Fertig, Michael

Working Paper

What Can We Learn From International Student Performance Studies? Some Methodological Remarks

RWI Discussion Papers, No. 23

Provided in Cooperation with:
RWI – Leibniz-Institut für Wirtschaftsforschung, Essen

Suggested Citation: Fertig, Michael (2004) : What Can We Learn From International Student Performance Studies? Some Methodological Remarks, RWI Discussion Papers, No. 23, Rheinisch-Westfälisches Institut für Wirtschaftsforschung (RWI), Essen

This Version is available at:
https://hdl.handle.net/10419/18574

RWI : Discussion Papers

Michael Fertig

What Can We Learn From International Student Performance Studies?

Some Methodological Remarks

No. 23

RWI
ESSEN

Rheinisch-Westfälisches Institut für Wirtschaftsforschung

RWI : Discussion Papers No. 23

Published by Rheinisch-Westfälisches Institut für Wirtschaftsforschung, Hohenzollernstrasse 1/3, D-45128 Essen, Phone +49 (0) 201/81 49-0

Editor: Prof. Dr. Christoph M. Schmidt, Ph.D.
ISSN 1612-3565 – ISBN 3-936454-38-8

RWI : Discussion Papers

No. 23

Michael Fertig

What Can We Learn From International Student Performance Studies?

Some Methodological Remarks

RWI
ESSEN

Bibliografische Information Der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.ddb.de abrufbar.

ISSN 1612-3565
ISBN 3-936454-38-8

Michael Fertig*

What Can We Learn From International Student Performance Studies? Some Methodological Remarks

Abstract
The determinants which are decisive for a successful accumulation of human capital and the transfer of these skills into the labor market are a contentious issue in the literature on the economics of education. Different studies on, for instance, the impact of school resources typically reach different conclusions even if they utilize the same dataset. The reason behind this is that each and every study decisively depends on a set of identification assumptions which are anything but innocuous for the results obtained. This paper aims at clarifying this point by embedding the discussion on the determinants of test success in international performance studies like PISA into a theoretical model of cognitive achievement and an empirical frame of reference.

JEL-Classification: I21

Keywords: PISA 2000, cognitive achievement, identification

---

* RWI Essen and IZA-Bonn. The author is grateful to Thomas K. Bauer, Jochen Kluve and Christoph M. Schmidt for helpful comments. All correspondence to Michael Fertig, Rheinisch-Westfälisches Institut für Wirtschaftsforschung (RWI Essen), Hohenzollernstr. 1-3, 45128 Essen, Germany, Fax: +49-201-8149-236, Email: fertig@rwi-essen.de.

1. Introduction

It is a widely accepted insight – not only among economists – that the source of a nation's wealth is the skills of its people. However, the consensus breaks down as soon as the discussion reaches the determinants which are decisive for a successful accumulation of human capital and the transfer of these skills into the labor market (see e.g. the debate by Hanushek 2003 and Krueger 2003 on the relative importance of school quality in a special issue of the *Economic Journal*).

Clearly, the way in which nations try to mold their young minds and talents into productive young adults differs widely across different countries and even within a specific country like Germany. A variety of education systems employs different organizational structures and educational tools with varying intensity. The extent to which this really makes a difference is a topic of perpetual interest. Within the United States, researchers intensely debate the role of school quality for educational attainment and subsequent success in the labor market (e.g. Card, Krueger 1992, 1996; Hanushek 1986; Carneiro, Heckman 2003 for an overview). International studies (e.g. Barro, Lee 2001) directly compare educational investment.

Yet, it is very difficult to compare the comparable across countries. While it might be straightforward to ascertain information about inputs and organizational approaches, and while it might also be a convincing identification assumption to presume identical distributions of inherited cognitive abilities of any cohort of newborns, it is the comparison of outcomes across economies that is so difficult. The world-wide "OECD Programme for International Student Assessment" (PISA 2000) held the promise to deliver the data for a meaningful international comparison. It was designed by eminent specialists in pedagogical issues with the aim of measuring practical knowledge in math, science and reading.

The results of the study (OECD 2002) induced quite different reactions throughout the participating countries. Whereas, for instance, the British were quite satisfied with the results of their students, Americans showed themselves rather disappointed and Germans were shocked. In the aftermath of the report, the PISA 2000 examination has initiated an intense discussion on the causes of these results and the consequences to be drawn.

