IFAU
Institute for Evaluation of Labour
Market and Education Policy

Inquiry-based learning put to test: long-term effects of the Swedish science and technology for children program

Erik Mellander
Joakim Svärdh

WORKING PAPER 2015:23

The Institute for Evaluation of Labour Market and Education Policy (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 and educational policies, studies of the functioning of the labour market 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

ISSN 1651-1166

Inquiry-based learning put to test: long-term effects of the Swedish science and technology for children program[a]

by

Erik Mellander[b] and Joakim Svärdh[c]

October 14, 2015

Abstract

We quantitatively evaluate the effects in grade 9 on content and process skills in sciences, from participation in the Swedish science and technology for children program, the NTA program. Students predominantly participate in this program during grades 1-6. Our outcome measures are scores and grades on nationwide tests, and course grades, in *biology*, *chemistry* and *physics*, 2009-2010. A nationally representative random sample of almost 16 000 test-taking students is coupled with multi-level information about the NTA, and background factors. Non-random selection into the program is addressed by propensity score analysis. The matched sample of pairs of NTA participants and non-participants, determined by the propensity scores, is quite well behaved, although there are significant, but small, differences for some of the matching covariates. We find significantly positive effects on test scores (average increase 16.4%) and test grades in *physics*, but not in *biology* and *chemistry*. No significant effects are found for course grades.

Keywords: Quantitative evaluation, national standardized tests, representative sample, non-random selection, propensity score analysis, multi-level modeling, post-matching multivariate regression.

[a] We gratefully acknowledge constructive comments from participants in seminars at the IFAU, the KTH and Gothenburg University, in particular Jan-Eric Gustafsson, Helena Holmlund, Ulrika Vikman, Christina Wikström, and Monika Rosén.

[b] Institute for Evaluation of Labour Market and Education Policy (IFAU), P.O. Box 513, SE – 751 20 Uppsala, Sweden; erik.mellander@ifau.uu.se

[c] KTH, Education and Communication in Engineering Science (ECE), Osquars backe 31, SE – 100 44 Stockholm, Sweden; jsvardh@kth.se

Table of contents

1 Introduction

For decades, the pros and cons of inquiry-based science learning have been debated in the scientific literature**.** To a considerable degree, disagreements are due to conceptual issues, relating to the content, properties and structure of inquiry-based learning. However, conflicting *empirical* evidence about the relative merits of inquiry-based learning has also been important in keeping the debate alive; see further Section 3.

The purpose with this paper is to make a contribution with respect to the empirical strand of this literature. While our analysis concerns Sweden we believe it to be of general interest. To begin with, *the science and technology for children* (STC) program that we study is a very international phenomenon. Originally developed in the U.S., different versions of the program today exists in a number of countries like, e.g., Chile, Thailand, China and Germany, as well as in Sweden. And, to the best of our knowledge, there is no previous large scale assessment of the effects of the STC. Undoubtedly, this is to a large extent due to a lack of appropriate and reliable outcome measures. This shortcoming is not limited to assessments of inquiry-based learning provided through the STC. Geier et al. (2008, p. 924) note: “The lack of student-level distal standardized test data to demonstrate achievement gains from standards-based inquiry science curricula remains a weakness in the literature.”, distal tests referring to statewide standardized tests, as opposed to local test [Geier et al. (op.cit., p. 923)].

In this study, we have access to results on high-stakes standardized tests in natural sciences. From 2009, these tests are taken by all Swedish students at the end of compulsory school, in grade 9. They provide information about science content knowledge as well as process skills. Our sample is also nationally representative.

Our analysis also incorporates methodological contributions. As we make use of observational data – rather than data generated in an experimental setting – we have to allow for non-random treatment assignment. Enrollment in the STC program being voluntary, there is reason to suspect participants to differ systematically from non-participants. We deal with this issue by means of propensity score analysis, a multi-variate method which by means of a scalar in the (0,1) interval, the propensity score, allows for differences across treated and non-treated individuals along a multitude of dimensions, cf. Guo and Fraser (2009). Moreover, among our outcome measures – test scores, test grades and course grades – the latter two are discrete variables. In our empirical analysis we explicitly account for this property.

The matching of participants and non-participants on the propensity score works well in that it yields pairs of NTA and non-NTA individuals that both cover the entire (0,1) propensity score range and exhibit small within-pair score differences. However, for some of the covariates used to estimate the propensity scores there are significant, but small, differences between the NTA and non-NTA groups in the matched sample.

With respect to effect estimates, we find statistically significant positive effects of the Swedish STC program on the standardized test results (test scores and test grades) for natural sciences in general. Disaggregating over subjects, we establish that this effect derives from positive and significant effects on the test scores and test grades in *physics* only; there are no significant positive effects for *biology* and *chemistry*. Moreover, we find no statistically significant effects at all when we employ course grades as outcome variables. This result provides empirical support to the claim in Geier et al. (op.cit.) about the importance of distinguishing between standardized and non-standardized outcome variables when assessing inquiry-based learning.

In the next section, we briefly describe the Swedish science and technology for children program. As a first step in our sample selection analysis, we also describe the process of joining the program. Section 3 contains a selective review of earlier, quantitative analyses of the effects of the STC and similar programs. In Section 4 we describe our data and Section 5 we consider the methodological challenges that we face in our analysis. Section 6 reports on the empirical results and Section 7 concludes.

2 The Swedish science and technology for children program

The Swedish science and technology for children program is an inquiry-based program encompassing teaching materials and teacher instructions, including teacher training sessions. In Sweden, the program is called *natural science and technology for all* (NTA). The program is intended for use in compulsory school, implying grades 1-9, with a focus on grades 1-6.

2.1 The U.S. original and the Swedish version of the program (the NTA)

The NTA program is an adapted and less extensive version of the U.S. STC program. Precursors to the STC program have a history dating back to the 1960s. The present, second generation, version has its origin in the National Science Education Standards, formulated in 1995 by the National Research Council (NRC) and the American Association for the Advancement of Sciences (AAAS).

The STC program is quite ambitious and far-reaching: in addition to teaching materials – "boxes" – and support for the implementation of these boxes, the program includes teacher training and support for school leaders. The program is divided into two parts. The first addresses teaching from pre-school to up to 6th grade. The second part is developed for teaching in grades 6-8. The boxes are structured by subject with a progression that follows age. The subjects are organized according to the themes *life science, earth science*, and *physical science & technology*.[1]

The NTA program was introduced in 1997; the Swedish Royal Academy of Sciences (KVA), together with the Swedish Royal Academy of Engineering Sciences (IVA), translated the STC and adapted it to the requirements of the Swedish curriculum (Lgr11). As of yet, NTA primarily addresses grades 1-6, although it has been used in grade 7-9. The NTA program does not provide as many boxes as the STC – 19, compared to 32 in the STC. No school leader support is offered within the NTA and the programs for teacher training are somewhat less extensive with respect to scope and depth than their STC counterparts. After participating in the training sessions – one per theme – the teacher utilizes a box for one semester. The work with the theme follows a template: the children formulate a hypothesis, conduct experiments, analyze the results, and, finally, document their work in writing. In 2011, NTA was employed by some 7 000 teachers in a third of Sweden's 290 municipalities. In total, about 100 000 students were involved in the program, making up 1 out of 8 students in grades 1-9.

2.2 How to join the NTA program

To assess the effects of the NTA it is important to know about the selection into the program. Both public and private schools can join the NTA.[2]

The first step in joining the NTA is to become a member of the NTA support organization. If the school is private, this decision is taken by the school itself. With respect public schools, the decision is taken by the municipality. The membership involves an entrance fee of approximately $ 2 200. If the school is private it has to stand

[1] About the STC, see further http://www.carolinacurriculum.com/STC/STC+Educational+Effectiveness.asp.

[2] Almost 11 percent of the Swedish pupils in grades 1-9 attended private schools in the school year 2009/2010.

the entire cost by itself. For a public school the cost may be lower than $ 2 200, as the municipality's entrance cost may be shared among several schools.

The next step is to decide if the school(s) should make use of the NTA program. Again, the private school decides itself whether to do so. With respect to public schools the decision may be taken at the municipality level or locally by school principals, or by individual teachers.

When running, the NTA program entails a variable cost of around $ 250 per box. In addition, a $ 3 fee is charged per student, semester, and box. In public schools these costs are sometimes covered by the municipality, sometimes by the schools themselves.

To sum up: selection into the NTA program does not take place at the individual level but at the municipality and school level. Both private and public schools can join the program. It is easier for private schools to do so, but also more costly. Still, for public and private schools alike the costs are negligible compared to teacher salaries.

3 Quantitative analyses of the effects of inquiry-based learning

In a recent meta-study Minner et al. (2010) conclude that out of 138 analyses of inquiry science instruction conducted during the 1984-2002 period, 51% showed positive impacts on student learning and retention. These mixed results are mirrored in the existence of strong advocates as well as strong opponents of inquiry-based learning. For example, a critical view is articulated by Kirschner et al. (2006) while Hmelo-Silver et al. (2007) express a favorable opinion. In a very short and selective literature review we consider some quantitative analyses of STC-like programs in the US, representing these opposing positions. We also describe a survey study of the NTA program.

3.1 US Studies

We first consider three studies reporting positive effects of STC(-like) programs.

Young and Lee (2005) analyze the effects STC and a similar related program by comparing about 200 participants in grade 5 with demographically matched students that had had traditional teaching. Pre- and post-tests were conducted. The results showed a significant positive difference between the program participants and the students attending traditional teaching.

Lynch et al. (2005) compared approximately 1 200 eight-grade students that had attended inquiry-based learning in *chemistry* with 1 000 non-participants of the same age. They found that both overall and with respect to sub-groups the inquiry students outperformed the non-participants.

Geier et al. (2008) showed significantly higher pass rates on high-stakes standardized tests in sciences (earth, physical and life science) for 1 800 students that had taken part in inquiry-based instruction in *physics* and *ecology/earth* science, compared to around 17 500 comparison students. The inquiry-based program also reduced the achievement gap experienced by Afro-American boys.

We next consider two examples of studies concluding that guided instruction outperforms inquiry-based learning.

First, Mayer (2004) in a meta-study reviewed studies conducted from 1950 to the late 1980s. Comparing analyses of unguided, problem-based instruction (inquiry-based approaches) with guided forms of instruction, he concluded that the "…. debate about discovery has been replayed many times in education but each time, the evidence has favored a guided approach to learning." (p. 18).

Second, Klahr and Nigam (2004) studied 112 third and fourth grade children that were randomly assigned to guided instruction and discovery learning, respectively. Two aspects on science learning were tested. First, if students learned more through discovery versus instruction and, second, whether the quality of learning differed with respect to the ability to transfer the learning to new contexts. In both respects they found that instruction involving considerable guidance produced superior results.

To us, the three studies supporting inquiry-based learning seem to be superior to the two studies in favor of guided instruction. For instance, the pro-inquiry based studies appear to have better research designs, employ larger samples and to be more updated. Still, we find it hard to dispute the argument put forward by Kalyuga et al. (2001). Novices in an area need extensive guidance in order to build up their long-term memory capacity – problem solving is an inefficient way to achieve this objective, they argue. Problem solving becomes superior to guided instruction only once learners have amassed a sufficient amount of experience. These arguments imply that the relative efficiency of inquiry-based learning and guided instruction to a large extent is a matter of timing – inquiry-based learning is unlikely to be effective if employed "too early" but might give good results if applied to students with the right basic knowledge. Since it generally will be difficult to know the right timing with respect to a given inquiry-based learning program, empirical effect evaluations will be necessary.

Finally, since we are investigating the long-term effects of the Swedish STC, studies with a longer time perspective are of interest. We have only found one: Bredderman (1983) compared students who participated in inquiry-based programs in low- and middle-school but received traditional science education in high school to students that had not attended any inquiry-based programs. No significant differences were found.

3.2 Swedish studies

While there are many qualitatively oriented studies of the NTA program [cf. Svärdh, (2013)], there is no previous analysis that employs quantitative methods. However, Anderhag and Wickman (2007) document the results of an interview study involving 80 students in grade 6, of which 40 were randomly chosen from classes that had and had not participated in the NTA program, respectively. According to their findings, the NTA-students had achieved a deeper understanding of natural science concepts and processes than students in the control group. The difference was most clearly visible in the tails of the skill distribution, i.e. with respect to low- and high-performing students.

4 The data

Our data originate from two random samples, drawn by researchers at Umeå university, from the students that took national tests in natural sciences in grade 9, in 2009 and 2010.[3] In both years, a 10 percent sample was to be drawn from the students that showed up at the test.

The standardized test in natural sciences was introduced in 2009. Actually, it is not one test but three different subject tests: one in *physics*, one in *chemistry* and one in *biology*. However, each student only takes one of the three subject tests. Which one (s)he writes is determined by the authority administering the tests, the Swedish national agency for education. This agency uses stratified random assignment to assign the subject tests across schools.[4]

[3] The samples were extracted by the department of applied educational science, the unit for educational measurement, which was also responsible for the construction of the national tests in the natural sciences.

[4] From 2010 and onwards the assignment of tests by subject is not entirely random since the Swedish national agency for education imposes a constraint on the assignment: a given school shall not be assigned tests within the same subject in two consecutive years.

We first consider how the samples that we use in our empirical analysis relate to the population(s) of grade 9 students and the samples drawn by Umeå university among the students taking the test. In the next subsection we explain our definition of NTA participation. We then go on to consider how the resulting sub-samples of NTA and non-NTA participants, respectively, are distributed across subject tests and how the corresponding test results compare. In the last subsection, we compare the characteristics (the "explanatory variables") of the NTA and non-NTA participants at four levels of aggregation: the individual level, the school level, the municipality level and the regional (county) level. It should be noted that the discussion in this section is based on "raw" data, i.e. before any attempt has been made to control for the non-random selection into the NTA program. In Section 6, we will provide the corresponding comparisons of the NTA participants with a matched sample of non-NTA participants, constructed by means of propensity score analysis. As the propensity score analysis is supposed to do away with the problem of non-random selection, the differences between the characteristics of the NTA and the matched non-NTA participants should be much smaller than the differences found in the raw data, i.e. the differences reported in this section.

4.1 The grade 9 students – from the population to the sample we analyze

Table 1 provides information about the Swedish populations of grade 9 students in the school years 2008/2009 and 2009/2010, the numbers of these that actually took the test, the random samples drawn from these individuals and, finally, the corresponding samples that we use in our empirical analysis.

Table 1. Student populations and the samples analyzed

School year	# students in grade 9	# (%) students taking the test	# (%) of test-taking students sampled	# (%) of sampled students in our study
2008/2009	118 032	88 491 (75.0)	8 028 (9.1)	6570 (81.8)
2009/2010	113 545	98 848 (87.1)	7 839 (7.9)	6396 (81.6)
Sum	*231 577*	*187 339 (80.9)*	*15 867 (8.5)*	*12 966* (81.7)

Sources: The Swedish national agency for education; the department of applied educational science, Umeå university; own calculations.

From the table it is quite clear that a considerable share of the students supposed to take the test did not show up to do so. Moreover, the share of no-shows varies over time, from 25 percent in the school year 2008/2009 to 13 percent the following school year.[5] The table also indicates that the intention to extract samples corresponding to 10 percent of those taking the test was not fulfilled, especially not in the school year 2009/2010.[6] The last column in table shows the number students included in our analysis. That these numbers are lower than the numbers in the next to last column is due to the fact that we were unable to classify about 18 percent of the students data as either NTA participants or as individuals that had not participated in the NTA program. This issue is the topic of the next sub-section.

