively. The outcome variable is $\Delta Y_t^\star = Y_t^\star - Y_0^\star$, i.e. the variation of student achievement from the beginning of the academic year to the generic month t of the same academic year.[1] The treatment variable is school quality, $Q_0^\star$, relative to the 2002/03 academic year (hence fixed at 0), which I assume unidimensional in the first instance, for simplicity. Note that $Q_0^\star$ is continuous, differently from the traditional studies of evaluation, where the treatment is binary.

The aim of my exercise is to estimate the causal effect of $Q_0^\star$ on the change in student achievement $\Delta Y_t^\star$. Define $\Delta Y_t^\star(q_0^\star)$ as the potential outcome corresponding to the level of quality $q_0^\star$ in the support $\mathcal{Q}$ of $Q_0^\star$, that is the variation of achievement which I would have observed if students had attended schools with quality $q_0^\star$. To evaluate the effect of school quality I would like to be able to compare the observed

[1] Of course, I am interested in the gain in student achievement over the entire academic year, that is from September to June. However, the reason for using this notation will be made clear in the next section, and this is mainly related to data availability issues.

outcome with all the potential outcomes in the counterfactual settings, different from the observed one, for all students. Unfortunately, there is a fundamental problem of missing data and only one state is observable for each individual (this is the "fundamental problem of causal inference", Holland [1986]), so that I can compare outcomes produced by different levels of quality solely taking different students.

The main problem is that students are nonrandomly distributed across different types of schools. For instance, the most able students are more likely to attend schools characterized by high quality, or parents more interested in children' education place them into better schools (at least accordingly to their perception). But students' ability and parents' behaviour towards sons are other two important determinants, in addition to school quality, of students' acquisition of knowledge. Therefore, the difference in the average outcome between students attending different schools is the effect not only of different school qualities, but also of different abilities and family backgrounds. In a few words, the impact of school quality on achievement estimated through the comparison of students attending schools with different levels of quality is contaminated by the effect of all the factors that influence school choice (and hence school quality) and achievement simultaneously.

My strategy to deal with this endogeneity issue is to control for the variables, relevant to learning, that are likely to drive school choice. Let X_0 be the variables that affect learning and school selection jointly; they determine the assignment process before the treatment occurs, so they are pre-determined with respect to $Q_0^\star$—they are not affected by the treatment—and we can consider them fixed at 0. My strategy relies on what Heckman and Robb [1985] call "selection on observables assumption". It asserts that conditionally on X_0 the assignment process of students into schools of different quality is random:

$$\Delta Y_t^\star(q_0^\star) \perp Q_0^\star | X_0 \text{ for all } q_0^\star \in \mathcal{Q}. \tag{4.1}$$

This is to say that students with the same X_0, but different values of $Q_0^\star$ (attending schools characterized by different levels of quality), would have behaved equivalently, in terms of $\Delta Y_t^\star$, if they had attended the same type of school (schools with the same quality $q_0^\star$, whatever is $q_0^\star$ in $\mathcal{Q}$).

Therefore, the quantity of interest can be written as the following partial derivative:

$$\frac{\partial \mathrm{E}(\Delta Y_t^\star | Q_0^\star, X_0)}{\partial Q_0^\star},$$

which indicates the effect of an increase of school quality on the gain in student achievement (the marginal change in $\mathrm{E}(\Delta Y_t^\star | Q_0^\star, X_0)$ when $Q_0^\star$ is increased by a small amount, more exactly), holding X_0 constant. This parameter corresponds to the well-known "average treatment effect" conditional on the observable variables that satisfy the "selection on observables assumption" in the simplest case of binary treatments (see Heckman, Lalonde and Smith [1999]).

The main advantage of the value-added formulation of the model, where I consider the change in student achievement $Y_t^\star - Y_0^\star$ rather than the level $Y_t^\star$ (see Hanushek [1979] for a discussion about value-added models), is clear-cut. These are the advantages of the difference-in-differences matching estimators over the cross-section matching estimators of program evaluation (see Smith and Todd [2005]). Suppose I had considered the specification on the levels, then (4.1) would have been replaced by $Y_t^\star(q_0^\star) \perp Q_0^\star | X_0$, which is a stronger assumption, since it requires the observation of a greater set of variables: in addition, all the variables affecting $Q_0^\star$ with the same impact on $Y_0^\star$ and $Y_t^\star$. In other words, the value-added approach allows for time-invariant differences in the variable $Y^\star$ among students who attend different types of school (such as student's "fixed effects"): the sorting of students into schools of different quality can arise from individual time-invariant unobservables that simplify each other taking the difference $Y_t^\star - Y_0^\star$. It is no more necessary to control for the past educational inputs that affect the initial level of achievement, at the beginning of the academic year, but have no impact on the value-added during the period considered—how much students get worse or improve from September to the generic month t.

I frame the approach described above into a fully parametric set-up. More precisely, I assume the following basic linear model for the gain in student achievement:

$$Y_t^\star - Y_0^\star = \beta_t' X_0 + \gamma_t Q_0^\star + \epsilon_t, \tag{4.2}$$

where (4.1) implies that

$$\mathrm{E}(\epsilon_t | X_0, Q_0^\star) = 0.$$

Since I am able to observe all the variables X_0 that affect $Y_t^\star - Y_0^\star$ and $Q_0^\star$ simultaneously, students who are the same in terms of the observable variables X_0 but who attend schools characterized by different values of $Q_0^\star$ do not differ, on average, in the unobserved component ϵ_t.

The parameter of interest is γ_t. If all quantities in (4.2) were known, a simple Ordinary Least Squares (OLS) regression would yield the estimate of γ_t, owing to assumption (4.1). Unfortunately, this is not my case, because both the outcome variable $\Delta Y_t^\star$ and the treatment variable $Q_0^\star$ are unobservable. In the following section I will address this problem.

4.2 The problem

The outcome variable of interest is the variation of student achievement during the entire academic year. Therefore, my goal is to estimate the impact of school quality on the change in student achievement from September 2002 ($t = 0$) to June 2003. Let $t = 3$ denote June 2003. The reference model is

$$Y_3^\star - Y_0^\star = \beta_3' X_0 + \gamma_3 Q_0^\star + \epsilon_3,$$

where

$$\gamma_3 = \frac{\partial \mathrm{E}(Y_3^\star - Y_0^\star | Q_0^\star, X_0)}{\partial Q_0^\star}$$

is the parameter of interest.

Hereafter, the conditioning upon X_0 will be left implicit, for simplicity. Hence consider, equivalently, the residuals—written in small letters—from the regressions of each variable on X_0:

$$y_3^\star - y_0^\star = \gamma_3 q_0^\star + \epsilon_3. \tag{4.3}$$

Then

$$\gamma_3 = \frac{\mathrm{Cov}(y_3^\star - y_0^\star, q_0^\star)}{\mathrm{Var}(q_0^\star)}.$$

I have put the superscripts "star" to school quality and achievement, because they are latent variables not directly observable. As a result, my evaluation presents two additional complications compared to classic evaluation case-studies: both the outcome and the treatment variables are unobservable. I clarify in the following the available information on these quantities, explaining the problems arising for the estimation of the parameter of interest.

4.2.1 Unobservable student achievement

In this section I deal with the identification problem that arises from not observing directly the outcome variable $y_3^\star - y_0^\star$. Throughout this section I will assume that the school quality variable $q_0^\star$ is instead observable. Two proxies for student achievement are available from PISA:

- the score in the PISA test[2], taken in April 2003 ($t = 2$), and
- the teacher's mark in mathematics of the last school report, in January 2003 ($t = 1$)[3].

[2] The PISA score is not simply of the type 'number of correct answers'. I have decided to use the mean of five plausible values for each student, which are five random draws from an ability distribution estimated for students with similar item response patterns and backgrounds by the modern techniques of the Item Response Theory. Please, refer to the manual OECD [2005a] for the construction of the plausible values.

[3] Exactly school reports are delivered between the end of January and the beginning of February. I have not been able to access additional information on the date of school reports. Throughout I will refer to January 2003, for simplicity, because it is reasonable to expect that even school reports in February derive from assessments in January.

Although these indicators are meant to proxy student achievement, they do not measure it directly. Moreover, it is unlikely that they describe the same dimension of student achievement. Indeed, the PISA test has been constructed to evaluate students' ability to apply their knowledge to the resolution of real-life problems[4], differently from the mark of the teacher, who evaluates, mainly, students' acquisition of curricular subjects.

To make the point of the identification problem arising from the available information, define the following measurement model for student achievement at each month t:

$$\begin{cases} \tilde{y}_t = \tilde{y}_t^\star + w_t & \leftarrow \text{mark} \\ y_t = y_t^\star + u_t & \leftarrow \text{PISA} \end{cases}, \tag{4.4}$$

where $\tilde{y}_t$ and y_t are mark and PISA, respectively, and $\tilde{y}_t^\star$ and $y_t^\star$ refer to the different measured latent factors. The disturbances w_t and u_t are assumed to be mutually independent and independent of all other variables (they are pure random noise). In practice I observe only $\tilde{y}_1$ and y_2.

I am interested to evaluate the impact of school quality on the variation of the latent factor measured by PISA ($y_t^\star$) from September 2002 ($t = 0$) to June 2003 ($t = 3$): $y_3^\star - y_0^\star$. What I can obtain is the difference between the PISA score in April 2003 ($t = 2$) and the teacher's mark in January 2003 ($t = 1$): $y_2 - \tilde{y}_1$.

I therefore face two fundamental problems:

1. I calculate the difference between two variables that measure different latent factors, only one of which is of interest;

2. the observation window January-April (1-2) is shorter than the period of interest September-June (0-3).

Let me consider in more detail the consequences of these problems on the estimation of γ_3. From equations (4.2)[5] and (4.4) we have that

$$y_2 - \tilde{y}_1 = (\gamma_2 - \gamma_1)q_0^\star + \{(y_1^\star - \tilde{y}_1^\star) + \epsilon_2 - \epsilon_1 + u_2 - w_1\}. \tag{4.5}$$

The second problem implies that I am able to estimate the effect of school quality on the gain in student achievement from January to April ($\gamma_2 - \gamma_1$), not from September to June (γ_3). But the first problem ($y_1^\star \neq \tilde{y}_1^\star$) implies that even $\gamma_2 - \gamma_1$ is not identified. Indeed, the OLS estimator of $\gamma_2 - \gamma_1$ in the above regression model is not consistent, because of the endogeneity of $q_0^\star$, which correlates with the error term through $y_1^\star - \tilde{y}_1^\star$:

$$\frac{\text{Cov}(y_2 - \tilde{y}_1, q_0^\star)}{\text{Var}(q_0^\star)} = (\gamma_2 - \gamma_1) + \frac{\text{Cov}(y_1^\star - \tilde{y}_1^\star, q_0^\star)}{\text{Var}(q_0^\star)}. \tag{4.6}$$

[4]See OECD [2004] for the definition of the concept measured by PISA.

[5]Taking the residuals of each variable from X_0.

How can I achieve a consistent estimator of $\gamma_2 - \gamma_1$? And, if possible, what can I say about the return to school quality over the entire academic year (γ_3)?

4.2.2 Unobservable school quality

Let me now consider the implications of school quality nonobservability. Leave aside the problem of the time-window and focus on the shorter period January-April (from 1 to 2). Suppose, hence, to be interested in the estimation of the parameter $\gamma_2 - \gamma_1$, instead of γ_3. Assume also to know the real variation of student achievement from January to April

$$y_2^\star - y_1^\star = (\gamma_2 - \gamma_1)q_0^\star + \{\epsilon_2 - \epsilon_1\} . \tag{4.7}$$

School quality, like student achievement, is a broad concept with a high degree of abstractness. If we define this concept as the ability of schools to increase achievement, school quality can be indirectly measured through a series of proxies, intended to perform with respect to this goal of promoting education (see Chapter 3, Section 3.2, for the available information on school quality). In the simplest case of only two proxies, the measurement model for $q_0^\star$ is

$$\begin{cases} q_1 = q_0^\star + v_1 \\ q_2 = \alpha_2 q_0^\star + v_2 \end{cases} ,$$

where the assumptions on the disturbances are the classic ones: v_1 and v_2 are uncorrelated with each other and with all other variables (they are pure random noise). The intercept terms do not enter the measurement equations because I am considering the residuals from the regressions on X_0, which include the intercept. Without loss of generality I may assign to $q_0^\star$ the scale of q_1 (otherwise the variance of $q_0^\star$ may be normalized to one).

What I can do is, for instance, to replace $q_0^\star$ with q_1 (the proxy with the same scale of $q_0^\star$) in equation 4.7, obtaining

$$y_2^\star - y_1^\star = (\gamma_2 - \gamma_1)q_1 + \{\epsilon_2 - \epsilon_1 - (\gamma_2 - \gamma_1)v_1\} , \tag{4.8}$$

and estimate $\gamma_2 - \gamma_1$ by OLS. The OLS estimator of $\gamma_2 - \gamma_1$ turns out to be inconsistent and downward biased:

$$\frac{\mathrm{Cov}(y_2^\star - y_1^\star, q_1)}{\mathrm{Var}(q_1)} = (\gamma_2 - \gamma_1)\left(1 - \frac{\mathrm{Var}(v_1)}{\mathrm{Var}(q_0^\star) + \mathrm{Var}(v_1)}\right)^{6} .$$

This is the first standard result which we encounter in the literature of measurement errors; see for example Wooldridge [2002] for a textbook treatment of this topic.

[6] The fact that I am conditioning on X_0—that is, I am considering the residuals from the regressions on X_0 of each variable—suggests to us the existence of a trade-off between the selection on observables assumption (4.1) and the attenuation bias (bias toward zero) within the brack-

4.2.3 Summary

The reference model is (4.3), where

$$\gamma_3 = \frac{\mathrm{Cov}(y_3^\star - y_0^\star, q_0^\star)}{\mathrm{Var}(q_0^\star)} \tag{4.9}$$

is the parameter of interest: the causal effect of school quality $q_0^\star$ on the gain in student achievement $y_3^\star - y_0^\star$ from September to June.

It follows from the discussion in the previous two sections that the available information in PISA permits to construct the empirical analogue of the following quantity:

$$\frac{\mathrm{Cov}(y_2 - \tilde{y}_1, q_1)}{\mathrm{Var}(q_1)} = (\gamma_2 - \gamma_1)\left(1 - \frac{\mathrm{Var}(v_1)}{\mathrm{Var}(q_0^\star) + \mathrm{Var}(v_1)}\right) + \frac{\mathrm{Cov}(y_1^\star - \tilde{y}_1^\star, q_0^\star)}{\mathrm{Var}(q_0^\star) + \mathrm{Var}(v_1)},$$

through the OLS regression of $y_2 - \tilde{y}_1$ on the generic proxy q_1.