This reaction is astounding, however. After all, the results presented by the OECD report consist by and large of country averages which do not control for any other covariate of individual student achievement. Specifically, whether education systems operate under similar or vastly different conditions regarding family background, intergenerational skill transmission and school inputs has not been explored in the report. Yet, the publicly available

background information (http://www.pisa.oecd.org) collected in PISA 2000, family and individual characteristics and a rich set of school-related variables, allows for a deeper analysis and induced a growing body of literature on the determinants of test success within a specific country and across two or more countries (e.g. Ammermueller 2004; Fertig 2002, 2003; Fertig, Schmidt 2002; Fertig, Wright 2003; Fuchs, Woessmann 2004; Jürges et al. 2004, Wolter, Vellacott 2002).

However, in the public debate on the results of the PISA study there is still quite a lot of confusion. Clearly, policy makers as well as the public are interested whether institutional details of the education system impinge upon cognitive achievement of students and how the education system should be reformed to foster higher achievement. For instance, in Germany these issues are currently on top of the political agenda and the putative results of the PISA study have to serve as a justification for almost every reform proposal.

Thus, the question what we can really learn from international performance studies like PISA still needs clarification. This paper aims at shedding some more light on this issue. To this end, we embed the scientific interest to provide empirical evidence on the determinants of students' cognitive achievement into a theoretical model and an empirical frame of reference. The ultimate objective of this endeavor is the clarification of the decisive *assumptions* which are necessary to infer on the determinants of cognitive achievement from datasets like the PISA 2000 study[1].

The remainder of this paper is organized as follows. Section 2 provides a theoretical model of cognitive achievement and section 3 delineates an empirical frame of reference. In section 4 we discuss the potentials and limitations of utilizing PISA data as well as the decisive assumptions which are necessary to generate empirical evidence on the determinants of cognitive achievement. Finally, section 5 offers some conclusions.

2. The Determinants of Cognitive Achievement – Theoretical Model

The theoretical model underlying the discussion in this paper is a slightly augmented version of the model by Todd/Wolpin (2003). This model describes cognitive achievement as a cumulative process of knowledge acquisition over time. During this process different agents provide inputs at different points in time. More specifically, for a specific child with inherited ability μ, cognitive achievement at time t is denoted by A_t. In every period t the school provides inputs S_t (e.g. by determining the size of the class a student has to attend or by providing special tutoring courses) and the family/parents of the child provide

[1] In principle, the following discussion also applies to comparable datasets which aim at measuring student achievement by internationally comparable standardized tests, like TIMSS.

inputs F_t (e.g. by the provision of homework help or by the degree of care offered to the student) in the knowledge production process. An important element is the timing of these inputs. It is assumed that the decision of the family on F_t follows the decision of the school on S_t, i.e. it is assumed that the family decides upon their investment in the child *conditional* on the school investment.

In the pre-school period $t = 0$ only the family/parents provide inputs. The school decides in each period on S_t conditional on the accumulated achievement of the child and its inherited ability. Furthermore, in this model the family does not only decide about F_t but also where they live and/or about the school they send their child. This decision in turn is assumed to depend on their permanent resources (the family's wealth), the achievement A_t of their child at the beginning of period t, its inherited ability μ and the available school inputs S_t. This implies again, that the family decides *conditional* on the school's decision.

Finally, in each period t the student decides on her motivation, her effort in learning etc. denoted by ε_t. This additional decision process is an augmentation of the original model. The student-specific input might depend on the family and/or school inputs, for instance if students' motivation is higher in smaller classes or learning effort increases with the degree of parental care. Altogether, this yields the following development of cognitive achievement and its determinants over time which is illustrated in Figure 1.

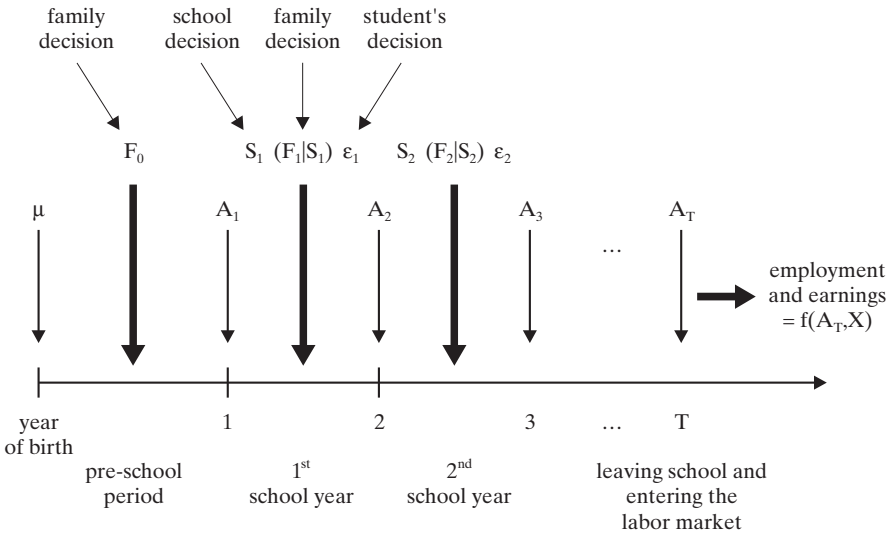
From this figure it becomes transparent that the process of knowledge acquisition is rather complex, depending on a variety of different inputs from different agents. In principle, we are interested in the complete process of knowledge acquisition and all its inputs. In more formal terms, this process can be summarized by

