IFAU – INSTITUTE FOR
LABOUR MARKET POLICY
EVALUATION

Empirical studies of public policies within the primary school and the sickness insurance

Erica Lindahl

DISSERTATION SERIES 2009:1

Presented at the Department of Economics, Uppsala University

The Institute for Labour Market Policy Evaluation (IFAU) is a research institute under the Swedish Ministry of Employment, situated in Uppsala. IFAU's objective is to promote, support and carry out scientific evaluations. The assignment includes: the effects of labour market policies, studies of the functioning of the labour market, the labour market effects of educational policies and the labour market effects of social insurance policies. IFAU shall also disseminate its results so that they become accessible to different interested parties in Sweden and abroad.

IFAU also provides funding for research projects within its areas of interest. The deadline for applications is October 1 each year. Since the researchers at IFAU are mainly economists, researchers from other disciplines are encouraged to apply for funding.

IFAU is run by a Director-General. The institute has a scientific council, consisting of a chairman, the Director-General and five other members. Among other things, the scientific council proposes a decision for the allocation of research grants. A reference group including representatives for employer organizations and trade unions, as well as the ministries and authorities concerned is also connected to the institute.

Postal address: P O Box 513, 751 20 Uppsala
Visiting address: Kyrkogårdsgatan 6, Uppsala
Phone: +46 18 471 70 70
Fax: +46 18 471 70 71
ifau@ifau.uu.se
www.ifau.se

This doctoral dissertation was defended for the degree of Doctor in Philosophy at the Department of Economics, Uppsala University, November 07, 2008. The first three essays have been published by IFAU as Working paper 2007:24, Working paper 2007:25 and Working paper 2008:21. A part of the fourth essay contains revised versions of research published by IFAU as Report 2008:8.

ISSN 1651-4149

Empirical studies of public policies within the primary school and the sickness insurance

Doctoral dissertation presented to the Faculty of Social Sciences 2008

Abstract

Lindahl, Erica, 2008, Empirical studies of public policies within the primary school and the sickness insurance, Department of Economics, Uppsala University, *Economic Studies 111*, 142 pp, ISBN 978-91-85519-18-7
urn:nbn:se:uu:diva-9304
(http://urn.kb.se/resolve?urn=urn:nbn:se:diva-9304)

This thesis consists of five self-contained essays.

Essay 1 (with Elly-Ann Johansson) estimates the effect of attending an MA-class (MA-class) during grades 4-6 on students' cognitive skills. Using a unique survey with information on students, parents and teachers, we are able to control for many factors that could otherwise bias the results. We find a negative effect on the short-run skills, as measured by grade 6 cognitive tests.

Essay 2 compares results on national tests with teachers' assessment of student performance, by using Swedish data of grade 9 students (16 years old). I examine whether there are systematic differences correlated with gender and ethnic background. The results show that girls are more generously rewarded in teachers' assessment compared to test results in all three subjects studied. Non-native students are more generously rewarded in teachers' assessment compared to test results in two out of three subjects studied.

Essay 3 investigates the importance of gender and ethnic interactions among teachers and students for school performance. School leaving certificates assigned by the teacher is compared with results on comprehensive national tests. I find that a student is likely to obtain slightly better test scores in Mathematics, when the share of teachers of the same gender as the student increases. Correspondingly, ethnic minority students, on average, obtain better test scores in Mathematics, when the share of ethnic minority teachers increases. The positive same-gender effect on test scores is counteracted by a *negative* assessment effect. That is, conditional on test scores, same-gender teachers are slightly less generous than opposite-gender teachers when assessing students' performance. In Swedish and English no statistically significant effects are found.

Essay 4 (with Per Johansson) evaluates a multidisciplinary collaboration programme with the aim to prevent long-term sickness. The selection of eligible candidates was mainly based on register information, implying a good prerequisite of estimating the effects by using the same information. In addition, we have run a small experiment. Both evaluation approaches provide the same result: the programme prolongs rather than shortens the sickness absence spell. The hazard of leaving a sickness absence spell is reduced by on average 22 per cent. Two potential and complementary explanations for these negative results are (i) inefficiency in the organization outside the programme (i.e., in the rehabilitation process) and (ii) moral hazard in the sickness insurance.

Essay 5 (with Per Johansson) evaluates a policy to call sick-listed individuals without permanent employment to information meetings about the rights and duties associated with the sickness insurance system. The evaluation is based on experimental data: a random displacement of *when* the call is sent out. Comparing the survival functions of the individuals who are called immediately and those whose calls are delayed (by about 30 days) makes it possible to estimate a lower bound of the effect of being called on the sickness absence duration. The result suggests that calling a sick-listed individual without permanent employment to an information meeting reduces the sickness absence duration by at least 23 per cent on average.

Acknowledgements

Several people have contributed to this thesis in various ways. First of all I would like to thank my supervisor Per Johansson. His patience with all my questions and his ability to explain econometrics in a way that I could grasp has meant a lot to me. Per, thank you for all meetings during these years!

My co-supervisor, Andreas Westermark, I would like to thank for valuable comments on manuscripts and nice and encouraging talks in the corridor. I also want to acknowledge Sören Blomquist for help in the beginning of my PhD period and for encouraging me to apply to the PhD programme. I am also grateful for useful comments and help given by Peter Fredriksson and other referees and editors at IFAU as well as colleagues at the department.

The friendly administration staff, Monica, Katarina, Eva, Berit and Kristina, also deserves a special acknowledgment as well as Åke for all computer support.

During one part of my thesis work, I have collaborated with researchers from the department of Public Health and Caring Sciences. I would like to thank Ingrid, Ingrid, Ulrika and Ann-Sophie for nice meetings.

The friendly atmosphere at the department has made my working days a pleasure most of the time. Especially, I would like to thank Karin for all great discussions, Elly-Ann for nice co-work, Cilla and Hanna for all coffee-breaks.

Thanks are also due to my friends outside the department as well as my parents, sisters with families and parents-in-law. Especially, I would like to thank my mother-in-law, Birgitta, for all practical support during these years. Without her help this thesis would not have been completed today.

Finally, I would like to thank my daughters, Lova and Irma, for helping me to bring problems back to their right proportions and my husband, Björn, for all love and support. I try to approach science (as well as every weekday) as seriously *and* relaxed as you do, Björn!

Erica Lindahl, Uppsala, September 2008

Introduction

Economic research has spread into different topics. This has lead to a discussion about the definition of modern economics. Levitt and Dubner (2006) stated that “Economics is what economists do”. Becker (1976) defines economics “by the nature of the problem to be solved”. A general and non controversial definition is in terms of allocation of scarce means. This definition relates to fundamental challenges in the welfare society as equality and efficiency. The challenge of realizing equality and efficiency is a practical political issue. However, empirical economics or econometrics can generate knowledge about the consequences of a certain policy, i.e., how well a certain policy fulfils its explicit aims.

This thesis consists of five self-contained essays, which studies policy relevant questions in the primary school and the sickness insurance. A common theme is the methodological approach: to estimate a casual relationship. This summary first gives a brief introduction to the challenge of estimating a casual relationship. Then the different policy evaluations are presented in their respective context.

Estimating a casual relationship

Policy evaluation is about establishing causal relationships: whether a change in one variable causes a change in another. Observed correlations do not necessarily imply causality. For example, the fact that individuals with many years of schooling tend to have high earnings does not necessary imply that an additional year of schooling would result in a higher wage. It could be the case that individuals with much schooling have an ability and/or motivation implying both much schooling and high wages (confounding variable). It could also be the case that higher earnings enable agents to get more schooling (reverse causality).

In order to determine whether a change in one variable *causes* a change in another, one has to hold all other (relevant) factors fixed, i.e., the notion of ceteris paribus. How to do that is the scientific challenge.

The ideal design of a study in order to establish a causal relationship is an experiment. In social sciences experiments are rare. The usual methodological solution is instead to explore register data and identify a strat-

egy that can imitate an experiment, i.e. a quasi experiment. There are pros and cons with both strategies. In an experimental study the identification is (often) straightforward but the external validity may be a concern. For example, if the experiment is small, the study sample may not be representative for the population of interest. On the other hand, in observational studies using register data over the whole population, the external validity is (often) straightforward but the identification can (often) be questioned.

In this thesis, the first three essays explore register data. In the remaining two experimental data is used.

Primary school in Sweden

Education interests most people and the school system is always debated. During the last years Swedish media has reported about a school in crisis but facts are that according to international comparisons (PISA, 2006), Swedish students perform at average (in Mathematics) or above (in Readings) students in the OECD countries. However, striking, also in the Swedish context, is the international phenomena that girls outperform boys and the persistent (over time) gap between native and non-native students (PISA 2003 and NAEP 2004). These gaps are the starting point for essay 2 and 3.

Another characteristic of the Swedish school system is turbulence. During the last two decades several reforms have changed the working conditions for both students and teachers. Decentralization, school choice, school competition and goal steering are new concepts.[1] In addition, the school has experienced new teaching modes, which sometimes have been introduced against the will of parents and teachers. One example is the rapid increase of mixed-age (MA) classes (a class with students of different ages). The effect of attending an MA-class is the focus of essay 1.

MA-class in Sweden

MA-classes have increased rapidly in Sweden during the 1980:s and 1990:s. The argument for mixing students of different ages has sometimes been the pedagogical belief that students of different ages learn from each other (Vinterek, 2003). Especially for low performing students MA-classes have been claimed to be superior. However, the concept of mixed-age grouping is not clearly defined and the arguments used to support them are often contradictory. For example, the students' cognitive

[1] See Björklund A, Edin P-A, Fredriksson P and Kreuger A B (2004) for a discussion of the impacts of these reforms.

and non-cognitive skills are often assumed to be enhanced by the greater heterogeneity in the classroom, either because this heterogeneity in itself is beneficial for student development, or because it allows ability grouping (more homogeneity) within the classroom. However, the policy-relevant question is: what are the consequences of attending an Ma-class?

The scientific evidence on the effects of mixed-age grouping is ambiguous.[2] Among the Swedish studies there is *no* study using representative samples. Many studies are simply questionnaires collected among teachers in MA-classes. By using a representative sample essay 1 investigates the effect of attending an MA-class during grade 4-6 (age 10-12) on grade 6 cognitive tests and on grade 9 (age 15) credits. We find a negative effect on the short-run cognitive skills but no statistically significant effect on grade 9 credits. The effect (5 percentile points decrease) is about the same when studying girls, low-performing students or students with a non-Nordic background separately. Thus, our results do not support the argument about pedagogical benefits sometimes used for introducing MA-classes.

Differences in school performance across sub groups

The gap in school performance between genders and between ethnic groups has interested Economists all over the world. A suggested explanation is the unbalances depending on gender and ethnic background among the teachers; male teachers are under-represented, in primary school in particular, and non-native teachers are under-represented in general. The idea is that students may perform better when the teacher is of the same gender or ethnicity.

A number of studies have estimated student-teacher interaction effects. Some authors use formal tests as outcome variable, while others focus on subjective evaluations of students' performance and behaviour. For example, Dee (2004) explored the experimental data generated by STAR[3] and showed that assignment to an own-race teacher significantly increased the Mathematics and Reading test achievement of both black and white students. Further, Dee (2005) showed that both the gender and the ethnicity of the student are important for teachers' perceptions of student behaviour. The general empirical evidence from this literature suggests that student-teacher interactions matter for student outcomes, although the estimated effects are small (around a few percentile points increase). However, there is little understanding of the behavioural mechanism be-

[2] See Veenman, 1995 for a discussion of the international evidence.

[3] The Student/Teacher Achievemnt Study, i.e., the large-scale longitudinal experimental study of reduced class size, performed in Tennesse 1985-1989.

hind; does student-teacher interaction matter because of a change in student or in teacher behaviour? This knowledge is not only interesting from a behavioural perspective. Better understanding of the mechanisms is needed for guiding policies aiming at eliminating the gaps in school performance across genders and ethnic groups. For example, teacher training programs would be relatively ineffective if the dominant effect stems from student behaviour. The aim of essay 3 is to estimate the importance of having a teacher of the same gender or ethnic background, using Swedish data, and to shed some light on the behavioural mechanisms behind student-teacher interactions. Before discussing the results of essay 3, I will say some words about the Swedish grading system and present essay 2.

Measuring school performance

In Sweden, school performance is mainly measured by school leaving certificates. Those are set by the teachers and should reflect the teacher's overall impression of the student's knowledge and skills in the subject. It is clearly stated that the school leaving certificates should not reflect attention in the class room, diligence and ambition (Skolverket, 2004). As guidance for grading, the teacher shall use comprehensive national test results. The national test results should be used as the primary consideration for the school leaving certificates.

When studying the performance gap between boys and girls and between natives and non-natives, it is decisive how school performance is measured. According to Swedish data on grade 9 students (16 years old), the gender and ethnic differences are larger in school leaving certificates than in national test results. The aim of essay 2 is to clarify this observation – to investigate if the relationship between school leaving certificates and national test results is systematically correlated with gender and ethnic background. The result is that if the student is a girl, she is on average better rewarded than a boy with the same result on the national test. This result has been observed in earlier studies in Education (Skolverket, 2006; Nycander, 2006; Wester-Wedman, Gisselberg, Mattsson and Wedman, 1988 and Emanuelsson and Fischbein, 1986). New evidence, as far as I know, is that also non-native born students on average are more generously rewarded in two out of three subjects studied.

Student-teacher interactions

Several reasons have been raised for why student-teacher interactions may matter. Dee (2004) distinguishes between passive teacher effects and active teacher effects. An example of the former is that teachers may

serve as role-models for their students while the latter refers to the idea that teachers treat (unconsciously or consciously) students differently depending on gender and ethnicity. This distinction is problematic since the term *interaction* suggests a mutual behavioural change appearing in the interaction. However, from a policy perspective it is interesting to distinguish if the interaction effect *mainly* stems from passive or active teacher behaviour. Studying the effect of student-teacher interactions on school leaving certificates and conditioning on the national test results, could say something about the importance of changed teacher behaviour in the interaction. I call the additional potential change in teacher behaviour captured by school leaving certificates an assessment effect.

The results of essay 3 suggest that student-teacher interactions matter both for student performance on national tests and for how teachers assess students. Students perform slightly better on the test when the teacher is of the same gender or ethnicity (in Mathematics). In the gender case, there seems also to be an assessment effect. Surprisingly this assessment effect seems to work in the opposite direction: teachers with the same gender are less generous when setting the school leaving certificate. The outlined theoretical prediction in earlier literature has rather suggested that the different parts of the student-teacher interaction effect work in the same direction.

Sickness insurance

Moral hazard

An important part of the welfare state is risk sharing. A substantial amount of literature in economics has focused on optimal insurance design. The crux is asymmetric information and moral hazard. The insured has more information about his/her actions than the provider of the insurance (asymmetric information). This situation implies that the insured may be tempted to behave differently from how he/she would behave if he/she was fully exposed to the risk (moral hazard). The practical solutions to mitigate these problems are, for example, experience rating, limited compensations levels (co-insurance) and monitoring and sanctions.[4]

In the case of sickness insurances (SI) many countries, such as Sweden, provide a compulsory and publicly provided insurance. Compulsory SI can be motivated by solidarity reasons and the free-rider

[4] See Larsson, L., Kruse, A., Palme, M. and Persson, M. (2005) for a discussion about a sustainable sickness insurance.

argument[5]. A public insurance can be motivated by, for example, advantages of economics of scale. In such a system the major concern is *ex post* moral hazard, i.e., the insured is tempted to claim more sickness cash benefit than he/she would if exposed to the full risk.[6]

The phenomenon of *ex post* moral hazard implies an excess use of the SI. Several studies confirm that moral hazard in the SI- context is a real problem (Broström, Palme and Johansson 2004; Henreksen och Persson, 2004 and Johansson and Palme, 1996; 2002 and 2005). The approach among these studies has been to investigate if sick-listed individuals respond to changes in the benefit caps. The general conclusion is that the lower the cost of being sick-absent, the more likely that the individual is on sick-leave.

Further evidence of moral hazard is related to interplay between the sickness- and the unemployment insurance. Larsson and Runeson (2007) conclude that the higher the sickness benefits are in relation to the unemployment benefits, the larger the probability of reporting sick if unemployed. Further, Larsson (2006) and Henningsen (2006) show that the probability of reporting sick among the unemployed increases drastically as the expiration date of the unemployment period approaches. The literature suggests that moral hazard is a real problem. The policy relevant question is hence how this problem should be mitigated.

The extent of moral hazard

Recent economic studies suggests that the extent of moral hazard in the SI- context is determined by at least two other components: i) the personal moral which in turn seems to be affected by the social context and ii) the degree of monitoring.

The importance of local social norms for explaining sickness absence behavior has been analyzed by Lindbeck and Persson (2006). They stipulate a theoretical framework: the disutility (the stigma) of being sickness absent is negatively associated with the number of sick-listed individuals. This idea is in accordance with the fact that the sickness absence rate varies across regions in Sweden, as demonstrated in Lindbeck, Persson and Palme (2004) and Palmer (2006). Briefly, there seems to be local differences in norms about how to use the SI.

During the last years a few papers have studied the importance of social interactions for explaining absence due to sickness (Hesselius, Jo-

[5] That is, individuals tend to *not* take an insurance because they believe others will pay for them in case of sickness.

[6] In contrast to *ex ante* moral hazard, i.e., the individual conducts a more unhealthy life than he/she would have done if fully exposed to the risk.

hansson and Vikström, 2008; Lindbeck, Persson and Palme 2007 and Ichino and Maggi, 2000). The hypothesis is that an individual's sickness absence behaviour is affected by others usage of the SI. In the Swedish context, the social interaction effect seems to be large: a one per centage exogenous increase in mean absence within the network would lead to an immediate increase in the individual hazard (from work absence to work) by 0.57 per cent, according to Hesselius *et al.* (2008). A policy implication of these studies is that if we change the norms in the target group, there is a spill-over. That is, people in the surrounding is probably also affected by the change in norms.

One study addresses monitoring in the SI- context. Hesselius, Johansson and Larsson (2005) show that postponing the first formal point of monitoring (a requirement for a doctor's certificate) during a sickness-absence spell increased short-term sickness absence by 6.6. per cent.

Swedish figures

Figure 1 summarizes the use of the sickness- and unemployment insurances in Sweden during the last decades.[7] The figure shows a long-run increase in the use of both insurances. This long-run trend is mainly due to an increase in the number of unemployed and individuals who are early retired. The sickness absence (with sickness benefit) rate has not increased significantly in a long-run perspective. However, notably is the large variability over time. This variability can hardly be explained by chocks in the health among the Swedes.

The figure also demonstrates a negative correlation between the use of the sickness- and unemployment insurances; the sickness absence (with sickness benefit) is low when the unemployment is high and vice versa.[8] There are two possible explanations for this pattern (Larsson *et al* 2005). The first focuses on the composition of employment: when the unemployment rate is low more people with health problems and with low work morale is employed. The second explanation stresses the disciplinary effect of unemployment: the incentives to turn up to work are weakened in times of low unemployment.

During the late 1990s the share on sickness benefits increased by more than 100 per cent, which mainly can be explained by an increase in long-term sickness (sick spell durations > 365 days) (Försäkringskassan,

[7] This figure is presented and discussed in Larsson *et al* (2005).

[8] At the same time there is a positive cross-section correlation between the *increase* in sickness absence in the first years of the 2000s and unemployment in the late 1990s (Larsson *et al.* 2005). This strongly suggests that sickness insurances have to a large extent been used as a form of unemployment insurance.

2006). The trend culminated in 2003. This change coincided with the point of time when a number of different programmes was launched by the Swedish Social Insurance Agency (SSIA) in order to reduce the sickness absence. For example, in 2003 SSIA introduced a new national programme in order to reduce costs associated with sickness absence. This programme involved new routines for handling sick-listed individuals. An explicit aim of this programme was to stress working ability rather than incapacity with respect to sickness (Försäkringskassan, 2007). That is, an explicit ambition of changing norms both among the personnel and among individuals in general.

The following sections present evaluations of two other programmes introduced by SSIA. The first addresses the increase of long-term sickness (essay 4). The second is about norms (essay 5).

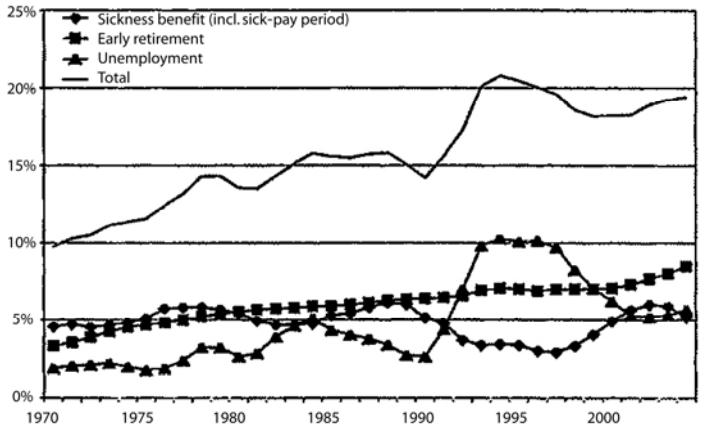


Figure 1: Share of the Swedish population (age 20-64) who received benefits from the sickness- and unemployment insurances (calculations based on Arbetskraftsundersökningen, AKU, Socialförsäkringsutredningen)

Long term sickness

Prevention of long-term sickness has been one of the most prioritized questions at the SSIA during the last years. The efforts made have been influenced by recommendations stated by the Swedish National Board of Health and Welfare (Socialstyrelsen) and the Swedish Council on Technology Assessment in Health Care (SBU). Before discussing these recommendations, I will say some words about the Swedish sick-leave process.

Claiming sickness benefits in Sweden involve several different local authorities. The medical doctor judges the working incapacity and has the

responsibility for the medical rehabilitation. The case-worker at SSIA decides upon sickness benefits based on a certificate provided by the medical doctor. In addition, the case-worker has the responsibility for the non medical rehabilitation (i.e., contacts with a potential employer). This process cannot work without an efficient collaboration between the actors involved. This process is not unique for Sweden. In fact, how this collaboration should be organized is an international research field.[9]

In 2003, SBU claimed that a crucial factor for the high sickness-absence rate, and the long-term sickness absence in particular, was the lack of an efficient collaboration between the primary health care organization and the social insurance office. Further, the SBU report stressed the importance of involving behavioural and physiotherapy competence in the decision about reporting sick. The reason is that the two most common diagnoses among long-term sick-listed individuals were (are) musculoskeletal disorders and/or mental problems.

In Uppsala County, the local office at SSIA put these ideas into practice. Between 2004 and 2006 newly sick-listed individuals who were assessed to run the risk of becoming long-term sick-listed were discussed in multi-disciplinary teams, named Resursteam (RT). The team consisted of a medical doctor, a case worker, a behaviourist and a physiotherapist. By using the combined skills of the members of the team, RT should suggest a suitable rehabilitation for the insured individual.

The aim of essay 4 is to evaluate the effect of RT on sickness absence duration. In contrast to the explicit aim of the programme, the result is that RT prolongs the sickness-absence period with on average 22 per cent or by approximately 60 days. How should this discouraging result be interpreted? There could be several different explanations specific for RT. For example, it may be inefficient to include professionals who do not personally meet the sick-listed individual, as was the case in RT.[10] Another explanation is connected to moral hazard and norms. RT may have implied a less severe monitoring of the individual's health status. The reasoning is as follows. RT was implemented in 2004 when the sickness absence had started to decrease (see Figure 1). If this drop was due to a general change in norms, there are reasons to believe that this change did not embrace RT. The target group for RT was individuals with health problems that were difficult to assess (mental disorders or psychological problems without a further specification). When an individual was initi-

[9] See Dowling, Powell and Glendinning (2004) for an overview of the situation in Great Britain and Schmitt (2001) for the US case.

[10] This explanation has support in a survey among the personnel, see Anderzén et al (2008).

ated into RT, the health problems became confirmed and the health eligibility criterion to be on sickness benefits was not further reviewed in the same way as for the comparison group. Important to note is that whether the individual *should* have sickness benefits or not from a health, social or economic perspective is a normative question beyond the scope of this study. The point here is that RT may have been an "easy solution" for the actors involved: the sick-listed individual *and* the medical doctor *and/or* the case-worker.

Norms and monitoring

The last essay is about norms and moral hazard. In Uppsala County, local case-workers started a program with the aim to reduce excess use of the SI. The program, called information meetings (IM), implied that sick-listed individuals were called to a meeting about the rights and duties associated with the SI. Attendance at the meeting was mandatory; the sick-listed individual had to report back or show up at the meeting. The target group was sick-listed individuals without permanent employment. The aim of essay 5 is to estimate the effect of receiving a call to an IM on sickness absence duration. The result suggests that sickness absence duration is reduced with at least 23 percent on average. What can we learn from this result?

The large magnitude of the effect should be interpreted with the target group (sick-listed without permanent employment) in mind. The reason is that there are economic incentives to be on sick-leave rather than on unemployment benefits.[11] During the last years there have been institutional changes aimed to harmonize the sickness- and unemployment insurances.[12] However, our result suggests that moral hazard is a real problem also in the present institutional setting.

The explicit aim of IM was to affect norms and thereby reduce excess use of the SI. In addition to any potential effect on norms induced by attending the meeting, the call could be seen as a "threat".[13] The sickness benefit payment was tied to participation in an information meeting and the obligation to participate might have induced a "threat" leading to less usage of the SI. The low attendance at the meetings (30 per cent) suggests that a significant part of the effect stems from a "threat".

[11] See SOU (2007) for a discussion about the interaction between sickness absence and unemployment.
[12] Probably due to the findings by Larsson, 2006 and Larsson and Runesson, 2007
[13] This term is used in the context of tying benefit payments to labour market programmes; compulsory programmme participation motivates individuals to leave the unemployment insurance (Geerdsen, 2006 and Black *et al* 2003).

Unfortunately we are not able to clearly disentangle the two components of the total effect: a “threat”-effect and a change in norms. This distinction would be interesting to do since the two parts probably have different long-run implications. If the main effect stems from a “threat”, the changed behaviour about using the SI is tied to compulsory attendance at the meeting implying that a changed behaviour would only be observed under a similar ”threat” (or monitoring). If instead the effect stems from attending the meeting, it is reasonable to believe that there are longer-run effects through changed norms about using the SI. Furthermore, recent studies on the importance of social interactions suggest that such a change in norms would have important spill-over effects through endogenous effects (Hesselius *et al.*, 2008; Lindbeck, *et al*., 2007; and Ichino and Maggi, 2000).

For guiding policies aiming to reduce moral hazard, better knowledge about the behavioural mechanisms associated with the usage of the SI is needed. However, the results from essay 5 suggest that it may be possible to prevent excess usage of SI with quite small measures. In a system with different levels of replacement rates in the sickness- and unemployment insurances some sort of control seems necessary for preventing unintended flows between the systems. Calling sick-listed to IM is both unexpensive and can hardly be regarded as insulting, which is a concern sometimes raised regarding control.

References

Anderzén, I., Demmelmaier, I., Hansson, A-S., Johansson, P., Lindahl, E. and Winblad, U.: 2007, Utvärdering av samverkan i resursteam inom Försäkringskassan och Landstinget i Uppsala län – ett samverkansprojekt för att minska sjukskrivningstiden, Rapport september 2007.

Becker, G. S.: 1976, The economic approach to human behaviour, The University of Chicago, Press Chicago and London.

Björklund, A., Edin P-A., Fredriksson P. and Kreuger, A. B: 2004, Education, equality and efficiency – An analysis of Swedish school reforms during the 1990s, IFAU, Report 2004:1.

Broström, G., Johansson, P. and Palme, M.: 2004, Economic incentives and gender differences in work absence behavior, Swedish Economic Policy Review, vol 11, 33-63.

Dee, T.: 2004, Teachers, race and student achievement in a random experiment, Review of Economics and Statistics, vol. 8681, 195-210.

Dee, T.: 2005, A teacher like me: does race, ethnicity or gender matter?, American Economic Review, vol. 95(2),158-65.

Emanuelsson, I. and Fischbein, S.: 1986 Vive la difference? A study on sex and schooling, Scandinavian Journal of Educational Research 30, 71-84.

