Evaluating schooling systems. The role of international assessments of children's learning

Dalit Contini

Large scale international assessments of student learning achievement, providing comparable measures of competencies, school variables, and indicators of family background across countries and schooling systems, have given a great impulse to the empirical educational research. Purpose of this paper is to review the contribution of these surveys to the evaluation of educational policies with respect to quality of the instruction and equality of opportunity between children with different family backgrounds, and review this line of research. The focus is on the main problems involved in the identification of the causal impact of policies and on the solutions employed in the empirical literature to overcome them.

Key words: school quality, equality of opportunity, educational production function, policy impact, peer effects.

1. Introduction

Recent years have witnessed the growth of an increasing number of large scale international surveys on students' learning achievement, designed to provide an information-based educational decision making. PISA (Programme for International Student Assessment). in particular, was launched by OECD with an explicit policy orientation, with the aim of helping governments to draw policy lessons with respect to quality and equity of educational systems.

Tests are developed in an international cooperative effort, aiming at cross-country comparability. PISA focuses on the competences of the 15-year-olds, covering the domains of reading, mathematical and scientific literacy. PIRLS (Progress in International Reading Literacy Study) and TIMSS (Trends in International Mathematics and Science Study), conducted by the International Association for the Evaluation of Educational Achievement (IEA), assess competences respectively in reading literacy in fourth grade and in math and science in fourth and eighth grade[1]. PISA focuses on the knowledge and skills

Per corrispondenza: Dalit Contini, Dipartimento di Statistica e Matematica Applicata "Diego de Castro", Università degli Studi di Torino, via Maria Vittoria 38, 10123 Torino (Italia). E-mail: dalit.contini@unito.it.

useful to solve real life problems, while assessments promoted by IEA measure curricular competences. In addition to the tests, plenty of contextual data on family background, attitudes towards learning and perception of the school climate, information on the organization of schooling, school policies and resources, are collected via questionnaires submitted to students, parents, teachers and school principals[2].

Performance scores are created with *item response models*. These scaling techniques are common in educational assessments. To limit testing time students are not confronted with the complete battery of items: items are divided into clusters and each student receives a number of them, so that the data from different tests can be linked, allowing proficiency to be measured on a single scale. Tests are randomly assigned to students; however, they never have exactly the same difficulty. Item response models control for this difference: the "proficiency" posterior distribution given the item response pattern is estimated under the assumption that the probability of a correct answer to each item follows a logit model depending on individual ability and the item's difficulty. A number of random draws, referred as "plausible values", are taken from this posterior distribution for every student, allowing measurement error to be considered and to derive properly standard error of the estimates[3].

Although the rigour of these international assessments is widely recognised, test scores are subject to some criticisms. Goldstein (2004) highlights that surveys lack of a systematic assessment of the assumptions underlying item response models, and claims that ability has a multi-dimensional structure and should not be measured on a single scale[4]. The specification of item response models is also questioned: exploiting the fact that a three-parameter model[5] was applied to data in TIMSS 1995 in addition to the usual single parameter version, Brown *et al.* (2007) show that findings are fairly robust for central tendency, but not for dispersion. Bonnet (2002) critiques the comparability of PISA scores across educational systems and cultures and discusses the difficulties with translation. International league tables of average student achievement are questioned also because they are claimed to reflect the different non-response behaviour and sample selection rules across countries (Rotberg, 1995); national testing authorities are allowed, for example, to exclude to some extent students with disabilities. However, Hanushek and Woessmann (2011a) provide evidence that selectivity does not greatly affect the conclusions of typical statistical models relating achievement to family, school and institutional inputs (see next section). Test validity is also criticised on the grounds that effort may be low in assessment carrying no direct consequences on the students' schooling careers (Goldstein, Thomas, 2008).

International educational surveys share complex two-stage clustered sampling designs. First stage units are schools; in the second stage PISA samples students while TIMSS and PIRLS sample classes. Consequently, a number of

estimation and data modelling problems have to be tackled. Rutkowski *et al.* (2010) point up that: (i) student, teacher or school weights should be used to adjust for the different unit selection probabilities; (ii) multilevel models should be employed to account for the lack of independence among observations; (iii) replication methods should be applied to obtain standard errors of the estimates, as these cannot be derived analytically.