I am simply replacing the latent variables with the proxies in (4.9), i.e. I am using q_1 instead of $q_0^\star$ and $y_2 - \tilde{y}_1$ instead of $y_3^\star - y_0^\star$, with the result of two components of bias in the value estimated for $\gamma_2 - \gamma_1$, which is not even the parameter of interest. Consider the regression model of $y_2 - \tilde{y}_1$ on q_1:

$$y_2 - \tilde{y}_1 = (\gamma_2 - \gamma_1)q_1 + \{(y_1^\star - \tilde{y}_1^\star) + \epsilon_2 - \epsilon_1 + u_2 - w_1 - (\gamma_2 - \gamma_1)v_1\}\,. \tag{4.10}$$

There are two sources of endogeneity of q_1:

- its correlation with $y_1^\star - \tilde{y}_1^\star$, because I am using the proxies y_2 and $\tilde{y}_1$, which measure two different dimensions of student achievement, and
- its correlation with v_1, because I am using the proxy q_1 for school quality.

Furthermore, student achievement is measured from $t = 1$ to $t = 2$, and not from $t = 0$ to $t = 3$.

4.3 Identification strategy

In Section 4.3.1 I will deal with the endogeneity problem arising from the unobservable student achievement, assuming to know school quality. Then, in Section 4.3.2, I will consider the opposite case of unobservable school quality, assuming to know student achievement. Finally, in Section 4.3.3, I will combine the two

ets: the more the covariates X_0 explain $Q_0^\star$ (school quality before taking residuals), the larger becomes the noise-to-signal ratio $\frac{\mathrm{Var}(v_1)}{\mathrm{Var}(q_0^\star)}$ (the covariates explain nothing of the error term of the proxy, whose variance does not change taking the residuals; on the contrary $\mathrm{Var}(q_0^\star)$ becomes smaller with a richer set of covariates X_0). See Black and Smith [2006] for a detailed discussion about this point in a similar context.

solutions in order to account jointly for unobservable quality and unobservable achievement. For the moment I leave aside the problem of the time-window and I focus on $\gamma_2 - \gamma_1$ instead of γ_3. The conditioning on X_0 continues to be left implicit.

4.3.1 Unobservable student achievement

Because of focusing on the restricted period January (1)-April (2), the reference model becomes (4.7), where I handle the nonobservability of $y_2^\star - y_1^\star$ and I assume to know $q_0^\star$. In Section 4.2.1 it has been shown that the regression of the observable quantity $y_2 - \tilde{y}_1$ on $q_0^\star$ does not yield a consistent estimate of $\gamma_2 - \gamma_1$ (see equation (4.6)). Two assumptions allow me to correct this estimate.

Assumption 1 (EXCLUSION RESTRICTION) There exists a variable z_0 such that $z_0 \perp (y_2^\star - y_1^\star)$ and $\mathrm{Cov}(z_0, y_1^\star) \neq 0$.

This is to say that there is a variable z_0[7] which does not depend upon how much students learn during the period January-April $(y_2^\star - y_1^\star)$, but that has an effect on the starting level of achievement in January $(y_1^\star)$. It is clear that the cumulated achievement in 1, resulting from the entire process of knowledge acquisition up to 1, depends on a greater additional number of variables than the gain in achievement from 1 to 2: all the school, peers and family inputs of students' history without any continuing effect on the value-added during the specific period considered; see Todd and Wolpin [2003] for a detailed explanation of the production functions for cognitive achievement.

Note that z_0 is one of the variables that need not be included as controls in X_0 thanks to the value-added formulation of the model. Considering the regression of $y_2 - \tilde{y}_1$ on z_0, we have that

$$\frac{\mathrm{Cov}(y_2 - \tilde{y}_1, z_0)}{\mathrm{Var}(z_0)} = \frac{\mathrm{Cov}(y_1^\star - \tilde{y}_1^\star, z_0)}{\mathrm{Var}(z_0)}. \tag{4.11}$$

Assumption 2 $\tilde{y}_1^\star = \theta y_1^\star + e$, e is random noise.

This assumption asserts a linear relationship between the latent factors measured by PISA and teacher's mark in t=1, where θ is simply a scale parameter—a coefficient which adjusts the difference in scales between the two latent factors. The error term e needs to be independent of z_0 and $q_0^\star$ to make work our identification strategy. By assuming $\mathrm{Cov}(y_1^\star, e)$=0, $y_1^\star$ and $\tilde{y}_1^\star$ are such that their covariance is regulated by θ:

$$\mathrm{Cov}(\tilde{y}_1, y_1) = \mathrm{Cov}(\tilde{y}_1^\star, y_1^\star) = \theta \mathrm{Var}(y_1^\star).$$

[7]The subscript 0 indicates that the variable is predetermined with respect to $q_0^\star$.

These are not strong assumptions conditionally upon the set of observables X_0: clearly, teacher's mark measures other latent factors in addition to pure achievement (the factor underlying the PISA score)—for instance student's application, behaviour, personality and so forth—but I can control for them through the conditioning on X_0.[8] However e could incorporate a systematic evaluation error by teachers, varying with their quality—and so with $q_0^\star$—and/or $y_1^\star$ itself. Anyway, the covariance $\mathrm{Cov}(e, y_1^\star)$ does not prevent identification.

It follows from Assumption 2 that the measurement equation relative to $\tilde{y}_1$ becomes

$$\tilde{y}_1 = \theta y_1^\star + \tilde{u}_1,$$

where the error term $\tilde{u}_1$, equal to $e + w_1$, is independent of all other variables (with the exception of $\tilde{y}_1$, of course). Hence

$$y_2 - \frac{1}{\theta}\tilde{y}_1 = (y_2^\star - y_1^\star) + (u_2 - \frac{1}{\theta}\tilde{u}_1)$$

is a measure with classical measurement error of the gain in achievement between January and April $(y_2^\star - y_1^\star)$.

Under Assumption 2, we can rewrite equations (4.6) and (4.11) as the first and second equation of the following system:

$$\begin{cases} \frac{\mathrm{Cov}(y_2 - \tilde{y}_1, q_0^\star)}{\mathrm{Var}(q_0^\star)} = (\gamma_2 - \gamma_1) + \left(\frac{1-\theta}{\theta}\right) \frac{\mathrm{Cov}(\tilde{y}_1, q_0^\star)}{\mathrm{Var}(q_0^\star)} \\ \frac{\mathrm{Cov}(y_2 - \tilde{y}_1, z_0)}{\mathrm{Var}(z_0)} = \left(\frac{1-\theta}{\theta}\right) \frac{\mathrm{Cov}(\tilde{y}_1, z_0)}{\mathrm{Var}(z_0)} \end{cases}.$$

This system can be solved uniquely for the two unknown parameters $\gamma_2 - \gamma_1$ and θ.

The estimate of θ derives from the second equation, equivalent to

$$\frac{1-\theta}{\theta} = \frac{\mathrm{Cov}(y_2 - \tilde{y}_1, z_0)}{\mathrm{Cov}(\tilde{y}_1, z_0)}. \tag{4.12}$$

[8] To be more clear, let me assume the following additive model for teacher's mark before taking residuals:

$$\tilde{Y}_1 = \mu + \underbrace{\theta Y_1^\star + \phi}_{\tilde{Y}_1^\star} + w_1$$

(cf. equation (4.4)), where the latent factor underlying teacher's mark $(\tilde{Y}_1^\star)$ is the sum of the rescaled factor measured by PISA $(Y_1^\star)$ plus other latent dimensions measured by teachers (students' characteristics such as behaviour, effort, etc.), incorporated in the unobservable variable ϕ, which could be correlated with $Y_1^\star$ (μ is simply the intercept). At the base of Assumption 2, there is a linear approximation of ϕ in terms of the observables X_0:

$$\phi = \delta' X_0 + e,$$

where e is random noise. Therefore, the conditioning on X_0 implies Assumption 2.

We recognize the expression of an Instrumental Variables (IV) estimator, where $y_2 - \tilde{y}_1$ is regressed on $\tilde{y}_1$, and z_0 is used to instrument $\tilde{y}_1$.

The first equation is instead equivalent to

$$\gamma_2 - \gamma_1 = \frac{\mathrm{Cov}(y_2 - \tilde{y}_1, q_0^\star)}{\mathrm{Var}(q_0^\star)} - \left(\frac{1-\theta}{\theta}\right) \frac{\mathrm{Cov}(\tilde{y}_1, q_0^\star)}{\mathrm{Var}(q_0^\star)}.$$

The first term on the right-hand side corresponds to the OLS regression of $y_2 - \tilde{y}_1$ on $q_0^\star$. The term multiplied by $\frac{1-\theta}{\theta}$ corresponds to the OLS regression of $\tilde{y}_1$ on $q_0^\star$.

Therefore, if I assume to know $q_0^\star$, an estimate of $\gamma_2 - \gamma_1$ can be achieved combining two OLS regressions and one IV regression.

4.3.2 Unobservable school quality

In this section I focus on the unobservable school quality and I assume to know $y_2^\star - y_1^\star$. The reference model is always (4.7), where $\gamma_2 - \gamma_1$ is the parameter of interest.

Consider the simplest case of only two proxies q_1 and q_2 (cf. Section 4.2.2). Following tha approach of Black and Smith [2006], let me look at the covariance matrix of the data

$$\begin{cases} \mathrm{Var}(y_2^\star - y_1^\star) = (\gamma_2 - \gamma_1)^2 \mathrm{Var}(q_0^\star) + \mathrm{Var}(\epsilon_2 - \epsilon_1) \\ \mathrm{Var}(q_1) = \mathrm{Var}(q_0^\star) + \mathrm{Var}(v_1) \\ \mathrm{Var}(q_2) = \alpha_2^2 \mathrm{Var}(q_0^\star) + \mathrm{Var}(v_2) \\ \mathrm{Cov}(q_1, y_2^\star - y_1^\star) = (\gamma_2 - \gamma_1)\mathrm{Var}(q_0^\star) \\ \mathrm{Cov}(q_2, y_2^\star - y_1^\star) = (\gamma_2 - \gamma_1)\alpha_2 \mathrm{Var}(q_0^\star) \\ \mathrm{Cov}(q_1, q_2) = \alpha_2 \mathrm{Var}(q_0^\star) \end{cases}.$$

This system of six equations in six unknown parameters ($\gamma_2 - \gamma_1, \alpha_2, \mathrm{Var}(q_0^\star), \mathrm{Var}(\epsilon_2 - \epsilon_1), \mathrm{Var}(v_1)$ and $\mathrm{Var}(v_2)$) can be solved uniquely for the unknown parameters. It follows immediately that

$$\gamma_2 - \gamma_1 = \frac{\mathrm{Cov}(y_2^\star - y_1^\star, q_2)}{\mathrm{Cov}(q_1, q_2)},$$

which we recognize to be the usual IV estimator, with q_2 as instrumental variable for q_1 in the regression of $y_2^\star - y_1^\star$ on q_1; see equation (4.8) for the corresponding regression model. Of course, I can use all the other proxies of $q_0^\star$ not inserted in the structural equation as instruments for q_1 in the general case of more than two proxies (Generalized IV estimator). What I am considering is the simplest situation of a single-equation model with only one explanatory variable measured with error; see Wooldridge [2002] for a textbook treatment of this topic.

Note that this procedure is equivalent to a one-factor analysis, where the reference equation (4.7) enters the measurement model for $q_0^\star$, i.e. the value-added $y_2^\star - y_1^\star$ plays the role of an additional indicator of school quality. Clearly its extension to the multiple factor case (more than one dimension of $q_0^\star$) turns out to be a multiple factor analysis. I refer to Chapter 5 for the generalization of the approach when $q_0^\star$ is multidimensional.

4.3.3 Joint solution

After having solved separately the problems of achievement and school quality nonobservability, the joint resolution is straightforward. We have just seen that the solution to school quality nonobservability is an IV estimator in the regression of the value-added $y_2^\star - y_1^\star$ (assumed known) on q_1, where we instrument q_1 with q_2. Hence, in the same way, consider now the regression of the observable quantity $y_2 - \tilde{y}_1$ on q_1. Since equation (4.10) becomes

$$y_2 - \tilde{y}_1 = (\gamma_2 - \gamma_1)q_1 + \left(\frac{1-\theta}{\theta}\right)\tilde{y}_1 + \left\{u_2 - \frac{1}{\theta}\tilde{u}_1 + \epsilon_2 - \epsilon_1 - (\gamma_2 - \gamma_1)v_1\right\} \quad (4.13)$$

under Assumption 2, we have that

$$\frac{\mathrm{Cov}(y_2 - \tilde{y}_1, q_2)}{\mathrm{Cov}(q_1, q_2)} = (\gamma_2 - \gamma_1) + \left(\frac{1-\theta}{\theta}\right)\frac{\mathrm{Cov}(\tilde{y}_1, q_2)}{\mathrm{Cov}(q_1, q_2)}$$

instrumenting q_1 with q_2. As before, when handling student achievement nonobservability, I must first estimate the value of θ, through which correct the IV estimate written above, in order to get at a consistent estimate of $\gamma_2 - \gamma_1$. Once again, I will identify θ from (4.12) thanks to the Assumption 1 of exclusion restriction.

We see that there is no need to proceed in two steps. In fact model (4.13) implies to regress the observable quantity $y_2 - \tilde{y}_1$ jointly on q_1 and $\tilde{y}_1$, in phase of estimation. Since both q_1 and $\tilde{y}_1$ are endogenous in (4.13) because of their correlation with v_1 and $\tilde{u}_1$ (their measurement errors) in the error term, respectively, an IV estimator with q_2 as instrument for q_1 and z_0 as instrument for $\tilde{y}_1$ allows me to achieve a consistent estimate of $\gamma_2 - \gamma_1$.

4.4 The time-window problem

I assume a linear growth for student achievement, i.e. a straight line approximation of the growth of $y_t^\star$ over time (t):

$$y_t^\star = A + Bt,$$

where A and B are random variables that allow changeable intercepts and slopes across students. This is the usual approach adopted in the measurement of change literature, especially in educational research based on datasets limited to only two waves; see Brandt, Rogosa and Zimowski [1982] for a detailed justification.