Försäkringskassan: 2006 The scope and financing of social insurance in Sweden 2004-2007, Försäkringsdivisionen utvärderingsavdelningen.

Försäkringskassan: 2007, De gemensamma metoderna i sjukförsäkringen – hur blev det?, Försäkringskassan redovisar 2007:8.

Henningsen, M.: 2006, Moving between welfare payments – The case of sickness insurance for the unemployed, Memorandum no 04/2006, Departement of Economics, University of Oslo.

Henrekson, M. and Persson, M.: 2004, The effects on sick leave of changes in the sickness insurance system, Journal of Labor Economics, Vol. 22, No. 1, 2004.

Hesselius, P., Johansson, P. and Vikström, J.: 2008, Monitoring and norms in sickness insurance: empirical evidence from a natural experiment, Working Paper 2008:08, IFAU.

Hesselius, P., Johansson, P. and Vikström, J.: 2005, Monitoring sickness insurance claimants: evidence from a social experiment, Working Paper 2005:15, IFAU.

Ichino, A. and Maggi, G.: 2000, Work environment and individual background: explaining reginal shirking differentials in a large Italian firm", The Quarterly Journal of Economics, 115 (3), 1057-1090.

Johansson, P. and Palme, M.: 1996, Do economic incentives affect work absence? Empirical evidence using Swedish micro data, Journal of Public Economics, 59, 195-218.

Johansson, P. and Palme, M.: 2002, Assessing the effects of a compulsory sickness insurance on worker absenteeism, Journal of Human Resources, 37:2, 381-409.

Johansson, P. and Palme, M.: 2005, Moral hazard and sickness insurance, Journal of Public Economics, 89, 1879-1890.

Larsson, L, Kruse A., Palme, M. and Persson, M.; 2005 Välfärdsrådets rapport 2005: En hållbar sjukpenningförsäkring, SNS Förlag, Stockholm.

Larsson, L.: 2006, Sick of being unemployed? Interactions between unemployment and sickness insurance, The Scandinavian Journal of Economics, vol. 108, s 97-113.

Larsson, L. and Runeson, C.: 2007, Moral hazard among the sick and unemployed: evidence from a Swedish social reform, Working Paper 2007:8, IFAU.

Levitt, S. D. and Dubner, S. J.: 2006, Freakonomics, Penguin book.

Lindbeck, A. and Persson, M.: 2006, A model of income insurance and social norms, Seminar paper No. 742.

Lindbeck A, Palme, M. And Persson, M.: 2004, Sjukskrivningar som ett socialt fenomen, Ekonomisk debatt 4, 50-62.

Lindbeck, A., Palme, M. and Persson, M.: 2007, Social interaction and sickness absence, IFN Working Paper No. 725.

NAEP 2004, Trends in academic progress – three decades of student performance in Reading and Mathematics, US Department of Education, Institute of Education Sciences, NCES 2005-464.

Palmer, E.: 2006, Sjukförsäkring, kulturer och attityder – fyra aktörers perspektiv, Försäkringskassan Analyserar 2006:16.

PISA 2003, Problem solving for tomorrow's world, First measures of cross-curricular competencies from PISA 2003, Programme for international student assessment, OECD.

PISA 2006, Science competencies for tomorrow's world, vol. 2 - data: presents the PISA 2006 full data set underlying Volume 1. Programme for international student assessment, OECD.

SBU-rapport 2003:167.: Sjukskrivning – orsaker, konsekvenser och praxis – en systematisk litteraturöversikt. Stockholm, Statens beredning för medicinsk utvärdering.

Skolverkets allmänna råd 2004, Allmänna råd och kommentarer, Likvärdig bedömning och betygsättning.

Skolverket: 2006, Könsskillnader i måluppfyllelse och utbildningsval, rapport nr 286.

SOU: 2007, Arbetslösa som blir sjuka och sjuka som inte blir arbetslösa – samtal om socialförsäkringen, nr 16, Socialförsäkringsutredningen, Statens offentliga utredningar.

Veenman, S.: 1995, Cognitive and non cognitive effects of multigrade and multi-age classes: a best-evidence synthesis, Review of Educational Research, 65 (4).

Wester-Wedman, A., Gisselberg, K., Mattsson, H. and Wedman, I.: 1988, Vilka gynnas vid betygsättningen?, Pedagogiska rapporter Nr. 21, Pedagogoiska institutionen Umeå Universitet.

Vinterek, M.: 2001 Åldersblandning i skolan - elevers erfarenheter, Doktorsavhandling i Pedagogiskt arbete (1), Umeå universitet.

Essay 1: The effects of mixed-age classes in Sweden⋆

Introduction

Mixed-age classes are a common phenomenon in schools both in Sweden and in other countries. In mixed-age classes (henceforth MA-classes), students from different grades are mixed into one class for two major reasons: either out of demographic and economic necessity (too few children in each grade to form a class) or because it is believed that these classes have pedagogical benefits. For example, it is argued that students of different age and school experience interact and learn from each other. This belief has contributed to the rapid increase of MA-classes in Swedish schools during the 1980:s and 1990:s.

However, the scientific evidence on the effects of mixed-age grouping is ambiguous. Among the Swedish studies used to motivate the introduction of MA-classes there is, to our knowledge, *no* study using representative samples. Many studies are simply questionnaires collected among teachers in MA-classes. International studies are available, but they are also of varying quality with very few studies using representative samples. The international studies often yield contradictory results, with both positive, zero and negative effects of MA-classes. However, most studies conclude that the effect, if any, is small in magnitude.

From an economic point of view, investigating the effect of MA-classes is important since it may be one possible way towards greater cost-efficiency within schools. If it is the case that MA-classes, as is often

⋆ Co-authored with Elly-Ann Johansson. We are grateful to Peter Fredriksson and Per Johansson for valuable guidance. We would also like to thank Mikael Elinder, Patrik Hesselius, Jenny Nykvist, Peter Skogman Thoursie, Andreas Westermark and seminar participants at the Department of Economics, Uppsala University, for valuable suggestions and comments. Åsa Ahlin is acknowledged for the research idea. The financial support from the Swedish Council for Working Life and Social Research, FAS (dnr 2004-1222) is also acknowledged.

claimed, are a less expensive way to organize students than traditional classes, and if the students in these classes perform equally well or better than students in traditional classes, introducing MA-classes in a larger scale would be an efficient way towards reduced costs and/or increased student performance. This is particularly interesting in relation to one of the most debated policies in the economic and educational literature during the last years, namely reducing class-size. In contrast to class-size reductions, introducing MA would imply practically no extra costs.[14]

Examining the effect of MA-classes also sheds light on the question of how knowledge is produced. Economic research has mainly focused on quantitative aspects of education – if and how much resources matter for student achievement. But equally important are more qualitative aspects of the educational production function, and the effect of MA-classes is one such aspect.

The purpose of this paper is to estimate the effect of MA-classes in Sweden on students' cognitive skills. We focus both on short-term effects on grade 6 cognitive tests and on long run effects on grade 9 credits. We also allow the effect of attending an MA-class to vary between different groups of students considered potentially important: girls, low performing students and students with non-Swedish background.

A rich and representative data set allows us to control for many potential selection problems. In addition to register data on important socio-economic variables, we have access to a unique survey with information on parents and teachers and their attitudes towards school related issues.

The results show a negative effect of attending an MA-class in grades 4–6 on the grade 6 cognitive tests. In addition, this effect is not statistically different for girls, low performing students or students with a non-Swedish background. The point estimate of the effect of MA-class-attendance on grade 9 credits is negative but not statistically significant.

Background

MA-classes in Sweden

MA-classes have always been present to some extent in most countries' school systems. Historically, it has often been the only available mean to form a class due to limited numbers of children of the same age in a particular area. In Sweden, as in most developed countries, MA-classes were abandoned in favour of single-age groups as the population grew larger.

[14] For reviews of the class-size literature, see: Krueger (2003) and Hanushek (1999).

Around 1980, however, the belief that MA-classes were pedagogically superior to single-age groups started to spread, and in the years to come, the number of MA-classes increased rapidly (Vinterek 2003). In 2000, approximately one third of all Swedish students in the first three years of school attended MA-classes and about one fourth of the students in grades 4 and 5. That is nearly twice as many as only five years earlier. The share of students attending an MA-class during the last three years of compulsory schooling in Sweden is still rather small; about 2 percent of all students in these grades were in mixed-age groups between 1996 and 1998.

MA-classes are introduced primarily out of two different reasons: economic or pedagogical. We do not know whether the rapid increase in the number of MA-classes is due to pedagogical reasons or economic reasons (Vinterek, 2003). There is some evidence that pedagogical motives dominated in the lower grades 1 to 3, whereas economic motives dominated in the higher grades 4 to 6 (Sandquist, 1994). In grades 7 to 9, mixed-age classes are scarce, and if they do exist, they tend to be motivated by demographic necessities (Sandquist, 1994). There is also evidence that mixed-age classes are more prevalent in schools with many low performing students (Vinterek, 2003). The initiative to start an MA-class has usually come from groups of teachers within a school, often supported by the school management (Vinterek, 2003). However, since the beginning of 1990 it seems to be the case that MA-classes have been introduced by politicians against the will of teachers and parents (Vinterek, 2003; Edlund and Sundell, 1999; Sundell, 2002 and Sandquist, 1994).

According to Sandqvist (1994) and Vinterek (2003), there is some evidence suggesting that students in MA-classes work more individually. One reason could be the large heterogeneity within the class, making cooperation between students and group activities more difficult since they are at different knowledge levels. This implies that learning takes place through quiet reading and writing more than through listening and speaking. (This is somewhat contradictory, since one common argument for MA-classes is that the larger heterogeneity within the class enhances learning through group activities.) There are also tendencies to grade-specific teaching also in MA-classes. However, there are large differences depending on subject. Social sciences are often taught to all grades simultaneously, and leave large possibilities for group work and a thematic organization of the subject. In subjects like Mathematics or Athletics teaching is more often done separately for each grade. This can be achieved in different ways. Sometimes one grade within the class works individually with one subject while other grade listens to the teacher lec-

turing in another subject. In other cases, the highest graders stay in school later in the afternoon and have time to learn more advanced Mathematics when their younger classmates have left for the day.

Arguments for and against MA-classes

There is no clear consensus in the literature about what mixed-age classes and mixed-age teaching really is. According to the educationalist Monika Vinterek (Vinterek, 2001), the arguments used in favour of MA-classes are mainly found in popular science magazines while the arguments against are found in scientific journals. The concept of mixed-age grouping is often not clearly defined. Sometimes it denotes all classes consisting of children of different ages, sometimes also the teaching needs to be of a special fashion (usually more group work). In addition, the arguments used to support mixed-age classes are often contradictory. For example, the students' cognitive and non-cognitive skills are often assumed to be enhanced by the greater heterogeneity in the classroom, either because this heterogeneity in itself is beneficial for student development, or because it allows ability grouping (more homogeneity) within the classroom.

In the following we give an overview of the most commonly used arguments for and against MA-classes. The literature is mainly concerned with the supposed benefits of MA-classes. In contrast, arguments against such classes are scarcer. Hence, also this exposition will focus mainly on the pro-arguments but the reader should keep in mind that this does not mean that they are supported scientifically.

Veenman (1995) discusses the following benefits of MA-classes: MA-classes are claimed to enhance the children's security and confidence as they form relationships with a wider variety of children. MA-classes also invite cooperation and children benefit from learning from and teaching each other. Furthermore, MA-classes are considered to have a more relaxed atmosphere, and be more stimulating as similar but not equally able children meet. It is also claimed that the self-concepts of slower, older students are specially enhanced when they are asked to tutor younger students.

In order to motivate the introduction of MA-classes in Sweden, the following arguments have been used by many local politicians in local school directives (Sandqvist, 1994). MA-classes enable greater adaptation to individual maturity in different subjects and generate greater social training since the group is more heterogeneous with respect to age. In

addition, mixed age grouping is claimed to give rise to more acceptance for deviating behaviour among classmates.

Some local politicians also refer to the pedagogical idea that students are assumed to be naturally curious and hungry for knowledge and that children spontaneously learn from each other and willingly teach each other.[15] Given this view of schooling and children, a more heterogeneous group is desirable. Another argument, connected to the former, is that the new post modern information intensive society requires knowledge about how to *search* for information. To work in project teams and to cooperate among students in order to search for information are new features in the school directives that is claimed to fit well with MA-teaching.

Sundell (1995) also describes the arguments used in directives from the former Swedish National Agency for Education (Skolöverstyrelsen). Among the arguments in favour of MA-classes, the supposed positive impact on students' cognitive development is claimed to stem from the teaching adapted to the individual that is connected with MA-classes, as well as the idea that younger students learn from their older peers. The reason for the former argument is that in an MA-class, working groups are formed in accordance with the individual child's mental maturity rather than its actual age.

Further, it is often claimed that the individually adapted teaching connected with MA-classes specially benefit low performing students. The reasons are several. First, it is argued that the individually adapted teaching results in more teaching time to those in special need. Second, teaching in an MA-class is to a higher degree organized in small groups, which benefit low performing students. Finally, as stated above, in an MA-class low performing students have the possibility to compare themselves with younger children and in this way they do not need to perform worst.

The arguments used against MA-classes are in many cases similar to the ones used in favour of them. For example, it is argued that MA-classes impose a greater workload on the teachers and that most teachers are not adequately prepared to deal with MA-groups (Veenman, 1995). This can be compared with a similar pro-argument: it is claimed that MA-classes are supposed to give the teachers a better working environment as only a share of the class is new every year (Sundell, 1995). Thus, there is no clear theoretical consensus about the mechanisms behind MA-classes.

[15] The Montessori pedagogy is mentioned in some local school directives (Sandqvist, 1994).

Earlier studies

The scientific evidence on the effects of MA-classes is ambiguous and many studies are of poor quality. For example, Veenman (1995) summarizes evidence from 56 international studies. There were no experimental studies at all, and virtually no studies based on representative samples of the student population with well-defined treatment- and comparison groups. Many studies did not even make any attempts to condition on initial differences between students in MA-classes versus traditional classes. The studies yield contradictory evidence, and when summarizing the results from the studies of best quality, the average effect of attending an MA-class becomes zero. The reason for this zero effect is discussed by Burns and Mason (1996). They argue that selection of better students and/or teachers into MA-classes are counteracted by less effective instruction in these classes.

Using Swedish data, Sundell (2002) estimates the effect of MA-class-attendance in grade 2 on a number of abilities. Important to note is that the 752 students included in his study are not randomly sampled. When controlling for social and pedagogical background as well as initial achievements, the results show that students in MA-classes performed worse than other students in 12 out of 13 dimensions. The MA-students had for example lower mathematical ability, a less developed vocabulary and were perceived as more shy and troublesome by their teachers. However, they did perform better in reading comprehension.

Data

In this section, we describe the data used and show the differences between MA- and traditional classes in terms of some important covariates.

Data sources

Our main data source is a representative and stratified panel data set, Student Panel 4, provided by Statistics Sweden[16] where one cohort of students is followed through grade 3 to 9. In the first stage 35 municipalities were selected. In the second stage a random sample of grade 3 classes within these municipalities were selected.[17] Within the selected classes,

[16] Participation in the study is voluntarily. About 4 percent of the originally sampled students were not able to or chose not to participate in the study.
[17] For more information about how the data was collected, see SCB (1996)

information from all students in grade 3 was collected. This means that for students in traditional classes, we have information on the whole class, while for students in MA-classes, we only have information on the part of the class that spends their third year in school in 1992. That is usually one half or one third of the class, depending on how the MA-class is constructed.

The sampling of grade 3 classes were done in 1992; hence most students are born 1982 and finish 9th grade in 1998. It is important to note that all students sampled in grade 3 are followed over time, regardless of whether they move or change class; hence, regarding these data there is virtually no attrition. The panel includes approximately 8500 individuals.

This panel data set is combined with additional register data from the data bases RAMS and LOUISE provided by Statistics Sweden. These data include socioeconomic background information such as parental education and immigrant status. Most of this information is measured in 1998. We focus on students who finish 9th grade the expected year 1998 or later.[18]

In addition, we have access to a survey with information on students, parents and teachers and their attitudes towards school-related issues. This information was collected when the students were in grade 6 by the Department of Education at Göteborg University. Parents were asked about their involvement in school issues and if they actively had chosen school or simply accepted the nearest one. Teachers were asked about their work experience, whether they had a formal degree, and their attitude towards homework. Results from grade 6 cognitive tests of the students were also collected (a description of these tests is given in Appendix).

Due to non-response, survey information is only available for a subsample of the original sample. Of the individuals in the original sample, 85 percent have undertaken the grade 6 test, and 54 percent has answered all of the survey questions we use. It is this reduced sample we use for our analyses. Table A1 in Appendix shows the difference between the raw register data, data with test results available (the basic sample), and data with all survey information available (our survey sample). The differences in means are very small when comparing the raw data and the basic sample. In 6 out of 27 cases there are statistically significant differences at the ten percent level and in these cases the magnitudes of the differences are small. Comparing the raw data with the survey sample, there are some additional differences. The survey sample seems to consist

[18] 16 students finished school one year earlier, but due to a changed grading system, we do not include these in our sample.

of a slightly more "privileged" group of students than the raw data. For example, students in the survey sample have higher average credits and grade 6 test results, they are more seldom given special help or mother tongue education in grade 3, and their parents are better educated.

Differences between MA-classes and traditional classes

Table 1 presents descriptive statistics for students in MA- and traditional classes in our survey sample. First of all, we can note that students in MA-classes have lower scores on the grade 6 cognitive tests. This could have two different explanations: one is that MA-classes are detrimental to student achievement; another is that we have negative selection into MA-classes. Regarding parental and student characteristics, the groups are relatively similar with two exceptions. Students in MA-classes have to a less extent mothers with university degree, and are more often given mother tongue education in grade 3.

Regarding teacher and class characteristics, the differences are more striking. MA-classes are usually smaller. The teachers in MA-classes are less experienced, have spent a shorter time in each class, and are more often on leave than teachers in traditional classes. The teachers' attitudes also differ.[19] Teachers in MA-classes put less emphasis on homework, basic knowledge and formal tests than teachers in traditional classes. MA-class-teachers also believe student influence to be more important than their colleagues in traditional classes. Hence, from these descriptive statistics it seems as if the pedagogical environment for students in MA-classes differs substantially from the environment in traditional classes.

[19] The attitude variables are measured on a 1-5 scale; the more important a teacher regards the issue, the higher the number, see Appendix for more details.

Table 1: Descriptive statistics of students attending an MA-class during grades 4–6 versus others, survey sample

	MA-class in grades 4–6		**Ordinary class in grades 4–6**	
Individual characteristics	*Mean*	*Sd*	*Mean*	*Sd*
Grade 9 credits	51.49	28.34	52.67	28.66
Grade 6 test results	46.99	28.23	52.20***	28.80
Female student	0.48	0.50	0.50	0.50
Early start	0.01	0.11	0.01	0.08
Late start	0.03	0.16	0.02	0.15
Birth month	6.11	3.43	6.27	3.35
Help in grade 3	0.16	0.37	0.19	0.39
Mother tongue in grade 3	0.11	0.31	0.08*	0.27
Non-Nordic student	0.07	0.25	0.06	0.24
Mother sec. educ.	0.46	0.50	0.46	0.50
Mother univ. educ.	0.26	0.44	0.33***	0.47
Father sec. educ.	0.37	0.48	0.40	0.49
Father univ. educ.	0.19	0.39	0.22	0.41
Father educ. miss	0.20	0.40	0.22	0.41
Mother educ. miss	0.06	0.23	0.06	0.23
Father non-Nordic	0.09	0.29	0.10	0.31
Mother non-Nordic	0.11	0.31	0.10	0.30
Birth country miss	0.00	0.06	0.00	0.03
Father birth country miss	0.03	0.18	0.03	0.16
Mother birth country miss	0.02	0.12	0.02	0.13
Parent attitude: active school choice	0.15	0.36	0.15	0.36
Parent attitude: parent help	1.87	0.89	1.91	0.95
Parent attitude: parent active	2.39	1.05	2.34	1.03

Table 1: Cont'd

	MA-class in grades 4-6		**Ordinary class in grades 4-6**	
Teacher and class characteristics[1]	*Mean*	*Sd*	*Mean*	*Sd*
International school	0.0000	0.0000	0.0007	0.0265
Confessional school	0.0032	0.0562	0.0028	0.0530
Special school	0.02	0.14	0.03	0.18
Grade 9 students	101.56	39.25	114.31***	40.42
Few grade 9 students	0.07	0.25	0.01***	0.11
Teacher experience	18.44	10.62	20.12***	9.65
Teacher not qualified	0.04	0.20	0.04	0.20
Class size	18.31	7.19	23.67***	5.91
Small class	0.13	0.33	0.01***	0.11
Large class	0.21	0.41	0.37***	0.48
Share boys	0.55	0.12	0.51***	0.11
Share Swe2 students	0.07	0.13	0.06	0.14
Teacher not full time	0.10	0.30	0.12	0.32
Teacher on leave	0.08	0.26	0.03***	0.17
Teacher year in class	2.56	1.28	2.80***	0.84
Teacher attitude: home works	3.56	0.94	3.83***	0.91
Teacher attitude: tests	2.66	0.77	2.96***	0.94
Teacher attitude: basic knowledge	4.42	0.81	4.67***	0.60
Teacher attitude: student influence	3.96	0.80	3.85**	0.86
Teacher attitude: student responsibility	4.71	0.63	4.77*	0.50
Number of students	317		4,267	

Note: Teacher and class information are collected at the individual level (the teacher has filled in one form for each student) and are treated as individual level information when calculating standard errors. The reason is that we cannot identify class in the data set. *(**,***) denotes that the difference in means is significantly different at 10% (5%, 1%).

Estimating the effect

Identification

The potential effect on cognitive skills of attending an MA-class may stem from two types of factors: (i) the effect of interactions between students of different age and school experience and/or (ii) effects from parent and teacher involvement (see Figure 1). In the literature on MA-classes, most arguments for the beneficial effects of MA-classes focus on the student interaction effects.

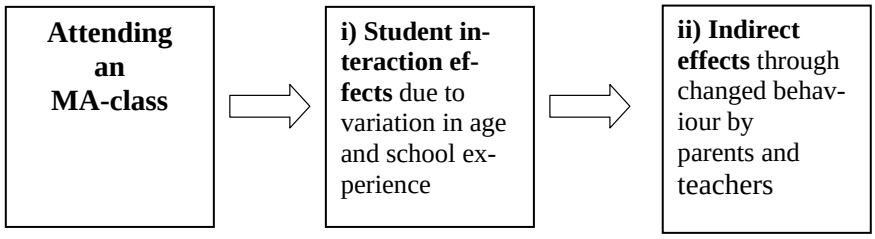


Figure 1: The different components of the MA-effect

Our purpose is to estimate the total effect of attending an MA-class, that is, both (i) and (ii). This is most interesting from a policy perspective and this is also what a randomized experiment would capture.[20]

The methodological problem that arises when one is to estimate the total effect using observational data is selection or sorting of students and teachers into MA-classes. There are several possible ways in which selection of students and teachers into class types might arise. First, there may be active choices on behalf of the families. If MA-classes are believed to have positive (negative) effects on students' cognitive or non cognitive development, it might be the case that informed parents actively choose an MA-class (or a conventional class) for their child. This can be

[20] Although it is not the purpose of our paper, measuring the effect of interactions between students of different ages only (i.e part of effect (i)) is relatively easily achieved. Given birth data on every student within each class, we could simply estimate the effect of age variance within a class on student outcomes. In our data set we only have information on all students within each class for the traditional classes, and our sample size is much too small for a precise estimation of this effect. In spite of that, we find that the point estimate of age variance in traditional classes on student achievement is negative.

achieved in different ways: people can choose schools within their area of living, or they can move to areas with or without MA-classes. Second, there could be active choices on behalf of the school management or teachers. Principals could place better (worse) students and teachers in MA-classes, or certain types of teachers could be overrepresented within MA-classes. Third, it is possible that areas with only MA-classes formed due to demographic necessity differ from areas offering both types of classes; the former is usually smaller municipalities far from larger cities. Fourth, it is also possible that some schools with special profiles, such as confessional or international schools, are forced to teach in a mixed-age fashion due to a limited number of students. These possible selection problems have to be dealt with in order to estimate the effect of MA-classes.

We use regression adjustment to control for the selection and sorting of students and teachers. We believe that this method is valid since we have extremely detailed information not only from register data but also from the survey data on the students, parents and teachers and their self reported attitudes towards different school issues.

However, we can note that the choice of control variables is not completely straightforward, partly because some of our control variables are measured in grade 6 (i.e. at the end of the MA-treatment), partly because the literature on MA-classes is rather vague when it comes to defining the MA-class concept. Although we view our variables to be controls for sorting and selection, it could be the case that some of them also reflect the indirect effect (ii) of attending an MA-class. One example is the variable attempting to measure how involved the parents are in school issues. Active parents may actively choose an MA-class (or traditional class) for their child (in which case the variable becomes an important control for selection) – but it could also be the case that parents in MA-classes (or traditional classes) are forced to become more actively involved in school issues (in which case the variable represents the indirect effect of MA).

Since our aim is to estimate the total effect (i+ii) of MA-classes, we do *not* want to include variables capturing the indirect effects in the estimations. We will estimate the effect of MA-classes both with and without the control variables considered potentially problematic. In the list of variables in Appendix we have distinguished between these two types of variables.

Estimation strategy

We estimate the following model using ordinary least squares (OLS):

$$y = \alpha + \beta ma456 + \delta X_1 + \gamma X_2 + m + \varepsilon$$

y denotes student achievement – either percentile ranked results from grade 6 cognitive tests or percentile ranked grade 9 credits. Our key explanatory variable, *ma456*, is a dummy variable for attending an MA-class all years in grades 4 to 6[21]. It is important to note that a class is defined as an MA-class only if it consists of students of both different ages and grades; this is not to be confused with traditional classes where some students happen to be born a different year than the others (for example, students with learning difficulties or especially skilled students). X_1 denotes the covariates used to control for selection bias. These include socioeconomic information such as parental education levels, immigrant status, gender and birth month of the student, and information on whether the student were given special help or mother tongue education in grade 3. For a complete list of all variables, see Appendix. When estimating the effect on grade 9 credits, we also control for the number of students in grade 9 at the school[22]. In addition, we have access to a variable indicating if the student attended an MA-class also during grades 7–9. This variable is included as a control in a separate estimation. X_2 denotes the variables used to control for selection, but where there is some uncertainty about whether or not they instead represent the indirect effects of MA-classes. These variables include the attitudes and behaviour of the teachers and parents. Finally, in all estimations we include municipality fixed effects, *m*.

We can also note that the two different measures of student outcomes, the grade 6 test results and the grade 9 credits, differ in two respects. Not only do they capture short- versus long run effects of attending an MA-class, they can also reflect slightly different types of skills. While the grade 9 credits are a weighted average of grades in different subjects, and as such could include not only the teachers' assessment of the student's skills but also to some extent the students' behaviour and diligence, the grade 6 tests are simply test results. The correlation between the two measures is also relatively low, with a correlation coefficient of 0.57.

[21] Using other definitions of the explanatory variable ma456, such as a dummy for attending an MA-class only in grade 4 or at least one year during grades 4–6 or a cumulative variable capturing the number of years spent in an MA-class does not change the results.
[22] We do not have information on the size of the school in grade 6.

Another thing to note is that we do not have information on whether the student attended an MA-class during grades 1–3. Since MA-class attendance in grades 1–3 is likely to be correlated with MA-class attendance in grades 4–6, it is possible that our dummy variable for MA-class attendance in grades 4–6 also partly captures the long run effects of earlier MA-class attendance.

With the estimation strategy above, we implicitly assume that the effect of attending an MA-class is equal for all groups of students. This may not be true – in fact, many of the arguments for or against MA-classes are concerned with how they affect different kinds of students. In particular, it is usually argued that attending an MA-class is especially valuable for students who do not perform as well as their peers. In many studies, it is shown that girls outperform boys in school and that immigrant students have lower school achievement than the average student. Hence, to relax the equal-effects assumption, we include interaction terms that allow the MA-effect to vary depending on gender, if the student has a non-Nordic background and if the student was low performing in grade 3 (measured by if the student were given special help in grade 3).