4.2 Making the definition of NTA participation operational

The definition of NTA and non-NTA participation has been made operational at the school level, according to a two-step procedure. In the first step, information from the NTA support organization made it possible to classify Sweden’s 290 municipalities as NTA or non-NTA municipalities. The classification was based on whether at least one of the municipality’s schools or none of its schools, respectively, had joined the NTA program in the year 2009. This resulted in 87 out of the 290 municipalities being defined as NTA municipalities.[7]

In the second step, local NTA coordinators, working in the NTA municipalities, were contacted and asked to assign the schools in their municipality that had pupils in grade 9 to one of the following five categories: [8]

i) All of the school’s students had participated in the NTA program up to 6th grade, implying that the average pupil participated 3-4 semesters.

ii) Like i) and, in addition, the school’s students had also participated in the NTA to some extent during grades 7-9.

iii) Some of the school’s students had participated in the NTA program

iv) None of the school’s students had participated in the NTA program

[5] According to the Swedish national agency for education a large number of schools choose not to take the test in 2008/2009, despite the fact that the Swedish national agency for education required them to do so.

[6] This was not related to the sample frame(s). Rather, it was due to i) Umeå university having no power to force the schools to report their test results and to ii) the fact that some schools refused to do the test. The second explanation was more important for the school year 2008/2009; cf. the previous footnote. The first explanation mattered more for 2009/2010; more schools choose not to report to Umeå university in that year, than in 2008/2009.

[7] Of these 87 municipalities, there were 6 in which only a single school participated in NTA. In each of these 6 municipalities the participating school was a private school.

[8] Not all Swedish schools offer all the grades 1 through 9. The local NTA coordinators therefore had to keep track both of which schools that participated in the NTA program and of what schools students attended before they enrolled at the school where they finished 9th grade and where they also did the national test in sciences.

v) No information was available about whether the school's students had participated or not.

In this study, a school is classified as an NTA school if it belongs to categories i) or ii). This results in almost 100 out of slightly more than 1 400 schools with grade 9 pupils, i.e. 7 percent, being classified as NTA schools. The classification of schools as NTA or non-NTA participants directly carries over to the corresponding classification of individuals: all students that in grade 9 attended a school classified as an NTA school are defined as NTA participants. This yields 1 121 NTA students, whereof 791 correspond to category i) and 330 to category ii).

A school is classified as a non-NTA if it belongs to category iv). This category contains 11 845 students.

Finally, schools belonging to categories iii) and v) have to be left out of the analysis because we cannot classify them as NTA or non-NTA schools. The number of students in this group is 2 901.

Table 2 summarizes our classifications and shows how our sample is related to the sample collected by Umeå university. Comparing with Table 1, we see that the number in the third column in Table 2 equals the last entry in the final column in Table 1, and that the number in the last column of Table 2 equals the last entry in the next to last column in Table 1.

Table 2. The relation between our sample and the Umeå university (UU) sample

NTA students	Non-NTA students	Our sample: NTA+non-NTA	Not classified students	NTA + non-NTA + not classified = UU sample
1121	11845	12966	2901	15867

It should be noted that the two-step procedure used to define students as NTA or non-NTA students is associated with a measurement error. Since students may move between schools, some of the grade 9 students in a given school may deviate from the typical schooling career of students in that school. If so, the deviating students may not have participated in NTA although having done so is typical of the students in the school. Similarly, if the typical students in a school have not participated in the NTA, there may be some students that have, in fact, participated in the program. However, as long as there is no systematic pattern in mobility across NTA and non-NTA schools the measurement error can be treated as random. As such, it will bias our estimates of the effects of NTA towards zero, making it more difficult to establish significant effects than in the absence of measurement error. Accordingly, positive estimated effects of NTA can be regarded as lower limit estimates of the true effects.

4.3 Distributions over science subjects in the national tests by NTA and non-NTA individuals

Before we turn to our outcome variables, i.e. the results on the national tests in the three different science subjects, we consider how the NTA and the non-NTA individuals in the random samples are distributed across the subject tests.

Table 3 shows that for the sample as a whole (i.e. for the school years 2008/2009 and 2009/2010 together), the non-NTA participants are evenly spread across subject tests while the NTA participants are somewhat under-represented in the *physics* test and slightly over-represented in *chemistry* and *biology*.

Comparing the individual school years, the differences in distributions across subjects is more marked for the NTA participants. This is to be expected, as the number of non-NTA individuals is over ten times larger than the number of NTA participants.

Table 3. Samples partitioned into NTA and non-NTA participants, by subject and year

	NTA participants			**Non-NTA individuals**		
	School year			School year		
Subject	**2008/2009** # (%)	**2009/2010** # (%)	**Sum** # (%)	**2008/2009** # (%)	**2009/2010** # (%)	**Sum** # (%)
Biology	165 (31)	224 (38)	389 (35)	2 005 (33)	1 974 (34)	3 979 (34)
Chemistry	198 (37)	190 (33)	388 (35)	2 089 (35)	1 955 (34)	4 045 (34)
Physics	175 (32)	169 (29)	344 (30)	1 937 (32)	1 884 (32)	3 821 (32)
Sum	*538 (100)*	*583 (100)*	*1121 (100)*	*6 031 (100)*	*5 813 (100)*	*11 845 (100)*

Sources: The department of applied educational science, Umeå university; own calculations.

We now proceed to look at the results for the NTA and non-NTA individuals, respectively. In so doing, we will, for expositional ease, only consider the results for the two school-years taken together, corresponding to columns denoted “Sum” in Table 2. However, the sensitivity analysis in section 6 will include a control variable for year of test.

4.4 Results for NTA and non-NTA individuals

The outcome variables available to us are: scores on the nation-wide standardized tests, grades on the standardized tests, and course grades. It should be kept in mind that the students take the test in only one of the subjects *biology*, *chemistry* and *physics*.

4.4.1 General information about the standardized science tests

Each test consists of two parts, conducted separately, at different points in time. One part relates to scientific content knowledge with a maximum test time of 120 minutes. The other part concerns scientific process skills, for which 90 minutes is allowed.

The questions making up the test are divided into groups according to subject matter and degree of difficulty. Regarding subject matter, one group of questions concerns, e.g., the understanding of natural science concepts, models and theories, while another group relates to natural sciences methods and procedures. Each group of questions is partitioned into sub-groups of increasing difficulty and the results of each of the sub-groups are assessed separately.

4.4.2 Test scores

The test score is the sum total of the number of points assigned to (the answers to) the questions making up the test. The test is here understood to mean both the part relating to scientific content knowledge and the part concerning scientific process skills.

The minimal test score is always zero whereas the maximum score varies by school year and subject. Specifically, for *biology* the maximum score was 33 in 2008/2009 and 39 in 2009/2010, while the corresponding numbers for *chemistry* were 44 and 43, and for *physics* 34 and 42. This means, of course, that to get the results comparable over time, the scores have to be normalized in some way. We have chosen the common approach of working with percentile ranked test scores.[9]

Another potential problem with the test scores relates to the assessment of the answers to individual questions. In that assessment no account is taken of the fact that the test is composed of questions corresponding to different degrees of difficulty; cf. section 4.4.1. Specifically, a correct answer to a question of mid-range complexity in the sub-group of questions of lowest difficulty may yield the same number of points as a correct answer to a question that is of mid-range difficulty within the most difficult sub-group of questions. However, as long as the examinees are aware of this structure and act rationally – begin with the easier problems and address the more complicated ones if time permits – this should not be a problem in practice.

[9] The test sore's percentile rank is computed according to: $[(c_\ell + 0.5f_i) / N] \times 100$ where c_ℓ is the number of scores less than the score of interest, f_i is the frequency of the score of interest and N is the total number of examinees.

Smoothed frequency distributions of the percentile ranked test scores are illustrated in Figure 1a-d; the original distributions (not shown) are too noisy to be informative. The smoothing has been carried out as follows. First, the frequencies have been aggregated by percentiles, yielding 100 frequencies. Second, moving averages involving 6 observations (i.e. percentiles), weighted equally, have been computed.

Figure 1 a shows that when the (percentile ranked) test scores for all subjects are lumped together the difference is small between the frequency distributions of the NTA participants and the non-NTA individuals. The NTA frequency distribution has somewhat larger mass on the percentile range 30-50 than the non-NTA distribution, however, while the non-NTA distribution's mass is slightly higher in the 80-100 percentile range. The difference in the means of the distributions is significant at the 5 percent level, indicating that non-NTA individuals do better at the test. That is to say, Figure 1 a is consistent with a *negative* impact of NTA on the results on standardized national tests in natural sciences. It should be remembered, though, that we are comparing "raw" data; we do not account for the non-random selection into the NTA program. If there are systematic differences between NTA students and non-NTA students the outcome in Figure 1 a might still be consistent with a positive effect of participation in NTA – or no effect at all. Accordingly, both the *t*-test and the effect size (Cohen's *d*) should be here viewed merely as descriptive statistics.[10]

[10] Like the *t* statistic, Cohen's *d* statistic has been computed using the non-smoothed percentile ranked data. The *d* statistic, defined as the mean difference, i.e. m^{NTA} - $m^{non\text{-}NTA}$, divided by the pooled standard deviation [Kenny (1987, p. 215)] is commonly used in educational science. Given knowledge about the number of observations on the NTA participants, N^{NTA}, and the number of observations on non-NTA individuals, $N^{non\text{-}NTA}$, the *d* statistic can be computed from the *t*-statistic according to

$$d = t / [(N^{NTA} \times N^{non\text{-}NTA}) / (N^{NTA} + N^{non\text{-}NTA})]^{0.5}$$

Thus, in Figure 1a, the *d*-statistic is t / $[(1121 \times 11845) / (1121 + 11845)]^{0.5}$ = -2,15 / 32 ≈ - 0.067.

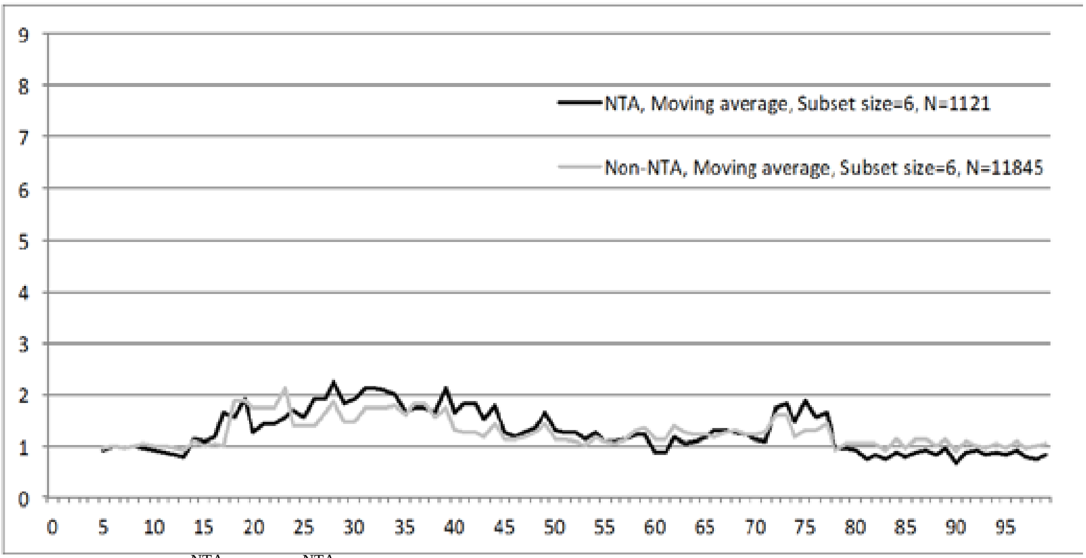
Regarding the interpretation of the *d* statistic, Hattie (2009, p. 9) suggests that, in terms of absolute value, $d = 0.20$ can be considered as a small effect, while $d = 0.4$ and $d = 0.6$ can be regarded as medium-sized and large effects, respectively. Thus, by Hattie's standard, $d = - 0.067$ should be considered a very small effect. This seems at odds with the fact that the *t* statistic is statistically significant.

The reason why the two statistics differ is that while the *t* statistic relates the difference in the means to the corresponding standard deviation, the *d* statistic relates the difference in the means to a weighted average of the standard deviations of the test scores – i.e. not to a function of the standard deviations of the *mean* test scores. As a consequence, the numerator and the denominator of the *d* statistic are not congruent.

The *d* statistic fails to account for the fundamental fact that an average of independently and identically distributed random variables has a standard deviation that is smaller than the standard deviation of one of the stochastic variables making up the average. Due to this property, the *d* statistic must *always* be interpreted as merely a descriptive statistic. This is in contrast to the *t* statistic, which under proper circumstances can be used for statistical inference.

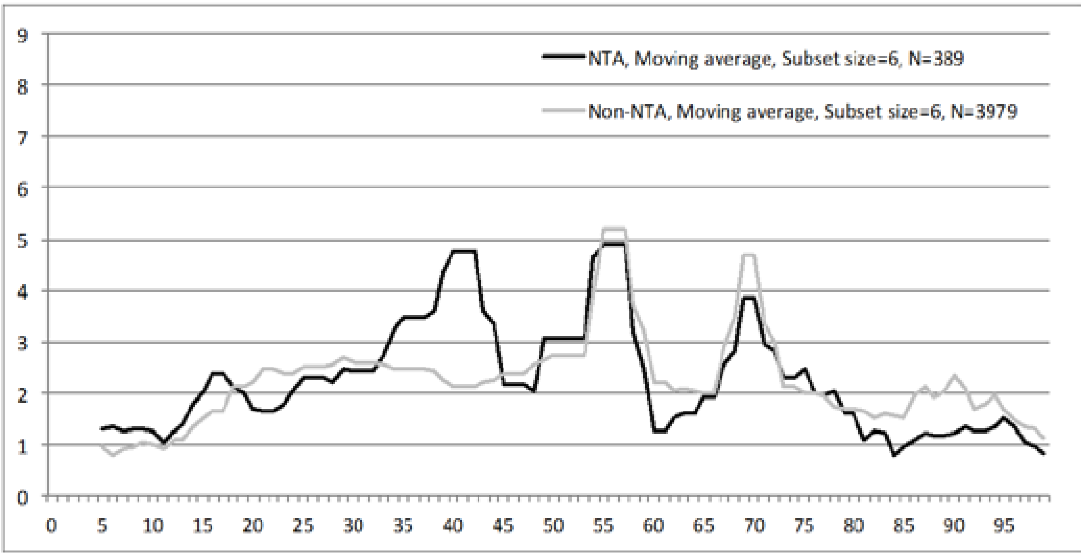
Figure 1

Figure 1 a. Frequency distribution (%) of percentile ranked test scores, by percentile; 6 percentiles moving averages, for NTA participants and non-NTA individuals, *raw data*, ***all subjects***



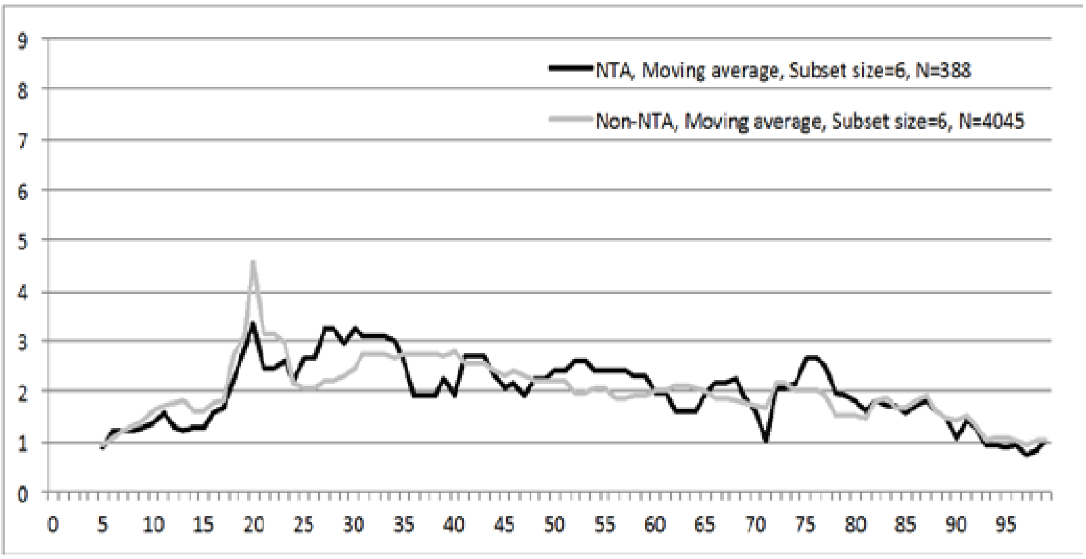
Mean difference: $m^{NTA} - m^{non\text{-}NTA}$ = 48.60 – 50.54; std. dev. (mean difference) = 0.9025; *t*-statistic = –2.15. Test for equality of means rejected at 5% level. Effect size (Cohen's *d*): – 0.067.