Hence, we have

$$y_3^\star - y_0^\star = c(y_2^\star - y_1^\star),$$

where c is a constant equal to the ratio of the time between the beginning and the end of the academic year (0-3) to the time between the delivery of school reports and the date of PISA (1-2). Really also c has changeable values across students, since the dates of PISA and school reports—which are not known from the dataset—vary from school to school. Anyway, I assume a constant value for c, approximately equal to 3, by placing the start of the academic year (0), the delivery of school reports (1), the administration of PISA (2) and the end of school (3) some time in September, January, April and June, respectively.[9]

Therefore, the impact of school quality on the outcome variable of interest $y_3^\star - y_0^\star$ (γ_3) is equal to the estimable impact on $y_2^\star - y_1^\star$ ($\gamma_2 - \gamma_1$) up to the constant $c = 3$:

$$\gamma_3 = 3(\gamma_2 - \gamma_1).$$

Further, we have that $3(y_2 - \frac{1}{\hat{\theta}}\tilde{y}_1)$ can be used as measure of the gain in student achievement from September to June ($y_3^\star - y_0^\star$), after having estimated θ by IV in the regression of $y_2 - \tilde{y}_1$ on $\tilde{y}_1$.

4.5 Results

In this section I report the results of the IV estimation of equation (4.13). Bear in mind that

- the estimate of the return to school quality, $\gamma_2 - \gamma_1$, refers to the shorter period January-April. Anyway, this effect is equal to that one over the entire period of interest, September-June, up to a constant on account of what just said in Section 4.4.

[9] The only drawback of this simple strategy is the short distance between January and April. Note that the reliability of the observed difference $y_2 - \frac{1}{\theta}\tilde{y}_1$, as measure of $y_2^\star - y_1^\star$,

$$\frac{\mathrm{Var}(y_2^\star - y_1^\star)}{\mathrm{Var}(y_2 - \frac{1}{\theta}\tilde{y}_1)} = \frac{\mathrm{Var}(B)}{\mathrm{Var}(B) + \frac{\mathrm{Var}(u_2 - \frac{\tilde{u}_1}{\theta})}{(t_2 - t_1)^2}},$$

decreases as the width of the observation-window $t_2 - t_1$, the temporal interval between PISA (t_2) and the delivery of school reports (t_1), decreases: in a short period of time the heterogeneity in individual growth could have no time to manifest itself (see Rogosa and Willet [1983]).

- The results are conditionally on the set of control variables X_0, in that I am considering the residuals of each variable from the regression on X_0. This means that the impact of school quality is netted out by the effects of the variables contained in X_0.

The variables which I control for are described in Section 3.2. They are individual and family background characteristics (sex, age, type of learning strategies, past and expected educational career, structure of family, parents' educational and occupational status, possessions at home), plus other school information in addition to school quality: type of school, location (city size, province) and school peer group characteristics (the school average value of some family background variables).

As regards the instruments (z_0) for teacher's mark ($\tilde{y}_1$), i.e. the variables assumed to affect only the initial level of achievement, but not the value added, I use the (i) loss of two or more consecutive months during primary/low secondary school and the (ii) change of school during primary/low secondary school. Whatever pair of proxies q_1 and q_2 is chosen, these variables pass the test of rank condition: they are jointly (and almost always individually) statistically significant in the first stage regression of the endogenous $\tilde{y}_1$ on the full set of instruments; they also pass the Hansen overidentification test to determine their uncorrelation with the error process.[10] For the sake of brevity I do not report these results in the following tables.

Finally, the estimated standard errors are corrected for the heteroskedasticity and correlation of the error terms in the model induced by the clustering of students in schools, via the Huber-White formula, which assumes a block diagonal covariance matrix of disturbances. I do not weight observations as recommended by PISA manuals, because I take a structural approach and the stratification occurring in the sampling procedure is not on the outcome variable[11].

Being the focus of research and political debate, I choose to show detailed results only for class size, which I measure through the ratio between the enrolment and the total number of classes in grade 10, in order to avoid potential within-school sorting bias, arising from compensatory policies placing low achievers into smaller classes (Akerhielm [1995]). Anyway, this endogeneity issue is recognized to be not a problem for the Italian educational system. Furthermore, I am conditioning upon the control variables X_0 to take into account individual ability, which also permits to control for the non-random sorting of students across schools (revise Section 4.1).

[10]See Baum, Schaffer and Stillman [2002] for a review of estimation and testing techniques with instrumental variables.

[11]See Cameron and Trivedi [2005] for a convincing justification.

In Panel A of Table 4.1 I use class size as "instrumented" proxy q_1, i.e. I regress the difference between PISA score and teacher's mark $(y_2 - \tilde{y}_1)$ on teacher's mark plus class size. In addition to the IV estimates of equation (4.13) introduced in Section 4.3.3, where z_0 (change of school and loss of months) and another generic proxy q_2 serve as instruments for $\tilde{y}_1$ and q_1 (*IV – joint solution*), I present "intermediate" (wrong) results, not solving for the endogeneity of both $\tilde{y}_1$ and q_1 (*OLS*) and only q_1 and $\tilde{y}_1$, separately (*IV – solution 1* and *IV – solution 2*). Precisely,

- in the first row I estimate by OLS the regression of $y_2 - \tilde{y}_1$ on q_1 (cf. equation (4.10)), so that the resulting value estimated for $\gamma_2 - \gamma_1$ is biased both because (i) I compute the difference between two different measures of achievement and (ii) I use a proxy of school quality.
- Instead, in the second row I estimate equation (4.13) by instrumenting only teacher's mark, so as to not remove the bias arising from (ii).[12]
- On the contrary, in the third row I do not take into account (i) and I regress the raw value added $y_2 - \tilde{y}_1$ on only q_1 using q_2 as instrument for q_1.

In Panel B of Table 4.1 I simply swap the role of the proxies and class size becomes the "instrumental" proxy q_2 for the proxy q_1 used as regressor. The other exploited proxies concern school size and quantity of teaching staff (SISSI total school enrolment and PISA student-teacher ratio; note that part-time teachers contribute 0.5 to the total number of teachers); school resources, specifically with respect to Information Communication Technology (proportion of computers connected to Internet, proportion of computers connected to a local network, pc-student ratio and monthly time spent in computer laboratory); school selectivity (freshman retention rate and proportion of successful freshmen); "type" of teaching staff (proportion of certified teachers, proportion of teachers with permanent contract, average teachers' seniority and proportion of full-time teachers). The reader is referred to Section 3.2 for more details about these variables.

I report the first-stage results only for q_1 regressed on q_2 and z_0 (*IV - joint solution*) in order to show that there are quite a few proxies which are not good instruments for class size and vice versa. Exactly I report the t statistic on the instrumental proxy q_2.

There is no evidence of an impact of school quality on students' gain in achievement and the estimated effect is at the most weakly significant. If we look at Panel A, we see that class size has a counter-intuitive effect when instrumented by school enrolment; on the contrary, the estimated impact becomes negative, but only at

[12] This is not the same as estimating by OLS the regression of the adjusted value-added $y_2 - \frac{1}{\theta}\tilde{y}_1$ on class size, because also θ comes out biased owing to the endogeneity of q_1.

the 10 percent level, when the proportion of teachers with a permanent contract is used as instrument.

Results change in Panel B, considering class size as instrumental proxy: for instance, the effect associated to the proportion of permanent teachers disappears, whereas the coefficient on the freshman retention rate turns to be statistically significant. Furthermore, there seems to be a (very weak) negative effect of the number of computers per student and a positive effect of the student-teacher ratio (accordingly to the unexpected effect of class size).

These inconsistent findings plus the fact that results are mostly insignificant and change inverting the role of the proxies suggest the existence of multiple dimensions of school quality, measured by different set of proxies, each one potentially affected by more than one dimension of quality. This could be, therefore, an explanation for the contradictory findings of literature, with respect to class size mainly (cf. Section 2.3.2).

If we wanted to comment these results all the same, what could we say about the counter-intuitive effect estimated for class size and the number of students per teacher? This is a typical result by analysing data from international surveys like PISA; consider for instance Woessmann [2003], and Hanushek and Luque [2003] for the International Mathematics and Science Study (TIMMS). Theoretically the effect should be negative, in the sense that students' achievement should increase in response to class size reductions, because cuts in class size translate into more teaching time per student. Researchers provide many justifications for the contradictory indications from real data. One is the Social Cognitive Learning Theory (Lazear [2001] and Dobbelsteen, Levin and Oosterbeek [2002]), which states that the indirect positive effect of larger classes works through the increased likelihood of having classmates of comparable or higher ability.

Then, the positive impact of being in a class with lots of students could be the effect of attending a school with a large enrolment, i.e. a popular school with higher demand, recognized to be at the top of quality, where mainly able students and/or students with a strong family background decide to enrol (Hanushek and Luque [2003][13]). This is also supported by my data, taking school size as proxy for class size and vice versa. But I am controlling for students' endowment and home environment via the conditioning upon X_0, which contains the size of the site where school is located as well (schools with lower enrolment could be simply rural schools). Furthermore, I have replicated all empirical analyses including school size among regressors and results do not change.

Finally, there could be a trade-off between students' quantity and teachers' quality, due to a dilution of teacher quality as a consequence of class size reductions, i.e. class size reduction policies force schools to hire less able and less

[13]See also Angrist and Lavy [2002], Woessmann [2003], and Jakubowski and Sakouski [2006].

qualified teachers (Addonizio and Phelps [1995]). Observe, in our case, the negative relationship between class size and the proportion of permanent teachers from the IV first stage results. Clearly job tenure could be a strong indicator of teaching quality, because permanent teachers not only have (i) more experience, but also are (ii) more motivated and can promote more education in better conditions of (iii) teaching continuity and work stability.

As regards the scale parameter θ, it is strongly statistically significant with a value between 0.015 and 0.02, approximately. I obtain the same value considering all the other pairs of proxies, whose results are reported in Table 4.2 only referring to the correct estimate of the parameter of interest $\gamma_2 - \gamma_1$ in the joint IV solution. For the sake of more clearness, rows are empty when proxies are not good instrument of each other (coefficient of the instrumental proxy insignificant at the 10 percent level in the IV first stage regression of q_1 on q_2 and z_0).

At a first glance, in addition to the indications provided by Table 4.1 (positive effect of school and class size), there seems to be a negative impact of computer use at school[14] (Angrist and Lavy [2002][15]) and, instead, a positive impact of teachers' characteristics such as type of contract (permanent and full-time), certification and experience. However, this evidence is likely to be misleading, because the assumption of only one dimension of school quality could be false.

As said above, this conclusion is supported by the fact that results change with the proxies I consider and are not the same by swapping their role[16]. Moreover, many pairs of proxies fail the t test in the first step of the IV estimation. Therefore, moving to a multiple factor model, where I assume a multidimensional school quality, is necessary for the success of my analysis. I address the general case of multidimensional school quality in the following chapter.

[14]And the number of computers per student.

[15]Other references are Leuven et al. [2004], Fuchs and Woessmann [2004], Golsbee and Guryan [2006].

[16]The values estimated for $\gamma_2 - \gamma_1$ are not comparable using different q_1, because they refer to school qualities defined on different scales—the scale of q_1.

Table 4.1 - Panel A: results using class size as "instrumented proxy"

q_2	enrolment	student-teacher ratio	prop. pcs connected to www	prop. pcs connected to lan	pc-student ratio	pc use
OLS						
$\gamma_2 - \gamma_1$	6.330 (2.977)	6.330 (2.977)	6.330 (2.977)	6.330 (2.977)	6.330 (2.977)	6.330 (2.977)
IV – solution 1						
$\gamma_2 - \gamma_1$	2.831 (1.361)	2.831 (1.361)	2.831 (1.361)	2.831 (1.361)	2.831 (1.361)	2.831 (1.361)
θ	.017 (5.759)	.017 (5.759)	.017 (5.759)	.017 (5.759)	.017 (5.759)	.017 (5.759)
IV – solution 2						
$\gamma_2 - \gamma_1$	22.649 (4.636)	16.028 (2.167)	92.953 (1.012)	-139.452 (-.795)	2.722 (.240)	-18.477 (-.367)
IV – joint solution						
$\gamma_2 - \gamma_1$	22.195 (3.820)	7.586 (1.070)	36.801 (.704)	-77.387 (-.745)	5.752 (.585)	42.570 (.566)
θ	.019 (5.663)	.018 (5.364)	.018 (5.181)	.015 (3.499)	.018 (5.408)	.018 (4.854)
q_2 t stat. 1st stage	6.964	5.573	.981	-.845	-3.282	-.850

* Cluster robust t statistics are given in parenthesis.

Continued on next page...

... Table 4.1 continued

q_2	freshman retention rate	prop. successful freshmen	prop. full time teachers	prop. certified teachres	prop. permanent teachers	teachers' seniority
OLS						
$\gamma_2 - \gamma_1$	6.330	6.330	6.330	6.330	6.330	6.330
	(2.977)	(2.977)	(2.977)	(2.977)	(2.977)	(2.977)
IV – solution 1						
$\gamma_2 - \gamma_1$	2.831	2.831	2.831	2.831	2.831	2.831
	(1.361)	(1.361)	(1.361)	(1.361)	(1.361)	(1.361)
θ	.017	.017	.017	.017	.017	.017
	(5.759)	(5.759)	(5.759)	(5.759)	(5.759)	(5.759)
IV – solution 2						
$\gamma_2 - \gamma_1$	28.031	-8.386	192.728	-24.376	-50.749	-2.816
	(2.027)	(-.365)	(.681)	(-.368)	(-1.752)	(-.134)
IV – joint solution						
$\gamma_2 - \gamma_1$	15.636	-10.027	89.485	-40.500	-69.037	-18.941
	(1.195)	(-.449)	(.544)	(-.539)	(-1.851)	(-.722)
θ	.018	.018	.020	.016	.017	.017
	(5.859)	(5.759)	(2.502)	(4.800)	(4.292)	(4.653)
q_2 t stat. 1st stage	2.797	-1.718	.642	-.715	-2.245	-1.558

* Cluster robust t statistics are given in parenthesis.

Table 4.1 - Panel B: results using class size as "instrumental proxy"

q_1	enrolment	student-teacher ratio	prop. pcs connected to www	prop. pcs connected to lan	pc-student ratio	pc use
OLS						
$\gamma_2 - \gamma_1$	8.495 (5.062)	5.658 (2.156)	3.920 (2.769)	5.108 (3.425)	-.382 (-.150)	.711 (.391)
IV – solution 1						
$\gamma_2 - \gamma_1$	8.260 (4.284)	2.223 (.879)	1.999 (1.260)	4.586 (2.744)	-.865 (-.398)	-4.674 (-2.195)
θ	.017 (5.822)	.018 (5.405)	.017 (5.926)	.0173 (5.827)	.017 (5.546)	.016 (5.626)
IV – solution 2						
$\gamma_2 - \gamma_1$	16.856 (3.230)	25.474 (2.572)	103.407 (1.001)	-110.170 (-.747)	-24.588 (-2.043)	-152.089 (-.816)
IV – joint solution						
$\gamma_2 - \gamma_1$	8.274 (1.681)	16.563 (1.834)	41.236 (.759)	-51.207 (-.645)	-15.404 (-1.542)	-41.146 (-.750)
θ	.018 (5.738)	.019 (5.066)	.018 (4.601)	.017 (4.992)	.018 (5.293)	.015 (4.594)
q_2 t stat. 1st stage	7.739	5.916	1.011	-.838	-3.727	-.866

* Cluster robust t statistics are given in parenthesis.