Educational economics has a long-standing tradition in addressing policy issues and as we discuss below, large scale international assessments offer new opportunities to this line of research, which is now being developed by scholars in different disciplines (economists, sociologists, applied statisticians). First, the availability of comparable measures of competencies for a large number of countries allows to compare educational systems in terms of their capacity to produce high quality instruction and to ensure equal opportunities to all children. Next, the availability of comparable background information on students, families and schools allows to relate achievement to a set of inputs in a cross-national framework; given the larger variability observed between than within nations, the added value is remarkable.

Official reports of PIRLS, TIMSS and PISA in particular, summarise a massive amount of relevant results from the perspective of the evaluation of educational systems with respect to *quality*, *inequality* and *inequality of opportunity* among children with different family background. Measures of central tendency and rankings allow to evaluate countries' relative standing; dispersion indexes and the share of students performing at different levels describe the extent to which competencies are unevenly distributed within country student populations. PISA shows between and within-school score variance as well, providing evidence of school stratification. All surveys relate scores to family background[6]. PISA also reports the percentage of variance explained by individual and school average socio-economic status, and the corresponding regression coefficients. A major empirical finding is that quality and equality are not competing policy objectives.

There are only few attempts to compare the results of PIRLS, TIMSS and PISA at the cross-country level. Haahr *et al.* (2005) overview the aggregate results to formulate recommendations on improved provision of education in the UE in relation to the Lisbon Strategy, focusing on characteristics of the educational systems and aspects of school management. Brown *et al.* (2007) reach the broad conclusion that there is considerable agreement between the different tests.

2. The education production function

A major contribution to educational policy evaluation comes from the educational economic research. The basic model is given by the *education*

production function, describing the productivity relationship between schooling inputs and test scores outcomes. The theoretical model (Todd, Wolpin, 2003) assumes that achievement Y_{ija} of child *i*, in school *j*, of age *a* is a function of family-supplied inputs $F(a)$ up to age *a*, current and past school-supplied inputs $S(a)$, innate ability μ_o and a random error.

$$Y_{ija} = \alpha_a + \beta_a F_{ij}(a) + \gamma_a S_{ij}(a) + \mu_{otj} + \varepsilon_{ija} \quad [1]$$

However, model estimation is typically hampered by missing data, as innate ability is unobserved and only data on contemporaneous inputs are generally available. Thus, a commonly used specification is:

$$Y_{ija} = \alpha_a + \beta_a F_{ija} + \gamma_a S_{ja} + \varepsilon_{ija} \quad [2]$$

where F_{ija} and S_{ija} are current inputs (the first often represented by social origin or income) and the residual term includes all omitted factors. Parameter estimates are unbiased if past inputs don't matter and if the residual component is not correlated with contemporaneous inputs. Yet, these assumptions will be too restrictive if performance at age *a* is the outcome of a continuous ability development process; moreover school inputs are generally not exogenous, as ability itself affects school choices.

If lagged performance scores are available, a *value-added* specification:

$$Y_{ija} = \alpha_a + \beta_a F_{ija} + \gamma_a S_{ja} + \lambda Y_{ija'} + \varepsilon_{ija} \quad [3]$$

can be estimated. Lagged performance $Y_{ija'}$ could be assumed to capture the dependence on past inputs and innate ability; in this case OLS produces consistent estimates of the causal effect of current inputs between age *a'* and *a*[7].

Production functions are employed to evaluate educational systems from quite different standpoints. In accountable systems *school* (or sometimes *teacher*) contributions to students' achievement are evaluated and publicly reported, the aim being to inform parental school choice and promote competition among schools, which in turn should foster higher quality instruction. Here the term "value added" denotes the school effect and is represented by the school component residual, estimated by the difference between the raw performance indicator and the expected outcome given the

inputs, employing multilevel versions of models [2] and [3] with school-specific random effects. Given the aim of the evaluation, these analyses typically rely on national data; an exception is Jürges and Schneider (2007) who exploit PIRLS German data to design a method for ranking teachers accounting for the school composition and measurement error in the achievement variable. For a stimulating discussion on school value-added assessment from the causal inference point of view, see Rubin *et al.* (2004).