Figure 1 b. Frequency distribution (%) of percentile ranked test scores, by percentile; 6 percentiles moving averages, for NTA participants and non-NTA individuals, raw data, **biology**



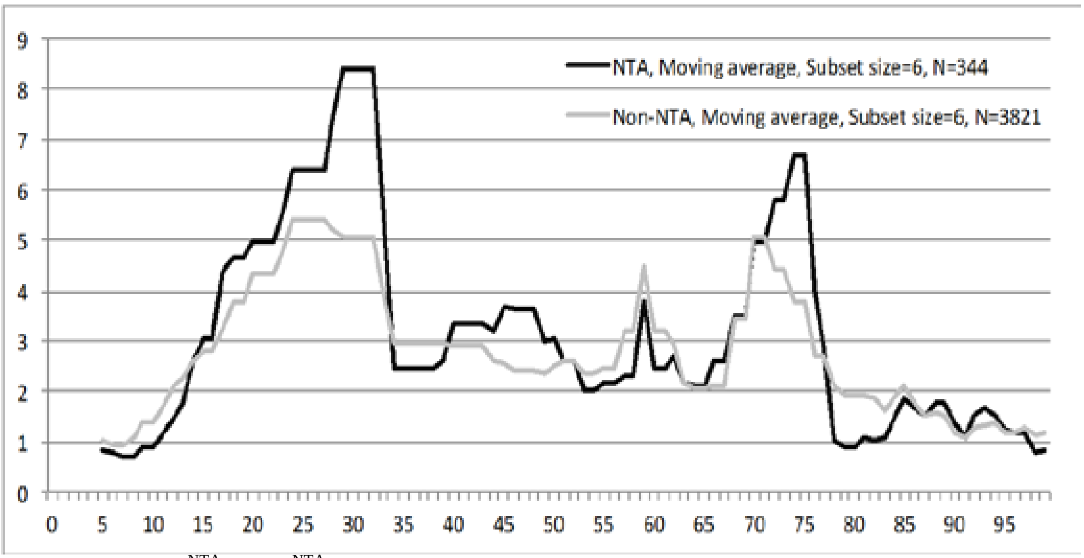
Mean difference: $m^{NTA} - m^{non\text{-}NTA}$ = 45.92 – 51.18; std. dev. (mean difference) = 1.5321; *t*-statistic = –3.44. Test for equality of means rejected at 1% level. Effect size (Cohen's *d*): – 0.183.

Figure 1 c. Frequency distribution (%) of percentile ranked test scores, by percentile; 6 percentiles moving averages, for NTA participants and non-NTA individuals, raw data, **chemistry**



Mean difference: mNTA – mnon-NTA = 51.01– 50.06; std. dev. (mean difference) = 1.5340; t-statistic = 0.62 Test for equality of means not rejected. Effect size (Cohen's d): 0.033.

Figure 1 d. Frequency distribution (%) of percentile ranked test scores, by percentile; 6 percentiles moving averages, for NTA participants and non-NTA individuals, raw data, **physics**



Mean difference: $m^{NTA} - m^{non\text{-}NTA}$ = 48.91 – 50.38; std. dev. (mean difference) = 1.6280; *t*-statistic = – 0.91. Test for equality of means not rejected. Effect size (Cohen's *d*): – 0.051.

The weak tendency towards *lower* test grades among NTA students for science in general is more pronounced with respect to *biology*; cf. Figure 1 b. For this subject the null hypothesis that the NTA and the non-NTA distributions are equal is decisively rejected. But again, it should be remembered that we are here considering the raw data.

For *chemistry* (Figure 1 c), the difference between the NTA and non-NTA students' mean percentile ranks is small and insignificant; the t-test cannot reject the null hypothesis that means are equal. The same holds for *physics* (Figure 1 d).

4.4.3 Test grades

Test grades are given according to the following scale: *fail (F), pass (P), pass with distinction (PD), and pass with special distinction (PSD).* Numbers – points, *p* – are associated with these grades according to: *F* = 0 *p*, *P* = 10 *p*, *PS* = 15 *p*, and *PSD* = 20 *p*. Figure 2a-d show the test grade distributions for NTA and non-NTA individuals, for all subjects taken together and by individual subjects. When comparing these distributions statistically we use a non-parametric test, the Mann-Whitney U test, which compares the medians of the two distributions, rather than the means.[11]

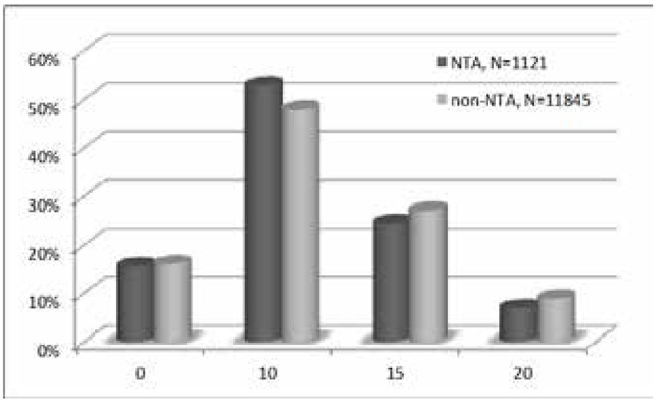
For all subjects taken together (Figure 2a) the differences between NTA and non-NTA individuals are small, just like with respect to the test scores. In analogy with the test score distributions, the NTA test grade distribution has a somewhat larger part of its mass to the left, on the grade pass (= 10) than the non-NTA distribution which, instead, has slightly more mass on pass with distinction (= 15) and pass with special distinction (= 20). Similar to the *t*-test, the Mann-Whitney U test rejects the null hypothesis that the grade distributions for the NTA and non-NTA individuals are equal, indicating that the median of the non-NTA distribution is the larger one.

Again, in *biology* the distinction between the NTA pupils and the non-NTA pupils is quite clear, cf. Figure 2b, to the advantage of the non-NTA pupils. The null hypothesis that the NTA and the non-NTA distributions are equal is rejected at the 1% level.

[11] There are two reasons why we do not use the (parametric) *t*-test. One is that we want to allow for the possibility that, for practical purposes, the grades are not measured on an interval scale. If so, the actual differences between the grades pass, pass with distinction, and pass with special distinction, might not all be equally large, as suggested by the points assigned to the grades, i.e. 10, 15, and 20, respectively. The second reason is that, in contrast to the *t*-test, the Mann-Whitney U test does not require the grade distributions of the NTA and non-NTA participants to be normal, which, clearly, they are not; cf. Figure 2a-d.

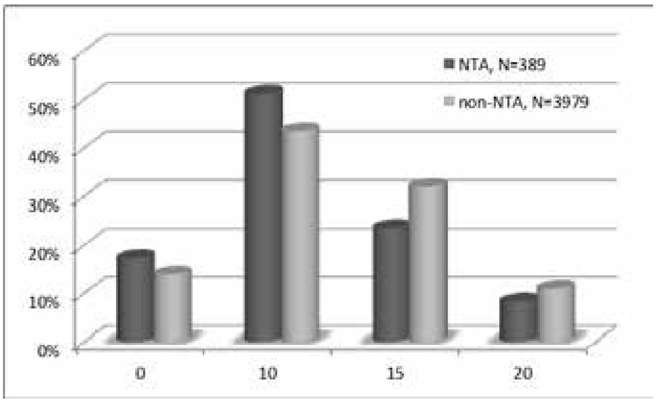
Figure 2

Figure 2a: Test grade distributions for NTA and non-NTA individuals, raw data, ***all subjects***



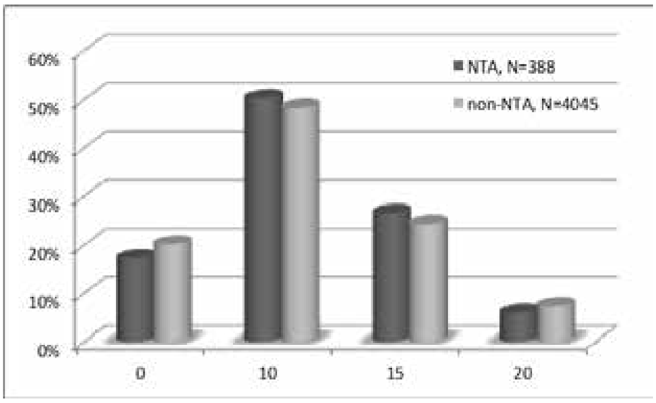
Mann-Whitney U test *rejects* equality of distributions at 5 % level of significance, *p* = 0.023

Figure 2b: Test grade distributions for NTA and non-NTA individuals, raw data, ***biology***



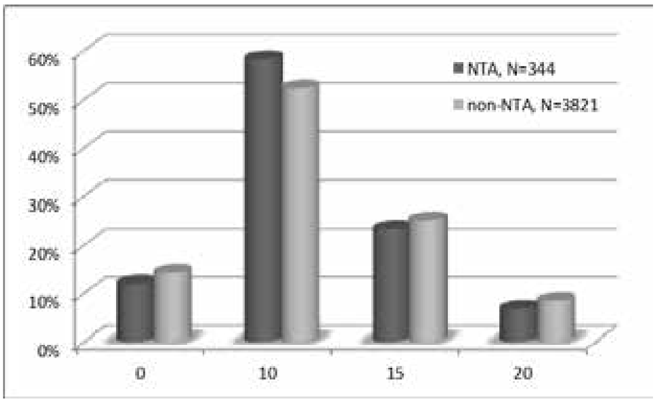
Mann-Whitney U test *rejects* equality of distributions at 1% level of significance, *p* = 0.000

Figure 2c: Test grade distributions for NTA and non-NTA individuals, raw data, ***chemistry***



Mann-Whitney U test *does not reject* equality of distributions at 5 % level of significance, *p* = 0.458

Figure 2d: Test grade distributions for NTA and non-NTA individuals, raw data, ***physics***



Mann-Whitney U test *does not reject* equality of distributions at 5 % level of significance, *p* = 0.551

Note: 0 = Fail (F), 10 = Pass (P), 15 = Pass with Distinction (PD), 20 = Pass with Special Distinction (PSD)

Consistent with the test score distributions, there are no significant differences between the test grade distributions in *chemistry* and *physics*.

4.4.4 Course grades

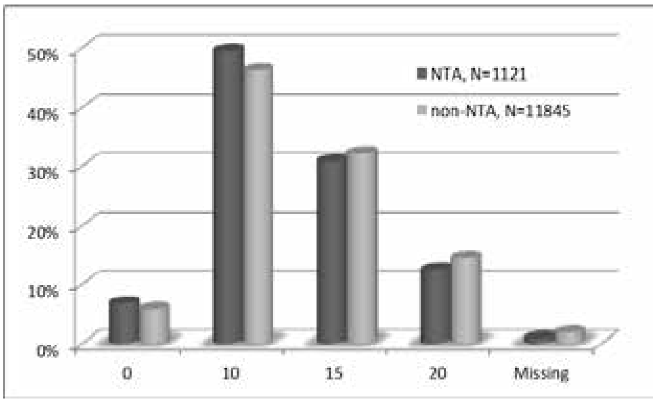
A student's course grade should take his/her grade on the standardized test into account and, in addition, reflect his/her results on other written and oral tests given during grade 9, as well as his/her performance in class and during lab experiments.

Figure 3a-d show that the tendency towards better results for non-NTA students that we observed with respect to the test scores and test grades is even more marked for the course grades.[12] Compared to the corresponding non-NTA course grade distributions, the NTA distributions exhibit *lower* frequencies for the grades pass with distinction (= 15) and pass with special distinction (= 20) in all but one case, namely Figure 3c, showing the course grades for *chemistry*. The Mann-Whitney tests support the visual impression: with the exception of *chemistry*, the hypotheses that the NTA and non-NTA course grade distributions are equal are decisively rejected.

[12] Figures 3a-d include a missing category. It would seem that this category should contain individuals that have not received course grades due to, e.g., to insufficient attendance at lectures and/or absence at written or oral tests beside the national standardized test. However, researchers at Umeå university have told us that the most likely reason simply is that some teachers forgot to fill in grades for some students. In any case, the students with missing course grades – which happen to be non-NTA individuals only – are not included in the empirical analysis reported in Section 6.