Continued on next page...

. . . Table 4.1 continued

q_1	freshman retention rate	prop. successful freshmen	prop. full time teachers	prop. certified teachres	prop. permanent teachers	teachers' seniority
OLS						
$\gamma_2 - \gamma_1$	3.921	.627	5.281	.648	15.991	.287
	(2.470)	(.345)	(3.030)	(.350)	(3.870)	(.155)
IV – solution 1						
$\gamma_2 - \gamma_1$	1.971	.439	5.171	1.530	22.267	1.307
	(1.107)	(.230)	(2.721)	(.729)	(4.218)	(.610)
θ	.018	.018	.018	.016	.018	.017
	(6.005)	(5.976)	(5.570)	(5.857)	(5.173)	(4.846)
IV – solution 2						
$\gamma_2 - \gamma_1$	44.109	-76.094	205.888	-155.807	-107.462	-45.847
	(2.391)	(-1.382)	(.660)	(-.655)	(-1.464)	(-1.343)
IV – joint solution						
$\gamma_2 - \gamma_1$	32.846	-52.402	38.705	-23.802	-56.426	-12.228
	(2.008)	(-1.329)	(.322)	(-.514)	(-1.089)	(-.621)
θ	0.019	0.019	0.020	0.017	0.016	0.016
	(5.448)	(5.188)	(2.193)	(5.161)	(4.676)	(4.795)
q_2 t stat. 1st stage	2.817	-1.677	.639	-.701	-1.809	-1.510

* Cluster robust t statistics are given in parenthesis.

Table 4.2 - Results for the other pairs of proxies.

	$\gamma_2 - \gamma_1$	t stat. $\gamma_2 - \gamma_1$	t stat. q_2 1st stage
q_1=school enrolment			
student-teacher ratio	13.738	1.250	3.491
class size	8.274	1.681	7.739
prop. pcs connected to www	23.751	1.015	1.721
prop. pcs connected to lan	55.074	1.661	1.751
pc-student ratio	7.542	1.199	-6.591
pc use	40.207	1.814	-2.164
prop. successful freshmen	-9.091	-.476	-1.976
freshman retention rate			
prop. full-time teachers	30.705	2.278	2.758
prop. certified teachers			
prop. permanent teachers			
teachers' seniority			
q_1=student-teacher ratio			
school enrolment	53.262	2.650	3.252
class size	16.563	1.834	5.916
prop. pcs connected to www	23.508	1.020	2.414
prop. pcs connected to lan			
pc-student ratio	3.188	.313	-4.409
pc use			
prop. successful freshmen			
freshman retention rate	31.972	.914	1.650
prop. full-time teachers			
prop. certified teachers			
prop. permanent teachers			
teachers' seniority			
q_1=prop. pcs connected to www			
school enrolment	62.358	1.524	1.702
student-teacher ratio	7.935	.613	2.280

Continued on next page...

... Table 4.2 continued

	$\gamma_2 - \gamma_1$	t stat. $\gamma_2 - \gamma_1$	t stat. q_2 1st stage
class size			
prop. pcs connected to lan	12.287	2.527	6.932
pc-student ratio			
pc use	-15.142	-1.841	4.544
prop. successful freshmen			
freshman retention rate			
prop. full-time teachers			
prop. certified teachers			
prop. permanent teachers			
teachers' seniority			
q_1=prop. pcs connected to lan			
school enrolment	67.682	1.639	1.745
student-teacher ratio			
class size			
prop. pcs connected to www	5.892	1.306	7.449
pc-student ratio			
pc use			
prop. successful freshmen			
freshman retention rate	-11.144	-1.016	-3.033
prop. full-time teachers			
prop. certified teachers			
prop. permanent teachers			
teachers' seniority			
q_1=pc-student ratio			
school enrolment	-30.955	-3.614	-5.398
student-teacher ratio	-5.635	-.682	-4.191
class size	-15.404	-1.542	-3.727
prop. pcs connected to www			
prop. pcs connected to lan			
pc use	-18.545	-1.759	3.779
prop. successful freshmen			

Continued on next page...

... Table 4.2 continued

	$\gamma_2 - \gamma_1$	t stat. $\gamma_2 - \gamma_1$	t stat. q_2 1st stage
freshman retention rate	-12.526	-1.130	-3.505
prop. full-time teachers			
prop. certified teachers			
prop. permanent teachers			
teachers' seniority			
q_1=pc use			
school enrolment	-79.240	-1.932	-2.099
student-teacher ratio			
class size			
prop. pcs connected to www	12.939	1.296	4.347
prop. pcs connected to lan			
pc-student ratio	-4.788	-.415	3.511
prop. successful freshmen			
freshman retention rate			
prop. full-time teachers			
prop. certified teachers			
prop. permanent teachers			
teachers' seniority			
q_1=prop. successful freshmen			
school enrolment	-69.638	-1.597	-1.951
student-teacher ratio			
class size	-52.402	-1.329	-1.677
prop. pcs connected to www			
prop. pcs connected to lan			
pc-student ratio			
pc use			
freshman retention rate	-22.448	-.973	-1.672
prop. full-time teachers			
prop. certified teachers			
prop. permanent teachers			
teachers' seniority			

Continued on next page...

... Table 4.2 continued

	$\gamma_2 - \gamma_1$	t stat. $\gamma_2 - \gamma_1$	t stat. q_2 1st stage
q_1=freshman retention rate			
school enrolment			
student-teacher ratio	23.411	.804	1.715
class size	32.846	2.008	2.817
prop. pcs connected to www			
prop. pcs connected to lan	-33.243	-1.628	-2.504
pc-student ratio	8.856	.687	-2.905
pc use			
prop. successful freshmen			
prop. full-time teachers	27.778	1.810	2.561
prop. certified teachers			
prop. permanent teachers			
teachers' seniority			
q_1=prop. full time teachers			
school enrolment	51.559	2.615	2.619
student-teacher ratio			
class size			
prop. pcs connected to www			
prop. pcs connected to lan			
pc-student ratio			
pc use			
prop. successful freshmen			
freshman retention rate	10.832	1.076	3.200
prop. certified teachers			
prop. permanent teachers	48.903	2.645	2.861
teachers' seniority			
q_1=prop. certified teachers			
school enrolment			
student-teacher ratio			

Continued on next page...

... Table 4.2 continued

	$\gamma_2 - \gamma_1$	t stat. $\gamma_2 - \gamma_1$	t stat. q_2 1st stage
class size			
prop. pcs connected to www			
prop. pcs connected to lan			
pc-student ratio			
pc use			
prop. successful freshmen			
freshman retention rate			
prop. full-time teachers			
prop. permanent teachers	26.550	2.625	3.732
teachers' seniority	5.610	.535	3.399
q_1=prop. permanent teachers			
school enrolment			
student-teacher ratio			
class size	-56.426	-1.089	-1.809
prop. pcs connected to www			
prop. pcs connected to lan			
pc-student ratio			
pc use			
prop. successful freshmen			
freshman retention rate			
prop. full-time teachers	89.304	2.188	3.059
prop. certified teachers	5.433	.276	2.644
teachers' seniority	10.316	.581	4.019
q_1=teachers' seniority			
school enrolment			
student-teacher ratio			
class size			
prop. pcs connected to www			
prop. pcs connected to lan			
pc-student ratio			
pc use			

Continued on next page...

... Table 4.2 continued

	$\gamma_2 - \gamma_1$	t stat. $\gamma_2 - \gamma_1$	t stat. q_2 1st stage
prop. successful freshmen			
freshman retention rate			
prop. full-time teachers			
prop. certified teachers	1.053	.115	3.627
prop. permanent teachers	20.415	2.864	5.021

Chapter 5

Multidimensional school quality

The insignificant and, at the most, weakly significant results of the previous chapter suggest that (i) school quality does not matter once we control for individual endowment, family background and peers (Coleman Report [1966]) or, more properly, (ii) the common measures of school quality are rather poor proxies not closely related to the true differences in quality among schools, which really exist (Hanushek, Kain and Rivkin [2005]). But the inconsistent findings, with respect to size and/or sign of estimated effects, looking at different pairs of proxies and/or inverting their role, plus the fact that many proxies are uncorrelated, indicate another plausible explanation, which literature pays no sufficient attention to: the possibility of multiple dimensions of school quality (Black and Smith [2006]), maybe correlated, measured by different groups of proxies, each one potentially explaining more than one aspect of quality.

Hence, in this chapter I extend our previous discussion to the general case of multidimensional school quality, considering all proxies jointly. First, in Section 5.1, I reformulate the estimation strategy for the simple case of one dimension of school quality in a factor analysis set-up, which is well suited for explaining the issues related to identification of the resulting multiple factor model (Section 5.2). Therefore, I use factor analysis mainly with an exploratory purpose to estimate the number of latent factors and give them an interpretation. Its indications along with subject matter theory from literature allow me to impose more structure in the model and estimate it in a convenient way, without resorting to the usual computationally heavy iterative procedures of factor analysis estimation.

I apply a straightforward generalization of the simple IV estimator employed in the previous chapter (Section 5.3). Proceeding in this way, I can also obtain directly, as a by-product, the standard errors of the parameters of interest, otherwise not available from the major part of statistical software for factor analysis estimation. Furthermore, I can correct these estimates for the heteroskedasticity and correlation of the error terms induced by the clustering of students in schools via the Huber-White formula.[1]

5.1 Factor analysis approach

5.1.1 Unidimensional case

The identification strategy presented in Chapter 4 amounts to a 2-steps estimation procedure, where

- first θ is estimated by IV in the regression of $y_2 - \tilde{y}_1$ on $\tilde{y}_1$, using z_0 as instrument for $\tilde{y}_1$:

$$y_2 - \tilde{y}_1 = \left(\frac{1-\theta}{\theta}\right)\tilde{y}_1 + \{\underbrace{(\gamma_2-\gamma_1)q_0^\star + \epsilon_2 - \epsilon_1}_{y_2^\star - y_1^\star} + u_2 - \frac{\tilde{u}_1}{\theta}\}, \tag{5.1}$$

  where $\mathrm{Cov}(y_2^\star - y_1^\star, z_0)$=0 under Assumption 1;

- and then the adjusted measure of gain in achievement, $y_2 - \frac{1}{\theta}\tilde{y}_1$, is regressed on the generic proxy q_1, instrumented by another proxy q_2, to estimate $\gamma_2-\gamma_1$:

$$y_2 - \frac{1}{\theta}\tilde{y}_1 = (\gamma_2-\gamma_1)q_1 + \left\{\epsilon_2 - \epsilon_1 + u_2 - \frac{\tilde{u}_1}{\theta} - (\gamma_2-\gamma_1)v_1\right\}.$$

As I have said in Section 4.3.1, $y_2 - \frac{1}{\theta}\tilde{y}_1$ is a measure with classical measurement error of the gain in achievement $y_2^\star - y_1^\star$, thus it can be viewed as a supplemental proxy, in addition to q_1 and q_2, for $q_0^\star$. Indeed, the IV estimator of $\gamma_2 - \gamma_1$ in the second step above can also be derived from solving the moment equations of a factor analysis model where $y_2 - \frac{1}{\theta}\tilde{y}_1$ plays the role of an indicator of $q_0^\star$, along with q_1 and q_2:

$$\left\{\begin{array}{l} y_2 - \frac{1}{\theta}\tilde{y}_1 = (\gamma_2-\gamma_1)q_0^\star + \left\{\epsilon_2 - \epsilon_1 + u_2 - \frac{\tilde{u}_1}{\theta}\right\} \\ q_1 = q_0^\star + v_1 \\ q_2 = \alpha_2 q_0^\star + v_2 \end{array}\right. ;$$

the assumptions made before are such that the error terms are uncorrelated with each other and with $q_0^\star$.

Notice that the factor analysis approach allows to adopt the normalization $\mathrm{Var}(q_0^\star)$=1 to fix the scale of $q_0^\star$, and so permits to obtain comparable estimates of the parameter of interest—now the effect produced by an increase of one standard

[1]The usual methods for factor analysis estimation, for instance Maximum Likelihood, assume independently and identically distributed observations.

deviation in $q_0^\star$—using different pairs of proxies.[2] Of course, we can use more than two proxies (other ones besides q_1 and q_2) as indicators of $q_0^\star$; in such case the system that relates the covariance matrix of the indicators to the unknown parameters is overidentified (GIV estimator in the other approach). Therefore, another advantage of the factor analysis approach is the possibility to estimate also the scale parameters and the error variances of the proxies, so as to evaluate their reliability and to rank them from the least to the most noisy one.

Finally, the factor analysis approach can be easily extended to the general case of $q_0^\star$ multidimensional. I address this issue in the following section.[3]

5.1.2 Multidimensional case

In this section I extend the discussion to the general case of $q_0^\star$ multidimensional. Clearly, it is more likely that schools have several dimensions of quality, concerning

[2]With this parameterization (α_1 free) the resolution of the moment equations gives

$$\begin{cases} \gamma_2 - \gamma_1 = \sqrt{\frac{\mathrm{Cov}(\Delta y, q_1)\mathrm{Cov}(\Delta y, q_2)}{\mathrm{Cov}(q_1, q_2)}} \\ \alpha_1 = \sqrt{\frac{\mathrm{Cov}(\Delta y, q_1)\mathrm{Cov}(q_1, q_2)}{\mathrm{Cov}(\Delta y, q_2)}} \\ \alpha_2 = \sqrt{\frac{\mathrm{Cov}(\Delta y, q_2)\mathrm{Cov}(q_1, q_2)}{\mathrm{Cov}(\Delta y, q_1)}} \\ \mathrm{Var}(\nu) = \mathrm{Var}(\Delta y) - \frac{\mathrm{Cov}(\Delta y, q_1)\mathrm{Cov}(\Delta y, q_2)}{\mathrm{Cov}(q_1, q_2)} \\ \mathrm{Var}(v_1) = \mathrm{Var}(q_1) - \frac{\mathrm{Cov}(\Delta y, q_1)\mathrm{Cov}(q_1, q_2)}{\mathrm{Cov}(\Delta y, q_2)} \\ \mathrm{Var}(v_2) = \mathrm{Var}(q_2) - \frac{\mathrm{Cov}(\Delta y, q_2)\mathrm{Cov}(q_1, q_2)}{\mathrm{Cov}(\Delta y, q_1)} \end{cases},$$

where $\Delta y = y_2 - \frac{1}{\theta}\tilde{y}_1$ and $\nu = \epsilon_2 - \epsilon_1 + u_2 - \frac{\tilde{u}_1}{\theta}$.