$$A_{i,j,k,t} = g_{t-1}(\boldsymbol{F}_{i,j,k,t}, \boldsymbol{S}_{i,j,k,t}, \mu_{i,j,0}, \varepsilon_{i,j,k,t}, \nu_{i,j,k,t}). \tag{1}$$

In equation (1) $A_{i,j,k,t}$ denotes an achievement measure (e.g. test scores) for student i in country j and school k at time t. The vectors **F** and **S** comprise the *history* of parental and school inputs up to year t, μ denotes the mental capacity of student i which is assumed to be inherited ($t = 0$) and the vector ε captures the history of student inputs in the knowledge acquisition process. Finally, the error term ν allows for measurement error in the achievement outcome.

One problem in all empirical studies is the fact that some of these inputs are missing which might induce serious omitted variable biases into estimation results. Thus, one has to find a way to overcome this problem. This is anything but trivial and it is helpful to conceptualize empirical research by a method-

Figure 1
Model of cognitive achievement as a cumulative process over time



ological frame of reference which is related to the decisive question of every empirical study: "What is the most convincing identification assumption?".

3. The Determinants of Cognitive Achievement – Empirical Frame of Reference

The principal challenge of each and every empirical analysis is the fact that the counterfactual situation is unobservable. The counterfactual situation is implied by a counterfactual question which can always be described by a "what would have happened, if…" question and which is at the heart of every empirical analysis. Examples for counterfactual questions in the context at hand are: "What would have happened to the performance of students in a standardized test, if they had attended another school?" or "What would have happened to the performance of students in this test, if the size of their classes had been reduced (by 10% or 50%)?"

The problem is obvious; this situation is unobservable or – in technical terms – not identified. One could observe a given student at a specific point in time only once, i.e. in a specific school with a specific number of classmates, but not

in both regimes. Consequently, without an observable counterpart to this unobservable situation a *causal* relationship can not be established[2]. Such an observable counterpart, however, can only be constructed by invoking suitable identification assumptions. These assumptions have to hold *a priori*, i.e. they are not testable in statistical terms, and can only be judged upon economic/theoretic reasoning. In other words, they cannot be right or wrong *a priori*, or proven correct or false *a posteriori*; they can only be more or less plausible, or more or less easily violated.

The validity of these assumptions, however, is decisive for the validity of the derived results. Typically, different identification assumptions yield different results. Therefore, the decisive task for every empirical study is to find suitable and convincing identification assumptions. Choosing the appropriate identification strategy, for a specific issue under investigation therefore involves the collection of relevant information that justifies the identification assumption. This information generally requires knowledge on the institutional details of the process under investigation.

In general, three conceptual challenges have to be met. Firstly, one has to find an adequate *outcome measure*. In the context of human capital accumulation a wide variety of outcome measures approximating different dimensions have been employed in the literature. Among the most prominent are the results of standardized tests, schooling or vocational degrees, years of education, indicators for attending college, university, further training etc. or wages and labor market status. The concrete choice of one of these approximations rests upon the research question under investigation and the available data material.

The second challenge comprises the consideration of *observable* heterogeneity among observation units. Individuals are typically heterogeneous with respect to observable socio-economic characteristics. While this is theoretically a solvable problem, in many practical instances not all relevant characteristics are available in a sample, i.e. some potentially important variables are missing. Furthermore, the analyst has to decide how the available information is combined in a specific functional form. Although this decision seems to be unproblematic, it is not. There is a lively discussion in the existing literature on the advantages and disadvantages of specifying the education production function in levels or first differences (the latter is known as value-added specification (e.g. Hanushek, Taylor 1990; Krueger 1999).