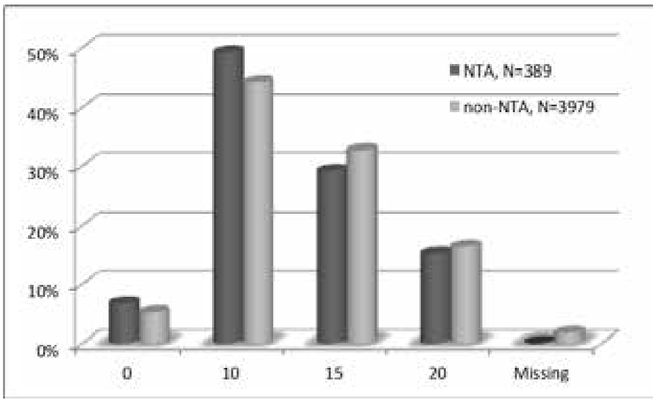
Figure 3

Figure 3a: Course grade distributions for NTA and non-NTA individuals, ***all subjects***



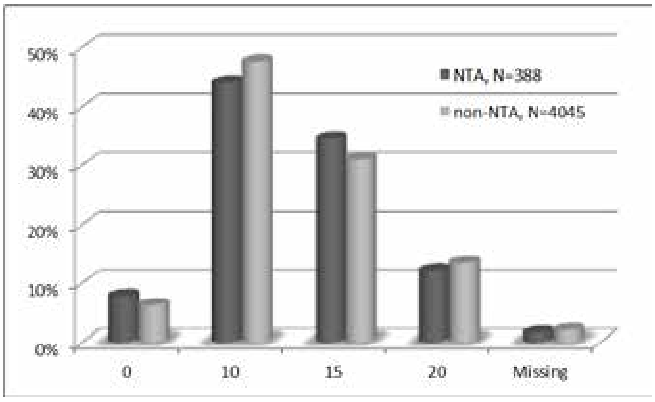
Mann-Whitney U test *rejects* equality of distributions at 1% level of significance, $p = 0.006$

Figure 3b: Course grade distributions for NTA and non-NTA individuals, ***biology***



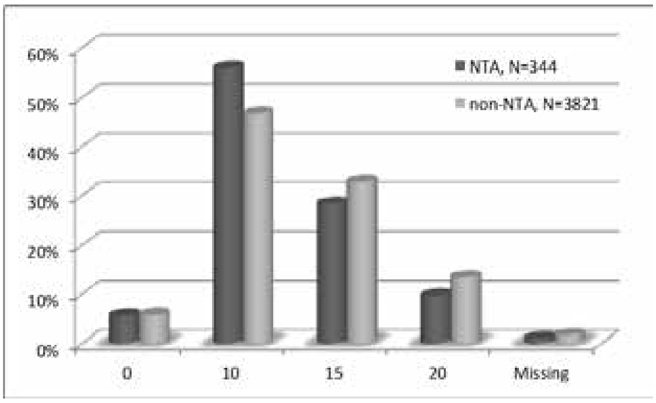
Mann-Whitney U test *rejects* equality of distributions at 5% level of significance, $p = 0.043$

Figure 3c: Course grade distributions for NTA and non-NTA individuals, ***chemistry***



Mann-Whitney U test *does not reject* equality of distributions at 5% level of significance, $p = 0.932$

Figure 3d: Course grade distributions for NTA and non-NTA individuals, ***physics***



Mann-Whitney U test *rejects* equality of distributions at 1% level of significance, $p = 0.004$

Note: 0 = Fail (F), 10 = Pass (P), 15 = Pass with Distinction (PD), 20 = Pass with Special Distinction (PSD)

4.4.5 Background characteristics of NTA and non-NTA individuals

In Table 4, NTA and non-NTA participants in the original sample are characterized in terms of individual level, school level[13], and municipality level data. Our individual level data come from the sample collected by Umeå university, while the school level and municipality level data have been obtained from a publicly available database administered by the Swedish association of local authorities and regions.

The last two columns of Table 4 show the mean differences between the groups of the NTA participants and non-NTA participants, respectively, with the corresponding standard deviations. In the column showing the mean differences we have also indicated if these differences are significant at the 10, 5, or 1 percent level of significance.

At the individual level (Panel A in Table 4), there are no differences between the NTA and the non-NTA participants that are significant at the 5 percent level of significance. This is in line with our discussion in Section 2.2 about the selection into the NTA program not taking place at the individual level. There is, however, a weak indication of boys being overrepresented in the NTA group; this difference is significant at the 10 percent level.

At the school and municipality level (Panels B and C, respectively, in Table 4), where the selection into NTA should occur, according to Section 2.2, several significant differences between the NTA and the non-NTA participants can be noted.

With respect to the school level, the two groups differ at the 1 percent of level of significance regarding the share of boys, the average educational level of the students' parents, the share of foreign-born students, and the share of grade 9 students with at least grade pass in all subjects. At the municipality level, the differences in population, median net earnings and incomes, average costs spent on grade 1-9 students, and the share of students with at least grade pass in all subject are all significant at the 1 percent level.

[13] The school level characteristics refer to the school that the pupils attended when they did the national tests in sciences, in grade 9. While it is natural to connect test results in a given school with the properties of the same school, the choice is less obvious when it comes to modeling the selection into the NTA program. For the latter purpose, the properties of the school that the pupils attended in grade 6 would have been more appropriate, as the majority of the pupils are enrolled in the NTA program only up until the 6th grade, cf. Section 4.2. Unfortunately, school level data are available for grade 9 only. However, this need not be a big problem, for three reasons. First, as noted in section 4.2, about a third of the NTA participants were enrolled in the NTA program also in grades 7-9. Secondly, some of the pupils that participated in the NTA program up to grade 6 attended the same school in grade 9 as in grade 6. Finally, among the pupils that participated in NTA until grade 6 and switched school between grade 6 and grade 9 quite a few switched to the same grade 9 school, implying that their (and their parents') characteristics will be reflected in the school level variables referring to the grade 9 school.

Table 4: Background characteristics of the NTA and non-NTA individuals, raw data

	NTA participants (N = 1 121) [1]		**non-NTA individuals (N = 11 845)** [1]		**Mean Differences**	
Variable	**Mean**	**Std. dev.**	**Mean**	**Std. dev.**	**Diff.**	**Std. dev.**
A. *Individual-level*						
Sex; female = 1	0.470	0.499	0.500	0.50	- 0.030*	0.016
Taking first language classes[2]	0.110	0.313	0.110	0.311	0.000	0.010
B. *School-level* [3]						
Share of boys	0.518	0.040	0.509	0.052	0.009***	0.002
Parents' educational level[4]	2.134	0.148	2.176	0.199	- 0.042***	0.006
Share of foreign-born students	0.065	0.061	0.074	0.083	- 0.009***	0.003
Share of students with foreign-born parents	0.062	0.096	0.061	0.085	0.001	0.003
Share of grade 9 students with at least pass in all subjects	0.740	0.100	0.775	0.104	- 0.035***	0.003
C. *Municipality-level* [5]						
Population in millions	0.081	0.135	0.140	0.211	- 0.058***	0.006
Share of population in densely populated areas[6]	0.844	0.113	0.836	0.139	0.008*	0.004
Median net earnings and incomes, millions of SEK[7]	0.180	0.023	0.177	0.019	0.003***	0.001
Share of teachers with tertiary pedagogical exam[8]	0.842	0.055	0.842	0.052	0.000	0.002
# students per 100 teachers[9]	0.124	0.008	0.124	0.008	0.000	0,000
Total costs for grades 1-9, millions of SEK, over # resident grade 1-9 students[10]	0. 072	0.006	0.071	0.007	0.001***	0.000
Share of grade 9 students with at least pass in all subjects	0.745	0.055	0.764	0.055	- 0.019***	0.002

Notes:

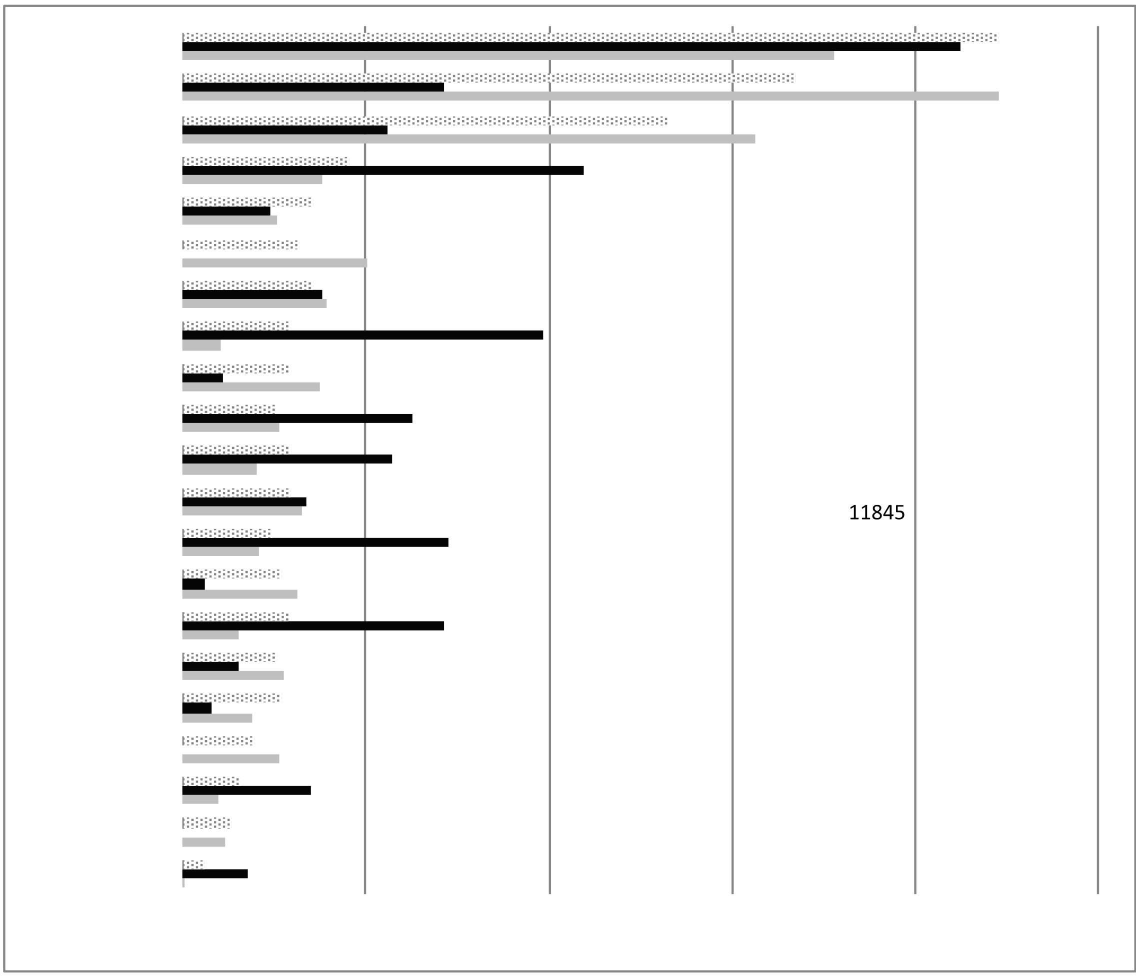
1. The number of observations differs slightly across variables, due to missing observations. For NTA participants numbers range between 1 112 and 1 121, while the range for the non-NTA participants is 11 756 to 11 845.
2. Binary indicator = 1 for students that haven't Swedish as mother tongue and attended lessons in their mother tongue in grade 9.
3. The school level variables are time averages, for the years 2004-2006 or, in a few cases, somewhat later.
4. The average of the educational levels of the student's parents, averaged over the school's students. Parents' educational levels are coded according to: 1 for 9 year compulsory school, 2 for upper secondary school, and 3 if the parent has at least a semester of tertiary schooling. If information on the education of one of the parents is missing the average level of education is set equal to the level of the parent for whom there is information available. Thus, this variable can take on the values 1, 1.5, 2, 2.5, and 3.
5. The municipality level variables refer to the year 2006.
6. Densely populated areas are defined as clusters of at least 200 inhabitants whose dwellings are not more than 200 meters apart.
7. Municipality median of earnings and incomes, net of taxes and negative transfers, for adults aged 20+.
8. The pedagogical exams considered here include pre-school teacher exams and exams for recreational pedagogues. The numbers of teachers are measured in terms of full-time equivalents.
9. The numbers of teachers are measured in terms of full-time equivalents.
10. Municipality costs are measured net of central government grants.
11. Significance levels denoted according to * = 10%, ** = 5% and *** = 1%.

It should be noted that with respect to the school-level and municipality level variables in Table 4, most of the significant differences between the NTA participants and the non-NTA individuals are quite small, both in absolute terms and relative to the corresponding mean values. Specifically, with respect to panels B and C in Table 4, the average of the absolute differences over the corresponding mean values of the NTA participants is 8.4 percent.

Geographically, there are large differences with respect to participation in the NTA program. Figure 4 provides information about the 21 Swedish counties. The counties are regional administrative entities responsible for health care and regional transport systems. The counties generally contain several municipalities. The county population shares are shown together with their shares in the NTA and non-NTA sub-samples.

Figure 4 shows that in a few counties the sample shares of the NTA and non-NTA individuals correspond quite closely to the county's share in the Swedish population, cf. the counties of Uppsala and Örebro, and Stockholm. However, given that the number of NTA participants in the sample is rather small one would in general not expect the NTA participants to be distributed across Sweden in accordance with the distribution of the Swedish population across counties. Figure 4 supports this conjecture. NTA participants are over-represented in the counties of Östergötland, Södermanland, Västernorrland, Dalarna, Kalmar, Gävleborg, Blekinge, and Gotland, and under-represented in the counties of Västra Götaland, Skåne, Halland, Värmland, Västmanland, Västerbotten, Kronoberg, and Jämtland. These over- and under-representations will be sustained when we balance our sample for the effect evaluation, as the balancing amounts to extracting a subset of the non-NTA participants that mimics the NTA participants.

Figure 4. Relative frequencies of the Swedish population and of the NTA and non-NTA, participants, respectively, in the Swedish counties, raw data



5 Methodological issues

The most important methodological issue in our empirical analysis is to account for the non-random selection into the NTA program. We do this by means of estimation of propensity scores, discussed in Section 5.1. The procedure applied to create the balanced (matched) sample by means of the propensity scores is described in section 5.2. section 5.3 considers how to allow for differences between NTA participants and NTA individuals that are not due to the selection into the program. Finally, Section 5.4 concerns the fact that some of our outcome measures – test grades and course grades – are discrete variables that only can take on a limited number of ordered values, namely *fail, pass, pass with distinction,* and *pass with special distinction*. Our estimates of the effects of NTA on grades will take this feature explicitly into account.

5.1 Sample selection and propensity scores

Participation in the NTA program is voluntary. This induces a problem with respect to the evaluation of the program's effects in so far as the characteristics of the program participants differ from the characteristics that would have been found in a group of participants randomly assigned to the NTA. Previous analyses suggest that this is indeed the case. The descriptive analysis in Svärdh (2013) indicates that NTA is over-represented among low-performing students with poor family background. It is thus conceivable that NTA has been used as a compensatory device, which would be consistent with the findings in Cuevas et al. (2005). In other words, there is reason to believe that the sample of NTA participants can be viewed as if it was *selected* in terms of certain properties, properties which are distinct from the population of grade 1-9 students at large. If so, the observed differences in test and course grades between the NTA and the non-NTA participants described in section 4 might be due to NTA and non-NTA individuals being intrinsically different, to begin with.

Ideally, to be able to infer causal effects of NTA participation we would like to compare the average results of the NTA participants with the average results they would have achieved had they not joined the program. Formally, let Y_{1i} denote the outcome of NTA participant i and denote by Y_{0i} the corresponding hypothetical result had (s)he not participated in the program. The parameter that we would like to estimate, τ, then is:

$$\tau = E(Y_{1i}) - E(Y_{0i}), \tag{1}$$

where E is the expectations operator.

Of course, the counterfactual outcome Y_{0i} cannot be observed – it is impossible to observe both Y_{1i} and Y_{0i}. Therefore, we have to resort to a second-best alternative, namely to compare participants with non-participants that as far as possible have the same properties as NTA participants, except that they did not participate in the NTA program. Propensity score matching is a method for identifying such non-participants.

The major advantage of propensity score matching is that it makes it possible to account for the fact that, in general, meaningful characterizations of individuals involve many dimensions. Examples of relevant dimensions in the present context are sex, family background, characteristics of the schools that the individual has attended (apart from whether the schools made use of the NTA program) and features of the individual’s hometown and the part of the country where that town was located. As these examples show, the dimensions may involve several levels of aggregation: the individual level, the school level, the community level and the regional level.