Quite a different issue is that of evaluating the impact of schooling policies and features of the educational system. The impact of *resource-based policies* has been the object of great interest. A major problem in the identification of the causal effect of resources is that school attributes are generally endogenous, as schooling decisions are affected by (often unobserved) performance at the time of the choice. For this reason many studies use resource variables at an aggregate level (for example, average class size, average teacher-pupil ratio), circumventing the selection problem (Hanushek, Woessmann, 2011b). Most studies rely on national surveys; yet, within-country variation is usually limited, so that international data can greatly improve statistical power. Hanushek (2003) reviews the evidence on the effect of class-size, teacher-pupil ratio, expenditure per pupil, teacher education, experience or salaries, and concludes that there is little evidence of a positive impact of the amount of resources on student performance. Exploiting TIMSS, Woessmann (2003) largely confirms these findings.

Particularly problematic is the evaluation at the within-country level of the impact of educational *system-level features*, as the whole student population is generally exposed to a common institutional setting. Identification is possible only if there is over time variation, as occurs if a reform is enacted and data before and after its enforcement is available[8]; an example is given by the evaluation of the British reform that took place during the 60's, when the secondary school system moved from early tracking to comprehensive education (Pischke, Manning, 2006). In this light, international assessments greatly enhance the prospects of evaluating the effect of institutional features that vary little within nations. Not without problems, as we will now discuss.

3. Evaluating institutional features of schooling systems

Providing comparable measures of competencies across countries, international educational surveys have contributed to raise questions about why some countries perform substantially better than others. In this perspective, some work is devoted to examine the performance gap observed across nations or regions, accounting for individual, family and schooling inputs. Attempting to explain the differential between Finland and Germany

with PISA, Ammermüller (2007) estimates separate country production function regressions and partitions the score gap into different components, using Oaxaca-type decomposition methods. He concludes that measurable characteristics of students and schools only account for a small part of the gap, suggesting that institutional features could be responsible for the poorer performance of German students. Similarly, Bratti *et al.* (2007) try to explain the dramatic regional differences between northern and southern Italy by complementing PISA data with administrative data to describe the socio-economic context at the sub-regional level, and conclude that territorial differences do not account for the whole differential. Note that in principle this issue could be addressed with national data, but standardized measures of achievement are not available in Italy[9], so that international educational surveys become an irreplaceable source of information.

Covering different schooling systems, international assessments have given a boost to the evaluation of institutional features by allowing the estimation of international education production functions with institutions included as explanatory variables. The basic specification is given by:

$$Y_{ijc} = \alpha + \beta F_{ijc} + \gamma S_{jc} + \theta I_{c} + \varepsilon_{ijc} \qquad [4]$$

where subscript *c* represents the country and *I* school design features, generally varying at the country level. Fuchs and Woessmann (2007) and Woessmann (2007) use PISA to assess the effect of external exit exams, standardized testing, school autonomy, share of public/private management. The latter uses German data, exploiting policy differences across Lander; the same data is used by Jürges *et al.* (2006), who evaluate the impact of central examinations and account for the policy potential endogeneity with propensity score matching. The main findings are that external exit exams and standardised testing are positively associated with student performance, and that the effects of school autonomy are positive in systems where external exit exams are in place.

A weakness of this approach is that the effect of within-country invariant features may be easily confounded with country-specific effects related to cultural traits or other features of the educational system. If the unobserved heterogeneity component at the country level is correlated to *I*, regression coefficient estimates will be biased. Furthermore, since the number of countries is inevitably small, only few system-level variables at a time can be considered. Note that the model cannot comprise country-specific dummies, as the effect of institutional features would no longer be identified.