Results

How does attending an MA-class affect student performance? In section 5.1 we estimate the average effect, while section 5.2 examine whether the effect varies by observed characteristics.

Main results

Table 2 and 3 show the effect of MA-classes on grade 6 cognitive tests and grade 9 credits, respectively. For the cognitive test results, there is a negative and statistically significant effect of attending an MA-class. The estimated effect on grade 9 credits is not statistically significant, although the point estimate is negative.

The magnitude of the effect is relatively large. Attending an MA-class in grades 4-6 reduces the cognitive test results by around 5 percentile points. This can be compared with the effect of class size reductions. In the Tennessee STAR experiment, reducing class size by one student increased student performance with almost one percentile point (Krueger, 1999).

The negative effect of attending an MA-class on grade 6 test results remains in about the same range regardless of the set of covariates used.

A comparison between column 2 and column 3 shows no large differences. Hence, the variables added in column 3, that we view as good controls for selection but that potentially also could capture the indirect MA-effects, do not seem to be important in explaining the difference in achievement between students in MA- and traditional classes. This is interesting since these variables include the parental and teacher attitudes towards school issues. In Section 3 above, we noted that the largest differences between MA- and traditional classes were in terms of these different parental and teacher attitudes. At the same time, they seem unimportant for explaining the negative effect of MA-classes.

Table 2: OLS-estimates of the effect of attending an MA-class during grades 4–6 on percentile ranked grade 6 test results, survey sample

	(1)	**(2)**	**(3)**
	Grade 6 test results	*Grade 6 test results*	*Grade 6 test results*
MA grades 4–6	-5.681	-4.524	-4.711
	(2.059)***	(1.763)**	(1.805)***
Including X_1	No	Yes	Yes
Including X_2	No	No	Yes
Constant	52.229	49.556	46.878
	(0.539)***	(4.259)***	(6.569)***
Observations	4584	4584	4584
R-squared	0.05	0.29	0.30
F-test if *added* parameters jointly equals zero		100.94	3.21
Probability>F		(0.0000)	(0.0026)

All models include municipality dummies, standard errors in parentheses are clustered on schools, * significant at 10%, ** significant at 5%, *** significant at 1%

Table 3: OLS-estimates of the effect of attending an MA-class during grades 4–6 on grade 9 credits, survey sample

	(1)	**(2)**	**(3)**	**(4)**
	Grade 9 credits	*Grade 9 credits*	*Grade 9 credits*	*Grade 9 credits*
MA grades 4–6	-2.579	-0.989	-1.169	-0.915
	(1.742)	(1.317)	(1.315)	(1.336)
Including X^1	No	Yes	Yes	Yes
Including X^2	No	No	Yes	Yes
Including MA grades 7–9	No	No	No	Yes
Constant	52.770	37.502	36.326	35.998
	(0.568)***	(6.178)***	(6.503)***	(6.514)***
Observations	4584	4584	4584	4584
R-squared	0.04	0.30	0.30	0.31
F-test if *added* parameters jointly equals zero		53.76	4.35	23.11
Probability>F		0.0000	0.0019	0.0000

All models include municipality dummies, standard errors in parentheses are clustered on schools, * significant at 10%, ** significant at 5%, *** significant at 1%

Heterogeneous effects

Table 4 shows the results from the heterogeneous effects estimations. Interestingly, we find no statistically significant differences for any of the subgroups studied. Girls seem to be equally affected as boys, and the same is true for low performing students compared to students without extra help in grade 3, and students with a non-Nordic background compared to Nordic students. This is in sharp contrast to the arguments commonly used in favour of MA classes – that MA-classes especially should benefit low performing students.[23]

[23] We have also studied the same heterogeneous effects on the grade 9 credits but find no statistically significant differences between groups.

Table 4: OLS-estimates of heterogeneous effects of attending an MA-class during grades 4–6 on percentile ranked grade 6 test results, survey sample

	Grade 6 test results
MA grades 4–6	-2.838
	(1.203)**
Female student	-1.377
	(0.351)***
(MA grades 4–6)* (Female student)	2.326
	(1.602)
Help grade 3	-11.197
	(0.436)***
(MA grades 4–6)* (Help grade 3)	-1.599
	(1.740)
Non-Nordic student	-3.059
	(1.127)***
(MA grades 4–6)* (Non-Nordic student)	2.722
	(2.793)
Including X_1	Yes
Including X_2	Yes
Constant	43.086
	(2.926)***
Observations	4584
R-squared	0.31

All models include municipality dummies, standard errors in parentheses are clustered on schools, * significant at 10% ** significant at 5%, *** significant at 1%

Internal validity

In this paper, we use a linear regression model and adjust it with observed covariates to control for potential selection. In the context of returns from schooling, Black and Smith (2004) discuss potential problems with linear regression models. They conclude that the result may be biased if it is driven by comparisons of "non-comparable" individuals, i.e. individuals outside the common support. In order to address this issue we employ a propensity score matching method. First, we estimate a probit regression model for the probability to enter an MA-class.[24] Second, we match (nearest neighbour without replacement) on these predicted values so that we for each MA-student get one comparable individual who has attended a traditional class. Using this more homogeneous sample we estimate the effect of MA without any covariates, i.e., simply compare the mean values. The results are presented in Table 5. With respect to grade 6 results the point estimate is close to the corresponding estimate received with a regression adjustment approach (see Table 3). With respect to grade 9 credits, the point estimate is somewhat lower but, also with this strategy, statistically insignificant.

Table 5: Matching approach, survey sample

	(1)	**(2)**
	Grade 6 results	*Grade 9 credits*
MA grades 4–6	-4.622	-0.281
	(1.79)	(0.13)
Constant	51.615	51.769
	(28.71)**	(30.90)**
Observations	634	634
R-squared	0.01	0.00

Standard errors in parentheses are clustered on schools, * significant at 10% ** significant at 5%, *** significant at 1%

External validity

Since our survey sample is a slightly selected group of students, it is relevant to ask how valid our estimates are for the wider population. One way to shed light on this issue is to compare the effect of MA in the survey sample with the effect found in the basic sample (register data with grade 6 test results available). Naturally, we can only utilize register covariates

[24] In the probit regression we include all X_1 and municipality dummies.

to capture selection and sorting for this comparison. The results are shown in Table A2 in Appendix. As is clear from the table, the effect of attending an MA-class is negative for student achievement in both samples, although the coefficient size is slightly larger in the survey sample.

Concluding remarks

Despite ambiguous scientific evidence, mixed-age classes are a common phenomenon in schools around the world. In some cases, it is because of demographic necessity; in other cases, it is because MA-classes are claimed to enhance student achievement. In Sweden these types of classes have been rapidly re-introduced and nowadays, around one fourth of all children attend an MA-class during grades 4–6.

In this paper, we present evidence that MA-classes have a negative effect on short-term cognitive skills, as measured by the grade 6 cognitive tests. This effect is robust to different definitions of the explanatory variable and it does not change significantly if we only focus on girls, low performing students or students with a non-Nordic background. The effect of attending an MA-class on grade 9 average credits is not statistically significant, although the point estimate is negative.

We have not been able to distinguish between MA-classes introduced out of pedagogical beliefs and MA-classes introduced out of economic and/or demographic necessity. Since the effect of MA-class attendance could differ between these two groups, this would be an interesting topic for future research.

Appendix

List of variables

Register data	
Variable name	Definition
Grade 6 test results	Percentile rank of the sum of the scores on the tests in number series and opposites given in grade 6
Grade 9 credits	Percentile rank of a summary measure of the student's 16 best credits in grade 9
MA grades 7–9	A dummy that equals 1 if the students attends an MA-class in any grade between grades 7-9
MA grades 4–6	A dummy that equals 1 if the students attends an MA-class during grades 4-6
Municipality dummies	One dummy for each municipality
Female student	A dummy that equals 1 if the student is female
Early start	A dummy that equals 1 if the student is born after 1982
Late start	A dummy that equals 1 if the student is born before 1982
Birth month	The student's month of birth
Help in grade 3	A dummy that equals 1 if the student has been given any form of special education intended for low performing students ("särundervisning", "anpassastudiegång" or "specialundervisning på annat sätt") in grade 3
Mother tongue in grade 3	A dummy that equals 1 if the student attended mother tongue education in grade 3
International school	A dummy that equals 1 if the school has an international profile
Confessional school	A dummy that equals 1 if the school has a confessional profile
Special school	A dummy that equals 1 if the school is not ordinary, for example schools at hospitals
Grade 9 students	The number of students in grade 9 at the school, collected in grade 9
Few grade 9 students	A dummy that equals 1 if the number of students in grade 9 at the school is smaller than 30

List of variables cont'd

Variable name	Definition
Non –Nordic	A dummy that equals 1 if the student is born in a non-Nordic country (missing values equals 1)
Mother secondary education	A dummy that equals 1 if the mother of the student has secondary education, at most 5 years in addition to compulsory schooling
Mother university education	A dummy that equals 1 if the mother of the student has university education, more than 5 years in addition to compulsory schooling
Father secondary education	A dummy that equals 1 if the father of the student has secondary education, at most 5 years in addition to compulsory schooling
Father university education	A dummy that equals 1 if the father of the student has university education, more than 5 years in addition to compulsory schooling
Mother education miss	A dummy that equals 1 if information about the mother's education is missing
Father education miss	A dummy that equals 1 if information about the father's education is missing
Mother non-Nordic	A dummy that equals 1 if the mother of the student is born in a non Nordic country
Father non-Nordic	A dummy that equals 1 if the father of the student is born in a non Nordic country
Birth country missing	A dummy that equals 1 if information about the student's country of birth is missing
Father birth country missing	A dummy that equals 1 if information about the father's country of birth is missing
Mother birth country missing	A dummy that equals 1 if information about the mother's country of birth is missing

List of variables cont'd

Survey data collected in grade 6		
Variable name	Question	Definition
	To teachers:	
Teacher experience	What is your teacher experience in years?	A variable ranging from 1 to 43 (measured in years)
Teacher not qualified	Do you have a certificate qualifying you to teach at this level?	A dummy that equals 1 if the answer is no
Class size	What is the number of girls and boys in the class?	The sum of boys and girls in the class - ranging from 0 to 60
Small class	Constructed from class size	A dummy that equals 1 if the class size is smaller than 10
Large class	Constructed from class size	A dummy that equals 1 if the class size is larger than 25
Share boys	Constructed from class size	The share of boys in the class
Share Swe2 students	What is the number of students in your class that take Swe2? Constructed from class size	The share of students in the class taking a special course in Swedish adapted for students who do not have Swedish as mother tongue
Teacher not full time	Do you work full time?	A dummy that equals 1 if the teacher work part time, 0 if full time
Teacher on leave	Have you been on leave during the last year?	A dummy that equals 1 if the teacher has been on leave full time or part time
Teacher year in class	How many years have you taught this class?	A variable ranging from 1 to 8 (measured in years)
Teacher attitude: home works (belongs to X_2, the extended set of covariates)	How important are home works and oral tests?	A variable ranging from 1 to 5 in the following way: Very important 5 Rather important 4 In between 3 Rather unimportant 2 Not at all important 1

List of variables cont'd

Survey data collected in grade 6		
Variable name	Variable name	Variable name
Teacher attitude: tests (belongs to X_2, the extended set of covariates)	How important are formal tests?	A variable ranging from 1 to 5 in the following way: Very important 5 Rather important 4 In between 3 Rather unimportant 2 Not at all important 1
Teacher attitude: basic knowledge (belongs to X_2, the extended set of covariates)	How important is the emphasis of basic skills?	A variable ranging from 1 to 5 in the following way: Very important 5 Rather important 4 In between 3 Rather unimportant 2 Not at all important 1
Teacher attitude: student influence (belongs to X_2, the extended set of covariates)	How important is student influence during planning?	A variable ranging from 1 to 5 in the following way: Very important 5 Rather important 4 In between 3 Rather unimportant 2 Not at all important 1
Teacher attitude: student responsibility (belongs to X_2, the extended set of covariates)	How important is it that the student takes own responsibility?	A variable ranging from 1 to 5 in the following way: Very important 5 Rather important 4 In between 3 Rather unimportant 2 Not at all important 1

List of variables cont'd

Survey data collected in grade 6	Survey data collected in grade 6	Survey data collected in grade 6
Variable name	Variable name	Variable name
	To parents:	
Parent attitude: active school choice (belongs to X_2, the extended set of covariates)	Have you chosen another than the closest school to your child?	A dummy that equals 1 if the answer is yes or yes we are going to and 0 if no or doubtful (probably not)
Parent attitude: parent help (belongs to X_2, the extended set of covariates)	Do you participate in your child's school work?	A variable ranging from 1 to 5 in the following way: Very often 5 Rather often 4 Sometimes 3 Rarely 2 Almost never 1
Parent attitude: parent active (belongs to X_2, the extended set of covariates)	How much do you participate in school activities?	A variable ranging from 1 to 5 in the following way: Very often 5 Rather often 4 Sometimes 3 Rarely 2 Almost never 1

The cognitive tests from grade 6

There are three test scores from grade 6 available. The tests represent verbal, spatial and reasoning abilities and are called: Opposites (motsatser), Number series (talserier) and Metal folding (platvik). In the test called Opposites, the child is asked to find the opposite of a given word among four choices (40 items, 10 minutes). In the Number series test the child is instead asked to complete number series (40 items, 18 minutes). In the last test, Metal folding, the child is asked to find the three-dimensional object among four choices that can be made from a flat piece of metal (40 items, 15 minutes).[25] The results on each of these tests are measured on a scale ranging from 0 to 40.

In this paper, we use the percentile ranked sum of two of the tests, Opposites and Number series. The correlation coefficient between each of these tests and the grade 9 credits is 0.51. The third test, Metal folding, involves tasks not regularly practised in schools, and it's correlation coefficient to the grade 9 credits is 0.36.

[25] A more detailed description of the test scores are given by Svensson (1964).

Descriptive statistics

Table A1: Register data, Basic sample and Survey sample

	Register data			**Basic sample**			**Survey sample**		
Variable	*Mean*	*Sd.*	*N*	*Mean*	*Sd*	*N*	*Mean*	*Sd*	*N*
Grade 9 credits [1)]	202.17	59.88	8,490	203.93*	58.56	7,234	209.16***	56.77	4,584
Grade 6 test results [2)]	44.24	12.68	7,420	44.32	12.66	7,234	45.13***	12.55	4,584
MA grades 4–6	0.08	0.26	8,531	0.08	0.27	7,234	0.07	0.25	4,584
MA grades 7–9	0.03	0.17	8,531	0.02**	0.15	7,234	0.02***	0.15	4,584
Covariates capturing selection:									
Female student	0.49	0.50	8,515	0.49	0.50	7,234	0.50	0.50	4,584
Early start	0.01	0.09	8,515	0.01	0.09	7,234	0.01	0.09	4,584
Late start	0.03	0.16	8,515	0.02	0.15	7,234	0.02	0.15	4,584
Birth month	6.28	3.36	8,515	6.27	3.36	7,234	6.26	3.35	4,584
Help in grade 3	0.21	0.40	8,531	0.20	0.40	7,234	0.19***	0.39	4,584
Mother tongue in grade 3	0.10	0.30	8,531	0.09	0.29	7,234	0.08***	0.27	4,584
International school	0.0011	0.03	8,360	0.0007	0.0263	7,234	0.0007	0.0256	4,584
Confessional school	0.0054	0.07	8,360	0.0043	0.0653	7,234	0.0028**	0.0532	4,584
Special school	0.04	0.19	8,360	0.04	0.19	7,234	0.03*	0.18	4,584
Grade 9 students	113.80	41.58	8,331	113.52	41.09	7,234	113.42	40.47	4,584
Few grade 9 students	0.03	0.16	8,331	0.02*	0.14	7,234	0.02***	0.13	4,584
Non-Nordic student	0.07	0.26	8,531	0.07	0.25	7,234	0.06**	0.24	4,584
Mother sec. educ.	0.45	0.50	8,531	0.45	0.50	7,234	0.46	0.50	4,584
Mother univ. educ.	0.30	0.46	8,531	0.30	0.46	7,234	0.32***	0.47	4,584

Table A1: Cont'd

	Register data			Basic sample			Survey sample		
Variable	*Mean*	*Sd.*	*Variable*	*Mean*	*Sd.*	*Variable*	*Mean*	*Sd.*	*Variable*
Father sec. educ.	0.38	0.48	8,531	0.39	0.49	7,234	0.40***	0.49	4,584
Father univ. educ.	0.20	0.40	8,531	0.20	0.40	7,234	0.22**	0.41	4,584
Father educ. miss	0.26	0.44	8,531	0.24*	0.43	7,234	0.22***	0.41	4,584
Mother educ. miss	0.07	0.26	8,531	0.07	0.25	7,234	0.06***	0.23	4,584
Father non-Nordic	0.12	0.33	8,531	0.11*	0.32	7,234	0.10***	0.30	4,584
Mother non-Nordic	0.12	0.32	8,531	0.11*	0.31	7,234	0.10***	0.30	4,584
Birth country miss	0.0014	0.04	8,531	0.0011	0.0332	7,234	0.0011	0.0330	4,584
Father birth country miss	0.03	0.18	8,531	0.03	0.17	7,234	0.03	0.16	4,584
Mother birth country miss	0.02	0.13	8,531	0.02	0.13	7,234	0.02	0.13	4,584
Teacher experience							20.00	9.73	4,584
Teacher not qualified							0.04	0.20	4,584
Class size							23.30	6.16	4,584
Small class							0.02	0.14	4,584
Large class							0.35	0.48	4,584
Share boys							0.51	0.11	4,584
Share Swe2 students							0.06	0.14	4,584
Teacher not full time							0.12	0.32	4,584
Teacher on leave							0.03	0.18	4,584
Teacher year in class							2.78	0.88	4,584
Parent attitude: active school choice							0.15	0.36	4,584

Table A1: Cont'd

	Register data	**Basic sample**	**Survey sample**		
Variable			*Mean*	*Sd.*	*Variable*
Covariates capturing selection and/or indirect MA-effects:					
Teacher attitude: home works			3.82	0.91	4,584
Teacher attitude: test			2.94	0.93	4,584
Teacher attitude: basic knowledge			4.65	0.62	4,584
Teacher attitude: student influence			3.86	0.85	4,584
Teacher attitude: student responsibility			4.76	0.51	4,584
Parent attitude: parent help			1.90	0.95	4,584
Parent attitude: parent active			2.34	1.04	4,584

Statistically significant difference compared to register data: * significant at 10%, ** significant at 5%, *** significant at 1%, note: 1) and 2) are not percentile ranked

The MA-effect in the basic sample versus the survey sample

Table A2: OLS-estimates of the effect of attending an MA-class during grades 4–6 on percentile ranked grade 6 test results, controlling for register covariates only, (1) estimated on basic sample, (2) estimated on survey sample

	(1)	***(2)***
	Grade 6 test results	*Grade 6 test results*
MA grades 4–6	-1.171	-2.141
	(0.558)**	(0.795)***
Including register covariates	Yes	Yes
Constant	44.518	44.672
	(0.533)***	(0.643)***
Observations	7234	4584
R-squared	0.31	0.30

Standard errors in parentheses are clustered on schools, all models include municipality dummies, * significant at 10% ** significant at 5%, *** significant at 1%

References

Mason, DeWayne A. and Burns, R. B.: 1996, Simply No Worse and Simply No Better' May Simply Be Wrong: A Critique of Veenman's Conclusion about Multigrade Classes, Review of Educational research, 66 (3), 307-322.

Edlund, A. C, and Sundell, K.: 1999 Åldersintegrerat eller åldersindelat? En jämförande studie av 1 111 elever i årskurs2, FoU-rapport 1999:9, Stockholm socialtjänstförvaltning: FoU-enheten.

Hanushek, E.: 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.

Krueger, A.: 1999, Experimental estimates of education production functions, Quarterly Journal of Economics, no. 114, 497-532.

Krueger, A.: 2003, Economic considerations and class size, Economic Journal, 113, 34-63

Sandqvist, K.: 1994, Åldersintegrerad undervisning: En kunskapsöversikt. Stockholm: HLS.

Sundell, K.: 1995, Åldersindelat eller åldersblandat? Forskning om ålderssammansättningens betydelse i förskola och grundskola, Studentlitteratur, Lund.

Sundell, K.: 2002, Är åldersblandade klasser bra för eleverna? En jämförande studie av 752 elever i årskurs 2 och 5, FoU-rapport 2002:7, Socialtjänstförvaltningen Stockholm stad.

Statistiska centralbyrån: 1996, Elevpanel för longitudinella studier, Elevpanel 4, Statistiska meddelanden U73, SM 9601. Örebro.

Svensson, A.: 1964, Sociala och regionala faktorers samband med över- och underprestation i skolarbetet: Pedagogisk-sociologiska studier jämte en beskrivning av skolstatistikens individualuppgifter. Rapporter från Pedagogiska institutionen, Göteborgs universitet.

Veenman, S.: 1995, Cognitive and Non cognitive Effects of Multigrade and Multi-Age Classes: A Best-Evidence Synthesis, Review of Educational Research, 65 (4).

Vinterek, M.: 2001, Åldersblandning i skolan - Elevers erfarenheter, Doktorsavhandling i Pedagogiskt arbete No.1, Umeå universitet.

Vinterek, M.: 2003, Åldersblandade klasser, Lärares föreställningar och elevers erfarenheter, Studentlitteratur, Lund.

Essay 2: Comparing teachers' assessments and national test results – evidence from Sweden·

Introduction

The Swedish educational system relies heavily on school leaving certificates. They are used as selection instrument for application to higher education as well as in job applications. The teacher alone is responsible for assigning school leaving certificates. When doing this, all available information should be taken into consideration. In many other countries, for example England and France, certificates or qualifications provided by national examination boards play a corresponding role.26 An argument for using teacher evaluations is that they are based on more infor-

♦ This paper has benefited from useful comments given by Per Johansson, Peter Fredriksson, Andreas Westermark, Tuomas Pekkarinen and Helena Holmlund. The author also thank Patrik Hesselius for inspiring discussions about the gender difference between school leaving certificates and test results, as well as seminar participants at the Department of Economics, Uppsala University, participants at the First Summer School of the Marie Curie Research Training Network, Padova, 16-18 June 2006 and participants at the COST meeting in Essen, 19-20 October 2006. The financial support from the Swedish council for working life and social research FAS (dnr 2004-1212) is acknowledged.

[26] For example the General Certificate of Secondary Education (GCSE) in Britain and "le baccalauréat" in France.

mation about the student than tests are able to capture. On the other hand, nationally provided tests are likely to be more objective, minimizing the risk of conscious or unconscious discrimination.

In Sweden, national tests are performed in order to enable equivalent and fair school leaving grades across the country. That is, the test results shall serve as a guideline for the teachers when they assign school leaving certificates. The national tests are graded according to nationally stated correcting instructions. Students' skills are tested in the three core subjects Swedish, English and Mathematics.

This paper investigates if there are systematic differences between school leaving certificates and national tests results. Specifically, the aim is to investigate if the relationship between school leaving certificates and national test results differs between girls and boys or between natives and non-natives.

The analysis is based on Swedish data on grade 9 students (16 years old). For each student, information is available about gender and country of birth as well as test results and school leaving certificates in Swedish, English and Mathematics. The results show that girls are more generously rewarded in school leaving certificates compared to test results than boys in all three subjects studied. Non-native students are more generously rewarded in Swedish and Mathematics but no statistically significant difference is found in English.

Related literature

There are persistent differences in school performance between the genders and ethnic groups. International studies show that girls, on average, outperform boys in Reading while the opposite is true in Mathematics (NAEP 2004; PISA 2003). In US, white students, on average, outperform black and Hispanic students (NAEP 2004) and in most European countries native students outperform non-native students (PISA 2003). In Sweden, according to school leaving certificates, girls outperform boys in almost all subjects (including Mathematics (Skolverket, 2006). The difference in school leaving certificates between non-natives and natives is also striking: 77 percent of the non-native students are qualified to Secondary school while the corresponding share for the natives is 91 percent (Skolverket 2005).

Studies from different countries show that, in comparison with test results, teachers assess girls' performance higher than boys' (Wester-Wedman, Gisselberg, Mattsson W Wedman, 1988 and Emanuelsson and

Fischbein, 1986). A suggested explanation to this observation is that the school environment might be adapted to traditionally female behaviour.[27] Skolverket (2006) and Nycander (2006) have analyzed in a descriptive way both school leaving certificates and national test results of grade 9 students in Sweden from a gender perspective. They conclude that the gender difference in favour of girls, observed on tests, is reinforced in school leaving certificates. That is, the gender gap is larger with respect to school leaving certificates compared to test results in Swedish, English and Mathematics. However, the gender difference in the difference between school leaving certificates and test results has not been tested in a formal setting including robust checks.

Exploring a natural experiment, Lavy (2005) presents evidence on discrimination against boys when teachers correct exams. The extent of the discrimination varies by subjects and type of tests and ranges from 5 to 25 percent of the standard deviation of the test score distribution. Regardless of the discrimination, girls outperform boys in most subjects but the gender gap in test results is reinforced by teachers' discrimination.

Although ethnic minorities or non-natives are often discussed in the context of discrimination, teacher assessments and test results have previously, to my knowledge, not been rigourously compared across ethnic groups.

The Swedish school system

The Swedish National Agency for Education formulates the criteria for different grade steps. In the latest curriculum for the Compulsory School System (Lpo, 94) it is stated that school leaving certificates should reflect skills and knowledge in the subject in accordance to the goals stated in the course syllabi. That is, school leaving certificates should not reflect attention in the classroom, diligence, ambition, home work and work during lesson, as long as it is not a prerequisite for attaining the goals (as in the case of laboratory work).

National tests are performed during the spring semester in ninth grade. The tests in languages (Swedish and English) measure writing and reading abilities as well as listening comprehension and verbal interaction. The tests in Mathematics include analysis and algebra and an oral part testing Mathematical reasoning. The tests are corrected at the school level but are graded according to nationally stated correcting

[27] See Emanuelsson and Fischbein for a more extensive review of this literature.

instructions. Teachers are encouraged to not correct their own students' exams, but they are allowed to do so.

Teachers *shall* use the nationally approved examinations when assigning the school leaving certificate (Skolverket 2004). However, in the individual case, the teacher is allowed to assign the school leaving certificate differently from the test result. The reason is that the student might be low performing on the test day due to occasional conditions. Further, the teacher should take into consideration all available information about the student's knowledge and ability in the subject. The tests are not guaranteed to capture *all* goals stated in the course syllabi, although the aim of the tests is to measure, as comprehensively as possible, the student's ability and knowledge in the subject. However, it is clear from the national directives that the tests should form an important basis for the school leaving certificates.

Both school leaving certificates and test results are assessed according to the same ordinal metrics: Fail (F), Pass (P), Pass with Distinction (PD) and Pass with Special Distinction (PSD). The aim of the Swedish school is that all students should attain at least "Pass". To be qualified to upper secondary school, the student has to attain at least "Pass" in the three core subjects: Swedish, English and Mathematics.

Data

Data sources

The main data source used is a register, provided by the Swedish Agency for Education (SAE), of school leaving certificates for all students in grade 9 in Sweden (årskurs-9-registret). From this register we also know the gender of the student, which year and which school the student attended in grade 9. To this register I have added information about test results, collected by SAE. Test results are available from 2001 up to 2005.[28] Between 2001 and 2002, test results from a random sample of 150 schools were collected. From 2003 and onwards, all schools were aimed to be collected. However, all subjects were not collected all years. Swedish and English were not collected in 2002 and 2003 and Mathematics is missing for year 2001. Since the availability of test results differs between the three different subjects, I construct one sample for each subject. These samples are restricted to include all covariates of

[28] A stratified sample of test results was collected between 1998 and 2001.

interest (no missing value of any covariate). All results presented are based on these restricted samples. The sample sizes used for the gender analysis are 112,648 in Swedish, 114,468 in English and 271,150 in Mathematics.

For the analysis of non-native students versus natives, an additional register (Louise) is used from Statistics Sweden on the country of birth. At present, this information is available for all students younger than 17 years of age in 2003. Students in Sweden are expected to reach the age of 16 in grade 9. Thus, country of birth is only available for students 2001 to 2004. This fact implies that exploring country of birth in the analysis reduces the samples further. The samples used in this case are: 9,492 in Swedish, 9,748 in English and 70,233 in Mathematics.