The first step in the construction of the propensity score is a multivariate logistic regression analysis. In this regression a binary indicator variable equal to 1 for NTA participants and 0 for non-participants is regressed on the sort of variables just discussed. The choice of variables should be guided by what is known about the enrollment in the program. Here, Section 2.2 tells us that participation in NTA is not determined by individual students (or their parents), but by schools or municipalities.

The propensity scores are simply the predicted values that the estimated model generates for the individuals included in the regression. This means that the propensity score is a weighted average of the values of the individual’s right hand side variables, the weights being given by the estimated regression parameters. The propensity score thus aggregates the multiple dimensions manifested by the regressors into a single scalar value. Being derived from a logistic regression, this scalar value will, by construction, belong to (0,1) interval, implying that it can be interpreted as the probability of participating in the NTA program. As the computation of the propensity score does not include information about whether the individual *actually* participated in the NTA program, propensity scores can be estimated for participants and non-participants alike. It has been shown that if all the variables relevant for the selection into treatment are included, then a treated (NTA) individual and a control (non-NTA) individual with equal propensity scores have the same distribution of the observed selection variables. Thus, if a NTA and a non-NTA individual have the same propensity score but different values on some of the selection variables, the latter differences are not systematic but due to chance; cf. Guo and Fraser, 2010, pp. 132-133).

When the propensity scores have been computed, the evaluation of the possible effect(s) of NTA can be conducted by comparing the average of the differences in test results between pairs of NTA and non-NTA participants that have, essentially, the same propensity score – participants and non-participants that have been matched on the propensity score. There are several ways to form the pairs of NTA and non-NTA participants, i.e. several matching procedures, cf. Section 5.2

The propensity score matching method rests on two important assumptions. The first is the so called Conditional Independence Assumption (CIA). In the present context this assumption states that conditional on a vector of *observable* covariates, **x**, enrolment in the NTA program is independent of the outcome of not participating and the outcome of participating. Formally,

$$(Y_0 ,Y_1)\ \hat{}\ \ T\text{ï}\,\boldsymbol{x} \tag{2}$$

where Y_0 and Y_1 denote outcomes of non-participation and participation, respectively, as above, the sign ^ denotes statistical independence, T is an indicator variable, equal to 1 for participants and 0 for non-participants and the sign ï denotes “conditional on”.

The interpretation of (2) is that once we control for the observable student characteristics we can treat the students as if their participation or non-participation in the NTA program was determined through a (fair) lottery, i.e. by random assignment. The key here is observability – the CIA rules out the possibility that there can be unobservable factors that matters for the participation in the NTA. In some contexts this assumption can be seriously questioned. For instance, if, hypothetically, the decision to participate in the NTA program had been taken by the individual student it is highly likely that factors like ambitions and attitudes would have mattered for enrolment. And as ambitions and attitudes often are very hard to measure they are, for practical purposes, usually unobservable. In our context, the decision to participate is, however, taken at the school or municipality level, cf. Section 2.2.

The second assumption underlying the propensity score approach is the *stable unit treatment value assumption* (SUTVA). Adapted to the present context, SUTVA requires that if a student participates in the NTA program his/her test results will not depend on i) how her/his participation in the NTA was decided and neither on ii) whether her/his fellow students participated. The first condition is essentially the same as the CIA discussed above. One interpretation of the second condition is that it rules out social interactions, cf. Heckman (2005). In our context, this requirement is somewhat odd, in the sense that single-student participation is not conceivable – either an entire class participates or no one in the class. Fortunately, this should not be a problem for our analysis, due to the fact that test-taking students in our analysis have been randomly sampled, cf. section 4. Accordingly, we can safely assume that the probability is (very close to) zero that any two of the NTA participating students in our sample have attended the same class.[14]

5.2 The matching procedure

Several matching procedures have been proposed, cf. Guo and Fraser (2010, Ch 5.4). We have used the simplest and most intuitive one, *nearest neighbor matching, without replacement,* which can be described as follows [Guo and Fraser (op.cit, p. 146)].

For each NTA participant, search for a non-NTA participant (a neighbor) whose propensity score is closest to that of the NTA participant. Once a neighbor is found the corresponding pair is set aside and the chosen non-NTA individual is removed from the set of potential neighbors, i.e. the set of all non-NTA individuals for which there are propensity scores. The process is repeated until all NTA participants have been assigned non-NTA neighbors. In the end, there be will as many pairs as there are NTA individuals.

An obvious issue here concerns the definition of close. Sometimes a constraint is imposed on the maximum distance between the treated individuals (here, the NTA participants) and the corresponding control individuals, i.e. their neighbors. One problem with this approach is that choice of the (value of) the constraint is quite arbitrary. A too "tight" constraint also runs the risk of strongly reducing the number of (pairs of) observations, which is unfortunate in studies using small samples.

5.3 Post-matching multivariate analysis

Also when the sample has been balanced with respect to participation in the NTA program, outcomes may differ across groups like, e.g., male and female students. Application of regression analysis makes it possible to control for such differences when estimating the effect of participation in the NTA program. For example, to control for gender differences with respect to (percentile ranked) test scores when estimating the effect of the NTA program, a linear regression is run with a gender dummy variable alongside with an NTA dummy variable.[15] More generally, several control variables may be included in the outcome regressions.

[14] However, the probabilities that students have attended the same school are non-negligible. This implies a problem to the extent that students not belonging to the same class but attending the same school do interact socially. While that seems quite likely, social interaction among school mates should be much less frequent and intensive than the social interaction among classmates and, hence, a less serious problem.

[15] It is important to note that this regression is run on matched data. If run on raw data, the effect estimate will be biased; due to selection bias, the NTA dummy will not be uncorrelated with the model's stochastic residual term.

Post-matching multivariate analysis can be viewed as a form of sensitivity analysis. This interpretation derives from the fact that the post-matching multivariate analysis shows if the effects are sensitive to controls for intrinsic differences between participants and non-participants that remain after the matching procedure.

5.4 Discrete outcome measures

Two of the outcome variables that we will consider are test grades and course grades. These are essentially discrete, ordinal, outcome measures. While the grades have numbers attached to them – *fail* = 0, *pass* = 10, *pass with distinction* = 15, and *pass with special distinction* = 20 – and these numbers are often added together it is in general not clear whether the grades are measured on an interval scale. If not, comparisons should not be based on differences in means but, rather, on differences in medians. The Mann-Whitney U Test employed in sections 4.4.3 and 4.4.4, in the comparisons of test grade and course grade distributions for NTA and non-NTA participants, respectively, is a non-parametric, median based, test.

Of course, the discussion in section 5.2 applies in the context of discrete outcome variables, too. To allow for the possibility that grades may differ across groups in the matched sample, control variables may be included in *ordinal* regressions, which account for the discrete and ordered nature of the grades.

The application of ordinal regressions raises two special statistical considerations, however. The first concerns the choice of functional form or, more correctly, the transformation applied to the cumulative probability corresponding to the ordinal outcome variable. When the probability distribution looks like the distributions in Figure 2 and Figure 3, with most of the mass on the low values of the outcome variable a so called *negative log-log* form is appropriate.[16] The other consideration concerns the modeling of the different values of the outcome variable, i.e. the different grades. The simplest alternative specifies that the different values of the outcome variable are all related in the same way to the treatment indicator and the control variables. Put differently, the regression model’s slope coefficients do not vary by grade. Instead, differences across grades are taken to be captured by threshold parameters that act like multiple intercepts (constants), separating the grades *fail* and *pass*, *pass* and *pass with*

[16] See, e.g., the SPSS manual, PASW Statistics Base, Chapter 18. In contrast, the standard logit form assumes that probabilities of the different outcome values are equally likely.

distinction, and *pass with distinction* and *pass with special distinction*. This so called *parallel lines* assumption is the one that we will work with. The primary reason is its simplicity: it allows the ordinal regression to be interpreted similarly to a standard multivariate regression. Moreover, it makes it possible to construct simple estimates of the *magnitudes* of the effects of NTA (if any). Specifically, we will propose estimates related to the "distances" between the different grades, as measured by differences between the threshold parameters.

6 Results

Section 6.1 reports on the logistic regression used to generate the propensity scores, the propensity scores of the NTA and non-NTA individuals, and the properties of the matched sample.

In section 6.2 we report the estimated effects of participating in the NTA program, obtained by means of the matched sample. Estimated effects are provided both for all natural science subjects taken together, and by subject.

Section 6.3 reports the results from the sensitivity analysis, i.e. the post-matching multivariate regression analysis controlling for differences between the NTA participants and the non-participants, over and above the differences captured by the matching procedure.

6.1 Propensity scores and properties of the matched sample

In section 6.1.1 the parameter estimates from the logistic regression are discussed, using the descriptive analysis in Svärdh (2013) as a background. The construction of the matched sample of NTA and non-NTA individuals by means of the propensity scores from the logistic regression is then considered, in section 6.1.2. In the final sub-section, section 6.1.3, we report the properties of the NTA individuals and the matched non-NTA nearest neighbors, in terms of individual level, school level, municipality level and regional level characteristics. The information for the two groups is structured in the same way as in section 4.4.5, making it easy to compare the properties of the matched data with properties of the corresponding raw data in Table 4 and Figure 4.

6.1.1 The logistic regression

From section 2.2 it is clear that the decision to join the NTA program is determined at the municipality and/or school level, and not by individual students (or their parents).

Therefore, we have modeled the selection into the NTA program by means of school level and municipality level variables. As the discussion in section 4.4.5 showed that there are large regional differences between NTA and non-NTA-participants (cf. Figure 4) we have also controlled for region, by means of county dummies.

The parameter estimates of our preferred logistic regression are reported in Table 5. Among the alternative regressions that we estimated this was the one yielding the largest common support region, i.e. the largest subset of the (0,1) interval containing propensity scores for both NTA participants and non-participants.

The parameter estimates in Table 5 correspond to the following partial derivative:

$$\partial \ln[P/(1-P)] \,/\, \partial x_{\mathrm{i}} \tag{3}$$

where P denotes the unknown probability of participating in the NTA program, $P/(1-P)$ being the odds ratio, and x_{i} is the variable associated with the parameter estimate. Qualitatively, the impact of x_{i} on the logarithm of the odds ratio is the same as its impact on the odds ratio itself, as the logarithmic transformation is monotonic.

In Table 5, the parameter estimates associated with the school level variables are all in agreement with the findings in the descriptive analysis conducted by Svärdh (2013), where NTA schools are compared to non-NTA schools within the same municipality, thus controlling for municipality characteristics. Accordingly, Svärdh's (op.cit.) interpretation that the NTA program may possibly be used as a device to compensate for disadvantageous learning conditions seems to apply here, as well. While this interpretation is straightforward with respect to the parameter estimates associated with the variables *parents' education level*, *share of students with foreign-born parents*, and *share of grade 9 students with at least pass in all subjects*, it needs some explanation with respect to the variables *share of boys* and *share of foreign-born students*. Regarding the result that a larger share of boys increases the likelihood that the school participates in the NTA program, we are thinking of the commonly made observation that classes with many boys are noisier and harder to manage than classes dominated by girls.[17] That a larger share of foreign-born students instead decreases the odds of the school participating in NTA does not seem unreasonable – with many foreign-born

[17] We are thus speculating that natural science experiments might catch the attention of the noisy boys and, thereby, calm them down.

students, improving the skills in natural sciences should be a second-order priority compared to learning Swedish.

With respect to the municipality level variables, we first note that the *share of grade 9 students with at least pass in all subjects* is negatively related to the probability of participation in the NTA program, just like at the school level, but the influence is much weaker at the municipality level. This is reasonable – if there is a negative impact it should be weaker at the municipality level, given that there is variation across schools.

For the municipality level variables, comparisons with the findings in Svärdh (op.cit) are less interesting than the above comparisons regarding the school level variables. The reason is that Svärdh's municipality level comparisons are unconditional, in contrast to his school level comparisons.[18] Accordingly, it should not be surprising that the results concerning municipality variables reported here, where we control for school level factors, (other) municipality variables, and fixed regional effects often differ from the corresponding finding in Svärdh (op.cit.)

Like Svärdh (op.cit.) we find that NTA participants are more likely to live in densely populated municipalities and municipalities with high median income. The latter result indicates that, *ceteris paribus*, the larger the municipality's financial resources the more likely that its schools participate in the NTA program, which stands to reason.[19]

Regarding the municipality's population, Svärdh (op.cit) observes that, on average, the NTA municipalities have considerably larger populations than non-NTA municipalities. In contrast, Table 5 says that an increase in municipality population reduces the likelihood of NTA participation. This difference can safely be attributed to the fact that an unconditional comparison of municipality population sizes will be influenced by a number of factors correlated with population size. Also, the result here agrees with the result for the municipality's median income – everything else held constant an increase in population means a reduction in per capita financial resources.

[18] As noted above, Svärdh (op.cit) compares NTA and non-NTA schools *within the same municipality*, which means that he controls for both municipality and regional characteristics.

[19] As explained in Section 2.2, participation entails some, albeit small, costs.

Table 5: Logistic regression; dependent variable 1 for NTA and 0 for non-NTA

Variable	Parameter estimate	Standard error	Significance level
School level:			
Share of boys	4.565	0.731	0.000
Parents' education level	- 0.833	0.310	0.007
Share of foreign-born students	- 3.269	0.711	0.000
Share of students with foreign-born parents	3.641	0.616	0.000
Share of grade 9 students with at least pass in all subjects	- 1.946	0.516	0.000
Municipality level:			
Population, in millions	- 4.921	0.342	0.000
Share of population in densely populated areas	2.865	0.447	0.000
Median net earnings and incomes, millions of SEK	52.223	3.105	0.000
Share of teachers with tertiary pedagogical exam	9.253	0.869	0.000
# students per 100 teachers	19.742	6.515	0.002
Total costs for grades 1-9 over # resident grade 1-9 students	114.351	8.318	0.000
Share of grade 9 students with at least pass in all subjects	- 0.112	0.010	0.000
County level:			
County dummies	Yes		

Notes:
1. For details on the variable definitions, cf. the notes to Table 4.
2. The regression's intercept has been suppressed, to increase the precision in the estimated slope coefficients.
3. To save space, the parameter estimates corresponding to the county dummies have not been included in the table. The estimates are, however, available from the authors, on request.

With respect to the variables *share of teachers with tertiary exam*, *# students per 100 teachers*, and *total municipality school costs for grades 1-9 divided by # resident grade 1-9 students*, Svärdh (op.cit.) finds no difference with respect to participation in the NTA program. However, our estimates, controlling for school level variables, other municipality level variables, and regional fixed effects, show that these factors all are significantly positively related to the likelihood of participating in the program. That higher municipality school costs increase the probability of NTA participation can be interpreted in the same way as the positive impact of an increase in the municipality median income, cf. above. Of the other two results, the positive relation between student/teacher ratio and the probability of NTA participation supports the interpretation suggested above that the NTA is used as a device to compensate for disadvantageous learning conditions. However, the result that a larger share of teachers with tertiary exam increases the likelihood of NTA participation appears to point in the opposite direction. It should be noted, though, that a high share of teachers with tertiary exams does not necessarily mean that the teachers teaching science are highly qualified in science subjects – it merely says that they are highly qualified in some subject; unfortunately there is no information about field of study.