A body of work is devoted to the measurement in a comparative perspective of inequality of opportunity (Woessmann, 2004) and to the assessment of

how specific institutional features affect inequality. Special attention has been devoted to the effect of tracking (differentiation of the curricula at some point in the educational career) with respect to comprehensive schooling[10]. Schuetz *et al.* (2005) estimate the model:

$$Y_{ics} = \alpha_c + \beta F_{ics} + \gamma S_{cs} + \theta F_{ics} I_e + \varepsilon_{ics} \qquad [5]$$

with TIMSS, where *c* and *s* stand for country and school respectively. θ is the parameter of interest, measuring how strongly institutional features affect the family background coefficient. Given the identification problems described above, they ignore the effect on average achievement, capturing country-specific effects with dummy variables α_c. The critical assumption is that, while the mean level of performance is influenced by various factors, cross-country differentials in the social origin effect depend only on the tracking policy. The empirical evidence is that social origin differentials are wider with early tracking. In a wide-ranging paper, Brunello and Checchi (2007) use similar models to investigate whether the reduction of equality of opportunity due to tracking observed at early ages persists over time, affecting educational attainment, labour market history and literacy. With respect to the latter, the authors use the International Adult Literacy Survey (IALS)[11]. While they confirm the findings that school tracking strengthens the impact of family background on school competencies, educational attainment and labour market outcomes, they do not report similar effects with respect to adult literacy.

Value-added models cannot be estimated with international assessments data as there are no previous measures of achievement. However, surveys run at different stages of the schooling career can be employed to assess with *difference-in-difference* methods the effect of schooling design features changing in between. Some scholars adopt this approach to estimate the impact of early tracking on inequality with PIRLS or TIMSS grade 4 (when students are still in comprehensive education) and PISA or TIMSS grade 8 (when in some countries tracking has taken place). In their influential paper, Hanushek and Woessmann (2006) estimate the impact of early tracking on score variability with country-level production functions. They show that in early tracking countries dispersion increases over time relative to late tracking countries and conclude that early tracking positively affects overall performance inequality. Addressing the issue of inequality of opportunity, Waldinger (2007) challenges the assumptions implicit in [5], arguing that it is unlikely that country-specific effects influence only average achievement and not the social origin-performance relation. In this light, he pools together the data from two surveys and allows the effect of social origin to vary across countries also before children are tracked. The estimated model is:

$$Y_{ijct} = \alpha_c + \delta t + \beta F_{ijct} + \gamma S_{jct} + \lambda_1 F_{ijct}\, t + \lambda_2 F_{ijct} I_c + \lambda_3 F_{ijct} I_c\, t + \varepsilon_{ijct} \qquad [6]$$

where t indicates the survey. The parameter of interest is λ_3, capturing the extent to which the social origin effect changes over time between early-tracking countries relative to late-tracking countries. Waldinger finds that family background is more important in early tracking countries, yet its effect does not increase after tracking has actually taken place. Consequently, he concludes that the stronger social origin effect in early tracking countries cannot be ascribed to tracking itself, but to other factors related to the tracking policy. Employing similar methods, Jakubowski (2010) reaches the same conclusion, while Ammermüller (2005) reports opposite findings.

4. Peer effects

Another relevant topic from the educational policy perspective is that of peer effects. Whether and how schoolmates influence students' academic outcomes provides some guidance on how students should be sorted into schools and classes, whether ability grouping and tracking are beneficial and for whom. In principle, there is no need for international data to address this issue, as peer effects can be best uncovered within homogeneous environments. In fact there is a flourishing literature based on national data, mainly from the US (Hoxby, 2000; Hanushek *et al.*, 2003); yet some countries lack of standardized achievement data so that international surveys, sampling more students per school or even whole classes (as in PIRLS), can be effectively used instead.

The theoretical model is an educational production function where performance Y is a function of individual effects, school characteristics, characteristics of the group of peers $X_{j(-i)}$ and average performance of peers $Y_{j(-i)}$:

$$Y_{ij} = \alpha + \beta X_{ij} + \gamma S_j + \delta Y_{j(-i)} + \lambda X_{j(-i)} + \mu_j + \varepsilon_{ij} \qquad [7]$$

Subscript $j(-i)$ indicates that the average has been calculated across schoolmates, excluding the individual himself. The error term comprises unobserved school effects and an individual random component. Although the effect of peers average performance and average characteristics cannot be separately identified because of the well known "reflection problem"[12] (Manski, 1993), the reduced form:

$$Y_{ij} = \alpha' + \beta' X_{ij} + \gamma' S_j + \lambda' X_{j(-i)} + \mu_j + \varepsilon_{ij} \quad [8]$$

is estimable. The parameter representing peer effects is λ' (function of the structural parameters δ and λ). The problem is to disentangle this effect, driven by social interactions, from potential spurious effects, induced by the fact that schools are not randomly assigned to students (as families are to a certain extent free to choose schools), so that peer characteristics are endogenous. Hence, the challenge is to handle the correlation between the school error component and the peer variables.