[3]Let me consider the simplest case of two dimensions of school quality ($p = 2$) in presence of a generic number K of proxies:

$$\begin{cases} y_2^\star - y_1^\star = \gamma_1 q_{01}^\star + \gamma_2 q_{02}^\star + \{\epsilon_2 - \epsilon_1\} \\ q_k = \alpha_{k1} q_{01}^\star + \alpha_{k2} q_{02}^\star + v_k \qquad \text{k=1, \dots, K} \end{cases}.$$

The exact replication of the IV procedure explained in Chapter 4, i.e. regressing the adjusted value added $y_2 - \frac{1}{\theta}\tilde{y}_1$ on the generic pair of proxies q_k and q_s, using other two proxies q_l and q_r as instruments, does not work in this case to estimate γ_1 and γ_2 without imposing any restrictions on the parameters. Indeed, in the regression model

$$y_2 - \frac{1}{\theta}\tilde{y}_1 = \frac{\gamma_1 \alpha_{s2} - \gamma_2 \alpha_{s1}}{\alpha_{s2}\alpha_{k1} - \alpha_{s1}\alpha_{k2}} q_k + \frac{\gamma_2 \alpha_{k1} - \gamma_1 \alpha_{k2}}{\alpha_{s2}\alpha_{k1} - \alpha_{s1}\alpha_{k2}} q_s + \nu,$$

where ν is a composite error term (which contains v_k and v_s, and correlates with q_k and q_s), the parameters of interest are not identified, being the coefficients relative to q_k ans q_s linear combinations of them.

Anyway, in Section 5.3, I will show that weak restrictions on the parameters (fixing the scale of $q_{01}^\star$ and $q_{02}^\star$, plus assuming that for each factor there exist a proxy completely saturated by that factor) allow to identify γ_1 and γ_2 by using IV.

different aspects of education, such as teaching, infrastructures, selectivity and the like. Therefore, I suppose there are p latent factors and K proxies for school quality. With the simplifying notation γ for $\gamma_2 - \gamma_1$, the model becomes

$$\begin{cases} y_2^\star - y_1^\star = \gamma_1 q_{01}^\star + \ldots + \gamma_p q_{0p}^\star + \{\epsilon_2 - \epsilon_1\} \\ q_1 = \alpha_{11} q_{01}^\star + \ldots + \alpha_{1p} q_{0p}^\star + v_1 \\ \vdots \\ q_K = \alpha_{K1} q_{01}^\star + \ldots + \alpha_{Kp} q_{0p}^\star + v_K \end{cases},$$

where the educational inputs potentially measure all school quality dimensions[4] and the disturbances $(\epsilon_2 - \epsilon_1, v_1, \ldots, v_K)$ are uncorrelated with each other and with the latent factors.

Generalizing the procedure for the one factor case,

- I first estimate θ by IV in the regression of $y_2 - \tilde{y}_1$ on $\tilde{y}_1$; see equation (5.1), where the nonobservability of $q_0^\star$, which is now a vector, is not a problem, because it enters the error term, that is uncorrelated with the instrumental variable z_0 owing to Assumption 1: whatever is the true dimension of $q_0^\star$, the estimate of θ can be obtained from (4.12); thus this first step, once it has been implemented for the unidimensional case, does not need to be replicated in the other cases.

- Then I apply a multiple factor analysis with $y_2 - \frac{1}{\theta}\tilde{y}_1$ as supplemental proxy for the multidimensional $q_0^\star$, in addition to $q_1, \ldots, q_K$:

$$y_2 - \frac{1}{\theta}\tilde{y}_1 = \gamma_1 q_{01}^\star + \ldots + \gamma_p q_{0p}^\star + \left\{\epsilon_2 - \epsilon_1 + u_2 - \frac{1}{\theta}\tilde{u}_1\right\}.$$

  We deal with a multiple factor model with $K+1$ indicators and p latent factors, whose identification issues are discussed in the following section; see Lawley and Maxwell [1971] for a detailed account of the identifiability of the factor analysis model.

5.2 Multiple factor model

5.2.1 Identification issues

The extended multiple factor model of the second step is

$$\begin{cases} y_2 - \frac{1}{\theta}\tilde{y}_1 = \gamma_1 q_{01}^\star + \ldots + \gamma_p q_{0p}^\star + \nu \\ q_1 = \alpha_{11} q_{01}^\star + \ldots + \alpha_{1p} q_{0p}^\star + v_1 \\ \vdots \\ q_K = \alpha_{K1} q_{01}^\star + \ldots + \alpha_{Kp} q_{0p}^\star + v_K \end{cases}, \tag{5.2}$$

[4] I continue to condition upon X_0, for simplicity.

where $\nu = \epsilon_2 - \epsilon_1 + u_2 - \frac{1}{\theta}\tilde{u}_1$. Note that I am assuming θ known; in practice I will replace it with the consistent estimate obtained in the first stage. Let Σ be the covariance matrix of the indicators (the observable variables, in the terminology of factor analysis). The covariance structure associated with the model is given by

$$\Sigma \equiv \Lambda\Phi\Lambda' + \Theta,$$

where

$$\Lambda = \begin{pmatrix} \gamma_1 & \gamma_2 & \dots & \gamma_p \\ \alpha_{11} & \alpha_{12} & \dots & \alpha_{1p} \\ \vdots & \vdots & & \vdots \\ \alpha_{K1} & \alpha_{K2} & \dots & \alpha_{Kp} \end{pmatrix}$$

is the $(K+1) \times p$ matrix of the factor loadings,

$$\Phi = \begin{pmatrix} \mathrm{Var}(q_{01}^\star) & & & & \\ \mathrm{Cov}(q_{02}^\star, q_{01}^\star) & \mathrm{Var}(q_{02}^\star) & & & \\ \mathrm{Cov}(q_{03}^\star, q_{01}^\star) & \mathrm{Cov}(q_{03}^\star, q_{02}^\star) & \mathrm{Var}(q_{03}^\star) & & \\ \vdots & \vdots & \vdots & \ddots & \\ \mathrm{Cov}(q_{0p}^\star, q_{01}^\star) & \mathrm{Cov}(q_{0p}^\star, q_{02}^\star) & \mathrm{Cov}(q_{0p}^\star, q_{03}^\star) & \dots & \mathrm{Var}(q_{0p}^\star) \end{pmatrix}$$

is the $p \times p$ covariance matrix of the latent factors and

$$\Theta = \begin{pmatrix} \mathrm{Var}(\nu) & & & \\ & \mathrm{Var}(v_1) & & \\ & & \ddots & \\ & & & \mathrm{Var}(v_K) \end{pmatrix}$$

is the $(K+1) \times (K+1)$ covariance matrix of the disturbances, which turns out to be diagonal.

It is well-known that there is a problem of indeterminacy (the "rotational freedom") in the factor model when $p > 1$, since we can replace Λ by ΛM and $q_0^\star$ by $M'q_0^\star$, where M is a $(p \times p)$ orthogonal matrix ($M' = M^{-1}$), without changing the covariance structure of the model:

$$\Sigma = (\Lambda M)(M'\Phi M)(M'\Lambda') + \Theta$$

(see Lawley and Maxwell [1971]). In the classic factor analysis this issue of non-uniqueness of Λ and Φ is solved by imposing p^2 independent constraints on the parameters of these matrices. Usually, in an exploratory context characterized by the absence of a priori information, it is convenient to choose these restrictions so that Φ is the identity matrix (uncorrelated factors with scale unity equal to one standard deviation; $\frac{1}{2}p(p+1)$ restrictions) and $\Lambda'\Theta^{-1}\Lambda$ is diagonal[5] ($\frac{1}{2}p(p-1)$ additional restrictions); see Joreskog [1967].[6] Consequently, the number of free parameters in the model becomes $(K+1)+p(K+1)-\frac{1}{2}p(p-1)$[7] and the degrees of freedom of the model, i.e. the difference between the number of nonredundant elements in the covariance matrix and the number of free parameters, is given by

$$df = \frac{1}{2}[(K+1-p)^2 - (K+1+p)].$$

Therefore, the model is identified, and we can obtain an estimate for the parameters of interest $\gamma_1, \ldots, \gamma_p$, only if the number of proxies (K) and the dimension of $q_0^\star$ (p) are such that $df \geq 0$.[8] In the case of $p = 2$ latent factors there must be at least $K = 4$ proxies for school quality and the model is overidentified ($df > 0$) also with the smallest possible number of proxies, $K = 4$ ($df = 1$). Increasing successively p by one ($p = 3, 4, \ldots$) the minimum number of necessary proxies becomes 5, 7 and so on ($df = 0, 2, \ldots$).

Anyway, the restrictions on Λ and Φ are imposed only in a first phase to make the loadings unique and estimate in a convenient way the model through the usual estimation techniques, such as the Maximum Likelihood method[9], assuming a multivariate normal distribution for observations. Then, in a second step, after the estimation of the model, the original set of factors is transformed into another set of factors in order to achieve a more meaningful interpretation. This transformation to a more simple structure is the result of the post-multiplication of Λ by an appropriate rotation (orthogonal) matrix M.[10] After the rotation, the new matrix $\Lambda^* = \Lambda M$ will show a more clear pattern: it will have as many elements close to zero as possible and each indicator will explain only a small subset of common factors (or, in other words, each common factor will affect only a small portion of indicators). Usually, the most meaningful interpretation is obtained by means of an "oblique" rotation, which produces a set of correlated factors; this is not a

[5] The restriction $\Lambda'\Theta^{-1}\Lambda$ diagonal translates into choose the factor loadings in such a way that the first factor explains the major part of variation in the indicators, the second factor explains the remaining greatest part, conditionally on being uncorrelated with the first one, and so forth.

[6] Alternatively, we can impose specific values for certain elements of Λ and Φ. Usually, instead of constraining $\Lambda'\Theta^{-1}\Lambda$ to be diagonal, $\frac{1}{2}p(p-1)$ elements of Λ are fixed equal to zero (exclusion restrictions), so that the chosen pattern of zero's is destroyed by any rotation, i.e. any multiplication of Λ by an orthogonal matrix M. This is the typical approach adopted in confirmatory factor analysis, where the researcher has a priori information about which indicators are involved with each factor from the subject matter theory.

[7] Before any restrictions, there are $(K+1)+p(K+1)+\frac{1}{2}p(p+1)$ parameters to estimate in the model—respectively in Θ, Λ and Φ.

[8] Exactly, this is a necessary but not sufficient condition for identification, because one or more residual variances in Θ could come out negative in phase of estimation (Heywood cases).

[9] The constraint $\Lambda'\Theta^{-1}\Lambda$ diagonal holds automatically maximizing the likelihood function.

[10] As I have just said, the covariance structure of the data does not change.

drawback because it is reasonable that the underlying latent variables, that now have a precise meaning, are correlated with each other[11].

In the present study, where I want to explain the proxies q_k in terms of a small number of hypothetical factors of school quality, I will choose the rotation procedure which allows me to achieve the most clear interpretation: for instance, in the case of two dimensions of school quality, the factors could describe teachers' quality (measured solely by proxies relating to teaching) and the selectivity of the school (measured solely by proxies relating to selectivity), respectively. Clearly, there could exit a correlation between these aspects of education.

5.2.2 The number of latent factors

The choice of the number of factors is another issue to be dealt with. Several techniques have been developed in explorative factor analysis, such as mine, to determine the number of latent factors. Usually this is viewed as a problem of model selection, where different models, with different numbers of factors, are compared on the basis of their fit to the data; at the end the factor model with the best fit, or with an acceptable fit but an easier interpretation, is chosen to describe the data. Note that to carry out this evaluation before or after the rotation of the factors is equivalent, because the fit of the model does not change after the rotation: the values of the goodness-of-fit statistics used to evaluate the model remain unchanged[12]. Other more applied rules to decide the value of p are the Kaiser criterion and the scree test, which the interested reader is addressed to in Wansbeek and Meijer [2000].[13]

---

[11]Among the most frequently used rotation methods that maintain and do not maintain the orthogonality of the factors, there are the *varimax* method and the *oblimin* method, respectively. The first chooses M and the final factor loadings in such a way to maximize a measure of simplicity of Λ^* given by the average of the variances of the squared loadings in each column. This produces both large and small loadings in absolute magnitude for each factor; see Lawley and Maxwell [1971]. The latter, instead, minimizes the correlation between columns of Λ^* with the result that each indicator explains only a small subset of factors; see Harman [1976].

[12]The goodness-of-fit statistics are measures of the discrepancy between the observed and the estimated values of the elements of Σ; these last ones remain the same choosing different matrices M. Clearly, the fit of the model is perfect when $df = 0$ (there are as many parameters as there are moment equations).

[13]The Kaiser criterion and the scree test are related to principal component analysis (PCA) and involve the computation of the eigenvalues of the correlation matrix of observations. Briefly, the first method suggests to choose p equal to the number of eigenvalues larger than one. Since in PCA each component corresponds to an eigenvalue, which is its variance, it would be silly to introduce components whose variance is smaller than the variance of the generic indicator (equal to one because of standardization). Instead, with the second method we have to represent on a plot the eigenvalues in order of size and connect the relative points with straight lines: the rank position of the eigenvalue after which the other eigenvalues decline only gradually determines

In my application, I will focus on the case $p = 3$, because I want to simplify as much as possible the model. Further, I will show that there is not a significant improvement in the fit of the model to the data increasing p by one; the simplest model with two factors presents, instead, a bad fit.

5.3 Estimation strategy

In phase of estimation I will apply factor analysis with an exploratory purpose to obtain some indications on the real model underlying the data, in terms of number of factors and their interpretation, i.e. which proxies are affected by each factor and in what measure they do. Hence, I will exploit this evidence plus my knowledge from the literature on school quality to make two weak constraints on the parameters (exclusion restrictions), which will allow me to estimate the model in a convenient way through a straightforward extension of the Instrumental Variables approach described in Chapter 4 for the one factor case.