---

2 Since the concept of counterfactual questions goes back to the literature of program evaluation, this problem is frequently referred to as the "evaluation problem" (e.g. Heckman et al. 1999; Kluve, Schmidt 2002). The causal model underlying this counterfactual notion of causality is commonly labeled "potential outcome model", since only one of the two outcomes required for causal inference is actually observable, and the second one is not, and thus a potential quantity. For further methodological details see e.g. Holland 1986, Kluve 2004.

The third and most challenging problem, however, is the treatment of *unobserved* heterogeneity. Some characteristics are by their very nature unobservable, e.g. inherited ability, motivation, learning effort etc. If specific groups of individuals differ with respect to unobservable differences, this might result in a serious bias of estimation results[3]. For instance, if schools decide to determine the size of classes conditional on the average learning effort or motivation of the students – e.g. smaller classes for on average less motivated students – the estimated effect of class size will comprise the impact of unobserved motivation and will therefore lead to fallacious conclusions on the true impact of smaller classes on cognitive achievement.

In the received literature two major lines of identification strategies exist; studies relying on social or natural *experiments* and *observational studies*. Social or natural experiments which are commonly perceived as the superior identification strategy utilize variation in the outcome measure and the respective variable(s) of interest which is induced by a process that individuals can not influence by their decisions. Prominent examples for (quasi-) experiments in the context of human capital are the randomization of students into classes of different size, e.g. in the so-called *Project STAR* in Tennessee (e.g. Krueger 1999), compulsory schooling laws in the United States (e.g. Angrist, Krueger 1991), the Vietnam draft lottery (e.g. Angrist 1990), the Chicago school lottery (e.g. Cullen et al. 2002), and Maimonides' Rule of class size (Angrist, Lavy 1999).

In sufficiently large samples this exogenous variation secures, on average, a balancing of all groups of individuals with respect to observable (**S** and **F** in our model) as well as unobservable (ε and μ) characteristics and allows an assessment of the impact of the variable(s) of interest. Experiments, however, are not without any problems. The most contentious issues concern the *internal and external* validity of experiments. That is, one has to assess whether the experiment truly worked like in a laboratory (internal validity) and if the derived results are indeed transferable to other populations than the one under study (external validity).

Furthermore, some observers fear that agents who are involved in an experiment change their behavior due to their involvement. This could result in more or less effort than outside the experimental situation and might introduce a bias into the estimated effect. This problem is known as *Hawthorne* vs. *John Henry* effects (e.g. Krueger 1999). Finally, the implementation of experiments is usually plagued by ethical and cost considerations since experiments are typically expensive and induce high administration efforts.

---

[3] Along the same lines, data problems like panel attrition, item non-response or sample selection may create comparable problems.

By contrast, observational studies can not rely on truly exogenous variation and have to cope with the impact of decisions and behavioral responses of individuals. These decisions can influence the outcome measure and the variable(s) of interest simultaneously. Therefore, the challenge in observational studies is to find suitable assumptions to *mimic* an experiment as good as possible (Rosenbaum, Rubin 1983; Rubin 1974, 1986). The work horse for almost all observational studies in the literature is some kind of regression model. Often, the central identification strategy is an instrumental variable approach (e.g. Angrist et al. 1996).

However, "traditional" instrumental variable approaches identify the effect of a specific variable of interest only if the effect of this variable is constant for individuals with the same value of covariates (Florens et al. 2002; Imbens, Angrist 1994). In the case of heterogeneous effects, i.e. the impact of e.g. an intervention like decreasing class size, varies over the population, the "traditional" instrumental variable approach identifies the mean effect of this intervention for the sub-population of the so-called compliers only. That is, for those individuals whose value of the treatment indicator changes in reaction to an exogenous change in the instrument[4].

The statistical and econometrics literature discusses a large number of alternative identification strategies based on different assumptions. One prominent and in the context of the evaluation of active labor market policy often applied strategy is some form of matching approach. The appeal of *matching* as an identification strategy originates from the fact that matching is a non-parametric approach, i.e. it does not require distributional assumptions which are at least to some extent quite arbitrary. Furthermore, no functional form assumptions are necessary when applying matching techniques. Rather, identification is based on the construction of "statistical twins".