Using PISA, Rangvid (2007) studies peer effects and the effect of heterogeneous classes in Denmark with an ordinary regression model for average achievement and quantile regression models to examine who gains and who loses from ability grouping; she claims to control for the peer variable endogeneity by including several variables describing parental attitudes in the model, which are available in PISA but not in typical databases. Ammermüller and Pischke (2009) study peer effects with PIRLS for a few European countries at the primary school level. They exploit the fact that more than one class per school is sampled: they assume that although families are allowed to choose schools, classes are randomly assigned: therefore, given the school, classes are exogenous. As a consequence, within-school variability in the peer variable can be taken as random, implying that peer effects are consistently estimated with within-group estimators. A similar strategy is used by Schneeweis and Winter-Ebmer (2007) who analyse PISA Austrian data, and formulate the (somewhat stronger) assumption that once the track is chosen, the school is randomly assigned. Overall, all studies report sizeable peer effects; however, while most of them find that mixed grouping affects disadvantaged groups positively but has no significant effects on the more privileged ones, Ammermüller and Pischke report no evidence of non-linear peer effects.

Some work has been done also on migrant background peer effects. Focusing on the effect of peer achievement, Entorf and Lauk (2008) use PISA to estimate "social multipliers", summarising the overall impact of exogenous changes in individual or school characteristics, directly affecting the performance of individuals and in turn that of other students' achievement. They compare migrant status peer effects between comprehensive and tracked schooling systems, finding stronger effects in the latter. Brunello and Rocco (2011) analyse whether a higher share of immigrant pupils affects the school performance of natives exploiting PISA data, aggregated at the country level. This approach has the advantage of removing school selection issues; however, since immigrants sort across countries and more developed countries generally have a higher share of immigrants, a different selection problem has to be tackled. Using data from all available waves, the authors

control for migration flows by conditioning on country fixed effects and on the country stock of immigrants at a given time, finding evidence of negative but small peer effects.

5. Conclusions

Despite some criticisms, there is no doubt that international educational surveys have provided high quality comparable measures of competencies across-countries and have revealed valuable information about quality and equity of educational systems across the world. Some countries are indeed performing much better than others, and the reasons laying behind the differentials cannot be ascribed to socio-economic factors alone. Some countries ensure much better opportunities to the disadvantaged segments of the society, and, very interestingly, quality and equity are shown not to be conflicting aims.

Analyses of international educational data have offered important insights on how institutional factors affect achievement and a valuable body of research has been produced. However, most models are still largely descriptive and the causal nature of the estimates can be questioned. Even though institutional features can often be thought as exogenous, their interplay with other country-specific and school-level factors must be dealt with. The former can sometimes be controlled for with difference-in-difference methods, analysing data from different surveys as if they were repeated surveys; however, conceptual framework differences across assessments are largely ignored.

If institutional and school attributes are related at the cross-country level, the omission of the latter gives rise to biased estimates. The inclusion of school variables at an aggregate level leads to error in variables; their inclusion as individual variables poses serious endogeneity problems that are difficult to solve, because ability, a major determinant of school choices, cannot be controlled for with cross-sectional data, and past inputs are not observed. Although the need to estimate policy effects in a causal inference perspective seems to be now fully understood by scholars of different disciplinary backgrounds, the possibility to actually do so is not warranted. International assessments are not designed to tackle specific research questions, but rather to provide a large multi-purpose database. Sometimes analysts may apply propensity scores, find appropriate instrumental variables or uncover sources of exogenous variation in the treatment variable and employ within-group methods, but this is not always the case. Longitudinal data would help to keep unobserved factors under control, under weaker conditions. In this perspective the need for a longitudinal design in international educational surveys has been strongly advocated by Goldstein (2004). Yet, at the moment this does not seem to be in the agenda of international organizations.