Variable definitions

In Sweden many individuals are born in another Nordic country than Sweden. These individuals speak Swedish well, look Swedish and know the country well. Non-natives are therefore defined as those born in a non-Nordic country.

For students with another mother language than Swedish and who are assessed to not be able to follow the ordinary course in Swedish, a special course is offered: *Swedish 2*. The *Swedish 2* course has about the same course syllabus as the ordinary course in Swedish but the teaching is adapted for students with another mother tongue than Swedish. Since 2001 the national test is the same for the course *Swedish 2* as for the ordinary course in Swedish. In this paper *Swedish 2* and Swedish are treated as one subject.[29] The focus is on the difference between school leaving certificates and test result and the type of course the student has taken within the subject should be of less importance.

The Swedish National Agency for Education summarizes the school leaving certificate (all final course grades) into a total Grade Point Average (GPA). This GPA is the instrument used for application to higher education. The values used for transforming the ordinal scale to a numerical scale in order to calculate the GPA are: 0, 10, 15 and 20; that is, Fail equals 0, Pass equals 10 and Pass with distinction and Pass with special distinction equals 15 and 20, respectively. All results presented in this study are based on grades transformed to these numerical values.

[29] Controlling for Swedish 2 does not affect any of the interaction estimates presented in this paper.

Sample selection

Test results are only available for a selected group of students. That is, if test results from the school are reported to SAE and if the student has completed all parts of the test. Test results shall be reported to SAE by the school and all students are requested to complete all parts of the test. Although these requests, we may have a selection of schools and/or students within schools. The Appendix presents descriptive statistics on school leaving certificates for unrestricted and restricted samples for each subject. The mean in school leaving certificates is somewhat higher in the restricted samples. A plausible explanation is that those students who did not complete all parts of the test, on average perform worse than those who did.

The conclusion is that this analysis is based on students who on average perform slightly above the average in Sweden.

Descriptive statistics

Table 1a and 1b show the differences between the average test results and the school leaving certificates for all students and for girls and boys separately (Table 1a) and for non-native- and native students separately (Table 1b). Figure 1a-f shows the distribution of school leaving certificates and test results for these sub-groups.

Girls outperform boys in Swedish and English according to test results as well as school leaving certificates (Table 1a). In Mathematics there is no difference between the genders according to test results but girls perform better according to school leaving certificates. Non-native students under perform native-students in all subjects according to both test results and school leaving certificates.

With respect to all students, school leaving certificates are, on average, significantly higher than test results in all three subjects. Thus, it seems that teachers overall are more generous compared to test results when assigning school leaving certificates, with the largest differences in Mathematics. For example in Mathematics, 11 percent of the students fail on the test, but only 4 percent fail according to school leaving certificates (Figure 1e).

The difference between school leaving certificates and test results is larger for girls and non-natives than for boys and natives, respectively. For example, in English, 15 percent of the boys qualified for PSD according to the test and also 15 percent received the highest school leaving certificates (Figure 1b). In contrast, 16 percent of the girls qualified for the highest grade according to test results, but 20 percent received the

highest school leaving certificate (Figure 1b). Among non-natives in Swedish, 14 percent score fail on the test but only 9 percent receive fail according to school leaving certificates (Figure 1b). The corresponding shares for native students in Swedish are 4 and 3 percent, respectively (Figure 1a).

Descriptive statistics show gender and ethnic differences between school leaving certificates and test results. However, in order to estimate the difference between school leaving certificates and test results across sub-groups, we need a formal model.

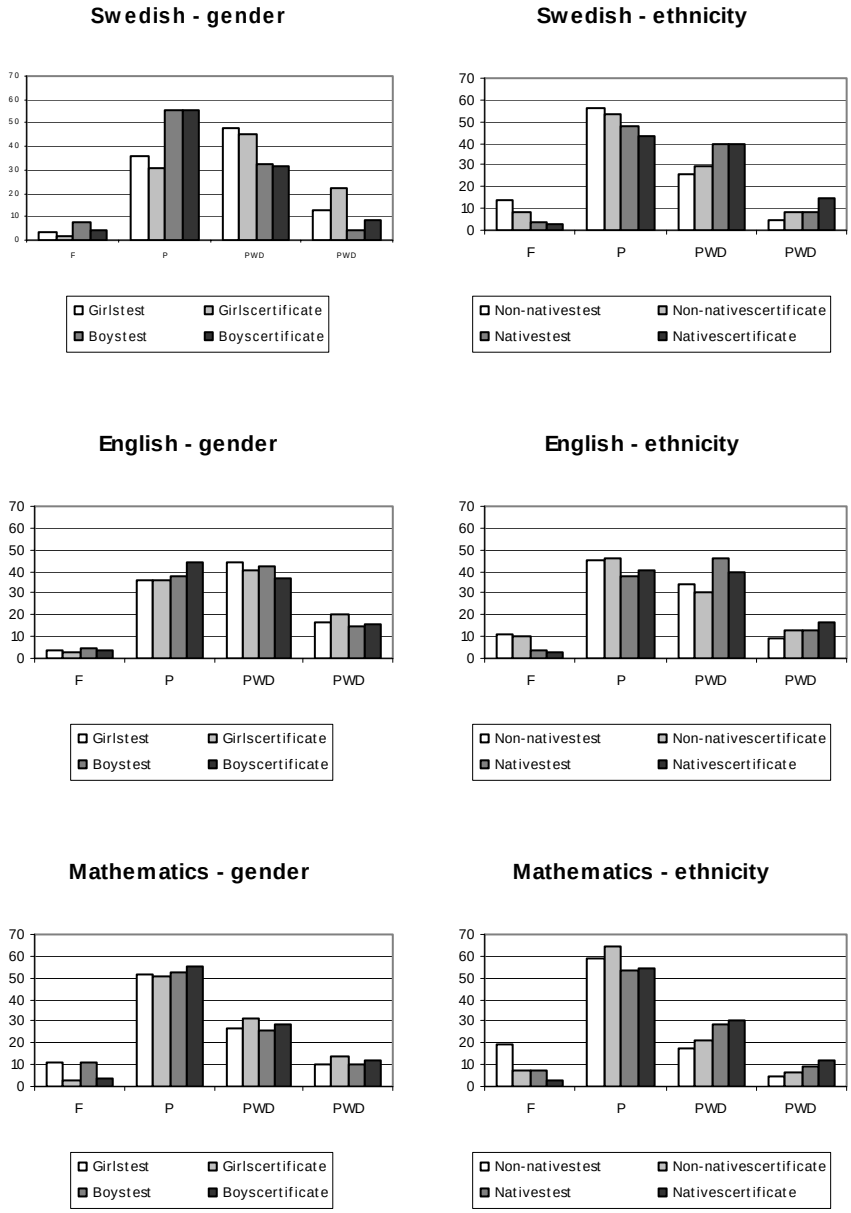
Table 1a: Test grade versus school leaving certificates, girls versus boys

Subject and group of students	**Number of observations**	**School leaving certificates**	**Test grade**	**Difference**
		Mean (*St dev)*	Mean (*St dev)*	Mean (*St dev)*
Swedish:				
All students	112,648	13.11	12.32	0.80
		4.25	*4.35*	*0.02*
Girls	55,288	14.28	13.38	0.90
		4.12	*4.10*	*0.02*
Boys	57,360	11.99	11.29	0.70
		4.06	*4.34*	*0.02*
English:				
All students	114,468	13.34	13.29	0.05
		4.44	*4.44*	*0.02*
Girls	56,139	13.76	13.47	0.29
		4.38	*4.35*	*0.03*
Boys	58,329	12.93	13.11	-0.18
		4.45	*4.52*	*0.03*
Mathematics:				
All students	271,150	12.50	11.28	1.22
		4.24	*5.16*	*0.01*
Girls	132,878	12.70	11.30	1.40
		4.24	*5.17*	*0.02*
Boys	138,272	12.31	11.26	1.05
		4.23	*5.15*	*0.02*

Table 1b: Test grade versus school leaving certificates, non-natives versus natives

Subject and group of students	**Number of observations**	**School leaving certificates**	**Test grade**	**Difference**
		Mean (*St dev)*	Mean (*St dev)*	Mean (*St dev)*
Swedish:				
All students[1]	9,492	13.00	12.23	0.77
		4.18	*4.17*	*0.06*
Non-natives	847	11.42	10.35	1.06
		4.69	*4.94*	*0.23*
Natives	8,645	13.16	12.41	0.74
		4.09	*4.04*	*0.06*
English:				
All students	9,748	13.23	13.04	0.19
		4.39	*4.31*	*0.06*
Non-natives	857	11.81	11.52	0.30
		5.29	*5.20*	*0.25*
Natives	8,891	13.37	13.19	0.18
		4.27	*4.18*	*0.06*
Mathematics:				
All students	70,233	12.38	11.45	0.93
		4.08	*4.76*	*0.02*
Non-natives	5,873	10.98	9.35	1.63
		4.32	*5.28*	*0.09*
Natives	64,360	12.51	11.64	0.87
		4.03	*4.67*	*0.02*

1) Note that this table is based on a different sample set than table 1a.



Figures 1a-f Distribution of test results and school leaving certificates, divided into girls and boys and non-native and native-students for the different subjects

Estimating the difference across groups

The grade steps in Sweden are ordinal. This fact suggests an *ordered probit* or an *ordered logit* model. However, I choose to estimate linear regression models. The reason is that the focus is on the average marginal effects of gender and ethnic background, respectively. Since numerical values exist for the ordinal scale and marginal effects are easily obtained from linear regressions, the following model is estimated with ordinary least squares:

$$Y_{ijt} = \beta_0 + \beta_1 S_{ijt} + \beta_2 T_{ijt} + \beta_3 (S*T)_{ijt} + \eta_t + \varepsilon_{ijt}$$

where Y_{ijt} is individual *i*'s type of grade *j* (school leaving certificate or test result) in year *t*. S_{ijt} is a dummy variable that equals 1 if the student is a girl and 0 if a boy. *T* is a dummy for the type of grade. *T* equals 1 if the grade corresponds to the school leaving certificate and 0 if the grade is the test result. The interaction term (*S***T*) consists of the gender- and the grade type dummy. Year-effects[30] are captured by η_t and finally ε_{ist} is assumed to be an idiosyncratic error term. The set-up of the equation above is analogous when natives and non-natives are compared. The only difference is that S_{ijt} equals 1 when the student is non-native.

With a difference in differences strategy we first remove student group fixed effects in the subject, as long as they have the same effect on school leaving certificates and test grades. Second, we remove the average difference between school leaving certificates and test result. This means that the interaction term captures the additional generosity associated with school leaving certificates and the fact that the student is a girls or non-native, respectively.

However, this interpretation of the interaction term does not hold if girls or non-native born students choose to attend certain schools *because* these schools are extraordinary generous with respect to school leaving certificates, more than their respective reference groups. If this scenario is true, the parameter in front of the interaction term captures this selection. In order to remedy this potential selection problem, we extend the model

[30]I estimate a static linear panel data model. However, the sample covers several years and some schools are observed several times. A year specific dummy is therefore also included to capture which year the individual is observed. Including year dummies imply that we do not have to bother about changes in grade policy (grade inflation) over time. Wikstrom and Wikstrom (2004) claim that grade inflation occurs in Sweden during the 1990s.

by including school dummies. School dummies capture school specific generosity in school leaving certificates.

When including school dummies, we explore the variation within schools over time. In those estimations we have to assume that the error term should not be correlated with the explanatory variables across time periods, in order to receive consistent estimates. This assumption is violated if, for example, the school, conditional on all covariates (including school dummies), assigns extraordinary high grades in year *t* and this in turn affect students' choices of school the following year. This could be a real problem if students choose to attend a particular school because this school had generous school leaving certificates, in comparison to test results, the previous year. However, it seems realistic to assume that school choice is based on average results from several years.

In order to increase the precision of the estimate of interest we also include the share of girls (non-natives) among the students at the school, the student's month of birth and, in the regressions comparing non-natives and natives, the gender of the student.[31] The parameter estimates of these additional covariates are not presented below.

In all model specifications, the inference presented is based on standard errors that allow for a common variance component at the school level. This is appropriate since the sampling unit with respect to test results is schools.

Results

The parameter estimates shown in Tables 3a and 3b confirm some of the patterns discerned from the descriptive statistics. Conditional on the type of grade, girls outperform boys in all subjects, except in Mathematics. Non-native students, on the other hand, exhibit lower performance than native students in all subjects. The coefficient in front of the type of grade dummy is positive in almost all cases. The interpretation is that teachers are in general more generous when assigning school leaving certificates than justified by test results. An exception is in English, where boys on average get a lower school leaving certificate than test result. However, non-native students, on average, perform below average in all subjects *and* they are more generously rewarded in school leaving certificates.

[31] The reason for *not* including the non-native dummy in the gender case is that this information would reduce the sample size. It is reasonable to assume that, β_3 in the gender case is orthogonal to the information about the student's country of birth. Thus, the country of birth information would not affect the parameter estimate, only the precision of the estimate.

The estimates of the interaction terms are statistically significant in all subjects with respect to gender. The interpretation is that, conditional on test results, girls are better rewarded than boys in terms of school leaving certificates. With respect to ethnic background, the estimates in front of the interaction terms are positive and statistical significant in Swedish and Mathematics, but not in English. In English it is almost zero and statistically insignificant. In Swedish and Mathematics, the interpretation is that teachers in general are more generous to non-natives than to others when assigning school leaving certificates, conditional on test results.

The results presented in this section seem robust. Adding school dummies and additional covariates (column 2) change neither the coefficients nor the precision substantially in any estimated model.

The effect sizes vary between subjects. In the gender case, the effect is largest in English and smallest in Swedish. In Mathematics, with respect to gender, the effect corresponds to 11 percent of the standard deviation of the grade difference distribution. Remember that the highest grade, "PSD", is given 20 points. The steps from "P" to "PS" and from "PS" to "PSD" correspond to 5 credit points each. Roughly speaking, in a class with 30 students (half of them girls) four girls, but only three boys, get a higher school leaving certificate than test result (Pass with distinction instead of Pass).[32]

With respect to ethnic background, the effect sizes in Mathematics correspond to 23 percent of the standard deviation of the grade difference distribution.

[32] The average difference per individual between school leaving certificates and test scores is 1.045 (Table 3a). In a class with 15 boys (boys are the reference in the model) the total difference between school leaving certificates and test scores is 15.675 (1.045*15) for boys and 21 ((1.045+0.355)*15) for girls.

Table 3a: Ordinary least squares estimation of the interaction between girl and type of grade, 1) basic 2) with school dummies and additional covariates

	Swedish		**English**		**Mathematics**	
	1	*2*	*1*	*2*	*1*	*2*
Female student (f)	2.089	2.071	0.362	0.323	0.034	-0.015
	(0.031) ***	(0.030) ***	(0.031) ***	(0.030) ***	(0.023)	(0.022)
Type of grade (T)	0.695	0.695	-0.179	-0.179	1.045	1.045
	(0.021) ***	(0.021) ***	(0.017	(0.017) ***	(0.019) ***	(0.019) ***
Interaction: f*T	0.206	0.206	0.469	0.469	0.355	0.355
	(0.021) ***	(0.021) ***	(0.016) ***	(0.016) ***	(0.014) ***	(0.014) ***
Constant	11.236	12.585	12.988	14.097	6.945	9.130
	(0.099) ***	(0.503) ***	(0.104) ***	(0.457) ***	(1.186) ***	(1.192) ***
Observations	225,296	225,296	228,936	228,936	542,300	542,300
R-squared	0.07	0.16	0.01	0.10	0.02	0.09

Standard errors in parentheses are clustered on schools, all models include year dummies, * significant at 10%; ** significant at 5%; *** significant at 1%, the number of observations is twice the number of students since the dataset is stacked; for each student there are two grades: school leaving certificates and test grades.

Table 3b: Ordinary least squares estimation of the interaction between non-native student and type of grade, 1) basic 2) with school dummies and additional covariates

	Swedish		**English**		**Mathematics**	
	1	*2*	*1*	*2*	*1*	*2*
Non-native student (n)	-1.871	-1.438	-1.371	-0.951	-1.618	-1.396
	(0.298)	(0.221)	(0.295)	(0.205)	(0.091)	(0.084)
	***	***	***	***	***	***
Type of grade (T)	0.744	0.744	0.181	0.181	0.870	0.870
	(0.050)	(0.050)	(0.049)	(0.049)	(0.031)	(0.031)
	***	***	***	***	***	***
Interaction: n*T	0.318	0.318	0.116	0.116	0.762	0.762
	(0.127)	(0.127)	(0.100)	(0.100)	(0.061)	(0.061)
	**	**			***	***
Constant	12.429	13.554	13.206	14.703	7.541	9.078
	(0.100)	(0.255)	(0.103)	(0.377)	(1.201)	(1.335)
	***	***	***	***	***	***
Observations	18,984	18,984	19,496	19,496	140,466	140,466
R-squared	0.03	0.12	0.02	0.12	0.05	0.12

Standard errors in parentheses are clustered on schools, all models include year dummies, * significant at 10%; ** significant at 5%; *** significant at 1%, the number of observations is twice the number of students since the dataset is stacked; for each student there are two grades: school leaving certificates and test grades.

Discussion of results

When interpreting the results it is important to keep in mind that school leaving certificates and test results are two *different* measures of school performance. Different qualities are captured in the two grades. For example, a test situation could be associated with more pressure than performance during regular class, which is reflected in the school leaving certificates. Further, accounts of home works (reflected in school leaving certificates) involve preparation and self-studies in order to succeed, while students cannot prepare specifically for the national tests. These differences can *per se* differ across genders and between natives and non-native students.

Descriptive statistics shows that teachers in general are more generous with respect to school leaving certificates than justified by test results. This is true for all sub-groups in all subjects, except for in English with

respect to boys[33]. The generosity in school leaving certificates is especially clear in Mathematics. In Mathematics, a large share of the students fail on the test while a significantly smaller share fail according to the school leaving certificates. Thus, teachers seem to be particularly generous toward students who fail on the test. If the share who fail on the test differs across sub-groups, the generosity towards students who fail on the test could explain, at least part of the interaction effects.

With respect to gender, a larger share of the boys fails on the test in Swedish and English (Figures 1a-b). In Mathematics, about the same share of the girls and the boys fail on the test (Figure 1c). Thus, in the gender case, this explanation seems not to hold. However, among non-natives, a significantly larger share of the students fail on the test compared to native students. This is true in all there subjects studied. Thus, in this case, the results could, at least partly, be explained by the fact that teachers are particularly generous toward students who fail on the test.

Conclusion

School performance could be measured in several ways. The Swedish school system relies heavily on teachers' assessment of student performance. In other countries, national tests play the corresponding role. This paper confirms earlier results that, in comparison to national tests, girls are better rewarded than boys in teachers' assessments. In addition, this paper shows, again in comparison to national tests, that non-native students are more generously rewarded than native students in teachers' assessments. With respect to gender, the result holds in all three subjects studied. In the non-native versus native case, the result holds in two out of three subjects studied.

Among non-natives, the results could partly be explained by the fact that teachers in general are more generous with respect to school leaving certificates towards students who fail on the test. In the gender case, a corresponding explanation does not seem to hold. Thus, the results raise a question for future research: why does the grade difference vary between sub groups of students? Does the school environment itself benefit certain groups of students or do teachers discriminate when assigning grades? Both non-native students and girls are often discussed in the context of discrimination. One possible explanation for why teachers are more gen-

[33] In this case, the opposite is actually true; boys receive on average a lower grade according to school leaving certificates than according to the test result.

erous to these two groups of students could be that teachers are *afraid* to discriminate them. As a consequence teachers perhaps over-compensate girls and non-native students when assigning school leaving certificates.

References

Emanuelsson, I. and Fischbein, S.: 1986, Vive la Difference? A study on Sex and Schooling, Scandinavian Journal of Educational Research 30, 71-84.

Lavy, V.: 2004, Do Gender Stereotypes Reduce Girls' Human Capital Outcomes?, Evidence from a Natural Experiment, NBER Working Paper 10678.

Lpo 94, Läroplan för det obligatoriska skolväsendet, förskoleklassen och fritidshemmet, Lpo 94.

NAEP 2004, Trends in Academic Progress – Three Decades of Student Performance in Reading and Mathematics, US Department of Education, Institute of Education Sciences, NCES 2005-464.

Nycander, M.: 2006, Pojkar och flickors betyg - En statistisk undersökning, Institutionen för lärarutbildning 2006.

PISA 2003, Problem Solving for Tomorrow's World, First Measures of Cross-Curricilar Competencies from PISA 2003, Programme for International Student Assessment, OECD.

Skolverkets allmänna råd 2004, Allmänna råd och kommentarer Likvärdig bedömning och betygsättning.

Skolverket: 2006, Könsskillnader i måluppfyllelse och utbildningsval, rapport nr 286.

Wester-Wedman, A., Gisselberg, K., Mattsson, H. and Wedman, I.M.; 1988, Vilka gynnas vid betygsättningen?, Pedagogiska rapporter nr 21, Pedagogiska institutionen Umeå Universitet

Wikström, C. and Wikström, M.: 2004, Grade inflation and school competition: an empirical analysis based on the Swedish upper secondary schools, Economics of education Review 24:309-322.

Appendix

Table 1a: School leaving certificates in respective subject for unrestricted and restricted samples, the gender case

Subject	**Unrestricted**			**Restricted**		
	Mean	*St dev*	*Obs*	*Mean*	*St dev*	*Obs*
Swedish	12.84	4.49	546,536	13.11	4.25	112,648
English	12.90	4.81	546,536	13.34	4.44	114,468
Mathematics	11.95	4.65	546,536	12.50	4.24	271,150

Table 1b: School leaving certificates in respective subject for unrestricted and restricted samples, the ethnic case

Subject	**Unrestricted**			**Restricted**		
	Mean	*St dev*	*Obs*	*Mean*	*St dev*	*Obs*
Swedish	12.84	4.49	546,536	13.00	4.18	9,492
English	12.90	4.81	546,536	13.23	4.39	9,748
Mathematics	11.95	4.65	546,536	12.38	4.08	70,233

Essay 3: Gender and ethnic interactions among teachers and students - evidence from Sweden·

Introduction

According to test scores, there are persistent differences in school performance between the genders and ethnic groups. International studies show that girls, on average, outperform boys in Reading while the opposite is true in Mathematics (NAEP 1999; PISA 2003). On Swedish national tests, girls score higher than boys in Reading (Swedish and English) while there are no significant differences in Mathematics (Skolverket, 2006; Nycander, 2006; Lindahl, 2007). Further, students that belong to the ethnic majority outperform students from ethnic minorities. For example, in the US, white students, on average, outperform black and Hispanic students (NAEP 2004). In Sweden, native students outperform immigrants in all core subjects according to national test scores (Lindahl 2007).

Another picture emerges when school performance is measured with school leaving certificates, which in Sweden are based on evaluations by the teachers. According to those, girls outperform boys in almost *all* subjects, including Mathematics. In Reading, the gender differences are even larger than what would be expected from corresponding test scores (Skolverket, 2006; Nycander, 2006; Lindahl 2007). Studies from different

· I am grateful for useful suggestions and comments given by Per Johansson, Peter Fredriksson, Tuomas Pekkarinen, Helena Holmlund and Andreas Westermark as well as seminar participants at the Department of Economics, Uppsala University, participants at the First Summer School of the Marie Curie Research Training Network, Padova, 16-18 June 2006 and participants at the COST meeting in Essen, 19-20 October 2006. The financial support from the Swedish council for working life and social research FAS (dnr 2004-1212) is acknowledged.

countries show that girls, on average, are better rewarded than boys in teacher evaluations, irrespective of their relative performance on the corresponding tests (Emanuelsson and Fischbein, 1986). A suggested explanation to this observation is that the school environment might be adapted to traditionally female behaviour.[34] Lindahl (2007) shows that teachers' evaluations differ from the corresponding tests, also with respect to the ethnic background of the students. In school leaving certificates, the difference between native and non-native students are less pronounced than in the national test scores.

The gender and ethnic differences in school performance could potentially be explained by a non-representative composition of the teacher staff with respect to gender and ethnicity. In Sweden, male teachers are under-represented in Reading, and non-native teachers are under-represented in general. Educationalists, for example Einarsson (1981) and Hultman (1981), have claimed that the teachers' gender affect how they devote time and attention to boys and girls (for a discussion see Emanuelsson and Fischbein, 1986). The same reasoning has been applied to ethnic interactions (Ferguson, 1998; and Casteel, 1998). Others have suggested that teachers serve as role-models for their students (Bettinger and Long, 2004). That is, students with the same gender or ethnicity as the teacher may more easily identify themselves with the teacher and thereby perform better in the school. A few studies on data from England and the US suggest that students with the same gender or ethnicity as the teacher perform better on test scores than others (Ammermueller and Dolton, 2006; Dee, 2005b; Dee, 2001). Further, evidence from the US suggests that same race, as well as same gender, of the teacher and the student, is important for the teacher's perception of the student's behaviour and performance (Ehrenberg, Goldhaber and Brewer, 1995; Dee, 2005a and 2005b).

The aim of this study is to examine if students perform better in school, when the share of teachers of the same gender or ethnic background as the student increases. That is, do boys perform better if the share of male teachers increases? Correspondingly, do ethnic minority students perform better if the share of ethnic minority teachers increases? As measurements of performance, I use both test scores *and* the corresponding assessments by the teacher, i.e., the school leaving certificates. Thus, I am able to compare how test scores are affected in comparison to assessments.

The analysis is partly based on Swedish data on grade 9 students (16 years old). For each student, information is available about gender and

[34] See Emanuelsson and Fischbein for a more extensive review of this literature.

country of birth as well as test scores and school leaving certificates in Swedish, English and Mathematics. This information is matched with school-level data on the gender- and ethnic composition of the teacher staff.

The following results are found in Mathematics. On average, students perform slightly better on the test when the share of the same-gender teachers increases. Correspondingly, minority students, on average, perform slightly better when the share of minority teachers increases. The effect sizes correspond to around 4.2 and 27 per cent standard deviations of the test score distribution with respect to gender and ethnicity, respectively. In school leaving certificates, the positive same-gender effect on test scores is counteracted by a *negative* assessment effect. That is, conditional on test scores, same-gender teachers are less generous than opposite-gender teachers in their assessment. This negative effect corresponds to 3 per cent of the standard deviations of the test score distribution in Mathematics with respect to gender. With respect to ethnicity, there is no evidence of an assessment effect. In Swedish and English, no statistically significant effects are found.

School leaving certificates and test scores

The Swedish educational system relies heavily on school leaving certificates. They are used as a selection instrument for higher education as well as in job applications. The teacher alone is responsible for assigning the school leaving certificate. However, the Swedish National Agency for Education formulates the criteria for different grade steps. In the latest curriculum for the Compulsory School System (Lpo, 94) it is stated that school leaving certificates should reflect skills and knowledge in the subject in accordance to the goals stated in the course syllabi. It is also clearly stated that school leaving certificates should *not* reflect attention in the classroom, diligence, ambition, home works and work during lesson, as long as it is not a prerequisite for attaining the goals (as in the case of laboratory work).

National tests are performed during the spring semester in grade 9. The tests in languages (Swedish and English) measure writing and reading abilities as well as listening comprehension and verbal interaction. The tests in Mathematics include analysis and algebra and an oral part testing mathematical reasoning. The tests are corrected at the school level but are graded according to national stated correcting

instructions. Teachers are encouraged to not correct their own students' exams, but they are allowed to do so.

Teachers *shall* use the nationally approved tests when assigning the school leaving certificate (Skolverket 2004). However, in the individual case, the teacher is free to assign the school leaving certificate differently from the test score. The reason is that the student might be low-performing on the test day due to occasional conditions. Further, the teacher should take into consideration all available information about the student's knowledge and ability in the subject. The tests do not capture *all* goals stated in the course syllabi, although the aim of the tests is to measure, as comprehensively as possible, the student's ability and knowledge in the subject. However, it is clear from the national directives that the tests should form an important basis for the school leaving certificates.