To make easier my explanation, let me consider the simplest case of two dimensions and four proxies of school quality. Replacing p by 2 and K by 4, the factor analysis model (5.2) becomes

$$\begin{cases} y_2 - \frac{1}{\theta}\tilde{y}_1 = \gamma_1 q^\star_{01} + \gamma_2 q^\star_{02} + \underbrace{\left\{\epsilon_2 - \epsilon_1 + u_2 - \dfrac{1}{\theta}\tilde{u}_1\right\}}_{\nu} \\ q_1 = \alpha_{11} q^\star_{01} + \alpha_{12} q^\star_{02} + v_1 \\ q_2 = \alpha_{21} q^\star_{01} + \alpha_{22} q^\star_{02} + v_2 \\ q_3 = \alpha_{31} q^\star_{01} + \alpha_{32} q^\star_{02} + v_3 \\ q_4 = \alpha_{41} q^\star_{01} + \alpha_{42} q^\star_{02} + v_4 \end{cases}.$$

As said above, we have to remove the intrinsic indeterminacy of the factor analysis model (the "rotational freedom") in order to identify the parameters. This means that we have to impose at least $p^2 = 4$ independent restrictions on the parameters, as I have explained in Section 5.2.1. My strategy is to (i) assume for each latent factor the existence of one proxy that measures only that factor and to (ii) assign to both factors the scale of the proxies completely saturated by them. For instance, in our exemplification, we can assume that q_1 and q_2 only measure and have the same scale of $q^\star_1$ and $q^\star_2$, respectively ($\alpha_{12} = \alpha_{21} = 0$ and $\alpha_{11} = \alpha_{22} = 1$).

Clearly, the second restriction does not cause any loss of generality. In explorative factor analysis fixing, instead, the variance of the latent factors to one,

---

the number of factors; the addition of each component whose eigenvalue is after the "jump" does not improve significantly the explanation of the total variance (the sum of the variances of the indicators is equal to the sum of the eigenvalues in PCA).

thereby standardizing them, is more usual. Here, the chosen parameterization implies that one unit change in the latent variables translates into one unit change in the correspondent saturated proxies (on average, of course). These two criteria are quite equivalent and interchangeable, and lead to the same solutions, even though with different values of the parameters, because of referring to latent factors measured on different scales. With respect to the first condition, it could be a rather strong assumption. But the evidence from factor analysis, especially after rotations, could suggest which proxy loads more strongly on each factor having only a minor loading on the other one.

The model simplifies to

$$\begin{cases} y_2 - \frac{1}{\theta}\tilde{y}_1 = \gamma_1 q_{01}^{\star} + \gamma_2 q_{02}^{\star} + \nu \\ q_1 = q_{01}^{\star} + v_1 \\ q_2 = q_{02}^{\star} + v_2 \\ q_3 = \alpha_{31} q_{01}^{\star} + \alpha_{32} q_{02}^{\star} + v_3 \\ q_4 = \alpha_{41} q_{01}^{\star} + \alpha_{42} q_{02}^{\star} + v_4 \end{cases},$$

where the first equation can be rewritten as

$$y_2 - \frac{1}{\theta}\tilde{y}_1 = \gamma_1 q_1 + \gamma_2 q_2 + \{\nu - \gamma_1 v_1 - \gamma_2 v_2\}.$$

At this point it is clear that we can proceed analogously to Chaper 4 for the unidimensional case and estimate γ_1 and γ_2 by IV using q_3 and q_4 as instruments for the endogenous q_1 and q_2 (these last ones being correlated with the error term through their measurement errors v_1 and v_2, respectively).

This method can be easily extended to the general case of K proxies and p factors. The only requirement is the presence of p "saturated" proxies (one for each factor) plus, at least, other p "instrumental" proxies [14], because of the *order condition* for identification of an equation with p endogenous regressors. Hence this estimation strategy needs $K \geq 2p$ proxies in order to work, differently from other usual methods for factor analysis estimation where a lower number of proxies is necessary—K such that $df = \frac{1}{2}[(K+1-p)^2-(K+1+p)] \geq 0$. Furthermore, this procedure allows to estimate only the parameters of interest γ_1 and γ_2. However, we can give an interpretation to the latent factors by means of the inspection of the first stage regressions of the IV estimator.

This estimation approach has the double advantage of not being iterative and giving a direct estimate of standard errors for the parameters of interest, in our case robust to the heteroskedasticity and correlation of the error terms induced by the clustering of students in schools, using the Huber-White formula. Finally,

[14] Clearly, there must be at least one "instrumental" proxy affected by each factor to make work the IV strategy.

also the scale parameter θ can be estimated in only one step along with γ_1 and γ_2. We have to regress the raw value added $y_2 - \tilde{y}_1$ on $\tilde{y}_1, q_1$ and q_2, jointly, and instrument $\tilde{y}_1$ as well:

$$y_2 - \tilde{y}_1 = \left(\frac{1-\theta}{\theta}\right)\tilde{y}_1 + \gamma_1 q_1 + \gamma_2 q_2 + \left\{\epsilon_2 - \epsilon_1 + u_2 - \frac{1}{\theta}\tilde{u}_1 - \gamma_1 v_1 - \gamma_2 v_2\right\}. \quad (5.3)$$

5.4 Results

Table 5.1 reports the values estimated for the parameters of a three-factor model applied to the adjusted value added (PISA score minus rescaled teacher's mark) and the whole set of proxies. Precisely, I use all proxies of Section 4.5, excepting the freshman retention rate and the proportion of successful freshmen, which I hold to be poorer measures of quality at the high school level than at the college level, especially in these last few years for the Italian educational system. In the light of previous findings, using two proxies at the time, I also show results conditionally upon school enrolment, according to the literature, where the effects of the typical school quality measures, such as class size, teachers' experience and education, etc., are evaluated leaving fixed the number of students in the school, that I include among the control variables X_0. The correlation matrix of the proxies is given in Appendix A, where furthermore I report the correlations between the residuals of the proxies from the regression on X_0 as well as X_0 plus school size, which are effectively analysed.

As already pointed out in Section 5.1, θ is estimated from the regression of the difference between PISA score and teacher's mark on the latter, using the variable loss of academic months before high secondary school as instrument for the endogenous teacher's grade. The value of the t statistic on the instrument is -6.21 in the first stage regression. The point estimate of θ is, instead, 0.017, which is significant ($t = 5.72$) and robust to an alternative simple estimation procedure described in Appendix B.[15]

The method of maximum likelihood has been applied to estimate the factor analysis model. The particular solution presented in Table 5.1 refers to the oblimin rotation of the factors[16], but the changes in the values estimated for the factor loadings are not sizeable choosing different orthogonal and oblique rotations, which means that the latent factors are likely to be uncorrelated. A confirm is given by the low values of the t statistics associated to the covariances of the factors.

[15] In Appendix C I also show the homogeneity of the parameter across different population subdivisions.

[16] The oblimin criterion rotates factors in such a way that the new loadings tends to be zero in one column if there is a high value in another column of the same row (Harman [1976]).

Table 5.1 – Maximum likelihood estimates for the three factor model

A – Factor loadings

variable	factor 1	factor 2	factor 3
value added	2.044	4.931	-.370
	.025	**.063**	**-.007**
	(.735)	(1.610)	(-.115)
class size	-.059	.050	.230
	-.145	**.090**	**.446**
	(-.475)	(.374)	(1.535)
student-teacher ratio	.056	.021	.274
	.087	**.033**	**.617**
	(.410)	(.139)	(1.769)
pc-student ratio	-.030	.154	-.208
	-.038	**.280**	**-.374**
	(-.208)	(.910)	(-1.172)
prop. pcs connected to www	.688	.0113	.059
	.794	**-.002**	**.044**
	(1.944)	(.040)	(.260)
prop. pcs connecte to lan	.354	.047	-.182
	.429	**.052**	**-.229**
	(1.250)	(.193)	(-.783)
pc use	.136	.079	-.031
	.217	**.124**	**-.059**
	(1.278)	(.751)	(-.306)
prop. of permanent teachers	.019	.118	-.001
	.072	**.578**	**-.021**
	(.280)	(1.990)	(-.029)
prop. of full-time teachers	.003	.122	-.040
	.002	**.162**	**-.056**
	(.031)	(1.111)	(-.347)
prop. of certified teachers	-.000	.290	.122
	-.031	**.493**	**.200**
	(-.003)	(2.020)	(.900)
teachers' seniority	-.049	.271	-.077
	-.079	**.410**	**-.121**
	(-.298)	(1.519)	(-.465)

* Standardized factor loadings are in bold. Cluster robust t statistics are given in parenthesis.

B – Correlations between factors

	factor 1	factor 2	factor 3
factor 1	1.000		
factor 2	.053 (.547)	1.000	
factor 3	.095 (.989)	.041 (.418)	1.000

*Cluster robust t statistics are given in parenthesis.

C – Goodness-of-fit tests

LRT(3 factors model vs. saturated model)
385.407
95 percentile
950.770

LRT(3 factors model vs. 4 factors model)
258.314
95 percentile
617.315

Standard errors have been computed via bootstrap, doing resampling over schools rather than students, to take into account the within-school dependence of observations[17]. Besides, bootstrap has been used to evaluate the fit of the model to the data and the improving in the fitting by the addition of a new factor, since the chi-square variable is no more suitable for describing the distribution of the likelihood ratio test (LRT) statistic[18]. We accept the null hypothesis of a three factor model underlying the data against the alternative hypotheses of both saturated model and model with four factors. Compare the LRT statistics and their

[17] The clustering of students in schools implies that the maximum likelihood estimator itself is wrong because the maximized likelihood function assumes independent observations. Anyway, results do not change with other estimation techniques and/or considering only one student per school.

[18] The model with only two factors is not taken into consideration because it presents a bad overall fit to the data.

95 percent percentiles from the simulated bootstrap distribution, for both tests, in panel C of Table 5.1. However small the significance level is, we do not reject the null hypotheses.

The three factors are rather easy to interpret by an inspection of factor loadings. Consider, for simplicity of interpretation, the standardized loadings, whose absolute values lie between 0 and 1, being associated to the standardized variables, i.e. to factor analysis applied to the correlation matrix instead of to the covariance matrix of the data. The first factor has high loadings on the proportions of computers connected to Internet and to a local network, and on pc use, thus it can be intended as the state and the effective use of computational resources at school[19], rather than their mere quantity, given by the number of computers per student, which has only a negligible correlation with the first factor[20]. Instead the second factor, which has larger loadings in correspondence of the proportion of permanent and certified teachers, and teachers' experience can be related to the quality of teaching. Finally, the first two proxies, class size and student-teacher ratio, as well as the number of computers per student, are mostly correlated with the third factor, which can be viewed as the amount of resources, both "teaching time" (human resources) and computers (physical resources), available to each student in the school[21].

After these considerations, to decide which proxies are "saturated" by each factor is straightforward. We see that the proportion of computers connected to Internet, the fraction of permanent teachers' and the student-teacher ratio have the highest standardized loading in correspondence of the first, second and third factor, respectively. They also have negligible loadings on the remaining factors. So these variables are those which we have to include as regressors in equation (5.3), in the specific case considered of three latent dimensions of school quality. Further, this choice for the "instrumented" proxies is supported by the bootstrap estimation of standard errors, which translate into higher values of the t statistics for these proxies. They have only very small values of the t statistic in correspondence of the other two factors.

There could be uncertainty whether to choose the proportion of permanent teachers or the proportion of certified teachers to fully explain the second factor of teaching quality. I opt for the first one, despite its lower significance, because the

[19]Perhaps, the first indicator could be identified with the conditions of all educational resources at school (libraries, science laboratories, etc.), having access to other proxies.

[20]In what way and how much students use computers at school, rather than how many computers there are at school, is important to evaluate (Rouse and Krueger [2004]). Indeed, in a school with many computers, students could use them with outdated or ineffective software and/or they could never go to computer laboratory.

[21]Pay attention that teachers per student and "minus" class size are indicators of quality with this interpretation. Note the negative sign of the computer-student ratio in the third column.

other two loadings (on the first and third factors) are zero and I believe it to be a better measure of teachers' quality, than teachers' certification, to my knowledge: as discussed in Section 4.5, teachers with a permanent contract may be more motivated and more effective in producing instruction, because of working in better conditions of teaching continuity and stability; they also have more experience, which has been found out to be an important determinant of students' achievement, contrary to teachers' education, in the literature (see Hanushek, Kain and Rivkin [2005])

As regards teachers' quality, it seems to have a positive effect on the value added, by looking at the bootstrap t statistic, significant at the 10 percent level. This finding somewhat agrees with the results reported in Table 5.3 for the IV estimation (following the procedure described in the previous section), which indicate the absence of any impact of school quality on the gain in achievement of students, with the exception, to some extent, of teaching quality, which is, however, insignificant at reasonable significance levels[22]. Anyway, also findings in literature demonstrate that if there is an effect of school quality, it comes from teachers rather than from other educational inputs (see Hanushek and Woessmann [2007] for references in this regard).

The proxies which come out statistically significant in the first stage regressions confirm the interpretation of the factors given above with factor analysis estimation. The variable 'loss of months' is a good instrument for teacher's mark and we again obtain a value of θ around 0.020, in a confidence interval lying between 0.010 and 0.030, roughly.

Without conditioning upon school enrolment, the IV estimation of the same model yields results fairly different, in that both the coefficients on the proportion of permanent teachers and the student-teacher ratio become significant[23], the latter to a more extent and in a counter-intuitive positive direction;[24] cf. panel A in Table 5.4. By assuming the existence of an additional dimension of school quality completely accounted for by school enrolment, the other three dimensions return to be insignificant and only school size positively affects students' learning; cf. panel B in Table 5.4 (student-teacher ratio, although insignificant, has the expected negative sign). Unfortunately, whether school size is a genuine school quality dimension additional to the factor "resources per student", or it is simply another measure of it or a confounding factor to control for, is not clear by my analyses conducted so far. More accurate investigations on the correct specification of the

[22]Note that the coefficients on the factors are not directly comparable with the two estimation procedures, because the factors are defined on different scales (see Section 5.3).

[23]In this case factor analysis shows that the school quality dimension "resources per student" is measured by school and class size, and student-teacher ratio, which is the most strongly correlated.

[24]See Section 4.5 for various justifications from literature.