In the estimation of the regional fixed effects the county of Stockholm, Sweden's largest county in terms of population, has been used as the reference county. The estimated county parameters (not shown in Table 5) are all negative, implying that the likelihood of participating in the NTA program is higher in the county of Stockholm than in all of the other counties, which is in line with Figure 4.

Regarding the overall properties of the regression, goodness-of-fit measures are not very meaningful as we have chosen to suppress the regression's intercept, to increase the precision of the slope estimates.[20] All goodness-of-fit measures for logistic regressions compare the estimated model with a model where the only nonzero parameter is the intercept. As we have suppressed the intercept this comparison is not informative. However, the capability of the model to predict the observed outcomes is of interest. Table 6 provides information about this property of the model.

Table 6: Numbers of observed vs predicted NTA and non-NTA individuals

Category	Indicator	Predicted outcomes			Percentage correct
		0	1	Total	
non-NTA	0	11 770	64	11 834	99.5
NTA	1	918	200	1 118	17.9
Total		12 688	264	12 952	92.4

Notes:
1. The cut-off value is equal to 0.5
2. There are missing values in both categories; 11 in the non-NTA category and 3 in the NTA category.

A first thing to note about Table 6 is that the model should not only be judged by the percentage of correct predictions. The reason is that there is a simple way to obtain a percentage of correct predictions that is quite close to number 92.4 percent given in the table: predict that none of the observed individuals participated in the NTA program – that would yield 91.4 percent correct predictions (11 834 / 12 952). However, such a model would not be very useful; it would be informative only about half of the possible range of the propensity score, i.e. the (0,5) subset of the (0,1) interval.

[20] The variability in several of the regressors is quite small, especially among the municipality level variables. This implies that these variables will be correlated with the intercept term, if that term is included, giving rise to multicollinearity and, hence, imprecisely estimated parameters.

For this reason, the model's capability to predict participation in the NTA program, i.e. to generate propensity scores > 0.5, is of particular interest. Table 6 shows that there are 264 such cases, corresponding to 2 percent of the total. Of these, 200 are correct predictions, i.e. predictions corresponding to individuals that participated in the NTA program. There are thus only 64 propensity scores larger than 0.5 corresponding to individuals that did not participate in the NTA program. This tells us that it will be rather difficult to find neighbors to NTA participants with propensity scores above 0.5.

We choose to address this problem by reducing the sub-sample of NTA participants with respect to observations with high propensity scores. A trade-off is involved here: fewer NTA observations with high propensity scores will enable better matching but it will also entail a loss of degrees of freedom and, hence, possibly less precise effect estimates.[21]

The results reported below have been obtained by means of a matched sample containing 1000 NTA participants, i.e. we have reduced the NTA sub-sample by eliminating 112 observations, or 10 percent, with propensity scores in the (0.5,1) interval. This leaves 88 NTA observations and 64 non-NTA observations in the (0.5,1) interval, enabling better (closer) matches in this range, than with the original NTA sub-sample. Essentially, this is equivalent to imposing a constraint on the maximum distance between the treated and the controls, cf. section 5.1.1. Qualitatively, the effect estimates reported below are exactly the same as the ones resulting from sample containing all the 1 112 NTA participants. However, with the smaller sample the effects are more precisely estimated (the parameter standard errors are smaller).[22] With respect to the trade-off noted in the previous paragraph, it thus appears that the improved matching dominated the loss of degrees of freedom.

6.1.2 The matching of NTA and non-NTA individuals by propensity scores

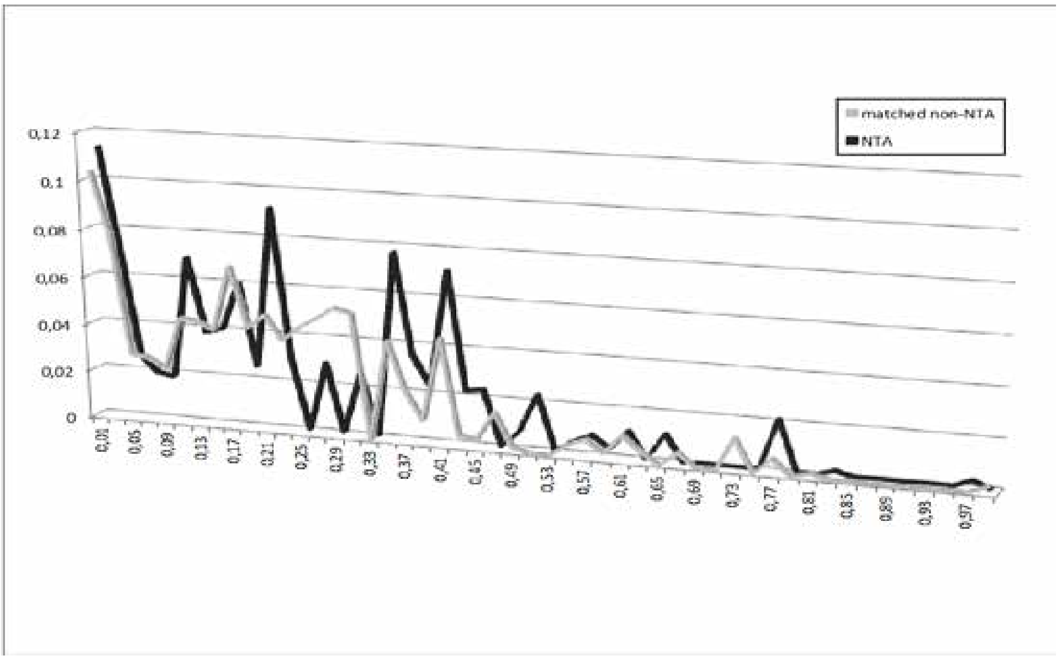
The nearest neighbor matching (without replacement) of the NTA and non-NTA individuals was carried out manually; neighbors to the NTA participants with the highest propensity scores were located first, followed by the identification of neighbors

[21] Another option is to match with replacement. We tried that, without success; as noted by Smith and Todd (2005), this approach has the drawback that it increases the variance of the effect estimator, through the creation of duplicate non-participants.

[22] The results obtained by means of the larger sample are available from the authors on request.

to NTA participants with successively lower propensity scores. The outcome of the matching process is illustrated in Figure 5.[23]

Figure 5. Propensity scores; relative frequencies for NTA and non-NTA individuals in the matched sample; 1000 pairs of NTA participants and nearest neighbors



Note: The correlation between the propensity scores of the NTA and non-NTA individuals is 0.991

Figure 5 shows that both the NTA participants and the matched non-NTA individuals effectively cover the entire (0,1) interval.[24] That is to say, the common support region is maximal. Nevertheless, the propensity scores of the NTA participants and the non-NTA individuals closely follow each other; this is clear both from visual inspection of Figure 5 and from the fact that the coefficient of correlation is 0.991 (cf. the note to Figure 5).[25]

[23] After having completed the empirical analysis, it came to our knowledge that recent versions of SPSS contain an algorithm for random matching, an approach advocated by Caliendo and Kopeinig (2008) as it makes the matching independent of the order in which the pairs of treated and controls are selected. We have compared the result of our manual matching procedure with that of the random matching procedure, accompanied by a constraint on the maximum distance between treated and controls set such that the resulting sample included almost exactly the same number of matched pairs as our sample, namely 1 002 pairs. The two matched samples are surprisingly similar. For example, disregarding the two pairs with the smallest propensity scores among the 1 002 pairs, we find that the correlation between the remaining 1 000 neighbors and the 1 000 neighbors from our manual matching procedure is 0.990. And the randomly matched pairs yield a diagram that is very hard to distinguish from Figure 5. Accordingly, we feel confident that our effect estimates do not hinge upon our chosen matching procedure.

[24] The propensity scores for the NTA individuals range between 0.002 and 0.973 while the scores for the non-NTA individuals range between 0.001 and 0.998.

[25] Given that the region of common support is allowed to be small, it is always possible to find propensity scores for the treated and non-treated individuals that are arbitrary close to one another.

6.1.3 Properties of the matched sample

We first note that with respect to the individual level characteristics that have *not* been employed in the matching procedure there are no significant differences between the NTA participants and the non-NTA individuals in the matched sample, cf. panel A in Table 7. This is what we should expect. Since we have argued, in section 2.2, that the selection into the NTA program is not affected by individual level variables, these variables should not exhibit significant differences across the two groups when selection is accounted for, just as they should not differ when selection is not accounted for.

With respect to the school level and municipality level variables (Panels B and C) that have been employed in the propensity score matching, most of the differences are still statistically significant. This means that the matched sample is not balanced with respect to the corresponding variables. However, the issue of overriding importance is whether the common support is large and the sample is well matched in terms of propensity scores. The previous subsection has shown that this is the case.

It should also be noted that, albeit significant, the mean differences in Table 7 are very small.[26] For nine of the twelve variables in panels B and C they are smaller than in the raw data in Table 4 – in spite of the fact that the differences between the unmatched variables were quite small to begin with, as noted in connection with Table 4.

A scalar measure of the decrease in the distances between the mean values of the NTA and non-NTA variables achieved by the matching is provided by the Euclidean norm, defined as the square root of the sum of the squared distances (the mean differences). When applied to the raw data mean differences in panels B and C in Table 4, the Euclidean norm is 0.0833. Applied to the corresponding matched data it is reduced to 0.0428, i.e. by almost half, or 48.6 percent.

We further note that the differences with respect to regions are much smaller in the matched sample than in the raw data.; compare Figure 6 below with Figure 4.

[26] That they, nevertheless, are significant is presumably partly due to the sample being relatively large and, thus, prone to produce significant *t*-statistics. Calculations (not reported) show that if the sample would consist of 2 × 250 observations, instead of 2 × 1000, *ceteris paribus*, only one of the differences would be significant at the 5 % level.

Table 7. Background characteristics of the NTA and non-NTA individuals, matched data

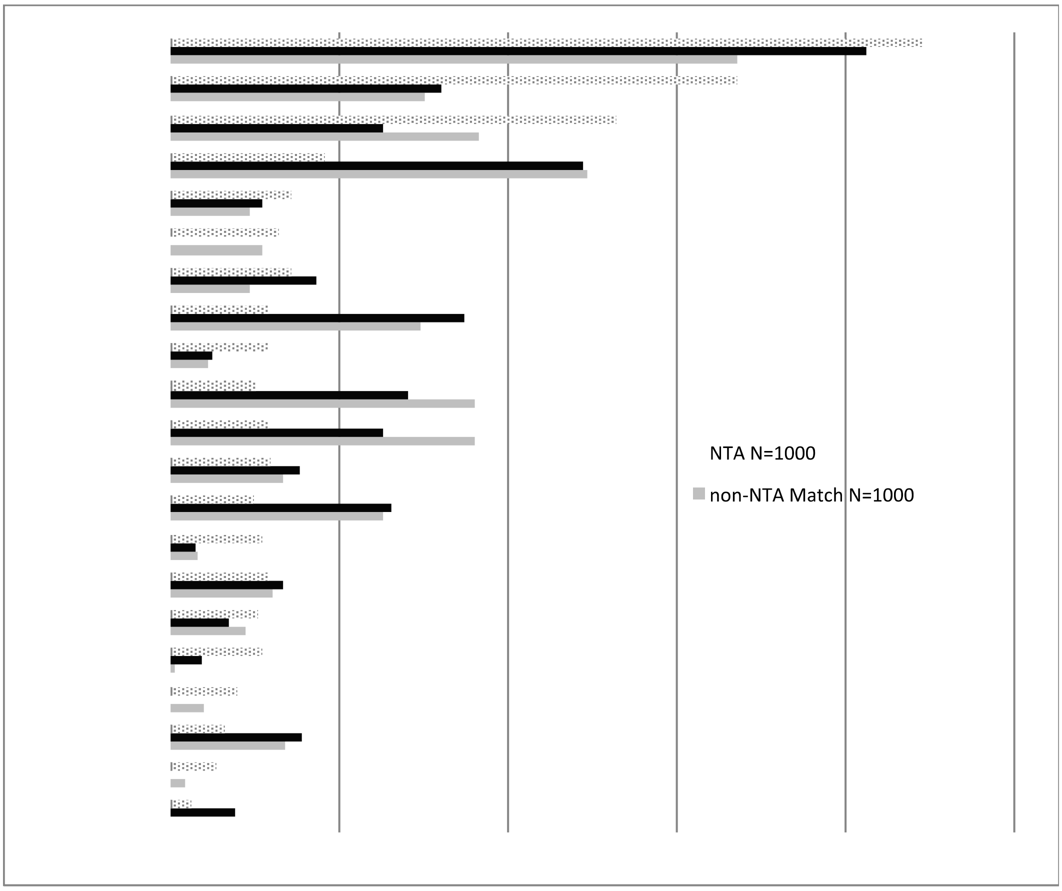
	NTA participants (N = 1 000)		**non-NTA individuals (N = 1 000)**		**Mean Differences**	
Variable	**Mean**	**Std. dev.**	**Mean**	**Std. dev.**	**Diff.**	**Std. dev.**
A. *Individual-level*						
Sex; female = 1	0.470	0.499	0.470	0.499	0.000	0.022
Taking first language classes[1]	0.120	0.324	0.130	0.338	0.010	0.015
B. *School-level* [2]						
Share of boys	0,517	0,041	0,516	0,052	0.001	0.002
Parents' educational level[3]	2.137	0.145	2.110	0.165	0.027***	0.007
Share of foreign-born students	0.066	0.062	0.063	0.073	0.003	0.003
Share of students with foreign-born parents	0.063	0.099	0.052	0.087	0.012***	0.004
Share of grade 9 students with at least pass in all subjects	0.742	0.102	0.738	0.107	0.004	0.005
C. *Municipality-level* [4]						
Population in millions	0.082	0.137	0.061	0.095	0.020***	0.005
Share of population in densely populated areas[5]	0.841	0.112	0.820	0.125	0.020***	0.005
Median net earnings and incomes, millions of SEK[6]	0.179	0.022	0.175	0.023	0.004***	0.001
Share of teachers with tertiary pedagogical exam[7]	0.843	0.057	0.849	0. 052	0.006***	0.002
# students per 100 teachers[8]	0.123	0.008	0.123	0.009	0.001*	0.000
Total costs for grades 1-9, millions of SEK, over # resident grade 1-9 students[9]	0.071	0.006	0.071	0.007	0.000	0.000
Share of grade 9 students with at least pass in all subjects	0.746	0.056	0.738	0.051	0.008***	0.002

Notes:

1. Binary indicator = 1 for students that haven't Swedish as mother tongue and attended lessons in their mother tongue in grade 9.
2. The school level variables are time averages, for the years 2004-2006 or, in a few cases, somewhat later, for data reasons.
3. The average of the educational levels of the student's parents, averaged over the school's students. Parents' educational levels are coded according to: 1 for 9 year compulsory school, 2 for upper secondary school, and 3 if the parent has at least a semester of tertiary schooling. If information on the education of one of the parents is missing the average level of education is set equal to the level of the parent for whom there is information available. Thus, this variable can take on the values 1, 1.5, 2, 2.5, and 3.
4. The municipality level variables refer to the year 2006.
5. Densely populated areas are defined as clusters of at least 200 inhabitants whose dwellings are not more than 200 meters apart.
6. Municipality median of earnings and incomes, net of taxes and negative transfers, for adults aged 20+.
7. The pedagogical exams considered here include pre-school teacher exams and exams for recreational pedagogues. The numbers of teachers are measured in terms of full-time equivalents.
8. The numbers of teachers are measured in terms of full-time equivalents.
9. Municipality costs are measured net of central government grants.
10. Significance levels denoted according to * = 10%, ** = 5 % and *** = 1%.