The basic idea of matching methods is to mimic a randomized experiment *ex post*. Utilizing information on a set of observable characteristics, matching constructs – from a pool of potential comparison units – a retrospective comparison group as similar or comparable as possible to the treatment group in terms of these observable characteristics. The comparison group thus substitutes for the experimental control group. The main difference is that, whereas randomized assignment in an experiment balances both observable and unobservable attributes across treatment and control groups, matching can only control for observable covariates. The identification assumption, which matching is based on, is commonly referred to as "conditional independence

4 This is also known as the concept of local average treatment effects (LATE; see Imbens, Angrist 1994). An alternative approach which is very similar to IV-estimation is the implementation of control functions (Heckman 1979). Due to the similarity of both identification strategies, the problem of LATE also applies to control functions.

assumption" (sometimes referred to as "ignorability" or "unconfoundedness"). Essentially, this assumption means that selection into treatment and comparison group is based on observables, and that, conditional on these observables, there is no difference between both groups in any aspect relevant for the outcome measure other than the "treatment" itself. In other words, there is no unobserved heterogeneity between both groups.

Typically, studies on the level of e.g. individual workers justify this assumption by controlling for the history of the outcome measure prior to the intervention (pre-treatment outcomes). If both groups differ in unobserved characteristics, this should be reflected in the values of the outcome measure prior to treatment as well. These pre-treatment outcomes can then serve as a proxy for unobserved characteristics provided that these characteristics remain persistent over time and thus repeated measurement of the outcome variable reveals information about them. Clearly, this approach requires data on the outcome measure under investigation for more than one point in time. Thus, matching does not seem to be a convincing identification strategy in studies using PISA data since PISA is only a cross-section and there is no information on historical test success.

In the next section we discuss the potentials and limitations of utilizing data from the PISA 2000 study to infer on the determinants of cognitive achievement. This discussion will be based on the theoretical model and the empirical frame of reference outlined above.

4. The Determinants of Cognitive Achievement – What Can We Learn From PISA?

Before we proceed to discuss the potentials and limitations of the PISA data in generating empirical evidence on the determinants of cognitive achievement, it is useful to take a brief look at the design of the PISA 2000 study. The PISA 2000 target population are 15 to 16 year old students enrolled in an educational institution at the time of the survey (the first half of 2000). The primary sample unit, however, were schools. In a second step, in every school a random sample of students from the target population was drawn resulting in a stratified cluster sample. The examination conducted among the students in the sample consisted of a reading, math and science literacy test. Furthermore, a wide variety of background information on the students was collected by student questionnaires. Among this individual information is the family background of the student, her attitudes towards visiting school, her learning strategy, a self-assessment of reading pleasure etc. Furthermore, the study also conducted interviews among the principals of the respective schools in order to collect information on the school resources, the number of teachers in the school, the principles of selecting students etc.

The particular test score of an individual student is not the direct share of correct answers. Rather, it is computed based on a procedure originating in "Item Response Theory" (e.g. Hambleton, Swaminathan 1989). Calculated scores are weighted averages of the correct responses to all questions of a specific category (e.g. reading literacy) with the difficulty of the question serving as weight (e.g. Warm 1989). These individual test scores are standardized in a subsequent step so that the unconditional sample mean of the PISA 2000 scores equals 500 and their unconditional sample standard error equals 100. These test scores typically serve as the outcome measure of empirical studies on the determinants of cognitive achievement.

Hence, the PISA data provides a measure of student achievement (i.e. test scores in reading, math and science) for the year 2000 together with information on parental as well as school inputs. Empirical investigation using PISA data for more than one country[5], therefore, typically utilize a regression model which takes the following generic form:

$$A_{i,j,k,2000} = g_{2000}(\alpha_j, \gamma_k, \tilde{\mathbf{F}}_{i,j,k,2000}, \tilde{\mathbf{S}}_{i,j,k,2000}) + \tilde{\nu}_{i,j,k,2000}. \quad (2)$$

In equation (2) $A_{i,j,k,2000}$ denotes test scores in one of the PISA 2000 examinations for student i in country j and school k at time $t = 2000$. α_j denotes a country-specific constant, which might be restricted to be equal across countries, i.e. $\alpha_j = \alpha \forall j$. The vectors $\tilde{F}$ and $\tilde{S}$ comprise contemporaneous parental and school inputs. $\tilde{\nu}$ is an additive error term. The coefficients measuring the impact of family and school inputs can be country-specific or restricted to be equal across countries. Finally, the education production technology g is typically assumed to be some linear function of all inputs.

Most empirical studies utilizing the individual-level data of the PISA 2000 study focus on the impact of one or more variables summarized in the vector $\tilde{S}$ which comprises school resources, since these are tangible aspects of the education system and thus policy relevant even in the short-run. In principle, family inputs captured by the vector $\tilde{F}$ might be addressed by policy measures as well, e.g. family income by public transfers. However, parental inputs comprise to a large extent aspects like parental care, the educational background of the parents etc., which are – at least in the short-run – difficult to be addressed by policy interventions.