Notes

[1] PISA is run every three years since 2000; each wave gives priority to a specific competence domain. PIRLS is run every five years since 2001, TIMSS every four years since 1995.

[2] Parents are interviewed only in PIRLS; teachers in both IEA studies.

[3] The idea is largely based on Rubin's multiple imputation method for handling missing data. Ordinary analyses are performed on each of the plausible values, and the results are combined to derive standard errors of the estimates that account for this source of uncertainty.

[4] Multidimensional models allow one group of students to be more able than others to answer correctly one sort of questions but not others. Goldstein (2004) argues that the development of subscales, measuring proficiency in different areas within each domain with a one-dimensional model (as done in PISA) is not a proper solution, because it involves no exploration of the dimensionality of the full set of items.

[5] Capturing the probability of guessing the right answer and the power of each item to discriminate between high and low ability students.

[6] Social background is measured in various ways. All surveys collect information on parental education, occupation, number of books at home (regarded as the most powerful *explanandum* of performance) and other cultural possessions; PISA also provides an overall index (*ESCS*) based on all of these variables.

[7] Although Todd and Wolpin (2003) show that the conditions underlying the value-added model are more restrictive than commonly presumed.

[8] The main empirical problem in this case is to disentangle the effect of the policy from the effect of other factors changing over time.

[9] Standardised assessments (prove INVALSI) have recently been introduced in primary and lower secondary school, so that in the near future there will be some achievement data at the national level.

[10] Age of tracking widely differs: in Germany students are divided at age 10, in Netherlands at 12, in Italy at 14, while in North Europe and UK schooling is comprehensive up to age 16 and in the US up to age 18.

[11] The International Adult Literacy Survey Database (IALS) was a seven-country initiative conducted for the first time in 1994. Its goal was to create comparable literacy profiles across national, linguistic and cultural boundaries. It investigated the prose, document and quantitative literacy of adults in a sample of OECD countries.

[12] The problem is given by the simultaneous determination of achievement for all classmates, with achievement of one student affecting the achievement of classmates and vice versa.

References

Ammermüller A.

2005 *Educational opportunities and the role of institutions*, ZEW Centre for European Economic Research, Discussion Paper 05-44.

2007 *PISA: What makes the difference? Explaining the gap in test scores between Finland and Germany*, in "Empirical Economics", 33, 2, pp. 263-87.

Ammermüller A., Pischke J. S.

2009 *Peer effects in European primary schools: Evidence from the PISA Study*, in "Journal of Labor Economics", 7, 3, pp. 315-48.

Bonnet G.

2002 *Reflections in a critical eye: On the pitfalls of international assessment*, in "Assessment in Education", 9, 3, pp. 387-400.

Bratti M., Checchi D., Filippin A.
2007 *Geographical differences in Italian students' mathematical competences: Evidence from PISA 2003*, in "Giornale degli Economisti e Annali di Economia", 66, 3, pp. 299-333.
Brown G., Micklewright J., Schnepf S., Waldmann R.
2007 *Cross-national surveys of learning achievement: How robust are the findings?*, in "Journal of the Royal Statistical Society", series A, 170, 3, pp. 623-46.
Brunello G., Checchi D.
2007 *Does school tracking affect equality of opportunity? New international evidence*, in "Economic Policy", 52, pp. 781-861.
Brunello G., Rocco L.
2011 *The effect of immigration on the school performance of natives: Cross country evidence using PISA test scores*, IZA Institute for the Study of Labor, Discussion Paper 5479.
Entorf H., Lauk M.
2008 *Peer effects, social multipliers and migrants at schools: An international comparison*, in "Journal of Ethnic and Migration Studies", 34, 4, pp. 633-54.
Fuchs T., Woessmann L.
2007 *What accounts for international differences in student performance? A re-examination using PISA data*, in "Empirical Economics", 32, 2, pp. 433-64.
Goldstein H.
2004 *International comparison of student attainment: Some issues arising from the PISA study*, in "Assessment in Education", 11, 3, pp. 319-30.
Goldstein H., Thomas S. M.
2008 *Reflections on the international comparative survey debate*, in "Assessment in Education: Principles, Policy and Practice", 15, 3, pp. 215-22.
Haahr J. H., Nielsen T. K., Hansen M. E., Jakobsen S. T.
2005 *Explaining student performance. Evidence from the international PISA, TIMSS, PIRLS surveys*, Danish Technological Institute.
Hanushek E. A.
2003 *The failure of input-based schooling policies*, in "Economic Journal", 113, 485, pp. 64-98.
Hanushek E. A., Kain J. F., Markman J. M., Rivkin S. G.
2003 *Does peer ability affect student achievement?*, in "Journal of Applied Econometrics", 18, 5, pp. 527-44.
Hanushek E. A., Woessmann L.
2006 *Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries*, in "Economic Journal", 116, 510, pp. 63-76.
2011a *Sample selectivity and the validity of international student achievement tests in economic research*, in "Economics Letters", 110, 2, pp. 79-82.
2011b *The economics of international differences in educational achievement*,