Both school leaving certificates and test scores are assessed according to the same ordinal metrics: Fail (F), Pass (P), Pass with Distinction (PD) and Pass with Special Distinction (PSD). The aim of the Swedish school is that all students should attain at least "Pass". To be qualified for upper secondary school, the student has to attain at least "Pass" in the three core subjects Swedish, English and Mathematics.

Data

Data sources

The main data source used is a register, provided by the Swedish Agency for Education (SAE), of school leaving certificates for all students in grade 9 in Sweden (årskurs 9 registret). This register contains the gender and age of the student, which year and which school the student attended in grade 9. This information is combined with an additional register about teachers (lärarregistret). The teacher register contains information on what school and in which subject each teacher teaches as well as the gender of the teacher. Finally, information about test scores, collected by SAE is added. I use test scores from year 2001 to 2005.[35] Between 2001 and 2002, test scores from a random sample of 150 schools were collected. From 2003 and onwards all schools were meant to be collected. However, all subjects are not collected all years. Swedish and English are not collected in 2002 and 2003 and Mathematics is missing in 2001.

[35] A stratified sample of test scores is collected between 1998 and 2001.

Since the availability of test scores differs between the three different subjects, one sample is constructed for each subject. These samples are restricted to include all covariates of interest (no missing value of any covariate). All results presented are based on these restricted samples. The sample sizes in the gender case are 109,204 in Swedish, 111,623 in English and 268,334 in Mathematics.

In the ethnic background case, an additional register (Louise) is used: a register from Statistics Sweden about country of birth. At present, this information is available for all students younger than 17 in 2003. Students in Sweden are expected to reach the age of 16 in grade 9. Thus, country of birth is only available for students who graduated before 2004. The samples used in this case are: 7,766 in Swedish, 8,579 in English and 69,149 in Mathematics.

Variable definitions

Unfortunately, the information about which subject each teacher teaches is not precise. The information is given by a code indicating if the teacher is teaching within a subject block. The subject blocks relevant in this study are: 1) Mathematics and Science (used for teachers in Mathematics), 2) Social Sciences and Swedish (used for teachers in Swedish) and 3) Swedish and languages (used for teachers in English). Thus, the information about which subject the teachers teach is more precisely measured in Mathematics than in Swedish and English.

The teacher information is matched with the student information via a school code. Thus, I only have information about the share of female and minority teachers at the school.[36]

In Sweden many individuals are born in another Nordic country than Sweden. Those individuals speak Swedish well, look Swedish and know the country well. Ethnic minority students are therefore defined as those born in a non Nordic country.

For students with another language than Swedish as mother tongue and who are assessed to not be able to follow the ordinary course in Swedish, a special course is offered: *Swedish 2*. The *Swedish 2* course has about the same course syllabus as the ordinary course in Swedish but the teaching is adapted for students with another mother tongue than Swedish. Since 2001, the national test is the same for the course *Swedish 2* as for the ordinary course in Swedish. In this paper *Swedish 2* and Swedish

[36] I have information about how many hours each teacher has taught at the school in respective subject. Thus, the share is adjusted for part-time working.

are treated as one subject.[37] The focus in this paper is on student and teacher interactions. Thus, the type of course the student has taken within the subject should be of less importance.

The Swedish National Agency for Education summarizes the school leaving certificate (all final course grades) into a total Grade Point Average (GPA). This GPA is the instrument used for application to higher education. The values used for transforming the ordinal scale to a numerical scale in order to calculate the GPA are: 0, 10, 15 and 20. Fail equals 0, Pass equals 10 and Pass with distinction and Pass with special distinction equals 15 and 20, respectively. All results presented in this study are based on grades transformed to these numerical values.

Sample selection

Except for test scores, all information used stems from register data containing all students in grade 9 in Sweden. However, test scores are only available for a selected group of students. That is, if test scores from the school are reported to SAE and if the student has completed all parts of the test. Test scores shall be reported to SAE by the school and all students are requested to complete all parts of the test. Despite these requests, I may have a selection of schools and/or students within schools. Table A in the Appendix presents descriptive statistics on all covariates used. Both unrestricted and restricted (the samples used in the analysis) are presented for all three subjects.

The most striking difference is that the mean in school leaving certificates is somewhat higher in the restricted samples. A plausible explanation is that those students who did not complete all parts of the test on average perform worse than those who did. Further, the share of unqualified teachers is lower in the restricted samples.

In the ethnic samples, there are differences in mean values also with respect to other covariates. The reason is probably that the number of schools that did not report test scores decreases between 2003 and 2005. In the ethnic background analysis, data until 2003 are used while data until 2005 are used in the gender analysis. The share of minority students is somewhat lower in the restricted samples in Swedish and Mathematics. Further, the teachers' mean age is lower in the restricted samples in Swedish and English.

[37] Controlling for Swedish 2 does not affect any of the interaction estimates presented in this paper.

The conclusion is that this analysis is based on students who on average perform (according to school leaving certificates) slightly above the average in Sweden, in comparison to the overall population.

Descriptive statistics

Table 1a-b presents the gender and ethnic differences in means according to school leaving certificates and test scores, respectively. The first column reports the share of female (minority) teachers in each subject. The fourth and the last column presents the correlation between the share of female (ethnic minority) teachers and the gender (ethnic) difference, at the school level, in test scores and school leaving certificates, respectively.

According to test scores, girls outperform boys in Swedish and English while there is no statistically significant difference in Mathematics. According to school leaving certificates, girls outperform boys in all subjects. Thus, in line with other studies, Table 1 shows that the gender gap in favour of girls seems to be reinforced in teachers' evaluations, compared to the corresponding test scores.

Female teachers dominate in Swedish and English; the shares of female teachers in these subjects are 62 per cent and 83 per cent, respectively. In Mathematics, the gender composition among teachers is almost balanced; 46 per cent of the teachers in Mathematics are female.

If girls perform better with a female teacher, we would expect the positive gender gap in favour of girls at the school level to be positively correlated with the share of female teachers at the school. Except for test scores in Swedish, there is a positive correlation between the share of female teachers and the difference between the performance of girls and boys. This correlation is largest in Swedish with respect to school leaving certificates.

Minority students on average perform below natives in all subjects. This is true both according to school leaving certificates and test scores. According to test scores, the gap is smallest in English while it is smallest in Swedish according to school leaving certificates. Thus, also in this case, the size of the gap depends on how school performance is measured.

The share of minority teachers is largest in English, 11 per cent, and smallest in Swedish, 2 per cent. In Mathematics the share is 7 per cent. Again, except for school leaving certificates in English, there is a positive correlation between the share of minority teachers and the difference in

performance between minority and majority students. This correlation is largest in Mathematics with respect to test scores.

The correlations reported in Tables 1a and 1b are suggestive of an association between the composition of teachers and the relative performance of students with certain characteristics. However, to estimate if the gender and ethnic balance among teachers has a *causal* effect on the performance of girls and minority students, we need a formal model.

Table 1a: Difference in tests scores and school leaving certificates between genders and the share of female teachers in respective subject

		Tests scores				**School leaving certificates**			
Subject	Share female teachers (Std. dev)	Average grade (Std. dev)		Difference (Std. dev)	Correlation: the share of female teachers and the gender gap at the school	Average grade (Std. dev)		Difference (Std. dev)	Correlation: the share of female teachers and the gender gap at the school
		Girls	*Boys*			*Girls*	*Boys*		
Swedish	0.62 (0.32)	13.40 (4.06)	11.29 (4.31)	2.11 (0.02)	-0.0275	14.30 (4.09)	11.99 (4.04)	2.31 (0.02)	0.0345
English	0.83 (0.24)	13.47 (4.31)	13.05 (4.48)	0.41 (0.02)	0.0017	13.77 (4.35)	12.92 (4.41)	0.85 (0.02)	0.0016
Math	0.46 (0.30)	11.24 (5.18)	11.22 (5.15)	0.02 (0.02)	0.0283	12.69 (4.22)	12.30 (4.22)	0.39 (0.02)	0.0131

Table 1b: Difference in tests grades and school leaving certificates between ethnic minority and ethnic majority students and the share of ethnic minority teachers in respective subject

		Tests scores				**School leaving certificates**			
Subject	Share minority teachers (Std. dev)	Average grade (Std. dev)		Difference (Std. dev)	Correlation: the share of minority teachers and the ethnic gap at the school	Average grade (Std. dev)		Difference (Std. dev)	Correlation: the share of minority teachers and the ethnic gap at the school
		Minority	*Majority*			*Minority*	*Majority*		
Swedish	0.02 (0.08)	10.82 (4.46)	12.48 (3.99)	-1.66 (0.10)	0.0107	12.06 (4.20)	13.25 (4.07)	-1.19 (0.10)	0.0092
English	0.11 (0.21)	11.93 (4.75)	13.13 (4.08)	-1.20 (0.10)	0.005	12.21 (4.83)	13.42 (4.14)	-1.22 (0.10)	-0.0011
Math	0.07 (0.17)	9.07 (5.37)	11.48 (4.76)	-2.41 (0.06)	0.0582	10.97 (4.31)	12.51 (4.03)	-1.55 (0.05)	0.0153

Econometric model

For ease of exposition I only refer to the gender case in the following. The same reasoning is applicable for the ethnic case -- just replace "female student" by "non-native student".

Identification

The correlation between the gender gap in school performance and the gender balance among teachers does not necessarily imply a *causal* relationship. That is, it is not necessarily the case that female teachers *cause* better performance of girls in comparison to boys. For example, it *might* be the case that girls inherently excel in readings. This fact *may* in turn imply that female teachers, to a higher degree than male teachers, choose to teach in reading.

In order to estimate a potential *causal* effect for a girl of having a female teacher, I have to control for gender-specific student and teacher characteristics that also influence school performance. In this study the information about teachers is at the school level. Thus, self-selection of students and teachers within schools is not an issue. However, I still might have sorting of both students and teachers between schools.

Both the gender of the student and the share of female teachers at the school may be correlated with the status of the school. In Sweden, there are large differences across schools both according to national test scores and according to school leaving certificates (Skolverket, 2007). Thus, schools may have an independent effect on test scores and school leaving certificates. Self-selection of students and teachers to schools may result in a positive association between the gender of the student and the gender of the teacher, although no such causal relationship exists. In order to take care of selection of students and teachers between schools, I apply a difference-in-differences strategy. I control for both the gender of the student and the share of female teachers at the school. The parameter of interest is then the parameter in front of the interaction between these two terms. This parameter can be interpreted as the additional effect for girls of an increase in the share of female teachers by one per centage point.

In order to examine the difference between any potential effect on test scores and any potential effect on school leaving certificates, I also include a dummy for type of grade (school leaving certificates or test scores). By interacting all independent variables also with type of grade, I

am able to study how a potential effect on school leaving certificates – an assessment effect – works in addition to a potential effect on test scores.

Estimation

The grade steps in Sweden are ordinal. This fact suggests an *ordered probit* or an *ordered logit* model. However, the results from a linear model are easier to interpret and established numerical values exist for the ordinal scale. Thus, a linear model is a good approximation. The following model is estimated with ordinary least squares.

$$Y_{ijst} = \beta_0 + \beta_1 f_{it} + \beta_2 T_{ijt} + \beta_3 f_{it} T_{ijt} + \beta_4 F_{st} + \beta_5 f_{it} F_{st} + \beta_6 F_{st} T_{ijt} + \beta_7 f_{it} F_{st} T_{ijt} + \eta_t + \varepsilon_{ijst}$$

where Y_{ijst} is individual *i*'s type of grade *j* (school leaving certificate or test score) in school *s*, year *t*; *f* is a dummy for being female. Furthermore, *T* is a dummy for the type of grade where *T* equals 1 if the grade corresponds to school leaving certificates and 0 if the grade is test scores. *F* denotes the share of female teachers in the subject. The term η_t captures year effects[38] and ε_{ist} is the assumed idiosyncratic error term.

The parameter estimates of interest are the ones in front of the interaction between the gender of the student (*f*) and the share of female teachers at the school (*F*). A student-teacher interaction effect on test scores is captured by β_5. If an increase in the share of female teachers positively affects girls' performance on the test, β_5 is positive. By interacting the term ($f_{it} F_{st}$) also with the type of grade, we capture an additional student-teacher interaction effect associated with school leaving certificates. That is, β_7 measures whether female teachers, in comparison with their male colleagues and conditional on any student teacher interaction effect on test scores, favour their own gender when assigning school leaving certificates. This is the potential assessment effect.

The model specification above takes care of potential selection between schools as long as girls do not self select to certain schools *because* there is a large share of female teachers (or the other way round, teachers choose school depending on the students). However, if a high share of girls at the school attracts female teachers (or vice versa) *and* if this selec-

[38] I estimate a static linear panel data model. However, the sample covers several years and some schools are observed several times. A year specific dummy is therefore included to capture which year the individual is observed. Including year dummies imply that we do not have to worry about changes in grade policy (grade inflation) over time. Wikstrom and Wikstrom (2004) claim that grade inflation occurs in Sweden during the 1990s.

tion is correlated with the outcome variable, β_5 and β_7 are not consistently estimated. With respect to ethnicity, this type of selection may exist due to preferences for teachers/students with a similar cultural background. In order to remedy this potential bias I add school dummies. With school dummies included in the model, I explore the variation within schools over time.

When the estimation stems from variation within schools over time, I have to assume that the error term should not be correlated with the explanatory variables across time periods, in order to receive consistent estimates. This assumption is violated if teachers, conditional on all covariates, assign extraordinary high grades in year *t* and this fact in turn affects students' or teachers' choices of schools the following year. This could be a real problem if students and/or teachers choose schools depending on the school results the previous year. However, it is presumably more realistic to assume that school choice is based on average results from several years, in which case this is not an issue.

In order to increase the precision of the estimates, I also add the following covariates: the share of female (minority in the ethnic case) students at the school, the student's age, the mean age of the teachers in the subject at the school, the share of unqualified and only generally unqualified teachers at the school.[39] In the ethnic case, I also include the share of female teachers. The estimates of these additional covariates are not presented.[40]

In all model specifications, inferences presented are based on standard errors that accommodate heteroscedasticity at the school level. This is appropriate since the sampling unit with respect to test scores is schools.

[39] Unqualified teachers are those who lack formal subject specific training while generally unqualified teachers are those who lack formal pedagogical training.

[40] The reason for *not* including the non-native dummy in the gender case is that this information would reduce the sample size. It is reasonable to assume that β_5 and β_7 in the gender case are orthogonal to the information about the student's country of birth. Thus, the country of birth information would probably not affect the parameter estimates, only the precision of the estimates.

Results

Table 2a presents the results for the gender case and Table 2b the results for the ethnicity case.

A statistically significant (at the 5 per cent level) student-teacher interaction effect is found in Mathematics on test scores. The effect remains when adding school dummies and additional covariates (column 2). The interpretation is that female students on average perform slightly better on the test when the teacher gender changes from male to female.[41] The effect size of this parameter (0.220 credit points) corresponds to around 4.2 per cent of the standard deviation of the test score distribution in Mathematics. Thus, the estimated effect is small; a change from a male teacher to a female would on average raise girls' test scores with 0.220 credit points, corresponding to around 4 percent of the grade step between "Pass" and "Pass with Distinction". This implies that 1 out of 23 female students would get one grade higher test result.

The result could also be interpreted from the perspective of boys, i.e., male students would on average perform slightly better on the test if the teacher changes from a female to a male.

The student teacher interaction effect associated with school leaving certificates – the assessment effect – is also statistically significant but *negative*. This effect is unaffected by the inclusion of school dummies and additional covariates (column 2). The *negative* assessment effect is more than half of the effect size found on test scores and corresponds to 3 per cent of the standard deviation of the test score distribution in Mathematics. Thus, when school performance is measured with school leaving certificates, the positive effect on test scores is counteracted by a negative assessment effect. It is therefore key how school performance is measured. A negative assessment effect associated with school leaving certificates, could explain why Holmlund and Sund (2006) and Skolverket (2006) did not find any positive effects when using school leaving certificates on Swedish data as the outcome variable.

In Swedish and English, neither the parameter intended to capture a student teacher effect on test scores, nor the parameter intended to capture an assessment effect, are statistically significant. The non-precisely estimated effects are close to zero, suggesting that no student teacher interaction effects exist in these subjects.

The other parameter estimates of the model also show interesting phenomena. Some of these results are discussed in Lindahl (2007) in which

[41] The variation in the variable F (share of female teachers at the school) varies between 0 and 1. Thus, one unit increase corresponds to a shift from 0 to 1.

the focus is on the gender difference when comparing school leaving certificates and test scores. However, in addition to the results presented in Lindahl (2007), the model in this study shows that a larger share of female teachers in Mathematics is associated with better test scores for male students. Since male students' test scores correspond to the reference category in the model, the estimate on the share of female teachers, *F*, show this.

With respect to ethnicity, the parameter estimate of the interaction effect on test scores in Mathematics becomes twice as large and statistically significant when school dummies are included in the model (compare column 1 and 2). Thus, in this case, school dummies probably take care of some selection. The interpretation of this result is that if the teacher in Mathematics changes from a native to a non-native, non-native students would perform on average 1.4 credit points better. This effect size corresponds to around 27 per cent of the standard deviation of the test score distribution in Mathematics.

Also in this case, a negative assessment effect is found. However, the estimate is not precisely estimated. Compared to the estimated assessment effect in Mathematics with respect to gender, the estimated effect size is much larger but the standard error is also significantly larger.

In Swedish and English none of the interaction parameters of interest are statistically significant. The non-precisely estimated effects on test scores are positive and about the same size as the corresponding estimated effect in Mathematics. However, the standard errors are larger. An explanation for the lack of precision could be the smaller sample sizes used in Swedish and English, compared to Mathematics. In addition, the information about which subject the teachers teach is less precisely measured in Swedish and English than in Mathematics.

Table 2a: Interacting share of female teachers at the school in the subject and female student, parameter estimates from ordinary least squares estimation: (1) Basic (2) Covariates and school dummies added

	Swedish		***English***		***Math***	
	(1)	*(2)*	*(1)*	*(2)*	*(1)*	*(2)*
Female student (f)	2.158	2.083	0.453	0.332	-0.083	-0.164
	(0.080)***	(0.080)***	(0.150)***	(0.143)**	(0.053)	(0.052)***
Type of grade (T)	0.723	0.723	-0.016	-0.016	1.122	1.122
	(0.054)***	(0.054)***	(0.081)	(0.081)	(0.043)***	(0.043)***
Interaction: f* T	0.174	0.174	0.475	0.475	0.432	0.432
	(0.051)***	(0.051)***	(0.069)***	(0.069)***	(0.033)***	(0.033)***
Share of female teacher (F)	0.061	0.285	0.413	-0.132	0.443	0.321
	(0.139)	(0.282)	(0.222)*	(0.272)	(0.126)***	(0.150)**
Interaction: F*f	-0.100	-0.079	-0.128	-0.108	0.241	0.220
	(0.121)	(0.121)	(0.175)	(0.167)	(0.102)**	(0.098)**
Interaction: T*F	-0.043	-0.043	-0.199	-0.199	-0.154	-0.154
	(0.081)	(0.081)	(0.095)**	(0.095)**	(0.079)*	(0.079)*
Interaction: T*F*f	0.053	0.053	-0.005	-0.005	-0.161	-0.161
	(0.078)	(0.078)	(0.082)	(0.082)	(0.062)***	(0.062)***
Observations	218408	218408	223246	223246	536668	536668
R-squared	0.07	0.16	0.01	0.10	0.02	0.09

Standard errors in parentheses are clustered on schools, all models include year dummies, * significant at 10%; ** significant at 5%; *** significant at 1%, the number of observations is twice the number of students since the dataset is stacked; for each student there are two grades: school leaving certificates and test scores

Table 2b: Interacting share of ethnic minority (non Nordic-born) teachers at the school in the subject and ethnic majority (non Nordic-born) born student, parameter estimates from ordinary least squares estimation: (1) Basic (2) Covariates and school dummies added

	Swedish		***English***		***Math***	
	(1)	*(2)*	*(1)*	*(2)*	*(1)*	*(2)*
Non Nordic-born student (n)	-1.666	-1.119	-1.487	-0.958	-1.704	-1.330
	(0.331)***	(0.227)***	(0.380)***	(0.268)***	(0.097)***	(0.093)***
Type of grade (T)	0.775	0.775	0.178	0.178	0.860	0.860
	(0.056)***	(0.056)***	(0.055)***	(0.055)***	(0.033)***	(0.033)***
Interaction: n* T	0.412	0.412	0.046	0.046	0.798	0.798
	(0.157)***	(0.157)***	(0.120)	(0.120)	(0.069)***	(0.069)***
Share of non Nordic-born teacher (N)	1.616	23.552	0.044	12.638	-0.130	-0.600
	(1.257)	(17.479)	(0.365)	(19.896)	(0.342)	(0.943)
Interaction: N*n	1.546	1.389	1.574	1.036	0.755	1.409
	(1.734)	(1.116)	(1.261)	(0.732)	(0.522)	(0.478)***
Interaction: T*N	-0.588	-0.588	0.155	0.155	0.275	0.275
	(0.457)	(0.457)	(0.257)	(0.257)	(0.227)	(0.227)
Interaction: T*N*n	0.228	0.228	0.469	0.469	-0.431	-0.431
	(1.233)	(1.233)	(0.471)	(0.471)	(0.312)	(0.312)
Observations	15532	15532	17158	17158	138298	138298
R-squared	0.03	0.17	0.02	0.13	0.05	0.12

Standard errors in parentheses are clustered on schools, all models include year dummies, * significant at 10%; ** significant at 5%; *** significant at 1%, the number of observations is twice the number of students since the dataset is stacked; for each student there are two grades: school leaving certificates and test scores

Conclusion

In Mathematics, girls and ethnic minority students on average would perform slightly better on the national test if the share of female or ethnic minority teachers, respectively, increases. The result could also be interpreted from the perspective of boys and native students, i.e., boys and ethnic majority students on average would perform slightly better on the national test if the share of male or ethnic majority teachers, respectively, increases. This effect suggests that teachers serve as role models or that the student-teacher interaction itself induces a positive effect on students' performance. However, when school performance is measured with school leaving certificates – assigned by the teacher – this positive effect is reduced. The reason is that school leaving certificates are associated with a *negative* assessment effect. In the gender case, this effect is statistically significant. That is, female teachers are on average less generous when assigning school leaving certificate to girls, compared to their male colleagues. Lindahl (2007) shows that conditional on test scores, girls are more generously rewarded in school leaving certificates than boys. In this study it is shown that female teachers in Mathematics seem to counteract this tendency.

References

Ammermueller, A. and Dolton, P.: 2006, Pupil-teacher gender interaction effects on scholastic outcomes in England and the USA, Discussion paper no. 06-060, ZEV, Centre for European Economic Research.

Bettinger, E. and Long, B.: 2005, Do faculty serve as role models? The impact of instructor gender on female students, The American Economic Review 95, 152-157.

Casteel, C. A.: 1998, Teacher-student interactions and race in intergrated classrooms, Journal of Educational Research 92: 115-120.

Dee, T.: 2001, Teachers, race and student achievement in a random experiment, Review of Economics and Statistics, vol.8681, 195-210.

Dee, T.: 2005a, A teacher like me: does race, ethnicity or gender matter?, American Economic Review, vol. 95(2), .158-65.

Dee, T.: 2005b, Teachers and the gender gaps in student achievement, National Bureau of Economic Research, Working Paper 11660.

Ehrenberg, R., Goldhaber, D. and Brewer, D.: 1995, Do teachers' race, gender, and ethnicity matter?, Evidence from NELS88, Industrial and Labor Relations Review 48(3): 547-561.

Emanuelsson, I. and Fischbein, S.: 1986, Vive la difference? A study on sex and schooling, Scandinavian Journal of Educational Research 30, 71-84.

Einarsson, J.: 1981, Report 5:1981 from the Department of subject methodology and theory (Swedish) at the university of Lund. Language and Sex Project.

Ferguson, R. F.: 1998, Teachers' perceptions and expectations and the black-white test score gap", in the black-white test score gap, Jencks, C. and Philips, M., editors, Brookings Institution Press Washington DC, 1998.

Holmlund, H. and Sund, K.: 2007, Is the gender gap in school performance affected by the sex of the teacher?, Labour Economics.

Hultman: 1981, Report 6:1981 from the Department of subject methodology and theory (Swedish) at the university of Lund. Language and Sex Project.

Lindahl, E.: 2007, Comparing teachers' assessments and national tests – evidence from Sweden, Working Paper 2007:24.

Lpo 94, Curriculum for the compulsory school system, the pre-school class and the leisure-time centre, Lpo 94.

NAEP 2004 Trends in academic progress – three decades of student performance in Reading and Mathematics, US Department of Education, Institute of Education Sciences, NCES 2005-464.

Nycander, M.: 2006, Pojkar och flickors betyg - en statistisk undersökning, Institutionen för lärarutbildning, Uppsala universitet.

PISA 2003, Problem solving for tomorrow's world, first measures of cross-curricular competencies from PISA 2003, Programme for International Student Assessment, OECD.

Skolverkets allmänna råd 2004, Allmänna råd och kommentarer, Likvärdig bedömning och betygsättning.

Skolverket: 2006, Könsskillnader i måluppfyllelse och utbildningsval, rapport nr 286.

Skolverket: 2007, Provbetyg - slutbeyg – likvärdig bedömning?, rapport nr 300.

Wikström, C. and Wikström, M.: 2004, Grade inflation and school competition: an empirical analysis based on the Swedish upper secondary schools, Economics of Education Review 24:309-322.

Appendix

Table 1a: Sample selection in Swedish, gender case

	Unrestricted data			***Restricted data***	
	Obser-vations	*Mean*	*St dev*	*Mean*	*St dev*
Female student	556,087	0.49	0.50	0.49	0.50
Student's month of birth	556,087	6.29	3.37	6.34	3.38
Student's year of birth	556,087	1987.06	1.67	1988.72	1.05
School leaving certificate	546,536	12.84	4.49	13.12	4.24
Share of female teacher	526,230	0.59	0.28	0.59	0.27
Mean teacher age in subject	526,230	41.33	7.27	41.89	6.96
Share unqualified teachers in subject	526,230	0.19	0.23	0.16	0.21
Share generally qualified teachers in subject	526,230	0.08	0.15	0.08	0.16
Test score[1]	113,568	12.31	4.38	12.33	4.35
Number of obser-vations, when no of above covari-ates are missing	109,204				

1) The restricted sample is almost defined by those students who have completed all parts of the national test. However, since there are some missing information also in register data, the sample of students for who national score results exist, is slightly reduced.

Table 1b: Sample selection in English, gender case

	Unrestricted data			***Restricted data***	
English Gender	*Obser-vations*	*Mean*	*St dev*	*Mean*	*St dev*
Female student	556,087	0.49	0.50	0.49	0.50
Student's month of birth	556,087	6.29	3.37	6.29	3.36
Student's year of birth	556,087	1987.06	1.67	1988.10	0.89
School leaving certificate	546,536	12.90	4.81	13.33	4.43
Share of female teacher	526,943	0.84	0.19	0.84	0.18
Mean teacher age in subject	526,943	43.25	7.07	43.54	6.66
Share unqualified teachers in subject	526,943	0.31	0.26	0.28	0.24
Share generally qualified teachers in subject	526,943	0.06	0.13	0.06	0.13
Test score[1]	115,361	13.28	4.46	13.28	4.44
Number of obser-vations, when no of above covari-ates are missing	111,623				

1) The restricted sample is almost defined by those students who have completed all parts of the national test. However, since there are some missing information also in register data, the sample of students for who national score results exist, is slightly reduced.

Table 1c: Sample selection in Mathematics, gender case

	Unrestricted data			***Restricted data***	
Mathematics Gender	*Obser-vations*	*Mean*	*St dev*	*Mean*	*St dev*
Female student	556,087	0.49	0.50	0.49	0.50
Student's month of birth	556,087	6.29	3.37	6.29	3.36
Student's year of birth	556,087	1987.06	1.67	1988.10	0.89
School leaving certificate	546,536	11.95	4.65	12.50	4.23
Share of female teacher in subject	539,823	0.46	0.24	0.47	0.23
Mean teacher age in subject	539,823	41.85	6.63	42.47	5.97
Share unqualified teachers in subject	539,823	0.26	0.23	0.24	0.21
Share generally qualified teachers in subject	539,823	0.08	0.14	0.08	0.13
Test score[1]	273,099	11.27	5.17	11.27	5.16
Number of observations, when no of above covariates are missing	268,334				

1) The restricted sample is almost defined by those students who have completed all parts of the national test. However, since there are some missing information also in register data, the sample of students for who national score results exist, is slightly reduced.