To compare Figure 4 and Figure 6, we apply the following arbitrary criterion: a difference between the NTA and non-NTA relative frequencies in a county is large if the non-NTA relative frequency is 1/3 larger or smaller than the corresponding NTA relative frequency. We then find that there are large differences for 17 counties in Figure 4 while the number of large differences in Figure 6 is equal to 8. Of these, 5 correspond to counties where NTA either is not represented at all (Halland, Kronoberg and Jämtland) or where NTA is used in all schools (Västerbotten and Gotland).

Figure 6. Relative frequencies of the Swedish population and of the NTA and non-NTA participants, respectively, in the Swedish counties, matched data



6.2 Effect estimates

In this section we compare the test score, test grade, and course grade distributions of the NTA participants and non-participants, respectively, in the matched sample. The comparisons are structured in the same way as the corresponding raw data comparisons in Sections 4.4.2 – 4.4.4. Thus, *t*-tests and effect sizes, and Mann-Whitney tests for equality of distributions are provided, for all subjects and by individual subjects.

Figure 7a-d show how the percentile ranked test score distributions compare in the matched sample. From Figure 7a it can be seen that for all science subject tests taken together the *t*-test rejects the null hypothesis that the means of the two distributions are equal, at the 1 % level of significance. This is due to the fact that the NTA participants have *higher* mean (percentile ranked) test scores than the matched non-NTA individuals. While it is quite clear that there is a significantly positive effect of NTA on the test scores, the size of the effect does not seem overly impressive when measured in terms of the descriptive Cohen's *d* statistic, however. As explained in note 11, this is due to the fact that Cohen’s *d* overestimates the standard error of the difference between the mean scores of the NTA participants and the non-NTA individuals.

An alternative measure of the size of the effect relates to the following thought experiment: How much higher, on average and in relative terms, would the non-NTA pupils have scored, had they participated in the NTA program? The answer is: $(m^{\text{NTA}} - m^{\text{non-NTA}}) / m^{\text{non-NTA}} = (47.90 – 44.21) / 44.21 = 8.3\%$, which certainly is non-negligible.

It is interesting to compare Figure 7a with Figure 1 a, showing the corresponding result for the raw data. In Figure 1 a the hypothesis of equality of means was rejected, too, but for a quite different reason, namely that the mean NTA test score was smaller than the mean test score of the non-NTA individuals. This goes to show that the matching procedure really matters, just as should be expected.

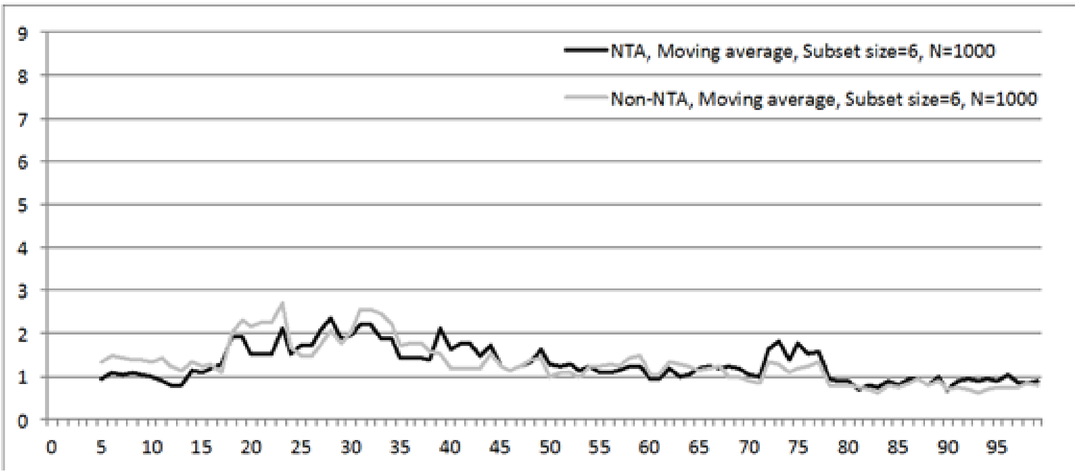
Figure 7b-d show what happens when the overall result is disaggregated by subject. From these figures it can be seen that the aggregate effect stems from a positive effect from NTA on the test scores in *physics*. According to Figure 7d, the *physics* test score frequencies of the NTA pupils are almost everywhere lower than the corresponding frequencies of the non-NTA pupils up to around the 40th percentile, while the ordering is reversed for the 40+ percentiles. Applying again the thought experiment proposed above we find that the average score on the *physics* test for a non-NTA pupil would have increased by 16.4% had (s)he participated in the NTA program, which we consider to be a large effect.[27]

No significant results are obtained with respect to *biology* and *chemistry*, cf. Figure 7b and Figure 7c. Just like for Figure 7a, the results reported in Figure 7b-d are in sharp contrast with findings based on the raw data, cf. Figure 1b-d, again stressing the importance of accounting for the non-random selection into the NTA program.

[27] $(m^{\text{NTA}} - m^{\text{non-NTA}}) / m^{\text{non-NTA}} = (48.25 – 41.44) / 41.44 = 16.4\%$

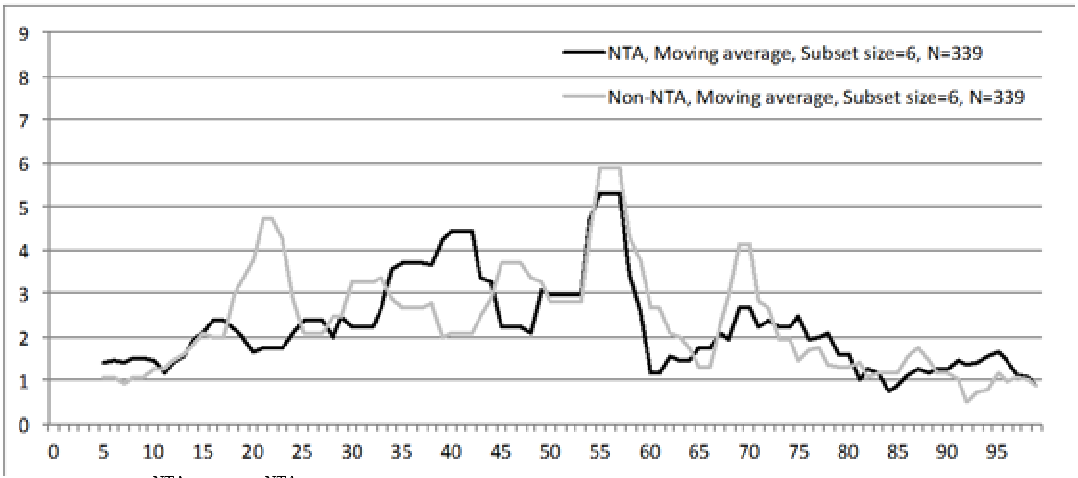
Figure 7

Figure 7 a. Frequency distribution (%) of percentile ranked test scores, by percentile; 6 percentiles moving averages, for NTA participants and non-NTA individuals, matched data, ***all subjects***



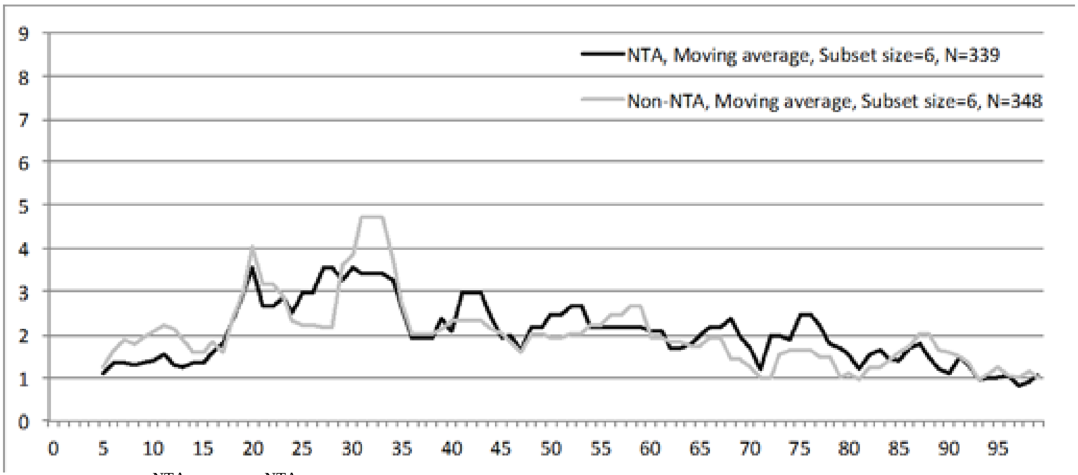
Mean difference: $m^{\text{NTA}} - m^{\text{non-NTA}}$ = 47.90 – 44.21; std. dev. (mean difference) = 1.2774; *t*-statistic = 2.89. Test for equality of means rejected at 1% level. Effect size (Cohen's *d*): 0.130.

Figure 7 b. Frequency distribution (%) of percentile ranked test scores, by percentile; 6 percentiles moving averages, for NTA participants and non-NTA individuals, matched data, ***biology***



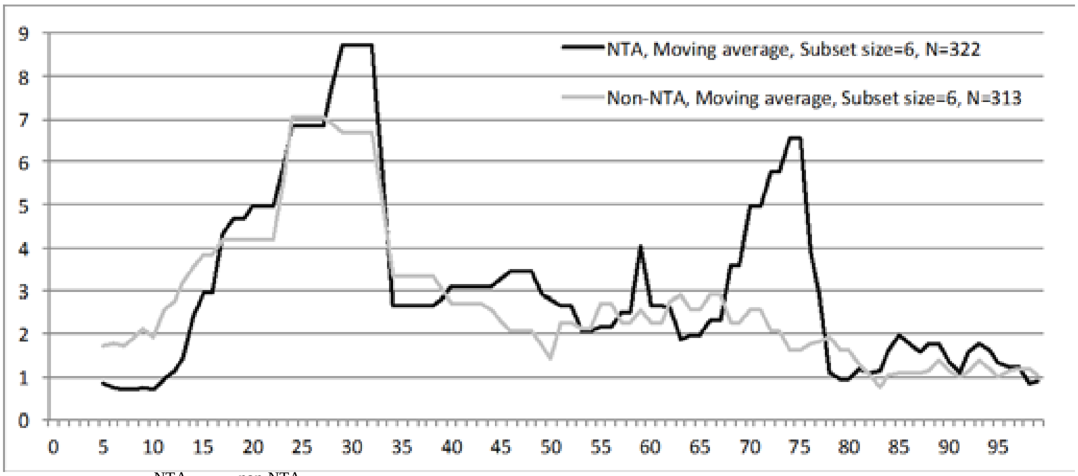
Mean difference: $m^{\text{NTA}} - m^{\text{non-NTA}}$ = 45.43 – 44.02; std. dev. (mean difference) = 2.1306; *t*-statistic = 0.66. Test for equality of means not rejected. Effect size (Cohen's *d*): 0.051.

Figure 7 c. Frequency distribution (%) of percentile ranked test scores, by percentile; 6 percentiles moving averages, for NTA participants and non-NTA individuals, matched data, ***chemistry***



Mean difference: $m^{\text{NTA}} - m^{\text{non-NTA}}$ = 50.03 – 46.90; std. dev. (mean difference) = 2.2019; *t*-statistic = 1.42. Test for equality of means not rejected. Effect size (Cohen's *d*): 0.109.

Figure 7 d. Frequency distribution (%) of percentile ranked test scores, by percentile; 6 percentiles moving averages, for NTA participants and non-NTA individuals, matched data, ***physics***



Mean difference: $m^{\text{NTA}} - m^{\text{non-NTA}}$ = 48.25 – 41.44; std. dev. (mean difference) = 2.2987; *t*-statistic = 2.96. Test for equality of means rejected at 1% level. Effect size (Cohen's *d*): 0.236.

Figure 8a-d compare the NTA individuals' test grade distributions with the test grade distributions of the non-NTA individuals in the matched sample. Qualitatively, these results are exactly the same as those shown in Figure 7a-d. Moreover, the significance levels at which equality of distributions are rejected are the same – cf. Figure 7a with Figure 8a and Figure 7d with Figure 8d – although both the nature of the outcome variable and the nature of the statistical test are different. Thus, the results based on the test scores and on the test grades corroborate one another.

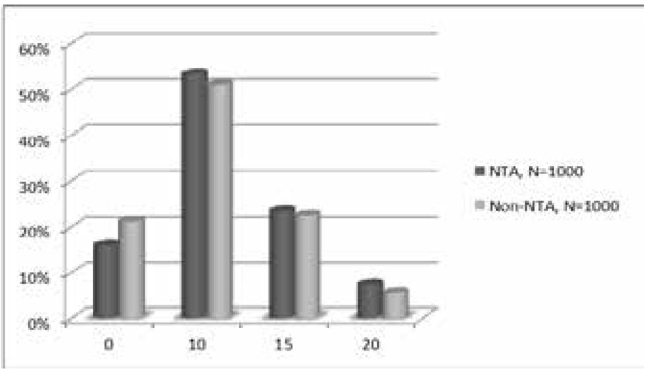
With respect to the course grade distributions the results differ markedly from the findings for the test scores and the test grades, cf. Figure 9a-d. There are no significant differences at all between the NTA distributions and the corresponding matched non-NTA distributions.[28]

In the introduction of this paper, we pointed to the fact that we have access to standardized national test data as an important advantage of this evaluation, compared to previous evaluations. The results just noted show that this feature of our evaluation is not only appealing on *a priori* grounds but of considerable empirical importance, too – had we based our assessment on course grades only our conclusions would have been entirely different, as we would have found no significant effects at all of the NTA program.

[28] Figures 9a-d are also strikingly different from the corresponding figures based on the raw, unmatched, data in Figures 3a-d. In Figure 3a-d equality of the distributions for the NTA and non-NTA individuals is rejected in three out of four cases, whereas that null hypothesis is not rejected in any of the four tests corresponding to Figure 9a-d.

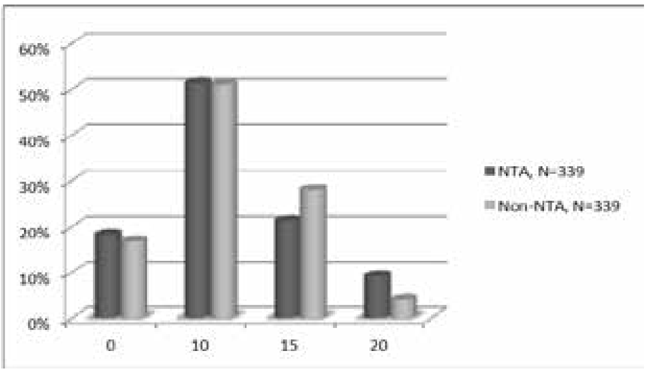
Figure 8

Figure 8a: Test grade distributions for NTA and matched non-NTA individuals, ***all subjects***



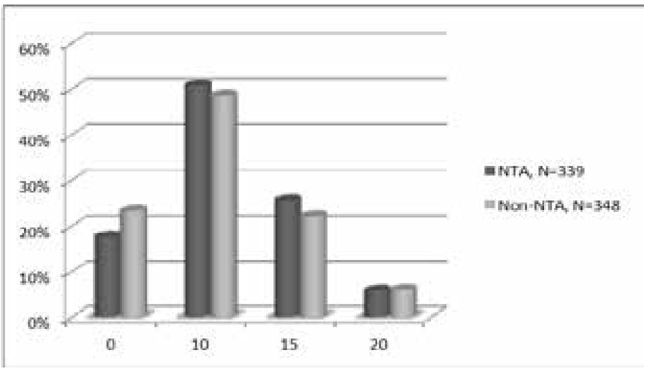
Mann-Whitney U test *rejects* equality of distributions at 1% level of significance.