in E. A. Hanushek, S. Machin, L. Woessmann (eds.), *Handbook of the economics of education*, 3, North Holland, Amsterdam, pp. 89-200.

Hoxby C.
2000 *Peer effects in the classroom: Learning from race and gender variation*, National Bureau for Economic Research WP 7867, Cambridge (MA).

Jakubowski M.
2010 *Institutional tracking and achievement growth. Exploring differ-rence-in-differences approach to PIRLS, TIMSS and PISA data*, in J. Dronkers (ed.), *Quality and inequality of education. Cross-national perspectives*, Springer, Dordrecht-Heidelberg-London-New York.

Jürges H., Richter W., Schneider K.
2006 *Teacher quality and incentives. Theoretical and empirical effects of standards on teacher quality*, in "FinanzArchiv: Public Finance Analysis", 61, 3, pp. 298-326.

Jürges H., Schneider K.
2007 *Fair ranking of teachers*, in "Empirical Economics", 32, 2, pp. 411-31.

Manski C. F.
1993 *Identification of endogenous social effects: The reflection problem*, in "Review of Economic Studies", 60, 3, pp. 531-42.

Pischke J. F., Manning A.
2006 *Comprehensive versus selective schooling in England and Wales. What do we know?*, National Bureau for Economic Research WP 12176.

Rangvid B. S.
2007 *School composition effects in Denmark: Quantile regression evidence from PISA 2000*, in "Empirical Economics", 33, 2, pp. 359-88.

Rotberg I.
1995 *Myths about test score comparisons*, in "Science", 270, 5241, pp. 1446-8.

Rubin D. B., Stuart E. A., Zanutto E. L.
2004 *Potential outcomes view of value-added assessment in education*, in "Journal of Educational and Behavioural Statistics", 29, 1, pp. 103-16.

Rutkowski L., Gonzales E., Joncas M., Von Davier M.
2010 *International large-scale assessment data: Issues in secondary analysis and reporting*, in "Educational Researcher", 39, 2, pp. 142-51.

Schneeweis N., Winter-Ebmer R.
2007 *Peer effects in Austrian schools*, in "Empirical Economics", 32, 2-3, pp. 387-409.

Schuetz G., Ursprung H. W., Woessmann L.
2005 *Education policy and equality of opportunity*, IZA Institute for the Study of Labor, Discussion Paper 1906.

Todd P., Wolpin K. I.
2003 *On the specification and estimation of the production function for cognitive achievement*, in "The Economic Journal", 113, 485, pp. 3-33.

Waldinger F.
2007 *Does ability tracking exacerbate the role of family background for students' test scores?*, London School of Economics, mimeo.

Woessmann L.
2003 *Schooling resources, educational institutions and student performance: The international evidence*, in "Oxford Bulletin of Economics and Statistics", 65, 2, pp. 117-68.
2004 *How equal are educational opportunities? Family background and student achievement in Europe and the United States*, IZA Institute for the Study of Labour, Discussion Paper 1284.
2007 *Fundamental determinants of school efficiency and equity: German states as a microcosm for OECD countries*, IZA Institute for the Study of Labour, Discussion Paper 2880.