Table 1d: Sample selection in Swedish, ethnic case

	Unrestricted data			***Restricted data***	
	Obser-vations	*Mean*	*St dev*	*Mean*	*St dev*
Female student	556,087	0.49	0.50	0.49	0.50
Student's month of birth	556,087	6.29	3.37	6.34	3.40
Student's year of birth	556,087	1987.06	1.67	1985.03	0.36
School leaving certificate	546,536	12.84	4.49	13.07	4.12
Share of non Nordic-born teachers	526,230	0.02	0.07	0.03	0.09
Mean teacher age in subject	526,230	41.33	7.27	35.22	6.79
Share female teachers in subject	526,230	0.59	0.28	0.61	0.32
Share unqualified teachers in subject	526,230	0.19	0.23	0.24	0.28
Share only generally unqualified teachers in subject	526,230	0.08	0.15	0.10	0.18
Non Nordic-born students	320,123	0.09	0.29	0.08	0.28
Test score[1]	113,568	12.31	4.38	12.28	4.09
Number of observations, when no of above covariates are missing	7,766				

1) The restricted sample is almost defined by those students who have completed all parts of the national test. However, since there are some missing information also in register data, the sample of students for who national score results exist, is slightly reduced.

Table 1e: Sample selection in English, ethnic case

	Unrestricted data			***Restricted data***	
	Obser-vations	*Mean*	*St dev*	*Mean*	*St dev*
Female student	556,087	0.49	0.50	0.49	0.50
Student's month of birth	556,087	6.29	3.37	6.32	3.40
Student's year of birth	556,087	1987.06	1.67	1985.03	0.34
School leaving certificate	546,536	12.90	4.81	13.23	4.36
Share of non Nordic-born teachers	526,943	0.10	0.16	0.11	0.22
Mean teacher age in subject	526,943	43.25	7.07	37.73	7.64
Share female teachers in subject	526,943	0.84	0.19	0.82	0.27
Share unqualified teachers in subject	526,943	0.31	0.26	0.47	0.36
Share only generally unqualified teachers in subject	526,943	0.06	0.13	0.05	0.17
Non Nordic-born students	320,123	0.09	0.29	0.09	0.28
Test score[1]	115,361	13.28	4.46	13.02	4.30
Number of observations, when no of above covariates are missing	8,579				

1) The restricted sample is almost defined by those students who have completed all parts of the national test. However, since there are some missing information also in register data, the sample of students for who national score results exist, is slightly reduced.

Table 1f: Sample selection in Mathematics, ethnic case

	Unrestricted data			***Restricted data***	
	Obser-vations	*Mean*	*St dev*	*Mean*	*St dev*
Female student	556,087	0.49	0.50	0.48	0.50
Student's month of birth	556,087	6.29	3.37	6.40	3.35
Student's year of birth	556,087	1987.06	1.67	1986.85	0.38
School leaving certificate	546,536	11.95	4.65	12.38	4.08
Share of non Nordic-born teachers	539,823	0.07	0.14	0.06	0.13
Mean teacher age in subject	539,823	41.85	6.63	42.28	6.06
Share female teachers in subject	539,823	0.46	0.24	0.45	0.23
Share unqualified teachers in subject	539,823	0.26	0.23	0.26	0.21
Share only generally unqualified teachers in subject	539,823	0.08	0.14	0.08	0.13
Non Nordic-born students	320,123	0.09	0.29	0.08	0.28
Test score[1]	273,099	11.27	5.17	11.44	4.76
Number of observations, when no of above covariates are missing	69,149				

1) The restricted sample is almost defined by those students who have completed all parts of the national test. However, since there are some missing information also in register data, the sample of students for who national score results exist, is slightly reduced.

Essay 4: Does collaboration between health care and the Social Insurance Agency help the sick?♦

Introduction

Several European countries have recently experienced high and rising costs of sickness insurance. In 2006, Swedish social insurance expenditure totaled 447 billion Swedish kronor (SEK)[42] or just below 16 per cent of GNP (Försäkringskassan, 2007b). One-third of this amount comprises payments to the sick and disabled. In response to rising sickness absence figures, the Swedish government has launched several actions in order to restrain the development. In 2003, the Swedish Council on Technology Assessment in Health Care (SBU) claimed that a crucial factor for the high sickness absence rate was the lack of efficient collaboration between the primary health care and the social insurance office, i.e., between the medical doctors, who assess the working capacity of insured individuals, and the case workers at the social security office, who make the decision on sickness benefits. In addition, the SBU report stressed the importance of involving behavioural and physiotherapy competence in the decision about reporting sickness. This conclusion is also stressed in a report from the Swedish National Board of Health and Welfare from 2004 (Socialstyrelsen, 2005).

♦ Co-authored with Per Johansson. The authors thank Lasse Einarsson at Försäkringskassan in Uppsala for helping us to conduct the experiment. The authors also thank seminar participants at Uppsala University for valuable comments. The financial support from the Swedish council for working life and social research FAS (dnr 2004-2005) is acknowledged.

[42] 46.6 billion euros. 1 euro = 9.6 SEK.

In Sweden, as well as in other countries, a large number of programmes have been launched in order to make the collaboration between medical doctors and case workers more efficient. Several of them have been evaluated from the perspective of the involved personnel's experiences and attitudes to the programmes (Statskontoret, 2006; Socialstyrelsen, 2001; Socialstyrelsen, 2000a; Socialstyrelsen, 2000b; Hultberg, 2005; Danemark and Kullberg, 1999). However, the overall aim of the programmes – to reduce the social insurance expenditure – has not been evaluated to the same extent.[43] This is the case for Swedish collaboration programmes as well as similar programmes in other countries (see Dowling, Powell and Glendinning, 2004, for an overview of the situation in Great Britain and Schmitt, 2001, for the US case).

The aim of this study is to determine if a collaboration programme between a social insurance office and the primary health care has succeeded in shortening the sick-spell durations of the individuals who received the programme. In the programme, named Resursteam (RT), the sick-listed individual's medical doctor, her case worker, a behaviourist and a physiotherapist meet regularly to discuss and assess the insured individual's need for rehabilitation. The target group consisted of individuals who were assessed by the medical doctor or the case worker to be at risk of becoming long-term sick. RT was in place in Uppsala County in Sweden, 2004–2007.

The evaluation is based on register information on individual sickness spells. In addition, we have conducted an experiment in which 50 eligible candidates were assigned randomly to RT. The randomization was performed by a random generator within the administration system.

Since the experiment is small, we extend the experiment with an observational study; we explore register information on all individuals who have been assigned to RT. A rich data set including individual (social background variables and sickness absence history) as well as sick-spell-specific information (for example, diagnosis at sick-spell start) allows us to recreate the selection for RT and to verify this strategy with rigorous sensitivity analyses.

In general, case workers select candidates for the programme based on register information on risk factors of long-term sickness. Medical doctors, on the other hand, make their decision after a personal meeting with potential candidates. Thus, it is reasonable to believe that potential selec-

[43] Two interesting exceptions are Hultberg (2005) and Kärrholm (2007). Hultberg did not find any positive effect on sickness absence. Kärrholm found that participants had 5.7 days fewer short-term sickness absence days per month than comparison individuals during a six-year follow-up period.

tion problems based on unobservable information are more severe for the group of individuals selected by medical doctors. We investigate if the estimated effect is sensitive to whether the individual is initiated by a case worker or a medical doctor. Further, we have access to repeated individual sick spells which allows us to take fixed unobserved individual heterogeneity into account.

The results from the experiment and the retrospective observational study suggest a negative effect of RT; on average RT prolongs the sick-spell duration rather than shortens it.

Sickness absence and institutions

Sickness absence in Sweden

In Western Europe, Sweden and Norway are the countries with the most costly public sickness insurances schemes relative to GDP. Sweden has a large number of older women in the workforce, which could explain part of the internationally high figures (Nyman, Palmer and Bergendorff, 2002). Another characteristic is that the sickness absence in Sweden has varied significantly over time. From an international perspective, the sickness absence among employees has increased significantly in Sweden, Norway and the Netherlands, compared with other European countries during the late 1990s (Nyman *et al.*, 2002).

During the last years, this negative trend has stopped. Relative to GDP, social insurance payments fell for the third consecutive year in 2006 (Försäkringskassan, 2007a). In September 2003, the number of days compensated for sickness and work disability was 43.3 days per year and insured. In December 2006, the corresponding figure was 39.9. The reason for this reduction in sickness absence has been debated and it is reasonable to assume that there is not just one reason. However, this change occurred at the same point of time as a number of different programmes were launched by the Swedish Social Insurance Agency (Försäkringskassan), and sanctioned by the Government, in order to reduce the sickness absence (for an overview: see Anderzén, Demmelmaier, Hansson, Johansson, Lindahl and Winblad, 2008). For example, in 2003, the Swedish Social Insurance Agency introduced a new national programme in order to reduce costs associated with sickness absence. This programme involved new routines for handling sick-listed individuals. An explicit aim of this programme was to stress working ability rather than incapacity with respect to sickness (Försäkringskassan, 2007b).

Sickness insurance

The aim of the sickness insurance in Sweden is to provide the population with protection against financial risk associated with illness and disability. All those who have reached the age of 16 and are resident in Sweden are in principle insured and registered with the Swedish Social Insurance Agency.

There are two main benefits in the Swedish sickness insurance: *sickness benefit* and *sickness compensation* or *activity compensation*. Sickness benefit replaces part of the income loss during temporary illness. The first 14 days of the illness period – the sick pay period – the employer pays the sickness benefits. If the illness lasts longer than 14 days, the employer will notify the Social Insurance Agency of the illness. If the insured individual is a student or unemployed, the Social Insurance Agency enters from the second day of the sickness period.

The Social Insurance Agency assesses the entitlement to sickness benefits. However, the benefit claimant needs a medical doctor's certificate to verify reduced working capacity due to sickness. This certificate is claimed after seven days of sickness absence. The number of days compensated depends on the length of time the working capacity is expected to be reduced (written on the certificate). In general, an insured individual cannot be compensated for a longer period without regular renewal of the doctor's certificate. The average number of days between two such renewals is 28 days (standard deviation is 91). The level of compensation is about 80 per cent of the sickness benefit qualifying income per year up to a maximum of 7.5 times the price base amount.[44]

Sickness compensation or activity compensation is paid to individuals whose work capacity is permanently reduced by at least one-quarter for medical reasons: activity compensation for those aged 19–29 and sickness compensation for those aged 30–64. If a temporary sickness period develops into a permanent work incapacity, the period with sickness benefits could be prolonged with a period of sickness compensation. The sickness absence duration of interest in this study is the total period of continuous sickness absence, irrespective of benefit type.

In addition to providing financial insurance, the Social Insurance Agency has the responsibility for providing vocational rehabilitation in the case of sickness absence. For medical rehabilitation, the County Council (Landsting) has the responsibility. The national rules for how these responsibilities should be carried out are vague and, hence, there is large regional variation in how these responsibilities are shared between

[44] The price base amount in 2007 was 40,300 SEK (4200 euros).

the involved actors. For example, there are no nationally stated criteria for when an individual qualifies for sickness compensation.

Collaboration via Resursteam

In March 2004, the Social Insurance Agency in Uppsala and the Uppsala County Council made an agreement to introduce RT in all local care centres in Uppsala County.[45] The multidisciplinary RT (i.e., a medical doctor, a case worker, a behaviourist and a physiotherapist) should intervene early in the insured individual's sickness absence. By using the combined skills of the members of the team, the hope was to obtain a better understanding of the reasons for sickness absence. Based on this holistic understanding, RT should suggest suitable rehabilitation for the insured individual.

The two most common diagnoses of sick-absent individuals are related to musculoskeletal disorders and/or mental problems. By including a behaviourist and a physiotherapist in the team, the needs of these two groups in particular are supposed to be satisfied. The team meets regularly from the date an eligible insured individual becomes initiated into the programme. Recommendations are followed up and discussed until the case is closed.

Both the medical doctor and the case worker initiate eligible individuals into the programme. Out of all individuals initiated into RT, about half of them have been initiated by case workers and the other half by medical doctors.

The Social Insurance Agency in Uppsala has summed up crucial criteria for running the risk of becoming long-term sick-listed. The case workers were supposed to base their decision about eligibility for RT on these criteria. The criteria could be summed up as follows:

1. **Medical basis**: Mental diagnosis or musculoskeletal disorder without further specification but 100 per cent reduced work capacity and/or recommended sickness absence for a longer period than two months.
2. **Sick-spell history**: Several sick spells in the last years.
3. **Work**: Demanding work or unemployment.
4. **Family**: Sickness absence among other family members, divorced and/ or having responsibility for disabled children.
5. **Motivation**: No future plan to return to work.

[45] Since October 2006, the collaboration via RT has continued on a voluntary basis.

The case worker, basically, had access to register information about the first four risk factors. Since they did not, in general, meet the sick-listed individuals, case workers lacked information on the sick-listed individuals' motivation to return to work.

The medical doctors, in general, personally met their patients but, on the other hand, they probably had less information about the sickness absence among other family members.

Data

The population is based on the database LOUISE.[46] From this database we sample all individuals over the age of 15 years living in Uppsala County in 2004. LOUISE contains social economic information. To this data, we match (using personal identifiers) register information from the Swedish Social Insurance Agency on sickness absence spells. This data contains all days reported sick to the Swedish Social Insurance Agency between 1996 and 2004. In addition, this register contains sick-spell-specific information such as, for example, the individual's medical diagnosis[47] at the sick-spell start. Furthermore, we add information provided by the Social Insurance Office in Uppsala about RT participation and whether the individual participated in the experiment or not.

Since we have register information on sickness absence, we have complete sickness absence spells for all spells except for those that are censored, i.e. those that have not yet ended when data on sickness absence were collected (20 March 2007). In the experimental study, we simply compare the sickness durations of the treated group with the sickness durations of the control group. In the observational study, we have to pick out a proper sick spell for comparing individuals. The sick spells during which individuals received RT, all began after 1 January 2004. For comparison individuals, we use the first sick spell beginning after 1 January 2004.

We have multiple sick spells for most of the individuals in our data allowing for a within-individual analysis. As comparison sick-spell in the within-individual analysis, we use the first previous (to the one used in the cross-sectional analysis) sick spell starting later than 1 January 2001.

[46] LOUISE is a longitudinal register database on education, income and employment managed by Statistics Sweden.

[47] The diagnosis information we explore is an ICD code indicating the caption of the diagnosis. In this study, we distinguish a diagnosis related to mental problems or musculoskeletal disorder, among others.

Among the RT individuals, 806 out of 1079 had an earlier sick-spell period that started before 1 January 2001.

In the observational study, the population consists of 1089 individuals who received RT. Among those, 13 individuals are dropped because they were not eligible (not sick-listed) when entered as RT participants in the register. Thus, the total number of (sick-listed) RT individuals is 1076. The number of comparison individuals is 37,938.

In the experimental study, five individuals are missed; one treated individual did not live in Uppsala in November 2004 (when background variables were collected) and it turned out that four control individuals were not eligible (not sick-listed) at the time of randomization. The reason for this last sample selection is simply that the case workers[48] did not have access to updated information about sick spells when selecting candidates. The number of individuals who are possible to use for our experimental study is, hence, 45 (21 treated and 24 controls).

The sick-spell duration is created from two different registers: the sickness benefit register and the sickness compensation register. Since a sick spell with sickness benefits can be extended with sickness compensation, we simply add the sick spell with sickness compensation to the sick spell with sickness benefits *if* the former spell started the day after the latter ended. Of all sick spells used, 3 per cent are extended. Among the RT individuals, the corresponding number is 5 per cent.

The experimental study

The randomization started on 13 April 2006 and included all individuals who were initiated into RT by case workers. When a case worker had recorded an eligible for RT in the register, a random generator decided if the individual would be a treated (initiated into RT) or a control (not initiated into RT). It is important to note that the randomization was done within the administrative computer system and could hence not be manipulated by the case workers.

Descriptive statistics

Since the number of individuals in the experiment is few, the control group and the treatment group may differ with respect to important explanatory variables for the sick-spell duration. Table 1 presents descriptive statistics for the treated and controls. There are some potentially im-

[48] The experiment was only conducted on individuals assigned by case workers.

portant differences between the groups. For example, the education level is on average higher in the control group. Further, the control group includes fewer divorced people with several children, a larger share of individuals with mental diagnosis but fewer with musculoskeletal disorder. Finally, the control group includes two unemployed individuals, while the treated group has none. However, it is important to note that the average difference in mental diagnosis is the only statistically significant (at the 10 per cent level) difference between the groups.

The variables presented in Table 1 are important determinants of long-term sickness absence according to the Social Insurance Agency in Uppsala. Thus, simply compare sick-spell durations between the groups may be misleading.

Table 1: Descriptive statistics (mean and standard error (se)) for treated and controls

	Treated		**Controls**	
Variable	*Mean*	*Se*	*Mean*	*Se*
Female	0.76	0.10	0.71	0.09
Age	46.57	2.18	45.21	2.55
Immigrant	0.14	0.08	0.21	0.08
Post upper secondary school education	0.24	0.10	0.29	0.09
Married	0.43	0.11	0.54	0.10
Divorced	0.24	0.10	0.13	0.07
Number of children	1.00	0.23	1.50	0.24
Divorced *Number of children	0.38	0.21	0.13	0.09
Earned income (SEK per year)	173,133	296,36	165,400	229,63
Number of days with S.B.* between 1997 and 2004	218.05	95.39	96.04	47.66
Number of days with S.B. between 1997 and 2004 of other family members	17.90	9.05	47.79	37.74
Number of days with S.C.** between 1997 and 2004	828.29	455.03	235.21	235.21
Number of days with S.C. between 1997 and 2004 of other family members	0.00	0.00	374.67	346.82
Sick-spell-specific information:				
Mental diagnosis	0.29	0.10	0.54*	0.10
Musculoskeletal disorder	0.48	0.11	0.29	0.09
Number of days sick-listed at the first decision[3]	12.00	0.00	32.21	14.88

Table 1 Cont'd

	Treated		Controls	
Variable	*Mean*	*Se*	*Variable*	*Mean*
Sick-listed for more than 60 days at the first decision	0.00	0.00	0.08	0.06
Number of days sick-listed at the first decision if a mental diagnosis	3.43	1.21	26.71	15.25
Number of days sick-listed at the first decision if a musculoskeletal disorder	5.71	1.34	3.50	1.14
Unemployed when sick spell starts	0.00	0.00	0.08	0.06
Number of observations	21		24	

Notes: *S.B. is sickness benefit. **S.C. is sickness compensation and activity compensation. [3]The number of days with S.B. until the first renewal of the certificate verifying working incapacity is required. The difference between the groups is: *significant at 10%; **significant at 5%; ***significant at 1%.

Result

Figure 1 presents estimated survival functions (see Kaplan and Meier, 1958) for the treated and the control group, respectively. The figure shows that, after approximately 100 days from the time of the randomization, a larger fraction of the treated (solid line) remains sickness absent in comparison with controls (dashed line). Before 70 days after the time of the randomization, the relationship was nearer the opposite.

In order to obtain a point estimate of the effects of RT, we use Cox regression models (Cox, 1972). Further, this allows us to control for potential confounders presented in Table 1. In Table 2 we present the estimated effect of RT (using exact maximum likelihood estimation). Column 2 presents the estimated effect without controlling for potential confounders and column 3 presents the estimated effect when we control for a subset of the risk factors considered to be the most important for becoming long-term sick-listed.[49] We find a negative but not statistically significant

[49] We control for sickness absence history of the individual herself and of the family members, diagnosis and number of days sick-listed until the first renewal of the certificate verifying working incapacity. Since the number of observations is small, the efficiency will decrease with the number of added covariates if the contribution in explaining survival time is low.

effect of RT in both specifications.[50] The interpretation of the statistically insignificant effect is that the probability of leaving a sickness absence spell is reduced by 14 per cent if the individual is initiated into RT.[51]

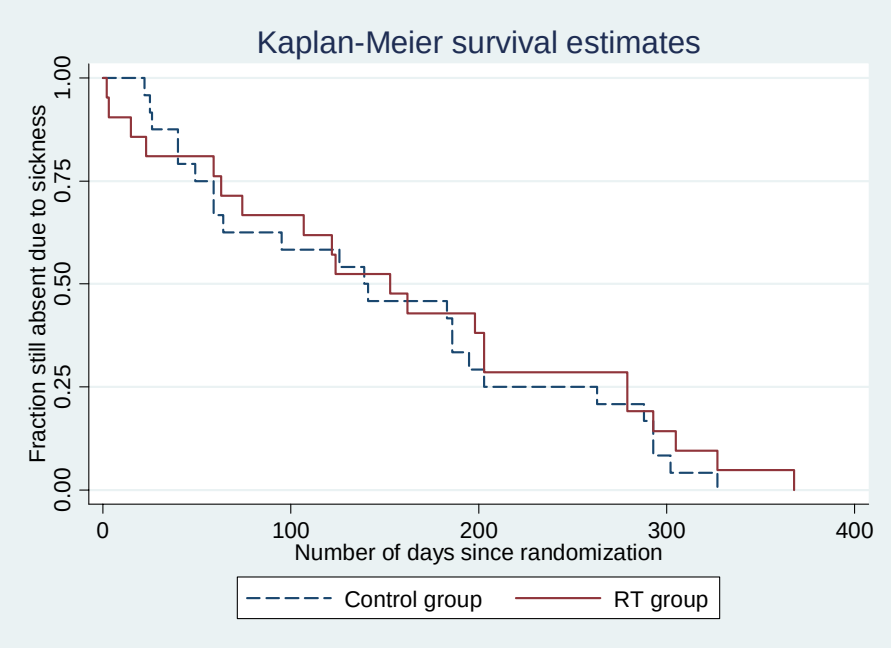


Figure 1: Fraction still absent due to sickness of the treated group (solid line) and the control group (dashed line).

Table 2: Results from the Cox regression estimation of the effects of RT of leaving a sick spell without covariate adjustments and with covariate adjustments

	Without covariates	**With covariates**
RT	-0.146	-0.148
Standard error	(0.331)	(0.483)
Observations	45	45

Notes: Standard errors in parentheses. *stat. significant at 10%; **stat. significant at 5%; ***stat. significant at 1%

[50] We have also estimated models with only one or two covariates. The negative and insignificant estimate is robust to various specifications.
[51] The percentage effect is obtained by 100*(exp(estimate)-1).

The observational study

The effect of RT is (again) estimated using Cox regression models. By using our detailed individual data, we control for the selection into RT. However, the effect of RT is measured with a time-varying step function. That is, we include a variable that is zero until the point of time when an individual enters RT. From this moment, it takes the value one. By estimating the effect of RT with a time-varying step function, we control for the fact that a prerequisite for being initiated into RT is that the individual still is reported sick.[52] It is worth noticing that the duration of a sick-spell may be correlated with individual attributes. Thus, by controlling for the duration to assignment, we also, at least partly, control for individual heterogeneity.

Figure 2 presents the development of the median number of days on sick leave of sick spells that started in a particular month in Uppsala County between 2004 and 2007.[53] The figure shows that we have some seasonal variation and a potential negative trend.[54] In order not to confound our estimations with any potential trend and/or seasonal variation, a flexible trend in sick-spell duration is allowed for by controlling for each month since 1 January 2001 (the first start of our sick-spell data).

[52] This type of selection is known as dynamic selection in the programme evaluation literature (a type of length bias sampling). See Fredriksson and Johansson (2008) for a thorough discussion of the problem.

[53] The sick-spell start is the start of the sickness benefit payments, i.e., on the 14th day of the sick leave.

[54] A caveat with Figure 1 is that the end of the study is 20 March 2007 and this biases the estimated trend downwards at least after 2006.

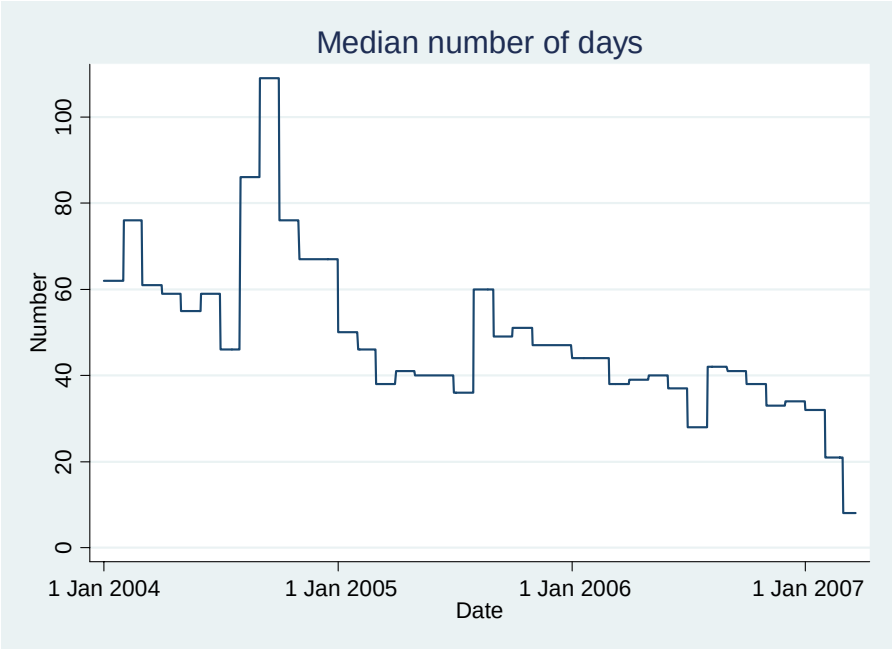


Figure 2: The median duration of sick spells started in Uppsala County between 2004 and 2007

Descriptive statistics

Table 3 presents descriptive statistics for the RT individuals and comparison individuals. This table shows that the individuals assigned to RT differ from the individuals not assigned to RT. The differences in averages are in line with what could be expected, given the eligible criteria for being initiated into the programme. The RT group includes a larger fraction of divorced individuals with many children and this group on average has a lower education level. The RT group also differs with respect to sickness absence history; on average, both the individual herself and the other family members have more days of sickness absence during the last seven years. Further, mental diagnosis and musculoskeletal disorder are both far more common among the RT spells as well as the number of days reported sick before a renewal certificate is requested. Thus, all these factors are important confounders to control for when estimating the effect of RT.

Table 3: Descriptive statistics (mean and standard deviations (St Dev)) for treated and comparison individuals

	Treated		**Comparison individuals**		
Individual-specific information	*Mean*	*St Dev*	*Mean*	*St Dev*	*Difference*
Female	0.65	0.48	0.61	0.49	0.04 ***
Age	47.14	10.75	46.01	12.12	1.12 ***
Immigrant	0.25	0.43	0.15	0.36	0.10 ***
Post upper secondary school education	0.21	0.41	0.28 ***	0.45	-0.07 ***
Married	0.49	0.50	0.46	0.50	0.02
Divorced	0.18	0.39	0.14 ***	0.34	0.05 ***
Number of children	0.92	1.16	0.76 ***	1.04	0.16 ***
Divorced * Number of children	0.12	0.49	0.07 ***	0.39	0.05 ***
Earned income (SEK per year)	1519.93	1182.38	1844.01 ***	1356.90	-324.08 ***
Number of days with S.B.* between 1997 and 2004	207.83	408.14	129.13 ***	331.19	78.69 ***
Number of days with S.B. between 1997 and 2004 of other family members	82.79	301.10	73.05	270.06	9.74

Table 3: Cont'd

	Treated		Comparison individuals		
Individual-specific information	*Mean*	*St Dev*	*Mean*	*St Dev*	*Difference*
Number of days with S.C.[**] between 1997 and 2004	329.61	2505.66	187.63***	1565.77	141.98***
Number of days with S.C. between 1997 and 2004 of other family members	119.65	847.39	87.82	877.25	31.83
Sick-spell-specific information					
Mental diagnosis	0.30	0.46	0.16***	0.37	0.13***
Musculoskeletal disorder	0.35	0.48	0.20***	0.40	0.15***
Number of days sick-listed at the first decision[3]	67.80	155.08	23.72***	67.06	44.08***
Sick-listed for more than 60 days at the first decision	0.16	0.37	0.04***	0.20	0.12***
Number of days sick-listed at the first decision if a mental diagnosis	21.31	91.97	5.53***	40.90	15.79***
Number of days sick-listed at the first decision if a musculoskeletal disorder	21.56	95.07	5.21***	34.77	16.35***
Unemployed when sick spell starts	0.17	0.38	0.11***	0.31	0.06***
Observations		1,076		37,938	

Notes: [*]S.B. is sickness benefit. [**]S.C. is sickness compensation or activity compensation (here treated as the same). [3]The number of days with S.B. until the first renewal of the certificate verifying working incapacity is required. The difference between the groups is: *significant at 10%; **significant at 5%; ***significant at 1%.