Table 5.3 – Instrumental Variables estimates for the three factor model

First stage regressions

student-teacher ratio	coef.	std. err.	t	p-value
class size	.195	.058	3.32	.001
pc-student ratio	-.145	.044	-3.26	.001
prop. pcs connected to lan	-.013	.041	-.33	.738
pc use	.018	.046	.39	.697
prop. full-time teachers	-.002	.038	-.07	.947
prop. certified teachers	.090	.067	1.35	.177
teachers' seniority	-.060	.048	-1.25	.213

adjusted R^2=0.1299
p-value F-test=0.0000

prop. permanent teachers	coef.	std. err.	t	p-value
class size	-.000	.029	-.00	.998
pc-student ratio	.040	.027	1.50	.136
prop. pcs connected to lan	.026	.016	1.64	.103
pc use	-.018	.020	-.89	.376
prop. full-time teachers	.037	.020	1.84	.067
prop. certified teachers	.087	.027	3.14	.002
teachers' seniority	.048	.023	2.06	.041

adjusted R^2=0.1664
p-value F-test=0.0001

prop. pcs connected to www	coef.	std. err.	t	p-value
class size	-.039	.117	-.34	.737
pc-student ratio	-.151	.099	-1.53	.128
prop. pcs connected to lan	.311	.067	4.59	.000
pc use	.220	.099	2.22	.028
prop. full-time teachers	.018	.083	.22	.829
prop. certified teachers	.051	.095	.54	.592
teachers' seniority	-.080	.090	-.90	.371

adjusted R^2=0.1386
p-value F-test= 0.0001

Second stage regression

value added	coef.	std. err.	t	p-value
student-teacher ratio	7.113	11.001	.65	.518
prop. permanent teachers	32.150	21.584	1.49	.136
prop. pcs connected to www	3.466	5.021	.69	.490

adjusted $R^2 = 0.349$
p-value Hansen stat. = 0.215
Number of obs. = 5092

Table 5.4

A – Instrumental Variables estimates for the three factor model without conditioning upon school enrolment

value added	coef.	std. err.	t	p-value
student-teacher ratio	25.726	10.130	2.54	.011
prop. pcs connected to www	4.202	5.251	.80	.424
prop. permanent teachers	48.743	24.241	2.01	.044

B – Instrumental Variables estimates for the four factor model without conditioning upon school enrolment

value added	coef.	std. err.	t	p-value
school enrolment	29.788	12.086	2.46	.014
student-teacher ratio	-30.232	24.289	-1.24	.213
prop. pcs connected to www	-1.661	7.277	-.23	.819
prop. permanent teachers	33.716	25.506	1.32	.186

model are necessary; they are left to the future by the use of other school quality proxies.

Appendix A

Correlations of the proxies

In the following pages I report the correlation matrix of the proxies exploited in the empirical analysis. The residuals from the regressions of the proxies on X_0 and X_0 plus school enrolment are also considered. See Section 5.4 and Section 3.2 for a description of the control variables contained in X_0.

Table A.1 – Correlation matrix of the proxies without taking residuals

	school enrolment	class size	student-teacher ratio	pc-student ratio	prop. pcs connected to www	prop. pcs connected to lan	pc use	prop. permanent teachers	prop. full-time teachers	prop. certified teachers	teachers' seniority
school enrolment	1.0000										
class size	0.5709*	1.0000									
student-teacher ratio	0.2192*	0.4630*	1.0000								
pc-student ratio	-0.4223*	-0.4498*	-0.3919*	1.0000							
prop. pcs connected to www	0.1074*	0.0838*	0.1338*	0.0107	1.0000						
prop. pcs connected to lan	0.0167	-0.0594*	0.0134	0.1160*	0.4195*	1.0000					
pc use	-0.1904*	-0.2186*	-0.3267*	0.4383*	0.1298*	0.0883*	1.0000				
prop. permanent teachers	0.2763*	0.3761*	0.2378*	-0.2082*	-0.0760*	-0.1285*	-0.1433*	1.0000			
prop. full-time teachers	0.2500*	0.1341*	0.0192	-0.1387*	-0.0134	-0.0599*	-0.0544*	0.1088*	1.0000		
prop. certified teachers	0.3465*	0.2638*	0.1823*	-0.2359*	-0.0069	-0.0163	-0.1444*	0.2482*	0.1441*	1.0000	
teachers' seniority	0.0594*	-0.0098	0.0045	0.1204*	-0.0022	0.0137	0.0542*	0.4872*	0.0947*	0.2389*	1.0000

*Correlations significant at the 5 percent level have the asterisk.

Table A.2 – Correlation matrix of the residuals of the proxies regressed on X_0

	school enrolment	class size	student-teacher ratio	pc-student ratio	prop. pcs connected to www	prop. pcs connected to lan	pc use	prop. permanent teachers	prop. full-time teachers	prop. certified teachers	teachers' seniority
school enrolment	1.0000										
class size	0.3876*	1.0000									
student-teacher ratio	0.1879*	0.2994*	1.0000								
pc-student ratio	-0.2957*	-0.2347*	-0.2532*	1.0000							
prop. pcs connected to www	0.0945*	0.0506*	0.1313*	-0.0063	1.0000						
prop. pcs connected to lan	0.0963*	-0.0483*	-0.0227*	0.0811*	0.3616*	1.0000					
pc use	-0.1084*	-0.0433*	0.0234*	0.2029*	0.2386*	0.0767*	1.0000				
prop. permanent teachers	0.0847*	-0.1459*	0.1086*	0.0280*	0.0140	0.0562*	0.0317*	1.0000			
prop. full-time teachers	0.1628*	0.0309*	-0.0148	-0.0030	0.0110	-0.0438*	0.0512*	0.1627*	1.0000		
prop. certified teachers	0.0596*	-0.0410*	0.0057	0.0551*	-0.0550*	0.0367*	0.0356*	0.3476*	0.0132	1.0000	
teachers' seniority	0.0116	-0.1149*	-0.0245*	0.0386*	-0.0719*	0.0436*	-0.0396*	0.3538*	-0.0214	0.2491*	1.0000

*Correlations significant at the 5 percent level have the asterisk.

Table A.3 – Correlation matrix of the residuals of the proxies regressed on X_0 plus school enrolment

	class size	student-teacher ratio	pc-student ratio	prop. pcs connected to www	prop. pcs connected to lan	pc use	prop. permanent teachers	prop. full-time teachers	prop. certified teachers	teachers' seniority
class size	1.0000									
student-teacher ratio	0.2625*	1.0000								
pc-student ratio	-0.1403*	-0.2173*	1.0000							
prop. pcs connected to www	0.0188	0.1219*	0.0105	1.0000						
prop. pcs connected to lan	-0.0807*	-0.0286*	0.1098*	0.3614*	1.0000					
pc use	-0.0164	0.0494*	0.1783*	0.2505*	0.0946*	1.0000				
prop. permanent teachers	-0.1960*	0.0991*	0.0645*	0.0046	0.0323*	0.0422*	1.0000			
prop. full time teachers	-0.0413*	-0.0435*	0.0401*	-0.0060	-0.0703*	0.0681*	0.1545*	1.0000		
prop. certfied teachers	-0.0514*	-0.0137	0.0899*	-0.0420*	0.0351*	0.0652*	0.3930*	0.0096	1.0000	
teachers' seniority	-0.1148*	-0.0109	0.0294*	-0.0727*	0.0260*	-0.0261*	0.3446*	-0.0120	0.2678*	1.0000

*Correlations significant at the 5 percent level have the asterisk.

Appendix B

Robustness checks

B.1 An alternative method for estimating θ

The estimate of θ that I have employed up until now, under Assumption 1 and Assumption 2 of Chapter 4, is the sample counterpart of

$$\frac{\mathrm{Cov}(\tilde{y}_1, z_0)}{\mathrm{Cov}(y_2, z_0)}$$

(cf. equation (4.12)). This yields a value of θ equal to 0.017 in a confidence interval between 0.010 and 0.030, roughly. In this section I present another method to identify θ when relaxing Assumption 1.

Define y_1 and u_1 as $\frac{1}{\theta}\tilde{y}_1$ and $\frac{1}{\theta}\tilde{u}_1$, respectively. The measurement equation for the teacher's mark becomes

$$y_1 = y_1^\star + u_1.$$

Since February is close to April, it is likely that $\mathrm{Var}(y_2^\star) = \mathrm{Var}(y_1^\star)$ and $\mathrm{Var}(u_2) = \mathrm{Var}(u_1)$, because the heterogeneity in individual growth has had no time to manifest itself ("assumption of dynamic equilibrium"; see Lord [1963] and Willet [1988]). Hence $\mathrm{Var}(y_2) = \mathrm{Var}(y_1)$ and

$$\theta = \frac{\sqrt{\mathrm{Var}(\tilde{y}_1)}}{\sqrt{\mathrm{Var}(y_2)}}.$$

The substitution of the sample counterparts leads to a point estimate of θ equal to 0.021, which lies in the confidence interval, written above, from the first strategy.

B.2 Heterogeneity of θ

The subjective teacher's evaluation seems to vary, in Italy, with type of school (lyceum, technical school and vocational school), geographical area (North, Centre and South plus Islands) and sex of the student. See, for instance, the Italian report *Rapporto regionale del Veneto OCSE-PISA 2003*, where descriptive statistics show that the teacher's evaluation is more severe in lyceums and technical schools, in North and for males; cf. Figure B.1.

I repeat the procedures described in the previous sections allowing for heterogeneity in θ with respect to gender, type of school and geographical area. This is to say that I assume a variable value of θ across a certain number S of strata:

$$\tilde{y}_1 = \theta_1 y_1^\star I_1 + \theta_2 y_1^\star I_2 + \ldots + \theta_S y_1^\star I_S + \tilde{u}_1,$$

where I_s is an indicator function equal to one if the student belongs to stratum s, $s = 1, \ldots, S$. This implies that our model becomes

$$y_2 - \tilde{y}_1 = \left(\frac{1-\theta_1}{\theta_1}\right) \tilde{y}_1 I_1 + \ldots + \left(\frac{1-\theta_S}{\theta_S}\right) \tilde{y}_1 I_S + (\gamma_2 - \gamma_1) q_0^\star + \left\{u_2 - \frac{\tilde{u}_1}{\theta} + \epsilon_2 - \epsilon_1\right\}.$$

So now $z_0 I_s$ is the instrument for the generic variable $\tilde{y}_1 I_s$.

Empirical results lead to accept the hypothesis of homogeneity of θ across strata. This follows from the fact that θ is simply a parameter which regulates the difference in scale between teacher's mark and PISA score, and the variables in question—sex, geographical area and school track—have an effect only on position, which is removed by taking the residuals on X_0, just including genre, province and type of school; cf. Figure B.2.

Figure B.1 – Fitted values from the regression of the PISA score on teacher's mark by genre, geographical area and type of school – original variables, before residuals

Figure B.2 – Fitted values from the regression of the PISA score on teacher's mark by genre, geographical area and type of school – residuals

Bibliography

Adams R.J., Wilson M.R. and Wang W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21, 1–24.

Addonizio M.F. and Phelps J.L. (2000). Class size and student performance: a framework for policy analysis. *Journal of Education Finance*, 25(2), 135–156.

Akerhielm (1995). Does class size matter? *Economics of Education Review*, 14(3), 229–241.

Altonji J.G. and Dunn T.A. (1996). Using siblings to estimate the effect of school quality on wages. *The Review of Economics and Statistics*, 78(4), 665–671.

Altonji J.G. and Pierret C.R. (2001). Employer learning and statistical discrimination. *Quarterly Journal of Economics*, 116(1), 313–350.

Angrist J. and Lavy V. (1999). 'Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics*, 114(2), 533–575.

Angrist J. and Lavy V. (2002). New evidence on classroom computers and pupil learning. *The Economic Journal*, 112, 735–765.

Armor D.J. , Conry-Oseguera P., Cox M., King N., McDonnell L., Pascal A., Pauly E. and Zellman G. (1976). *Analysis of the school preferred reading program in selected Los Angeles minority schools*. Rand Corp., Santa Monica, CA.

Baum C.F., Schaffer M.E. and Stillman S. (2002). Instrumental variables and GMM: estimation and testing. Boston College Economics Working Paper, 545.

Behrman J.R., Rosenzweig M.R. and Taubman P. (1996). College choice and wages: estimates using data on female twins. *The Review of Economics and Statistics*, 78(4), 672–685.

Bettinger E. and Long B.T. (2004). Do college instructors matter? The effects of adjuncts and graduate assistants on students interets and success. NBER Working Paper, 10370.

Bettinger E. and Long B.T. (2005). Help or hinder? Adjunct professors and student outcomes. Mimeo.

Betts J.R. (1995). Does school quality matter? Evidence from the National Longitudinal Survey of Youth. *The Review of Economics and Statistics*, 77(2), 231–247.

Betts J.R. (1996). Do school resources matter only for older workers. *The Review of Economics and Statistics*, 78(4), 638–652.

Black D., Daniel K. and Smith J. (2005). College quality and wages in the United States. *German Economic Review*, 6, 415–443.

Black D. and Smith J. (2004). How robust is the evidence on the effects of college quality? Evidence from matching. *Journal of Econometrics*, 121, 99–124.

Black D. and Smith J. (2006). Estimating the returns to college quality with multiple proxies for quality. *Journal of Labor Economics*, 24, 701–728.

Bollen K. (1989). *Structural equations with latent variables*. New York, Wiley.

Borjas G. (2000). Foreign-born teaching assistants and the academic performance of undergraduates. NBER Working Paper, 7635.

Borland M.V. and Howsen R.M. (1992). Student academic achievement and the degree of market concentration in education. *Economics of Education Review*, 11(1), 31–39.

Bowen W.G. and Bok D. (1998). The shape of the river: long-term consequences of considering race in college and university admissions. Princeton University Press, Princeton, NJ.

Brand J. (2000). Matching on propensity scores to estimate the effects of graduating from an elite college on early- end mid-career outcomes. Unpublished MS thesis, University of Wisconsin, Madison.

Brandt J.E. and Halaby C.N. (2003). Regression and matching estimates of the effects of elite college attendance on labor market outcomes. University of Wisconsin, Madison.

Brandt D., Rogosa D.R. and Zimowsky M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin*, 93, 726–748.

Bratti M., Checchi D. and Filippin A. (2007). Territorial differences in Italian students' mathematical competences: evidence from PISA 2003. IZA Discussion Paper, 2603.

Brewer D., Eide E. and Ehrenberg R. (1999). Does it pay to attend an elite private college? Cross-cohort evidence on the effects of college type on earnings. *Journal of Human Resources*, 34(1), 104–123.

Brunello G. and Capellari L. (2005). The labour market effects of Alma Mater: evidence from Italy. IZA Discussion Paper, 1562.

Brunello G. and Checchi D. (2003). School quality and family background in Italy. IZA Discussion Paper, 705.

Cameron A.C. and Trivedi P. (2005). *Microeconometrics*. Cambridge MA, Cambridge University Press.