By comparing equation (2) with the general specification of our model in equation (1), one observes that the PISA data does not provide any direct information on inherited ability μ and student inputs ε (e.g. motivation and

5 For studies analyzing the determinants of test success in one country separately, most points of the following discussion apply analogously. However, for reasons of space limitations, we focus the following discussion on the more frequent case of cross-country studies.

learning effort). Furthermore, it becomes transparent that family and school inputs are measured contemporaneously. In other words, the dataset does not contain any directly observable measure of historical inputs.

Hence, inference on the determinants of cognitive achievement based on estimating the model in equation (2) has to invoke the following *first set* of assumptions:

(i) Contemporaneous inputs capture the entire history of inputs, i.e. parental and school inputs are time-invariant[6].

(ii) Inherited mental capacity and contemporaneous inputs by family and school are uncorrelated.

(iii) Student and parental inputs are uncorrelated. The same holds for student and school inputs.

Against the background of the theoretical discussion in section 2 these assumptions are obviously difficult to justify. However, regarding assumption (i), each and every study which utilizes PISA data rests upon this assumption. Since this is inevitable, given the cross-sectional nature of the dataset, it has to be borne in mind for the interpretation of the derived results. With respect to assumptions (ii) and (iii) things are a little bit more complicated. The likelihood with which these assumptions are violated depends to a large extent on the concrete implementation of the empirical specification of equation (2). This will be discussed in some more detail below.

Before proceeding this way it is worth noting that every concrete empirical application has to cope with additional problems, which necessitate further assumptions. The specific set of additional assumptions invoked by a specific empirical application depends on the decisions of the involved researchers and, thus, provides some scope for discretion.

Additional assumptions are necessary for several reasons. Firstly, observable school inputs comprise only tangible school-specific inputs, like class size, teacher shortage etc. School inputs which are related to institutional aspects of the education system, like central exit examinations, the practice of class repetition or spending per student, are not directly observable in the data. If the specification contains a country-specific intercept α_j, these institutional aspects will be captured by the country-specific fixed-effect as long as there is systematic variation across countries. Unsystematic variation will be captured by the error term of equation (2).

However, α_j captures the true impact of institutional aspects only if this impact is *time-invariant*. In other words, the effect of any change in institutional

6 Alternatively, one has to assume that *only contemporaneous* inputs matter.

aspects of the education system in the past will be ignored. Furthermore, since the education systems of countries typically differ in more than one aspect, it is impossible to identify the driving force behind differences in country-specific fixed-effects. However, restricting the intercept to be equal across all countries, i.e. assuming that $\alpha_j = \alpha \forall j$, is equivalent to assume that there are no systematic differences of the impact of institutional factors on students' cognitive achievement across countries. Clearly, this is hardly realistic.

Secondly, additional assumptions are necessary because family inputs in the education production process can only be approximated by observable characteristics like parental education, labor market status of parents, number of books and durable goods in household etc. However, proxy variables are not without problems, especially if they approximate the desired variable only crudely. Todd/Wolpin (2003) provide examples for the case when the inclusion of a rather crude proxy variable confounds the interpretation of the impact of observed inputs and might thus lead to a greater bias in the estimated coefficients.

Finally, additional assumptions enter the analysis due to the concrete choice of explanatory variables entering equation (2). This brings us back to assumptions (ii) and (iii). The problem here is the potential endogeneity of explanatory factors incorporated in $\tilde{S}$ and $\tilde{F}$. Alternatively, unobserved heterogeneity induced by either unobserved mental capacity μ or unobserved student inputs ε might cause correlation between elements of $\tilde{S}$ or $\tilde{F}$ and $\tilde{\nu}$. In other words, assumption (ii) and/or (iii) will be violated which leads to biased estimates. Clearly, this is not a problem which is confined to studies using datasets like PISA. In empirical applications using panel data, it is possible to solve – or at least alleviate – this problem by estimating a model with individual-specific fixed-effects. If the source of unobserved heterogeneity is time-invariant, individual-specific fixed-effects eliminate the problem. In the case at hand, however, this approach is not possible since the PISA study is only a cross-section. Thus, the problem of unobserved heterogeneity is especially severe.

Finally, in studies using data on more than one country identification is achieved by some form of cross-sectional estimator. This implies, that all slope coefficients in equation (2) are restricted to be equal across countries and that the impact of a specific variable of interest, e.g. class size, is identified by using variation across countries in test scores and class size. The counterfactual employed by this approach is, therefore, that the average achievement of students in country *A* would have been equal to that of comparable students (in terms of family inputs and all other schools inputs) in country *B*, if e.g. the average class size of students in country *A* had been equal to that of country *B*. Implementing this identification strategy requires the following two assumptions to hold in addition to assumptions (i)–(iii):

(iv) Inherited ability is equally distributed across countries.