Estimation strategy

The model we estimate with Cox regression is the following:

$$h_i(t) = h_{0j}(t)\exp(\beta RT_i(t) + X_i'\gamma),$$

where *t* is the duration of sickness absence, *i* is an index for the individuals (both treated and controls), *j* is an index for the monthly calendar time, *j* = 1 in January 2001, $h_{0j}(t)$ is the base-line hazard at *t* for a sickness spell that started in month *j*, $RT_i(t)$ is the step function (the variable that takes the value one after being assigned to RT and zero for all control individuals and all RT individuals before this time period) and X_i represents the control variables (presented in Table 1). The effect of RT on the hazard of sickness absence is measured by β. The parameters are estimated with maximum likelihood.

The eligibility criterion for becoming initiated into RT is the risk (assessed by a medical doctor or a case worker) of becoming long-term sick-listed. To control for this selection of individuals, we include individual specific covariates but also sick-spell-specific information about, for example, diagnosis. These two types of variables are introduced sequentially in order to distinguish between potential individual-specific and sick-spell-specific selection.

Results

The results from the estimated Cox regression model are presented in Table 4. The table contains three different specifications. All estimations of the effect of becoming initiated into RT are negative and statistically significant. The interpretation is, thus, that RT prolongs rather than shortens the average length of sickness absence.

The first column (1) presents the result when we only control for any calendar time effects of the inflow of sick-spells (we stratify on month when the sick spell starts). In the second column (2), we have added individual-specific information. The estimate remains about the same, suggesting that individual-specific covariates are not that important for the result. The reasonable explanation is that the time-varying function (capturing dynamic selection), also controls for individual heterogeneity, since the pre-RT duration probably is correlated with relevant risk factors (like previous sickness etc.).

In the third column (3), sick-spell-specific information is added. The estimate is now somewhat lower. Thus, selection into RT is to some extent based on sick-spell-specific information.

The interpretation of the estimates is that the hazard of leaving a sickness absence spell is reduced by on average 22 per cent if an individual received RT.

Table 4: Estimated effects of RT on the hazard of leaving a sick spell using a stratified (on calendar time) Cox regression.

	(1)	**(2)**	**(3)**
RT step function	-0.325	-0.310	-0.247
	(0.032)***	(0.033)***	(0.033)***
Stratify on date (month and year) of spell start	Yes	Yes	Yes
Individual-specific control variables	No	Yes	Yes
Spell-specific control variables	No	No	Yes
Observations	39,025	39,025	39,025

Notes: Standard errors in parentheses. *stat. significant at 10%; **stat. significant at 5%; ***stat. significant at 1%.

Sensitivity analysis – observational study

In this section, we perform three different sensitivity analyses. The first analysis studies the problem with sample selection. In the second, we study heterogeneous treatment effects and finally we examine the functional form assumption used when estimating Cox regression models.

Sample selection

The identification is based on the assumption that we can control for the assignment to RT by using observed variables. This is quite a strong assumption. However, from a survey (see Anderzén *et al*., 2008) among the personnel who had experience of RT, we know that 75 per cent of the personnel had never experienced that a patient or insured individual refused to be assigned to RT and 67 per cent had never experienced that a patient or insured individual had proposed RT. Thus, self selection stemming from the sick-listed individual does not seem to be a problem. However, there could still be selection problems if the assignment is based on something that we do not observe in our data.

Within individual analysis

A within individual analysis allows us to control for potential individual-specific selection. In this analysis, we use the variation in spell length between two subsequent sick spells. In this approach, we cannot stratify on the inflow data; instead, we control for any trend in sick-spell duration by including dummy variables for the month and year the individual started the sick-spell. Further, notice that we partly control for the potential trend and seasonal variation in sick-spell duration by including spell-specific control variables in the analysis.

The results from this within analysis are presented in Table 7. The estimates of RT are still negative and larger in magnitude than before (the precision is, however, less good). The change in magnitude could be explained by the sample restriction.

Table 7: Estimated separate effects of RT on the hazard of leaving a sick spell using a stratified Cox regression estimator.

	Within analysis
RT	-0.522
	(0.127)***
Date (month and year) of spell start	Yes
Spell-specific control variables	Yes
Stratified on individuals	Yes
Observations	35,313

Notes: Standard errors in parentheses. *stat. significant at 10%; **stat. significant at 5%; ***stat. significant at 1%

Case-workers versus doctors

Another strategy to investigate whether the assignment is based on unobservable characteristics is to investigate if the estimated effects differ between individuals initiated by case workers and medical doctors. Remember that, in contrast to the case workers, the medical doctors base their decision on a personnel meeting with the patient. Thus, the risk of selection bias due to omitted variables is likely to be much smaller with respect to those initiated by case workers. In Table 5, we present descriptive statistics for the two groups assigned by case workers and medical doctors, respectively. The first impression is that the two groups are remarkable similar. However, the group assigned by the case workers has on average more days on sick-leave the last years. An explanation could be that case workers have access to this information from the registers. Another difference is that there are more individuals with a psychological diagnosis assigned by the doctors.

Table 5: Descriptive statistics (mean standard deviation (St. Dev.)) of the RT individuals subdivided into the groups assigned by case workers and medical doctors, respectively.

	Case workers		**Medical doctors**		
	Mean	*St Dev*	*Mean*	*St Dev*	*Difference*
Individual-specific information					
Female	0.66	0.47	0.64	0.48	0.02
Age	47.84	10.61	46.37	10.86	1.47**
Immigrant	0.23	0.42	0.28	0.45	-0.05*
Post upper secondary school education	0.22	0.41	0.20	0.40	0.02
Married	0.49	0.50	0.48	0.50	0.01
Divorced	0.18	0.39	0.19	0.39	0.01
Number of children	0.95	1.21	0.88	1.11	0.07
Divorced * Number of children	0.12	0.49	0.12	0.49	0.00
Earned income (SEK per year)	1595.88	1204.49	1425.45	1145.57	170.43**
Number of days with S.B.* between 1997 and 2004	214.30	408.35	198.86	408.20	15.44
Number of days with S.B. between 1997 and 2004 of other family members	83.99	279.22	81.73	327.93	2.26
Number of days with S.C.** between 1997 and 2004	350.15	1969.68	311.08	3047.07	39.07
Number of days with S.C. between 1997 and 2004 of other family members	110.48	817.15	119.29	840.64	-8.81
Sick-spell-specific information					
Mental diagnosis	0.32	0.47	0.27	0.44	0.05*
Musculoskeletal disorder	0.35	0.48	0.35	0.48	0.00
Number of days sick-listed at the first decision[3]	59.79	137.52	78.40	174.56	18.61

Table 5 Cont'd

	Case workers		Medical doctors		
	Mean	*St Dev*	*Mean*	*St Dev*	*Difference*
Sick-listed for more than 60 days at the first decision	0.15	0.36	0.18	0.39	0.03*
Number of days sick-listed at the first decision if a mental diagnosis	18.75	74.77	24.72	109.76	-5.97
Number of days sick-listed at the first decision if a musculoskeletal disorder	16.44	75.75	27.94	114.50	-11.5
Unemployed when sick spell starts	0.16	0.37	0.19	0.39	-0.03
Observations	582		485		

Notes: [*]S.B. is sickness benefit. [**]S.C. is sickness compensation or activity compensation (here treated as the same). [3]The number of days with S.B. until the first renewal of the certificate verifying working incapacity is required. The difference between the groups is: *significant at 10%; **significant at 5%; ***significant at 1%. In 9 cases, the information about by whom the individual is initiated is lacking, explaining why the total of individuals in this table is 1067.

In Table 6, we present the estimated effects of RT on the hazard of leaving a sick spell for those initiated by the medical doctors (column 1) and case workers (column 2), separately. As control variables, we include both individual-specific and spell-specific information and we stratify on calendar time.

The estimated effect of RT is negative and statistically significant for both groups. Compared with the estimate based on all individuals, the effect is somewhat less negative for those initiated by case workers. However, important to remember is that the effect is still significantly negative, also for those initiated by case-workers. The difference in effects between the two groups is .12. This difference is not statistically significant at the 5 per cent level but statistically significant at the 10 per cent level. We cannot exclude that there may be selection problems associated with those assigned by a medical doctor. Another explanation could be heterogeneous effects depending on diagnoses. Remember that mental diagnoses were more common among those initiated by case-workers. This issue is investigated in the next section.

Table 6: Estimated separate effects of RT on the hazard of leaving a sick spell for those initiated by medical doctors and case workers, respectively, using a stratified (on calendar time) Cox regression estimator

	(1)	**(2)**
	Medical doctors	*Case workers*
RT	-0.292	-0.173
	(0.047)***	(0.044)***
Stratify on date (month and year) of spell start	Yes	Yes
Individual-specific control variables	Yes	Yes
Spell-specific control variables	Yes	Yes
Observations	39,023	39,023

Notes: Standard errors in parentheses. *stat. significant at 10%; **stat. significant at 5%; ***stat. significant at 1%.

Heterogeneous treatment effects

In the previous section, we found that the treatment effect was smaller for the individuals assigned by the medical doctors than for individuals assigned by the case workers. This difference may be an effect of chance or selection on unobservable information, but may also be that these two groups respond differently to RT, i.e., heterogeneous treatment effects. If the heterogeneous effects depend on, for us, observed variables and if there are differences with respect to these observed variables between these two groups, it is possible to test for heterogeneous treatment effects.

By studying Table 5, we can see that there are some observed differences between the two groups of individuals. For example, the fraction with a mental diagnosis is significantly larger in the group assigned by the case workers. This difference is 5 percentage points and is highly statistically significant. Hence, if the effect of RT depends on diagnosis, then the difference between the estimated effects of being assigned by medical doctors and case workers may be due to this difference. To test this hypothesis, we estimate two Cox regression models. First, we estimate separate effects for individuals with a musculoskeletal diagnosis and a diagnosis related to mental problems (presented in column (1) in Table 8). Second (presented in column (2) in Table 8), we allow for heterogeneity in these effects depending on whether the individual has been initiated by a medical doctor or a case worker.

The most striking result is that there seems to be no negative effects for the patients with a mental diagnosis. When controlling for the diagnosis, the difference in estimates between those initiated by medical doctors and case workers is reduced to -.10 (previously -.12) and it is far from statistically significant. Thus, we conclude that the difference in effects

between those initiated by medical doctors and case workers is simply a random event.

Table 8: Estimated effects of RT: (1) depending on type of diagnosis and (2) subdivided into those assigned by medical doctors (MD) and case workers, respectively.

	(1)	**(2)**
RT	-0.380	-0.329
	(0.053)***	(0.073)***
RT × musculoskeletal	0.039	0.016
diagnosis	(0.075)	(0.102)
RT × mental diagnosis	0.463	0.465
	(0.078)***	(0.104)***
RT × MD		-0.101
		(0.104)
RT × musculoskeletal		0.044
diagnosis × MD		(0.147)
RT × mental diagnosis		-0.020
diagnosis × MD		
		(0.154)
Individual-specific control variables	Yes	Yes
Spell-specific control variables	Yes	Yes
Observations	39,023	39,023

Notes: Standard errors in parentheses. *stat. significant at 10%; **stat. significant at 5%; ***stat. significant at 1%..Estimation is performed using a stratified partial maximum likelihood estimator.

Functional form

We have used Cox regression to remove the dynamic selection problem and observed covariates to control for the remaining potential selection problems. The Cox regression models build on an assumed functional form of the hazard rates from a sickness-absence spell. One may be concerned that this functional form is wrong, implying a biased estimate of the treatment effect. In order to test whether the assumed functional form is correct, we have used the non-parametric matching estimators suggested by Fredriksson and Johansson (2008) and further developed by De Luna and Johansson (2007). The real advantage with this estimator is that one can allow for non-monotonous treatment effects. That is, for example, allowing the estimate first to be a negative effect (for example, due to an initial locking-in effect when waiting for rehabilitation) and, in a second stage, a positive effect (on the outflow by the health improvement

from the rehabilitation).[55] The results from the non-parametric matching estimation concur with the results using the Cox regression model: The time in sickness absence is prolonged by about 15 per cent if assigned to RT and the effect from RT is monotonous, i.e. the effects on the hazard are about the same irrespective of when in the sickness absence duration the effect is evaluated.

Concluding discussion

This paper contributes to the literature about the effects of collaboration in multidisciplinary teams on preventing long-term sickness. The results from the experiment and the retrospective observational study suggest a negative effect of RT; on average, RT prolongs the sick-spell duration rather than shortens it. The interpretation of this result is that an individual prolongs her duration by 22 per cent or by approximately 60 days.[56] The negative result of RT is robust to several sensitivity controls.

The effect is heterogeneous depending on diagnosis of the sick-listed individual. The two most common diagnoses in the target group are related to mental problems and musculoskeletal disorders. When estimating separate effects for these two groups, it turned out that there is no effect for the group of sick-listed individuals with a diagnosis related to mental problems. For the other group (with a musculoskeletal disorder) the effect is about the same magnitude as the average effect. This result is surprising since the aim of RT is to assist individuals with mental problems (as well as musculoskeletal disorders), in particular.

In a Swedish report, Anderzén *et al.* (2008), we present the results from a survey among the personnel with experience of RT and among the sick-listed individuals in the experiment. The general finding from the survey among the personnel is that they were rather positive towards the collaboration: 60 per cent of the case workers and 44 per cent of the medical doctors thought that RT facilitated their daily work "rather" or "very much" (instead of "rather little" or "not at all"). A majority (among the personnel) believed that RT speeded up the individual's return to work (only 7 per cent did *not* believe in a positive effect). These results are in line with the general conclusion from several similar surveys

[55] For information on estimation and the results, see Anderzén *et al.* (2008). The reason for not presenting the results in this paper is that the details on the matching estimator are quite extensive, implying a much longer paper without increasing the subject knowledge.

[56] The calculation of the effects on days is done under the assumption of a constant hazard rate. The mean sickness-absence duration in the study population is 272 days. Thus: $0.22*272 \approx 60$.

among the personnel participating in different collaboration programmes (for example, Hultberg, 2005; Dowling *et al.*, 2004; Schmitt, 2001; Socialstyrelsen, 2001). In the case of RT, it is astonishing that the personnel were rather positive towards the programme since the effect on sickness absence was strongly negative. This observation stresses the difficulty of evaluating the effect of a programme by using survey information from the collaborating personnel. However, the survey information can help us *explain* the negative result of RT.

A result from the survey among the sick-listed is that some individuals had been waiting for treatments suggested by the RT. Thus, inefficiency in the organization outside RT could be a reason for an inefficient rehabilitation process.

Another explanation is moral hazard problems in the Swedish sickness insurance (see e.g. Johansson and Palme, 2005 and Hesselius *et al.*, 2005). It is possible that the health eligibility criterion to be on sickness benefits was not reviewed as hard for the RT group as for the comparison group. The reason is as follows.

The target group for RT was individuals with health problems that are difficult to assess (musculoskeletal disorders or psychological problems without a further specification). Thus, it might have been difficult, for both the doctor and the case worker, to decide whether the sick-listed was qualified for further sickness benefits or not.[57] The decision of not allowing an individual to be on sickness benefits is difficult to take and takes time from the doctor's main activity.[58] It is also most likely a difficult decision for the case worker to inform both the doctor and the sick-listed individual that the certificate is not valid.

These observations are in line with the hypothesis that when an individual was initiated into RT, the health problems became confirmed and the health eligibility criterion to be on sickness benefits might not have been reviewed in the same way as for the comparison group. From our survey, we also found that the personnel were positive towards RT since it made "the daily work easier" (see Anderzén *et al.*, 2008). Whether the individual *should* have been sick-listed or not from a health, social or economic perspective is a normative question beyond the scope of this study. The point here is that RT may have been an "easy solution" for the

[57] The doctor writes the certificate that proves reduced working capacity due to sickness. However, the case-worker judges the certificate and decides whether sickness benefits should be allowed or not. It is rare that the case-worker does not follow the doctor's recommendations.

[58] It has been documented that doctors rarely object against an individual who would like to be sick-listed even though they believe that this individual has health that does not prohibit work (see e.g., Arrelöv, Edlund and Goine, 2006; Englund, 2001).

actors involved: the sick-listed individual *and* the medical doctor *and/or* the case worker.

A final comment is that collaboration between different actors in the rehabilitation process seems necessary. The question is how this collaboration should be designed and organized. Up to the present, there is little empirical evidence about the effects of different collaboration solutions on sickness absence. The design of the collaboration evaluated in this paper, RT, is in accordance with some of the recommendations by the Swedish National Board of Health and Welfare (Socialstyrelsen, 2005) and the Swedish Council on Technology Assessment in Health Care (SBU-rapport, 2003). The result in this study does not support these recommendations. However, it is important to keep in mind that each collaboration programme has its own singularities, which motivate further research about the effects of different collaboration programmes.

References

Arrelöv, B., Edlund, C. and Goine, H.: 2006, Grindvakterna och sjukförsäkringen – samspel och motspel i SKA Projektet: Sjukförsäkring, kulturer och attityder. Edward Palmer (ed.), Analyserar 2006:16, Stockholm: Försäkringskassan.

Anderzén, I., Demmelmaier, I., Hansson, A-S., Johansson, P., Lindahl, E. and Winblad, U.: 2008, Samverkan i Resursteam: effekter på organisation, hälsa och sjukskrivning. IFAU rapport 2008:8.

Cox, D.R.: 1972, Regression models and life tables, Journal of the Royal Statistical Society (34), 187–220.

Danemark, B. and Kullberg, C.: 1999, Samverkan. Välfärdsstatens nya arbetsform. Lund: Studentlitteratur.

De Luna, X. and Johansson, P.: 2007, Matching estimators for the effect of a treatment on survival times. IFAU Working Paper 2007:1.

Dowling, B., Powell, M. and Glendinning, C.: 2004, Conceptualising successful partnership, Health and Social Care in the Community, 12(4): 309–17.

Englund, L.: 2001, Förändringar i distriktsläkarnas sjukskrivningspraxis mellan åren 1996 och 2001 i ett svenskt landsting. Falun: Centrum för Klinisk Forskning Dalarna.

Fredriksson, P. and Johansson, P.: 2008, Dynamic treatment assignment – the consequences for evaluations using observational data, Forthcoming in Journal of Business and Economic and Statistics.

Försäkringskassan: 2007a, The scope and financing of social insurance in Sweden 2005–2000, Försäkringsdivisionen Utvärderingsavdelningen.

Försäkringskassan: 2007b, De gemensamma metoderna i sjukförsäkringen – hur blev det?, Försäkringskassan redovisar 2007:8.

Hesselius, P., Johansson, P. and Larsson, L., 2005, Monitoring sickness insurance claimants: evidence from a social experiment. IFAU 2005:15.

Hultberg, E-L.: 2005, Co-financed collaboration between welfare services – effects on personnel and patients with musculoskeletal disorders, Göteborgs universitet.

Johansson, P. and Palme, M.: 2005, Moral hazard and sickness insurance, Journal of Public Economics (89), 1879–1890.

Kaplan, E.L. and Meier, P.: 1958, Nonparametric estimation from incomplete observations, Journal of American Statistics Association (53), 457–481.

Kärrholm, J.: 2007, Co-operation among rehabilitation actors for return to work life, Thesis for doctoral degree 2007, Department of Public Health Sciences, Division of Rehabilitation Medicine, Karolinska Institutet, Stockholm, Sweden.

Nyman, K., Palmer, E. and Bergendorff, S.: 2002, Den svenska sjukan – sjukfrånvaron i åtta länder, Expertgruppen för studier i offentlig ekonomi (ESO), Finansdepartementet, Ds 2002:49.

SBU-rapport, 2003, Sjukskrivning – orsaker, konsekvenser och praxis – en systematisk litteraturöversikt, Stockholm: Statens beredning för medicinsk utvärdering.

Schmitt, M.H.: 2001, Collaboration improves the quality of care: methodological challenges and evidence from US health care research, Journal of Interprofessional Care, (15), 47–66.

Socialstyrelsen: 2001, Socsam. Ett försök med politisk och finansiell samordning – en slutrapport. Finansiell samordning 2001:1: Socialstyrelsen, Riksförsäkringsverket.

Socialstyrelsen: 2000a, Ekonomisk analys av Frisam – en studie i fyra lokala områden, Socialstyrelsen, 2000:7, Kommunforskning i Västsverige.

Socialstyrelsen: 2000b, Identifiering av hinder och framgångsfaktorer för samverkan, Socialstyrelsen, 2000:6.

Socialstyrelsen: 2005, Sjukskrivningsprocessen i primärvården – återföring av tillsynsbesök 2004 Stockholm, Socialstyrelsen: nr 2005-109-2.

Statskontoret: 2006, Fortsättning med Finsam – målgrupper, insatser och arbetsformer, Statskontoret 2006:6, bilaga 3.

Essay 5: Better under threat?♦

Introduction

Health status is imperfectly observed by the provider of sickness insurance implying an asymmetric information situation. Thus, the extent of monitoring and sanctions as well as norms is probably important for understanding sickness absence behaviour.

Today, there is quite a large body of empirical studies addressing *ex post* moral hazard[59] in sickness insurance. The evidence from these studies suggests that economic incentives are important when the replacement rate is high.[60] The evidence of the importance of monitoring and norms is still rather limited. The objective of this study is to add empirical evidence of the two latter. In particular, we evaluate a policy with the explicit aim to affect norms about the usage of the sickness insurance. The empirical analysis is based on data from a randomized experiment that we have conducted at the local social insurance office in Uppsala, Sweden.

The policy, called information meetings (IM), implied that sick-listed individuals were called to a meeting about the rights and duties associated with the sickness insurance (SI). Attendance at the meeting was mandatory; the sick-listed individual had to report back or show up at the meet-

♦ Co-authored with Per Johansson. The authors thank Kerstin Brindbergs, Kertie Lindh, Elisabet Högsten, Tove Meissner and other collaborators at Försäkringskassan in Uppsala who have helped us to conduct the experiment. The authors also thank Andreas Westermark and other seminar participants at Uppsala University for valuable comments. The financial support from the Swedish council for working life and social research FAS (dnr 2004-2005) is acknowledged.

[59] That is, the insured is tempted to claim more sickness cash benefit than he/she would if exposed to the full cost.

[60] See e.g. Allen (1981), Johansson and Palme (1996, 2002 and 2005), Henreksson and Persson (2004) and Barmby et al. (2002).

ing. The target group was sick-listed individuals without permanent employment.

The aim of this policy was to affect norms and thereby reduce excess use of the sickness insurance. However, in addition to any potential effect on norms induced by attending the meeting, the call could be seen as a "threat".[61] In this context, the sickness benefit payment was tied to participation in the information meeting. The obligation to participate might have induced a "threat" leading to less usage of the sickness insurance.

The main focus of this study is to estimate the combined effect of a potential threat and/or a potential change in norms after attending the meeting, i.e., an intention to treat (ITT) parameter. From a behavioural point of view, it would be interesting to disentangle the two effects but that is not the purpose of this paper. However, the combined effect is policy relevant: can the use of the sickness insurance be reduced by calling individuals to IM?

The design of the experiment is as follows. Case-workers each month call (via letters) all individuals who flow into sick-listing. When the case-worker has selected eligible individuals for IM, half of them receive a call as planned and the calls to the other half are postponed until the next occasion, about one month later. The division into calls sent as planned and calls postponed is made by a random generator. Thus, the identification is based on a randomized displacement of *when* an individual received the call; we analyse if the group that received the call early (in their sickness spell) left their sickness period faster than those who received it later. With this experimental design, the case-workers could continue to work as usual once the experiment was in place. This fact has two important implications. First, it was easy to perform the experiment.[62] Second, this design does not imply heavy ethical considerations.[63] A further advantage with this experiment is that the participants have not been informed about the study, implying that we do not have to worry about potential Hawthorn effects, i.e., that the participants change their behaviour due to the fact that they know that they are participating in an experiment. On the negative side, since all individuals eventually receive

[61] This term is used in the context of tying benefit payments to labour market programmes; compulsory programme participation motivates individuals to leave the unemployment insurance (Black *et al*., 2003; Geerdsen, 2006).

[62] In social sciences, experimental studies are rare. One reason is that they are in general associated with practical inconveniences and high organizational costs. With this design, practically no extra costs were imposed on the case-workers.

[63] Unequal treatment is an argument often used not to implement an experiment. With this design, all individuals finally received the same treatment.

the treatment, we can only identify a lower bound of the total effect, i.e., the effect of being called early in comparison with later on in a sick-spell.

The result suggests a negative effect on the duration of sickness absence: Receiving a call to IM early in contrast to one month later reduces the duration on average by 23 per cent. This effect provides a lower bound of an effect of receiving the call against not receiving it. A low attendance at the meeting (30 per cent) suggests that a significant part of the effect stems from the "threat".

Sickness absence and institutions

The development

In Western Europe, Sweden and Norway are the countries with the most costly public sickness insurance schemes relative to GDP. One explanation is that Sweden has a large proportion of older women in the workforce (Nyman *et al.*, 2002).

Another characteristic for Sweden is that the sickness absence has varied significantly over time. During the late 1990s, the sickness absence increased dramatically in Sweden compared with most other European countries. However, during the last years, this trend has stopped. Relative to GDP, social insurance payments fell for the third consecutive year in 2006 (Försäkringskassan, 2007). The reason for this change is difficult to grasp. However, this change occurred at the same point of time that a number of different programmes were launched by the Swedish Social Insurance Agency in order to reduce the sickness absence.[64]

The sickness insurance

The aim of the SI in Sweden is to provide the population with protection against the financial risk associated with illness and disability. All those who have reached the age of 16 and are resident in Sweden are in principle insured and registered with the Swedish Social Insurance Agency (SSIA).

There are two main benefits in the Swedish SI: *sickness compensation* or *activity compensation* and *sickness cash benefits*. Sickness compensation compensates individuals whose work capacity is permanently reduced. Sickness cash benefit replaces part of the income loss during tem-

[64] See Anderzén *et al.* (2007) for an overview.

porary illness. In this study, the focus is on sick-listed individuals with sickness cash benefits.

The first day of sickness is uncompensated. The benefit from the second day and forward is about 80 per cent of the sickness cash benefit qualifying income to a maximum of 10 times the price base amount.[65] The benefits for unemployed insured people are based on the wage before unemployment. Thus, unemployed persons without any employment history do not receive any sickness cash benefit.

The employer pays the sick pay for the first 14 days of the illness period. If the illness lasts longer, the SSIA pays the sickness cash benefit. If the insured is a student or unemployed, the SSIA enters from the second day of the sickness period.

Within a week, at the very latest on the eighth day of the sickness spell, the claimant must verify eligibility by showing a doctor's certificate that proves reduced working capacity due to sickness. The public insurance office then judges the certificate and decides upon further sickness benefit. It is very rare that the certificate is *not* approved. The certificate contains an expectation of the length of reduced working capacity. In general, a sick-listed individual needs to renew the certificate regularly in order to prove continued reduced working capacity.

When an individual has been granted sickness cash benefit, a rehabilitation plan is made. The medical doctor has the responsibility for the medical rehabilitation. The employer has the responsibility for rehabilitation associated with the workplace and the work conditions. If the insured is unemployed, the SSIA has the responsibility for the non-medical rehabilitation. The SSIA co-ordinates the different rehabilitation plans.