Card D. (1995). Using geographic variation in college proximity to estimate the return to schooling. Aspects of Labour Market Behavior: Essays in Honour of John Vanderkamp, Edited by L. N. Christofides, E. Kenneth Grant, Robert Swidinsky, and John Vanderkamp, University of Toronto Press Incorporated, Toronto, 201–222.

Card D. and Krueger A.B. (1992a). Does school quality matter? Returns to education and the characteristics of public schools in the United States. *Journal of Political Economy*, 100, 1–40.

Card D. and Krueger A.B. (1992b). School quality and black-white relative earnings: a direct assessment. *Quarterly Journal of Economics*, 107, 151–200.

Card D. and Krueger A.B. (1996). Labor market effects of school quality: theory and evidence. In Burtless G. (eds) *Does money matter? The effect of school resources on student achievement and adult success*, Washington DC, Brookings Institution Press.

Checchi D. (2004). Da dove vengono le competenze scolastiche? L'indagine PISA 2000 in Italia. *Stato e Mercato*, 72, 413–453.

Chevalier A. and Conlon G. (2003). Does it pay to attend a prestigious university?. CEP Discussion Paper, LSE.

Coleman J.S., Campbell E.Q., Hobson C.J., McPartland J., Mood A.M., Weinfeld F.D. and York R.L. (1966). *Equality of Educational Opportunity.* Washington DC, U.S. Government Printing Office.

Dale S. and Krueger A. (2002). Estimating the payoff to attending a more selective college: an application of the selection of observables and unobservables. *Quarterly Journal of Economics*, 117(4), 1491–1528.

Daniel K., Black D. and Smith J. (1997). College quality of young men. Unpublished manuscript, Department of Economics, University of Maryland.

Dearden L., Ferri J. and Meghir C. (2002). The effect of school quality on educational attainment and wages. *The Review of Economics and Statistics*, 84(1), 1–20.

Dobbelsteen S., Levin J. and Oosterbeek H. (2002). The causal effect of class size on scholastic achievement: distinguishing the pure class size effect from the effect of changes in class composition. *Oxford Bulletin of Economics and Statistics*, 64, 17–38.

Dustman C., Rajah N. and van Soest A. (1998). School quality, exam performance and career choice. Mimeo, London, Institute for Fiscal Studies.

Dustmann C., Rajah N. and van Soest A. (2003). Class size, education and wages. *Economic Journal*, 113, F99–F120.

Ehrenberg R.G. and Zhang L. (2005). Do tenured and tenure-track faculty matter? *Journal of Human Resources*, 40(3), 647–659.

Feinstein L. and Symons J. (1999). Attainment in secondary schools. *Oxford Economic Papers*, 51(2), 300–321.

Finn J.D. and Achilles C.M. (1990). Answers and questions about class size: a statewide experiment. *American Educational Research Journal*, 27(3), 557–577.

Fuchs T. and Woessmann L. (2004). Computers and student learning: bivariate and multivariate evidence on the availability and use of computers at home and at school. CESifo Working Paper, 1321.

Goldhaber D.D. and Brewer D.J. (1998). Why don't schools and teachers seem to matter? Assessing the impact of unobservables on educational productivity. *The Journal of Human Resources*, 32(3), 505–523.

Goolsbee A. and Guryan J. (2006). The impact of Internet subsidies on public schools. *The Review of Economics and Statistics*, 88(2), 336–347.

Gould E., Lavy V. and Paserman D.M. [2004 . Immigrating to opportunity: estimating the effect of school quality using a natural experiment on Ethiopians in Israel. *The Quarterly Journal of Economics*, 119(2), 489–526.

Griliches Z. (1977). Estimating the returns to schooling: some econometrics problems. *Econometrica*, 45, 1–22.

Griliches Z. and Mason W.M. (1972). Education, income and ability. *Journal of Political Economy* 80, S74–S103.

Grogger J. (1996a). Does school quality explain the recent black-white wage trend? *Journal of Labor Economics*, 18(4), 231–253.

Grogger J. (1996b). School expenditures and post-schooling earnings: evidence from high school and beyond. *Review of Economics and Statistics*, 78(4), 628–637.

Hambleton R.K. and Swaminathan H. (1984). *Item Response Theory – Principles and Applications*. Boston, Kluwer.

Hanushek E.A. (1971). Teacher characteristics and gains in student achievement: estimation using micro data. *American Economic Review*, 60(2), 280–288.

Hanushek E.A. (1979). Conceptual and empirical issues in the estimation of educational production functions. *Journal of Human Resources*, 14, 351–388.

Hanushek E.A. (1986). The economics of schooling: production and efficiency in public school. *The Journal of Economic Literature*, 24, 1141–1177.

Hanushek E.A. (1992). The trade-off between child quantity and quality. *Journal of Political Ecoonomy*, 100(1), 84–117.

Hanushek E.A. (1999). Some findings from an independent investigation of the Tennessee STAR experiment and from other investigations of class size effects. *Educational Evaluation and Policy Analysis*, 21(2), 143–163.

Hanushek E.A. (2002). Publicy provided education. NBER Working Paper, 8799.

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

Hanushek E.A. (2005). The economics of school quality. *German Economic Review*, 6(3), 269–286.

Hanushek E.A. and Harbison R.W. (1992). *Educational performance of the poor: lessons from rural northeast Brazil*. New York, Oxford University Press.

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

Hanushek E.A. and Kimko D.D. (2000). Schooling, labor force quality and the growth of nations. *American Economic Review*, 90(5), 1184–1208.

Hanushek E.A. and Luque J.A. (2003). Efficiency and equity in schools around the world. *Economics of Education Review*, 22(5), 481–502.

Hanushek E.A., Rivkin S.G. and Taylor L.L. (1996). Aggregation and the estimated effects of school resources. *The Review of Economics and Statistics*, 78(4), 611–627.

Hanushek E.A. and Woessmann L. (2007). The role of school improvement in economic development. NBER Working Paper, 12832.

Hanushek E.A. and Zhang L. (2006). Quality consistent estimates of international returns to skill. NBER Working Paper, 12664.

Harman H.H. (1967). *Modern Factor Analysis*. Chicago, University of Chicago Press.

Harman H.H. (1976). *Modern Factor Analysis*. Chicago, 2nd ed., University of Chicago Press.

Harmon C., Oosterbeek H. and Walker I. (2003). The returns of education: microeconomics. *Journal of Economic Surveys*, 17(2), 115–155.

Heckman J.J., Lalonde R. and Smith J. (1999). The economics and econometrics of active labour market programs. In Ashenfelter A. and Card D. (eds) *Handbook of Labour Economics*, vol. 3, Amsterdam.

Heckman J., Layne-Farrar A. and Todd P. (1995). Does measured school quality really matter? An examination of the earnings-quality relationship. NBER Working Paper, 5274.

Heckman J., Layne-Farrar A. and Todd P. (1996). Human capital pricing equations with application to estimating the effect of schooling quality on earnings. *The Review of Economics and Statistics*, 78(4), 562–610.

Heckman J.J., Lochner L.J. and Todd P.E. (2006). Earnings functions, rates of returns and treatment effects: the Mincer equation and beyond. In Hanushek E.A. and Welch F. (eds) *Handbook of Labour Economics*, Amsterdam, 307–458.

Heckman J.J. and Robb R. (1985). Alternative methods for evaluating the impact of interventions. In Heckman J. and Singer B. (eds) *Longitudinal Analysis of Labor Market Data*, 156–245, New York: Cambridge.

Hedges L.V., Laine R.D. and Greenwald R. (1994) Interpreting research on school resources and student achievement: a rejoinder to Hanushek. *Review of Educational Research*, 66(3), 411–416.

Hoffman F. and Oreopoulos P. (2006). Professor qualities and student achievement. NBER Working Paper, 12596.

Holland P.W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81, 945–960.

Hoxby C.M. (1996). How teachers' unions affect education production. *Quarterly Journal of Economics*, 111(3), 671–718.

Hoxby C.M. (2000). The effects of class size on student achievement: new evidence from population variation. *Quarterly Journal of Economics*, 115(3), 1239–1285.

INVALSI (2005). *Gli studenti quindicenni nel Veneto: quali competenze? Rapporto regionale del Veneto OCSE-PISA 2003.* Ufficio Scolastico Regionale per il Veneto.

Jakubowski M. and Sakowski P. (2006). Quasi-experimental estimates of class size effect in primary schools in Poland. MPRA Paper, 4958.

Johnson G. and Stafford F. (1973). Social returns to quantity and quality of schooling. *Journal of Human Resources*, 8, 139–155.

Joreskog K.G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika*, 32, 443–482.

Kramarz F., Machin S. and Ouazad A. (2006). What makes a grade? The respective contributions of pupils, schools and peers in achievement. IZA Prize Conference, November 2006.

Krueger A.B. (1999). Experimental estimates of education production functions. *Quarterly Journal of Economics*, 114(2), 497–532.

Krueger A.B. (2003). Economic considerations and class size. *Economic Journal*, 113, 34–63.

Lawley D.N. and Maxwell A.E. (1971). *Factor analysis as a statistical method.* London, 2nd ed., Butterwoth & Co.

Lazear E.P. (2001). Educational production. *Quarterly Journal of Economics*, 113(3), 777–803.

Lazear E.P. (2003). Teacher incentives. *Swedish Economic Policy Review*, 10(3), 179–214.

Leuven E., Lindahl M., Oosterbeek H. and Webbink D. (2004). The effect of extra funding for disadvantaged pupils on achievement. IZA Discussion Paper, 1122.

Levaic R. and Vignoles A. (2002). Researching the links between school resources and student outcomes in the UK: a review of issues and evidence. *Education Economics*, 10(3), 313–331.

Loeb S. and Bound J. (1996). The effect of measured school inputs on academic achievement: evidence from the 1920s, 1930s and 1940s birth cohorts. *The Review of Economics and Statistics*, 78(4), 653–664.

Long M.C. (2005a). Secondary school characteristics and early adult outcomes. Daniel J. Evans School of Public Affairs, Working Paper 2006-06.

Long M.C. (2005b). College quality and early adult outcomes. Paper presented at the American Education Finance Association, Louisville, KY.

Lord F.M. (1963). Elementary models for measuring change. In C.W. Harris (Ed.), *Problems in measuring change*, Madison, University of Wisconsin Press.

Loury L.D. and Garman D. (1995). College selectivity and earnings. *Journal of Labor Economics*, 13(2), 289–308.

Margo R.A. (1990). Race and schooling in the south 1880–1950: an economic history. Chicago, The University of Chicago Press.

Mincer J. (1970). The distribution of labor incomes: a survey with special reference to the human capital approach. *Journal of Economic Literature*, 8(1), 1–26.

Mincer J. (1974). Schooling experience and earnings. *NBER*, New York.

Moffit R.A. (1996). Symposium on school quality and educational outcomes: introduction. *The review of Economics and Statistics*, 78(4), 559–561.

Monks J. (2000). The return to individual and college characteristics: evidence from the National Longitudinal Survey of Youth. *Economics of Education Review*, 19(3), 279–289.

Mosteller F. (1995). The Tennessee study of class size in the early school grades. *The Future of Children*, 5(2), 113–127.

Mulligan C.B. (1999). Galton versus the human capital approach to inheritance. *Journal of Political Economy*, 107(6), S184–S224.

Murnane R.J. (1975). *Impact of school resources on the learning of inner city children*. Ballinger, Cambridge, MA.

Murnane R.J. and Phillips B. (1981). What do effective teachers of inner-city children have in common? *Social Science Research*, 10(1), 83–100.

Murnane R.J., Willet J.B., Duhaldeborde Y. and Tyler J.H. (2000). How important are the cognitive skills of teenagers in predicting subsequent earnings? *Journal of Policy Analysis and Management*, 19(4), 547–568.

Nickell S. (2004). Poverty and worklessness in Britain. *Economic Journal*, 114, C1–C25.

OECD (2004). *Learning for tomorrow's world – First results from PISA 2003*. Paris, OECD.

OECD (2005a). *PISA 2003 Technical Report*. Paris, OECD.

OECD (2005b). *PISA 2003 Data Analysis Manual. SPSS Users*. Paris, OECD.

Orazem P. (1987). Black-white differences in schooling investment and human capital production in segregated schools. *American Economic Review*, 77, 714–723.

Peltzman S. (1993). The political economy of the decline of American public education. *Journal of Law and Economics*, 36, 331–370.

Psacharopoulos G. and Patrinos H.A. (2004). Returns to investement in education: a further update. *Education Economics*, 12(2), 111–134 .

Rockoff J.E. (2004). The impact of individual teachers on student achievement: evidence from panel data. *American Economic Review*, 94(2), 247–252.

Rogosa D.R. and Willet J.B. (1983). Demonstrating the reliability of the difference score in the measurement of change. *Journal of Educational Measurement*, 20(4), 335–343.

Rouse C. and Krueger K. (2004). Putting computerized instruction to the test: a randomised evaluation of a ‘scientifically-based’ reading program. *Economics of Education Review*, 23(4), 323–338.

Schmidt S. (1995). Do school inputs matter? Historical evidence from New York State. Unpublished manuscript, Massachusetts Institute of Technology.

Smith J. and Petra T. (2005). Does matching overcome LaLonde’s critique of nonexperimental estimators? *Journal of Econometrics*, 125, 305–353.

Strayer W. (2002). The return to school quality: college choice and earnings. *Journal of Labor Economics*, 20(3), 475–503.

Tiebout C.M. (1956). A pure theory of local expenditures. *Journal of Political Economy*, 64, 416–424.

Tobias J. (2003). Are returns to schooling concentrated among the most able? A semiparametric analysis of the ability-earnings relationship. *Oxford Bulletin of Economics and Statistics*, 65, 1–30.

Todd P.E. and Wolpin K.I. (2003). On the specification and estimation of the production function for cognitive achievement. *Economic Journal*, 113, F3–F33.

Wansbeek T. and Meijer E. (2000) *Measurement error and latent variables in econometrics*. Bliss C.J. and Intriligator M.D. (Eds), Amsterdam.

Warren J.R., Hauser R.M. and Sheridan J. (2002). Occupational stratification across the life course: evidence from the Wisconsin Longitudinal Study. *American Sociological Review*, 67, 432–455.

Willet J.B. (1988). Questions and answers in the measurement of change. *Review of Research in Education*, 15, 345–422.

Woessmann L. (2003). Schooling resources, educational institutions and student performance: the international evidence. *Oxford Bulletin of Economics and Statistics*, 65(2), 117–166.

Woessmann L. (2005). Educational production in Europe. *Economic Policy*, 20, 446–504.

Wooldridge J. (2002). *Econometric analysis of cross section and panel data*. MIT.