(v) Student inputs are equally distributed across countries.

While assumption (iv) might be rather convincing, assumption (v) does not seem to be very persuasive *a priori*. It might well be the case that student inputs like learning effort and motivation might e.g. depend on the overall economic prospects of the specific country. That is, students living in poorer countries might display higher learning efforts than their peers in richer countries. Employing a school-fixed effect γ_k into the specification eliminates this problem if and only if student inputs depend solely on school inputs, are on average equal for all students attending a specific school and are time-invariant.

5. Conclusions

The determinants which are decisive for a successful accumulation of human capital are a contentious issue in the literature on the economics of education. Different studies on, for instance, the impact of school resources typically reach different conclusions even if they utilize the same dataset. The reason behind these conflicting results is the fact that each empirical study depends on a set of identification assumptions which are decisive for the results obtained. Different identification assumptions typically lead to different results.

In the public – and sometimes also in the academic – discussion on the results of the PISA 2000 study this fact is often ignored. Thus, this paper provides a theoretical model and an empirical frame of reference for a discussion of the central assumptions which are necessary to infer on the determinants of cognitive achievement from PISA data. From this discussion it should have become transparent that a large set of *rather strong* assumptions are inevitable. The primary reason for this is that the PISA study is only a cross-sectional dataset which does not provide any retrospective information on achievement in the past or on historical inputs in the education production process.

Furthermore, in studies using PISA data the identification of institutional-specific effects which are the most policy-relevant issue completely rests upon variation across countries[7]. Since many institutional issues of the education system are either unobservable or shared within the same country it is almost impossible to derive reliable evidence on the impact of different educational institutions.

[7] Ideally, identification would rely on variation across countries together with variation within a specific country across time. Unfortunately, the next wave of the PISA study (PISA 2003) does not involve the same students, not even the same schools. Thus, there will be no panel structure and studies using PISA 2003 data will have to cope with basically the same problems.

Thus, one has to be very careful in drawing structural conclusions or providing strong policy advice from empirical results derived by utilizing PISA or comparable data. PISA 2000 is only a snap-shot and the long-term impact of a specific institutional feature must necessarily remain an unresolved issue[8]. Consequently, the PISA study is only able to provide a small contribution to closing our knowledge gap on the determinants of cognitive achievement. Most European countries and to some extent even the US are clearly in need of more empirical evidence on the impact of tangible aspects of the education system on educational attainment. Progress on this issue can only be expected if policy makers are willing to put all policy interventions to test and to evaluate their success scientifically.

References

Ammermueller, A (2004), PISA: What Makes the Difference? Explaining the Gap in PISA Test Scores Between Finland and Germany. ZEW Discussion Paper 04–04. ZEW, Mannheim.

Angrist, J.D. (1990), Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records. *American Economic Review* 80: 313–336.

Angrist, J.D. and A.B. Krueger (1991), Does Compulsory School Attendance Affect Schooling and Earnings? *Quarterly Journal of Economics* 106: 979–1014.

Angrist, J.D., G.W. Imbens and D.B. Rubin (1996), Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association* 91: 444–455.

Angrist,.J.D. and V.C. Lavy (1999), Using Maimonides' Rule to Estimate the Effect of Class Size on Student Achievement. *Quarterly Journal of Economics* 114: 533–575.

Barro, R.J. and J.-W. Lee (2001), Schooling Quality in a Cross-Section of Countries. *Economica* 68: 465–488.

Card, D. and A.B. Krueger (1992), Labor Market Effects of School Quality: Theory and Evidence. In G. Burtless (ed.), Does Money Matter? *The Effect of School Resources on Student Achievement and Adult Success*. Brookings Institution, Washington, DC, 97–140.

Card, D. and A.B. Krueger (1996), Does School Quality Matter? Returns to education and the Characteristics of Public Schools in the United States. *Journal of Political Economy* 100: 1–40.

Carneiro, Pedro and James J. Heckman (2003), Human Capital Policy. *NBER Working Paper* No. 9495, Cambridge, MA.

Cullen, J., B. Jacob and St.D. Levitt (2002), Does School Choice Attract Students to Urban Schools: Evidence from over 1,000 Randomized Lotteries. Working Paper University of Chicago.

[8] An example in this context is the effect of repeating a class on educational outcomes in Germany; see findings in Fertig (2004) vs. the results derived with PISA data (Schümer et al. 2004).