The UI versus the SI

In this study, the focus is on sick-listed individuals without permanent employment, implying that most of the individuals are unemployed, temporarily employed or involuntarily part-time working. Thus, it is relevant to consider the economic incentives in the unemployment insurance (UI) and how these differ from the ones in the SI.[66]

The benefit cap within the SI is higher than the benefit cap within the UI. Briefly, the SI compensates (at most) 80 per cent of a qualifying (monthly) income of SEK 33,083. The UI compensates (at most) 80 per cent of a qualifying (monthly) income of SEK 20,075 during the first 100

[65] The price base amount in the year 2007 is 40,300 SEK .
[66] This section is based on SOU (2007), in which further details about the interplay between the sickness and unemployment insurance systems can be found.

days and thereafter SEK 18,700 per month. Furthermore, there is a difference in how the benefit qualifying income is calculated. Within the UI system, the benefit level is based on historic income levels whereas the benefit level within the SI is based on expected (or actual) income levels. Further, in contrast to the SI, the UI does not follow any index; instead, the cap is decided *ad hoc* by the government. Finally, there are differences in the maximum duration of compensation: unemployment benefits are limited to 300 working days, whereas sickness benefits (payment or compensation) have no time limit.

The Information meetings

IM have been held between 2005 and 2006 by the local social insurance office in Uppsala, which we for simplicity denote the office in the following. To each meeting, 20–30 sick-listed individuals were called by a letter sent out about two weeks ahead. In this letter, it was clear that the meeting implied *information* only. The letter was short but it was mentioned that the information was about *rehabilitation* and *rights and duties* of the sickness benefit claimant, potential employers and the Office. It was also stated that the meeting would last for about an hour and that participation was compulsory – the sick-listed individuals had to call back if he/she was not able to attend. There was no "threat" about withdrawing sickness benefits in the call, but sick-listed individuals who did not show up at the meeting or present a valid reason for not coming were contacted again.[67]

The Office in Uppsala is divided into different divisions. For example, there are special divisions for the insured who are public employed, private employed and unemployed. Sub-areas, outside the city of Uppsala, also have their special divisions for the insured living in the particular area. The experiment was conducted at the division for the unemployed – the division administrating those without permanent employment – and also in a small sub-area: Gimo.

Division for the unemployed

An individual belongs to the division for the unemployed if the office lacks information about an employer with rehabilitation responsibility. The office obtains this information from the insured's application for

[67] According to the case-workers who conducted the programme, they had the authority to withdraw the sickness cash benefit if the sick-listed individual did not co-operate. However, we do not know if that was put into practice or not.

sickness benefit. This application is renewed regularly during a longer period of sickness absence. If an individual at the beginning of a sickness period is employed but becomes unemployed during the sickness period, he/she is transferred to the division for the unemployed. Another reason for transferring an individual to the division for the unemployed is the employer not taking responsibility for rehabilitation, although it is against the rules. Between January 2005 and December 2006, the *inflow* of all sick-listed individuals *to* the division for the unemployed was called to IM. Exceptions were if the case-worker had information indicating that it would not be appropriate to call (for example if the individual was in hospital). The calls and the meetings were held about once a month. To each meeting around 30 individuals were called. In total, around 400 individuals per year have been called by this division.

Division Gimo

The division Gimo administrates all kinds of sick-listed individuals. IM were held during 2006. Eligible candidates were those whose sickness benefit qualifications were unclear. In practice, many of them lack permanent employment. The procedure for calling as well as how the meetings were organized were similar to the division for the unemployed. [68]

The experiment

When a case-worker had recorded an eligible individual in a digital register, a random generator decided whether the individual should be called at once or whether the call should be postponed until the next occurrence, about 30 days later. The random generator was built into the computer and the two outcomes (called immediately or postponed) had equal probability. This procedure was repeated every time an IM was planned to be held. Thus, at the next occasion, the case-worker called half of all the new candidates and all of the individuals whose calls had been postponed on the previous occasion. In total, the experiment included eleven different randomizations: eight at the division for the unemployed and three at the local office in Gimo.

It is important to note that the randomization creates an exogenous variation in *when* in a sickness absence spell a sick-listed individual received the call and *not if* the individuals received the call. We will ana-

[68] Excluding Gimo from the analysis does not change the main conclusions.

lyse if the group that received the call early (in their sickness spells) left their sickness period faster than those who received it later.

Data

The experimental data were collected between May and December 2006 by the Office in Uppsala. These data include detailed information on the experiment: the date for each randomization, when the call was sent out, when the meeting was held, if the insured individual participated in the meeting or not etc. In this study we focus on the date when an individual was actually called.

We have matched (via a personal identification number) the experimental data with a register, provided by the National Social Insurance Agency, about sickness absence: the start and end of individual sick spells. From this register, we have picked out the sick spell during which a case-worker identified the individual as eligible for the first time, i.e., the sick spell during which the randomization took place. A prerequisite is of course that the individual was on sick leave at this moment. The experimental data include 352 individuals. Out of them, 275 were on sick leave when the first randomization took place.[69] The reason why 20 per cent are dropped is that case-workers did not have access to updated information about sick-spell ends when picking out eligible candidates. Thus, a significant part of all the individuals called was in practice not eligible. The following analyses are based on the reduced sample of 275 individuals.

In order to describe the study population and to check the robustness of our results, we have added detailed register information on social background and the sick spell (for example diagnosis at the sick-spell start).

Study population

Table 1 presents descriptive statistics on the individuals who have been called and, as a comparison, all other sick spells in Uppsala County during the same period: between June 2006 and December 2006. As the table shows, our study population differs in several aspects from the average

[69] One individual is excluded since he/she has been on sick leave since 2003. It was noticed in the register that this individual had been picked from the stock and should, hence, not have been called.

insured individual who reported sick during the study period. Since we are studying individuals who are weakly established on the labour market, it is not surprising that the study population consists of more immigrants, less educated individuals and individuals with a lower average income. The study population also consists of somewhat fewer women, the average age is lower and fewer are married. Finally, among all the sick-listed individuals, the two most common reasons for sickness absence are related to mental problems or musculoskeletal disorder (the diagnosis at the time of inflow into sickness). In the study population, the fraction with a mental diagnosis is larger, compared with the corresponding share among all the sick-listed individuals.

Table 1: Descriptive statistics of the study population and other individuals on sick leave in Uppsala County during the study period

	Experimental data		**All sick spells**	
Variable	*Mean*	*St Dev*	*Mean*	*St Dev*
Female	0.53***	0.50	0.62	0.49
Age	44.21**	11.54	46.53	11.73
Immigrant	0.27***	0.44	0.17	0.38
Post upper sec. education	0.20**	0.40	0.27	0.44
Married	0.37***	0.48	0.46	0.50
Divorced	0.16	0.37	0.16	0.36
Number of children	0.67	1.01	0.75	1.05
Earned income (SEK 100,000 per year)	1.36***	2.41	1.68	1.35
Mental diagnosis	0.30***	0.46	0.20	0.40
Musculoskeletal disorder	0.23	0.42	0.24	0.43
Observation	275		18 001	

Notes: * stat. sign at 10 per cent level, ** stat. sign at 5 per cent level, significance levels based on t-test.

A look at the experimental data

Before analysing the experimental data, we present, in Table 2, descriptive statistics of the background variables for the treated and the controls, i.e., those with extended waiting time for treatment. As expected, there is no statistically significant difference between the groups with respect to any background variable.

Table 3 presents descriptive statistics on the variable of interest: the number of days on sick leave before and after the randomization. The difference is statistically insignificant *before* the randomization but not *after*.

Before randomization individuals in both groups have been on sick leave for, on average, about 100 days. The reason for this quite long duration is that the randomizations took place when a sick-listed individual flowed into the division for the unemployed (or was picked from the stock in Gimo). It is also striking that the variation in the number of days on sick leave before randomization is large. Descriptive statistics for each randomization separately show that this variation is rather constant across randomizations and hence is not explained by seasonal variation. However, since the number of days on sick leave before randomization is probably important for the outcome of interest, we address this issue by controlling for the different randomizations in the formal analysis.

After randomization, control individuals (with extended waiting for treatment) continue to be on sick leave for on average 18 days longer than those treated immediately. The difference in median between the groups is 20. This number indicates an effect of IM on sickness absence: the probability to leave a sick spell increases when an individual is called to an IM. However, this number should be interpreted with some caution since (i) the durations are right censored and (ii) the experiment consists of eleven different randomizations with potentially different distributions of individuals at each randomization.

Table 2: Descriptive statistics of treated and controls (with the extended waiting)

	Treated			**Controls (extended waiting)**			
Variable	*Mean*	*Q50*	*Se*	*Mean*	*Q50*	*Se*	*t-test*
Female	0.55	1.00	0.50	0.51	1.00	0.50	-0.52
Age	45.57	46.00	11.85	44.84	45.00	11.24	-0.52
Immigrant	0.28	0.00	0.45	0.25	0.00	0.44	-0.43
Post upper sec. education	0.20	0.00	0.40	0.21	0.00	0.41	0.21
Married	0.40	0.00	0.49	0.34	0.00	0.48	-1.04
Divorced	0.18	0.00	0.38	0.14	0.00	0.35	-0.80
Number of children	0.65	0.00	1.01	0.69	0.00	1.01	0.40
Earned income (100,000 SEK/year)	1.21	0.87	1.20	1.51	1.03	3.29	1.02
Mental diagnosis	0.27	0.00	0.45	0.33	0.00	0.47	1.06
Musculoskeletal disorder	0.24	0.00	0.43	0.21	0.00	0.41	-0.64
Observations		141			134		

Notes: t-test is with respect to differences in means * stat. sign at 10 per cent level, ** stat. sign at 5 per cent level.

Table 3: Descriptive statistics on the number of days on sick leave before and after randomization in the respective group

	Treated		**Controls**		**Equality of**	
					means	**medians**
Days on sick leave:	*Mean*	*Q50*	*Mean*	*Q50*	*p-value*	*p-value*
Before randomization	104.30 (85.08)	77.00	111.82 (96.34)	73.00	0.687	0.672
After randomization	134.53 (84.01)	118.00	152.16 (91.02)	138.00	0.048	0.051

Notes: Equality of means is tested with standard t-tests (standard errors are displayed within parenthesis). A Pearson chi-squared test is performed for the equality of the medians.

Identification and estimation

The experiment randomly divides the study population into two populations: one that is treated immediately ($T = 1$) and the other that is treated with a delay ($T = 0$). Normally, the untreated population is used to estimate the counterfactual hazard or the survival function (i.e., the hazard or the survival function the treated population would have had in the absence of the call to IM). In our study, the initially untreated individuals are treated eventually. A traditional control group only exists for about 30 days, i.e., the time period between the randomization and the occasion for the subsequent IM. Thus, we cannot estimate the counterfactual survival function for longer durations than 30 days for the initially treated. If there is an effect of IM, it is not reasonable to believe that it is realized within 30 days. The reason is the procedure of certificates proving reduced working capacity. These certificates need to be renewed after a certain number of days and a potential effect is probably not realized before such a renewal.[70] In this context, we estimate survival functions for the two groups ($T = 1$ and $T = 0$) and calculate the difference between them, i.e., we estimate a lower bound of an effect of receiving the call against not receiving it. The survival functions for the immediately treated and the delayed are given as:

[70] The individual has the option to end the sickness absence spell early but this happens very seldom in practice.

$$S(t \mid T=1) = \prod_{s=1}^{t} (1 - h(s \mid T=1)) \text{ and}$$

$$S(t \mid T=0) = \prod_{s=1}^{t} (1 - h(s \mid T=0)),$$

where $h(t \mid T=j)$ is the hazard of population *j* and *t* is the time (days on sick leave) since randomization. Under the null hypothesis (no effect), we have that $h(t \mid T=1) = h(t \mid T=0)$ for all *t*. Under the alternative hypothesis $h(t \mid T=1) > h(t \mid T=0)$, at least for all t < 30,[71] implying $\Delta(t) = S(t \mid T=1) - S(t \mid T=0) < 0$, for all *t*.

We can follow the sick-spell durations of the treated at most 300 days after randomization. A lower bound of the total reduction in days absent due to sickness is calculated as the sum of the differences between the groups from the first day after randomization until 300 days later:

$$\Delta = \sum_{t=1}^{\max t} \Delta(t).$$

Results

Figure 2 presents the difference between the survival functions of the treated and the controls, evaluated up to 300 days since randomization. Initially, there is no difference between the groups. After about 50 days there is a non- statistically significant negative difference and after about 185 days it becomes statistically significant (at the 5 per cent level).

During the follow-up period, the sum of the differences between the groups is 33 days and the mean sickness absence duration is 143 days. Thus, if an individual has been called to an IM early, the sick-spell duration decreases by at least 23 per cent on average (33/143=23).

The main focus of this paper is the total effect of receiving a call, i.e., an ITT parameter. However, the collected data include information about whether the individual has participated in the meeting or not. Despite the obligation to show up at the meeting, only 30 per cent of all the individuals called actually attended a meeting. This low number suggests that a significant part of the total effect stems from simply receiving the call, i.e., a "threat" effect.

[71] Note that any potential effect is probably not realized before a renewal of the doctor's certificate.

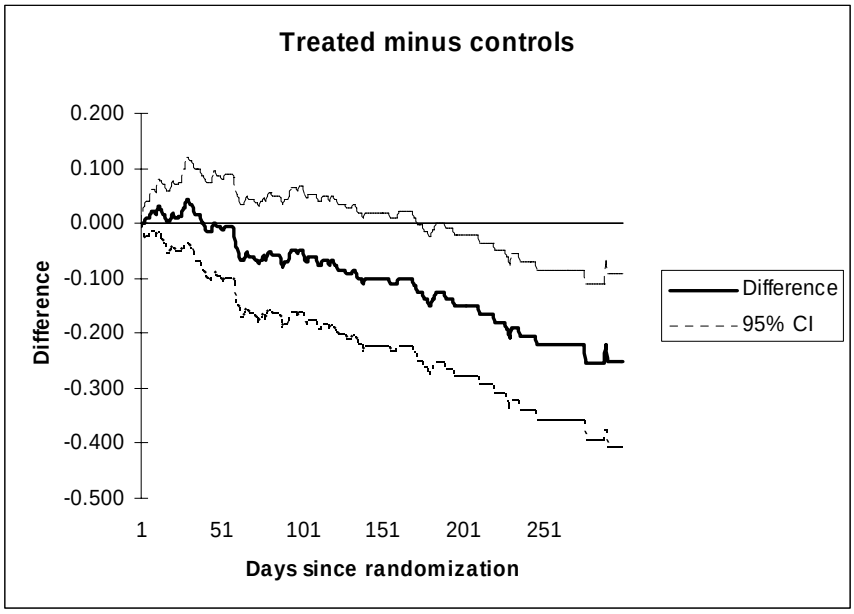


Figure 2 Difference in survival functions between treated and controls with 95 per cent confidence interval (CI) calculated according to Greenwood (1926)

Sensitivity analysis

In order to test for any differences in the distribution of individuals across randomizations, we perform two different sensitivity analyses.

First, we perform a log-rank test (Mantel and Haenszel, 1959) for equality of survivor functions and stratify on randomisations. The advantage of this test is that we can control for the different randomisations without making any functional form assumptions. The disadvantage is that we cannot verify if the magnitude of the effect is sensitive to stratification on randomizations.

Second, we estimate Cox regression models (Cox, 1972), which enable us to control for the different randomizations, the month of the sickspell start and individual observed heterogeneity using the covariates presented in Table 2. The drawback to this approach is that we have to assume that the effect is instantaneous and monotonous on the hazard.

Log-rank test

Table 4 presents the results from the log-rank test stratified on the different randomisations. As can be seen from the table, the number of ended sick spells (events observed) is higher (lower) than expected in all the randomisations except for in one. Due to the small sample sizes in each randomization, the difference is statistically significant only in one randomization when analysed separately. However, when combining the differences from all the randomisations into a single overall statistic, the result is statistically (p-value 0.0130) significant.

Table 4: Stratified log-rank test for equality of survivor functions

Randomization	**Treated**		**Controls**		**Chi(2)**
Date	*Events*[1] *expected*	*Events observed*	*Events expected*	*Events observed*	*Statistica*
24 May	7.83	10	8.17	6	1.19
29 May	8.87	10	11.13	10	0.27
04 Jul	8.87	11	9.13	7	1.03
02 Aug	6.60	7	8.40	8	0.04
01 Sep	6.75	10	6.25	3	3.41*
04 Oct	8.19	11	8.81	6	1.95
02 Nov	6.88	8	6.12	5	0.39
27 Nov	2.90	2	3.10	4	0.54
8 Sep, GIMO	3.29	4	4.71	4	0.26
12 Oct, GIMO	2.46	3	1.54	1	0.31
27 Nov, GIMO	4.17	5	1.83	1	0.56
Total	66.81	81	69.19	55	6.16**

1) Event refers to the number of individuals who became declared fit. * stat. sign at 10 per cent level, ** stat. sign at 5 per cent level

Cox regressions

Table 5 presents the estimation results of four different Cox regression models. In the first column, we present the estimate without control variables. In the second column, we have stratified on randomizations and in the third, we have added a dummy for the month of the sick-spell start. Finally, in the fourth column we have added the set of covariates presented in Table 2.

All estimations of the effect are statistically significant (at 5 per cent) and about the same magnitude irrespective of model specification. Thus, the estimated result seems not to be sensitive to stratification on randomizations.

Table 5 Estimation of the lower bound of the effect of receiving a call to an IM

	(1)	**(2)**	**(3)**	**(4)**
Called immediately	0.430	0.431	0.362	0.419
	(0.175)**	(0.176)**	(0.183)**	(0.186)**
Stratified by randomization	No	Yes	Yes	Yes
Month of sick-spell start	No	No	Yes	Yes
Covariates	No	No	No	Yes
Observations	275	275	275	275

Notes: * significant at 10 per cent; ** significant at 5 per cent; *** significant at 1 per cent; standard errors in parentheses

Discussion

Several studies confirm that moral hazard in the SI context is a real problem (Johansson and Palme, 1996, 2002 and 2005; Henreksson and Persson, 2004). The approach among these studies has been to investigate whether sick-listed individuals respond to changes in the benefit caps. The general conclusion is that the lower the cost of being sick absent, the more likely that the individual is on sick leave.

Further evidence of moral hazard is related to the interplay between the sickness and the unemployment insurance. Many countries have complex social insurance systems and different parts of them sometimes overlap in a way that can generate unintended flows between them (see e.g. Kreuger and Meyer, 2002; Henningsen, 2006; European Economic Advisory Group, 2007). In the Swedish context, this phenomenon has been analysed by Larsson (2006) and Larsson and Runeson (2007). The

general conclusion is that the extent of moral hazard (interpreted as responsiveness to changes in the replacement rate) seems to be larger among the unemployed compared with the employed (Larsson and Runeson, 2007). The suggested reason is the institutional differences between the insurances (see section 2).

The large effect found in this study should be interpreted with the target group in mind: sick-listed without permanent employment. Individuals in this group were in most cases unemployed, temporarily employed or involuntarily part-time employed. During the last years (probably as a result of previously mentioned findings), there have been institutional changes aimed to harmonize the sickness and the unemployment insurance. However, there are still economic incentives for being sick-listed rather than unemployed. The large estimated effect in this study suggests that moral hazard is still a problem among unemployed sick-listed individuals also in the present institutional setting. The policy-relevant question is hence how this problem should be mitigated.

It seems reasonable that the extent of moral hazard diminished when monitoring increased. An empirical study by Hesselius *et al*. (2005) supports this reasoning. The extent of moral hazard also depends on the personal moral, which seems to be correlated with the social context (Lindbeck *et al*., 2004; Palmer 2006) and affected by the social network (Ichino and Maggi, 2000; Lindbeck *et al*., 2007; Hesselius *et al.*, 2008).

The explicit aim of the programme (policy) studied in this paper was to affect norms. However, the low (30 per cent) attendance at the IM suggests that a significant part of the estimated effect stems from the "threat". Relating to the reasoning and findings in Black *et al.* (2003) and Geerdsen (2006)[72], our result suggests that there is also a "threat" effect of compulsory programmes in the SI context.

Unfortunately, we are not able to disentangle clearly the two obvious potential components of the total effect: a "threat" effect and a change in norms. This distinction would be interesting to make since the two parts probably have different long-run implications. If the main effect stems from a "threat", the changed behaviour about using the SI is tied to compulsory attendance at the meeting, implying that a changed behaviour would only be observed under a similar "threat" or monitoring. If, instead, the effect stems from attending the meeting, it is reasonable to believe that there are longer-run effects through changed norms about

[72] In which the "threat" from compulsory labour market programs is discussed. Geerdsen (2006) has shown that the pre-treatment effect associated with the compulsory participation is comparable in magnitude to the effect of benefit exhaustion found in studies of American UI systems. The odds of finding employment have increased by up to 145 % when compulsory programme participation approaches.

using the SI. Furthermore, recent studies on the importance of social interactions suggest that such a change in norms would have important spill-over effects through endogenous effects (Ichino and Maggi, 2000; Lindbeck *et al.*, 2007; Hesselius *et al.*, 2008).

For guiding policies aiming to reduce moral hazard, better knowledge about the behavioural mechanisms associated with the usage of the SI is needed. However, this study shows that it may be possible to prevent excess usage of SI with quite small measures. In a system with different levels of replacement rates in the unemployment and the SI, some sort of control seems necessary for preventing unintended flows between the systems. Calling sick-listed individuals to IM is both inexpensive and can hardly be regarded as insulting, which is a concern sometimes raised regarding control.

Conclusion

This study shows that calling sick-listed individuals without permanent employment (sick-listed individuals weakly established on the labour market) to information meetings about the rights and duties associated with the SI significantly reduces the sickness absence period. The estimated effect suggests a reduction of the time of sickness by at least 23 per cent.

The explicit aim of the policy was to reduce excess use of the SI. However, we are not able to disentangle the different components of the total effect: i) a “threat” effect due to compulsory attendance at the meeting and ii) changed norms after attending the meeting. However, the low attendance rate at the meetings suggests that the former is the most important part. If that is the case, norms about the usage of the SI may not have been affected from a long-run perspective.

References

Allen, S. G.: 1981, An empirical model of work attendance, Review of Economics and Statistics vol. 63, no. 1, 77–87.

Anderzén, I., Demmelmaier, I., Hansson, A-S., Johansson, P., Lindahl, E. and Winblad, U,: 2007, Utvärdering av samverkan i resursteam inom Försäkringskassan och Landstinget i Uppsala län – ett samverkansprojekt för att minska sjukskrivningstiden, Rapport, September 2007.

Black, D. A., Smith, J. A., Berger, M. C. and Brett, J. N.: 2003, Is the threat of reemployment services more effective than the services themselves? Evidence from random assignment in the UI system, American Economic Review, vol. 93, no. 4, 1313–1327.

Cox, D.: 1972, Regression models and life-tables with discussion, Journal of the royal statistical society, vol. 34, 187–220.

European Economic Advisory Group: 2007, Scandinavia today: an economic miracle?, Chapter 4 in report on the European Economy 2007, Ifo, Institute for Economic Research.

Försäkringskassan: 2007, The scope and financing of social insurance in Sweden 2005-2008, Försäkringsdivisionen Utvärderingsavdelningen.

Geerdsen, L. P.: 2006, Is there a threat effect of labour market programmes? A study of ALMP in the Danish UI system, Economic Journal, vol. 116, no. 513, 738–750.

Greenwood, M.: 1926, The natural duration of cancer, Reports on public health and medical subjects, vol. 33, 1–26, His Majesty's Stationery Office: London.

Henningsen, M.: 2006, Moving between welfare payments. The case of sickness insurance for the unemployed, Memorandum no. 04/2006, Department of Economics, University of Oslo.

Henreksson, M. and Persson, M.: 2004, The effects on sick leave of changes in the sickness insurance system, Journal of Labor Economics, vol. 22, no. 1, 87–113.

Hesselius, P., Johansson, P. and Larsson, L.: 2005, Monitoring sickness insurance claimants: evidence from a social experiment, Working Paper 2005:15, IFAU.

Hesselius, P., Johansson, P. and Vikström, J.: 2008, Monitoring and norms in sickness insurance: empirical evidence from a natural experiment, Working Paper 2008:08, IFAU.

Ichino, A. and Maggi, G.: 2000, Work environment and individual background: explaining regional shirking differentials in a large Italian firm, The Quarterly Journal of Economics, vol. 115, no. 3, 1057–1090.

Johansson, P. and Palme, M.: 1996, Do economic incentives affect work absence? Empirical evidence using Swedish micro data, Journal of Public Economics, vol, 59, 195–218.

Johansson, P. and Palme, M.: 2002, Assessing the effects of a compulsory sickness insurance on worker absenteeism, Journal of Human Resources, vol. 37, no. 2, 381–409.

Johansson, P. and Palme, M. 2005, Moral hazard and sickness insurance, Journal of Public Economics, vol. 89, 1879–1890.

Kreuger, A. and Meyer, B. 2002, Labor supply effects of social insurance, in A. Aurebach and M. Feldstein (ed.), Handbook of Public Economics, vol. 4, North-Holland/Elsevier, Amsterdam.

Larsson, L.: 2006, Sick of being unemployed? Interactions between unemployment and sickness insurance, The Scandinavian Journal of Economics, vol. 108, 97–113.

Larsson, L. and Runeson, C.: 2007 Moral hazard among the sick and unemployed: evidence from a Swedish social reform, Working Paper 2007:8, IFAU.

Lindbeck, A., Palme, M. and Persson, M.: 2004, Sjukskrivningar som ett socialt fenomen, Ekonomisk debatt, vol. 4, 50–62.

Lindbeck, A., Palme, M. and Persson, M.: 2007, Social interaction and sickness absence, IFN Working Paper no. 725.

Mantel, N. and Haenszel, W.: 1959, Statistical aspects of the analysis of data from retrospective studies of disease, Journal of the National Cancer Institute, vol. 22, 719–748.

Nyman, K., Palmer, E., Bergendorff, S.: 2002, Den svenska sjukan – sjukfrånvaron i åtta länder, Expertgruppen för Studier i Offentlig Ekonomi (ESO), Finansdepartementet, Ds 2002:49.

Palmer, E.: 2006, Sjukförsäkring, kulturer och attityder – fyra aktörers perspektiv, Försäkringskassan Analyserar 2006:16.

SOU: 2007, Arbetslösa som blir sjuka och sjuka som inte blir arbetslösa – samtal om socialförsäkringen nr 16, Socialförsäkringsutredningen, Statens offentliga utredningar.

Publication series published by the Institute for Labour Market Policy Evaluation (IFAU) – latest issues

Rapporter

2009:1 Hartman Laura, Per Johansson, Staffan Khan and Erica Lindahl, ”Uppföljning och utvärdering av Sjukvårdsmiljarden”

2009:2 Chirico Gabriella and Martin Nilsson ”Samverkan för att minska sjukskrivningar – en studie av åtgärder inom Sjukvårdsmiljarden”

2009:3 Rantakeisu Ulla ”Klass, kön och platsanvisning. Om ungdomars och arbetsförmedlares möte på arbetsförmedlingen”

Working Papers

2009:1 Crépon Bruno, Marc Ferracci, Grégory Jolivet and Gerard J. van den Berg “Active labor market policy effects in a dynamic setting”

2009:2 Hesselius Patrik, Per Johansson and Peter Nilsson “Sick of your colleagues’ absence?”

2009:3 Engström Per, Patrik Hesselius and Bertil Holmlund “Vacancy referrals, job search and the duration of unemployment: a randomized experiment”

2009:4 Horny Guillaume, Rute Mendes and Gerard J. van den Berg ”Job durations with worker and firm specific effects: MCMC estimation with longitudinal employer-employee data”

2009:5 Bergemann Annette and Regina T. Riphahn “Female labor supply and parental leave benefits – the causal effect of paying higher transfers for a shorter period of time”

Dissertation Series

2009:1 Lindahl Erica “Empirical studies of public policies within the primary school and the sickness insurance”

2008:1 Andersson Christian “Teachers and student outcomes: evidence using Swedish data”