Fertig, M. (2002), Educational Production, Endogenous Peer Group Formation and Class Composition – Evidence From the PISA 2000 Study. RWI Discussion Paper 2. RWI, Essen.

Fertig, M. (2003), Who's to Blame? The Determinants of German Students' Achievement in the PISA 2000 Study. RWI Discussion Paper 4. RWI, Essen.

Fertig, M. (2004), Shot Across the Bow, Stigma or Selection? The Effect of Repeating a Class on Educational Attainment. RWI Discussion Paper 19. RWI, Essen.

Fertig, M. and Ch.M. Schmidt (2002), The Role of Background Factors For Reading Literacy: Straight National Scores in the PISA 2000 Study. IZA Discussion Paper 545. IZA, Bonn.

Fertig, M. and R.E. Wright (2003), School Quality, Educational Attainment and Aggregation Bias. RWI Discussion Paper 9. RWI, Essen.

Florens, J.-P., J.J. Heckman, C. Meghir and E. Vytlacil (2002), Instrumental Variables, Local Instrumental Variables and Control Functions. Cemmap Working Papers 15/02. Institue for Fiscal Studies, UCL-London.

Fuchs, Th. and L. Woessmann (2004), What Accounts for International Differences in Student Performance? A Re-Examination Using PISA Data. IZA Discussion Paper 1287. IZA, Bonn.

Hambleton, R.K. and H. Swaminathan (1989), *Item Response Theory – Principles and Applications*. Boston: Kluwer.

Hanushek, E.A. (1986), The Economics of Schooling: Production and Efficiency in Public Schools. *Journal of Economic Literature* 24: 1141–1177.

Hanushek, E.A. (2003), The Failure of Input–Based Schooling Policies. *Economic Journal* 113: 64–98.

Hanushek, E.A. and L. Taylor (1990), Alternative Assessments of the Performance of Schools. *Journal of Human Resources* 25: 179–201.

Heckman, J.J. (1979), Sample Selection Bias As A Specification Error. *Econometrica* 47: 153–161.

Heckman, J.J., R.J. LaLonde and J.A. Smith (1999), The Economics and Econometrics of Active Labor Market Programs. In O. Ashenfelter and D. Card (eds.), *Handbook of Labor Economics,* Vol. III. Amsterdam et al.: North-Holland, 1865–2097.

Holland, P.W. (1986), Statistics and Causal Inference. *Journal of the American Statistical Association* 81: 945–970.

Imbens, G. and J. Angrist (1994), Identification and Estimation of Local Average Treatment Effects. *Econometrica* 62: 467–475.

Jürges, H., K. Schneider and W. Richter (2004), Teacher Quality and Incentives: Theoretical and Empirical Effects of Standards on Teacher Quality. CESifo Working Paper 1296. CESifo, Munich.

Kluve, J. (2004), On the Role of Counterfactuals in Inferring Causal Effects. *Foundations of Science* 9: 65–101.

Kluve, J. and Ch.M. Schmidt (2002), Can Training and Employment Subsidies Combat European Unemployment? *Economic Policy* 35: 409–448.

Krueger, A.B. (1999), Experimental Estimates Of Education Production Functions. *Quarterly Journal of Economics* 114: 497–532.

Krueger, A.B. (2003), Economic Considerations and Class Size. *Economic Journal* 113: 34–63.

OECD – Organisation for Economic Co-Operation and Development (ed.) (2002), *Knowledge and Skills for Life: First Results from PISA* 2000. Paris.

Rosenbaum, P.R. and D.P. Rubin (1983), The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 70: 41–55.

Rubin, D.B. (1974), Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology* 66: 688–701.

Rubin, D.B. (1986), Which Ifs Have Causal Answers? *Journal of the American Statistical Association* 81: 961–962.

Schümer, G., K.-J. Tillmann and M. Weiß (Eds.). (2004), *Die Institution Schule und die Lebenswelt der Schüler* – vertiefende Analysen der PISA 2000-Daten zum Kontext von Schülerleistungen. Wiesbaden: Verlag für Sozialwissenschaften.

Todd, P.E. and K.I. Wolpin (2003), On The Specification and Estimation of the Production Function for Cognitive Achievement. *Economic Journal* 113: 3–33.

Warm, T.A. (1989), Weighted Likelihood Estimation of Ability in Item Response Theory. *Psychometrika* 54: 427–450.

Wolter, St. and M.C. Vellacott (2002), Sibling Rivalry: A Look at Switzerland with PISA Data. IZA Discussion Paper 594. IZA, Bonn.