Figure 8b: Test grade distributions for NTA and matched non-NTA individuals, ***biology***



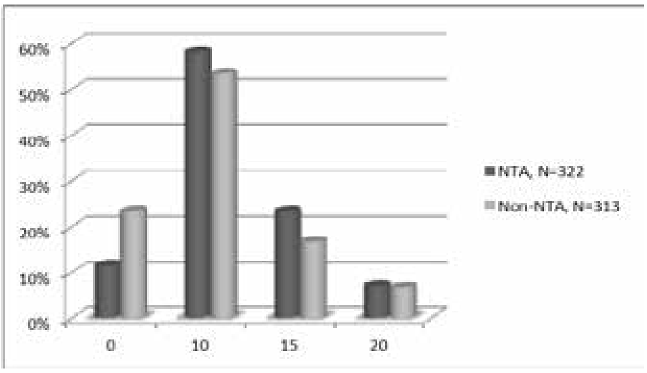
Mann-Whitney U test *does not reject* equality of distributions.

Figure 8c: Test grade distributions for NTA and matched non-NTA individuals, ***chemistry***



Mann-Whitney U test *does not reject* equality of distributions.

Figure 8d: Test grade distributions for NTA and matched non-NTA individuals, ***physics***

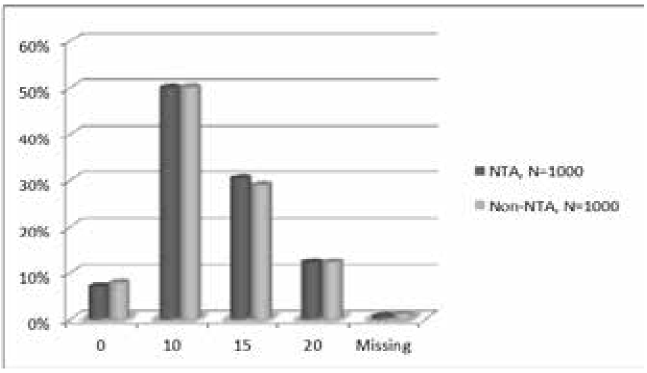


Mann-Whitney U test *rejects* equality of distributions at 1% level of significance.

Note: 0 = Fail (F), 10 = Pass (P), 15 = Pass with Distinction (PD), 20 = Pass with Special Distinction (PSD)

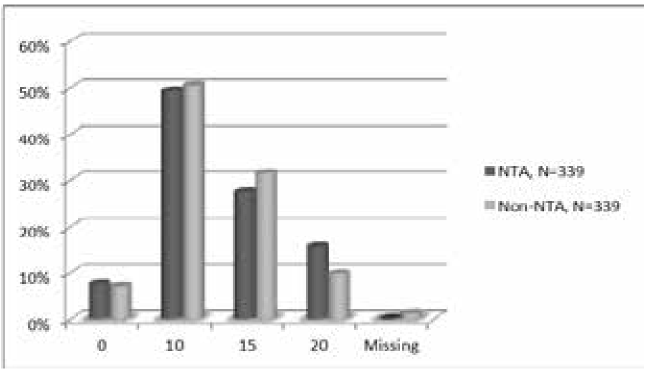
Figure 9

Figure 9a: Course grade distributions for NTA and matched non-NTA individuals, ***all subjects***



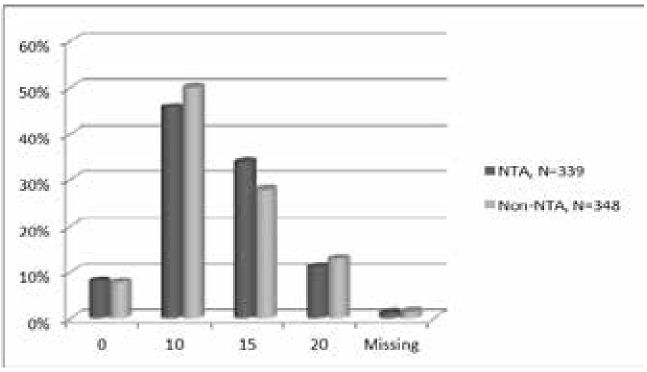
Mann-Whitney U test *does not reject* equality of distributions.

Figure 9b: Course grade distributions for NTA and matched non-NTA individuals, ***biology***



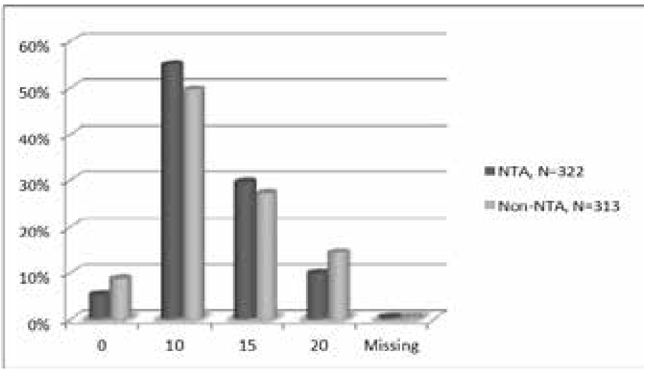
Mann-Whitney U test *does not reject* equality of distributions.

Figure 9c: Course grade distributions for NTA and matched non-NTA individuals, ***chemistry***



Mann-Whitney U test *does not reject* equality of distributions.

Figure 9d: Course grade distributions for NTA and matched non-NTA individuals, ***physics***



Mann-Whitney U test *does not reject* equality of distributions.

Note: 0 = Fail (F), 10 = Pass (P), 15 = Pass with Distinction (PD), 20 = Pass with Special Distinction (PSD)

6.3 Sensitivity analysis – adding control covariates

In this section, we report the results of post-matching multivariate analyses, structured in the same way as the non-parametric tests in the previous section. Accordingly, OLS regressions for percentile ranked test scores and ordinal regressions for test grades and course grades have been carried out for all subjects together and by individual subjects. Regarding the choice of covariates, the following two considerations have been made.

First, while individual level factors should not have impacted on the selection into the NTA program (cf. Section 2.2) and, hence, not on the estimates of the NTA effects, it is likely that they matter for the results on the standardized tests and for the course grades. If so, they should increase the precision in the estimated effects. We therefore include the two individual level variables to which we have access, namely the student's gender and whether (s)he has a foreign background. The latter is made operational by means of a dummy variable equal to 1 if the student t has attended first language classes in another language than Swedish.

Second, a time dummy, DT, has been included to account for possible differences between the years 2009 and 2010 due to changes in the schools taking the national tests and in the test-taking schools reporting their results to Umeå university; cf. section 4.1.

It turns out that the variable signaling foreign background ($D_{\text{First language class}}$) is significantly negatively related to all of the outcome variables, and across all subjects, cf. Table 8, Table 9 and Table 10. The results for the student's gender, on the other hand, show a mixed pattern. To the extent that the estimates are significant, they are (with one exception) positive, indicating that female students tend to score higher and get higher grades than male students. Regarding the issue of whether there were systematic differences between (the assessments of) the results on the national tests in 2009 and 2010, our results are inconclusive. While Table 8 shows that no significant difference is found with respect to test scores, it appears that test grades and course grades differed between the two years. In particular, *chemistry* grades were higher in 2010 than in 2009 while the opposite was true with respect to grades in *physics*, cf. Table 9 and Table 10.

Regarding the estimated effects of the NTA program, the results reported in Table 8, Table 9 and Table 10 are very close to the corresponding results in Figures 7a-d, 8a-d, and 9a-d, respectively, as expected.

With respect to the test score regressions, reported in Table 8 the results are very similar indeed to the results reported in Figure 7a-d. Thus, the estimated effects are positive and significant for all subjects taken together and for *physics* while insignificant for *biology* and *chemistry*. With respect to magnitudes, the estimated effects are somewhat, but not much, larger in Table 8. Specifically, the effect estimates in Table 8 are 3.90 and 7.46 percentiles for all subjects and for *physics*, respectively. The corresponding estimates in Figure 7a and Figure 7d are 47.91 – 44.21 = 3.69 and 48.25 – 44.41 = 6.81; cf. section 6.2. And, as expected, the OLS estimates are somewhat more precise but, again, the differences are small. For the all subjects effect the standard error of OLS estimate is 1.26 percentiles compared to 1.28 in Figure 7a, and for *physics* the standard error of the OLS estimate is 2.24 percentiles while the corresponding estimate in Figure 7d is 2.30. Accordingly, we conclude that controlling for individual-level covariates and time heterogeneity barely affects our basic effect estimates but make them marginally more precise.[29]

Table 8. OLS regressions; dependent variable: test scores

	All subjects	***Biology***	***Chemistry***	***Physics***
	(N = 2000)	(N = 678)	(N = 687)	(N = 635)
Constant	45.674*** (1.242)	42.889*** (2.075)	47.172*** (2.107)	46.937*** (2.262)
Individual level controls:				
D_{Sex} : Female = 1	0.565 (1.256)	6.220*** (2.087)	– 0.185 (2.176)	– 4.445** (2.243)
$D_{\text{First language class}}$: Taking class = 1	– 11.577*** (1.896)	– 12.948*** (2.943)	– 13.322*** (3.368)	– 7.455** (3.597)
Time effect:				
D_{T} : 2010 = 1	– 0.086 (1.258)	0.181 (2.113)	2.966 (2.173)	– 3.278 (2.237)
Effect of NTA				
D_{NTA} : Participant = 1	3.904*** (1.258)	1.148 (2.108)	2.876 (2.176)	7.461*** (2.239)

Note: Significance levels denoted according to * = 10%, ** = 5%, *** = 10%.

Turning to the ordinal regressions we first note that in these the constants are replaced by so called threshold values. These parameters acts as delimiters between the four different grades; the first threshold defines the border between *fail* (below the threshold) and *pass* (above the threshold) and so on. As noted in section 4.4.3, the grades have been assigned points (p) according to: *fail* (F) ¬ 0p, *pass* (P) ¬ 10p, *pass with*

[29] It should be noted that both the increases in the point estimates and the reduction in the standard errors contribute to higher precision, as measured in terms of the *t*-statistic.

distinction (PD) ¬ 15p, *pass with special distinction* (PSD) ¬ 20p. If these points correspond to an interval scale, which implicitly appears to be the case, then the ‘distances’ between the grades P, PD and PSD should be equally long, as the differences between the corresponding points is always 5p. This property can be checked by means of the estimated thresholds. Moreover, the potential NTA effects can be assessed in terms of the differences between the thresholds, answering the question: how much does the NTA contribute to increasing the student’s grade by, say, one level?

Consistent with the relationship between Figure 7a-d and Figure 8a-d, the test grade regressions in Table 9 are qualitatively equivalent to the test score regressions in Table 8. The only notable difference is that, except for *biology*, test grades differ between 2009 and 2010 – no such time variation was noted with respect to the test scores.

Table 9. Ordinal regressions; dependent variable: test grades

	All subjects (N = 2000)	***Biology*** (N = 678)	***Chemistry*** (N = 687)	***Physics*** (N = 635)
Threshold values:				
Fail / pass	– 0.388***	– 0.528***	– 0.208**	– 0.494***
	(0.053)	(0.091)	(0.090)	(0.096)
Pass / pass with distinction	1.215***	1.045***	1.362***	1.269***
	(0.063)	(0.105)	(0.109)	(0.114)
Pass with distinction / pass	2.864***	2.752***	3.130***	2.750***
with special distinction	(0.100)	(0.169)	(0.177)	(0.173)
Individual level controls:				
D_{Sex} : Female = 1	0.086	0.233***	– 0.003	0.030
	(0.052)	(0.090)	(0.091)	(0.094)
$D_{\text{First language class}}$:	– 0.444***	– 0.577***	– 0.496***	- 0.316**
Taking class = 1	(0.083)	(0.136)	(0.150)	(0.154)
Time effect:				
D_{T} : 2010 = 1	0.187***	0.124	0.610***	– 0.197**
	(0.052)	(0.091)	(0.092)	(0.097)
Effect of NTA				
D_{NTA} : Participant = 1	0.137***	– 0.091	0.111	0.410***
	(0.052)	(0.091)	(0.091)	(0.097)

Notes:
1. The regressions have been estimated using the log-log link.
2. Significance levels denoted according to * = 10%, ** = 5%, *** = 10%.

Regarding the threshold estimates in Table 9, the differences between adjacent thresholds are roughly equal, covering the range 1.5 – 1.8, and, thus, approximately consistent with the (differences in the) number of points assigned to the grades. Assessing the NTA effects in terms of thresholds, we can say that for *all subjects* the effect corresponds to about slightly more than 8 percent of the distance between two grade levels, while for *physics* the effect amounts to 23 percent of the distance between *pass* and *pass with distinction* and almost 28 percent of the distance between *pass with distinction* and *pass with special distinction*.

Finally, the ordinal course grade regressions, reported in Table 10, are entirely in line with the results in Figure 9a-d. That is to say, none of the estimated effects of participation in the NTA program are anywhere near significant.

Table 10. Ordinal regressions; dependent variable: course grades

	All subjects (N = 2000)	***Biology*** (N = 678)	***Chemistry*** (N = 687)	***Physics*** (N = 635)
Threshold values:				
Fail / pass	– 0.937*** (0.053)	– 0.854*** (0.091)	– 0.876*** (0.089)	– 1.126*** (0.100)
Pass / pass with distinction	0.640*** (0.057)	0.727*** (0.099)	0.661*** (0.096)	0.537*** (0.101)
Pass with distinction / pass with special distinction	2.067*** (0.078)	2.138*** (0.134)	2.144*** (0.134)	1.932*** (0.138)
Individual level controls:				
D_{Sex} : Female = 1	0.141*** (0.051)	0.300*** (0.089)	0.135 (0.088)	– 0.008 (0.092)
$D_{\text{First language class}}$: Taking class = 1	– 0.420*** (0.078)	– 0.371*** (0.125)	– 0.559*** (0.139)	– 0.360** (0.148)
Time effect:				
D_{T} : 2010 = 1	0.016 (0.051)	0.090 (0.090)	0.201** (0.088)	– 0.245*** (0.092)
Effect of NTA				
D_{NTA} : Participant = 1	0.028 (0.051)	– 0.010 (0.089)	– 0.008 (0.088)	0.088 (0.092)

Notes:
1. The regressions have been estimated using the log-log link.
2. Significance levels denoted according to * = 10%, ** = 5%, *** = 10%.

7 Summary and discussion

Given the finding in Minner et al. (2010) that out of 138 studies conducted between 1984 and 2002 almost exactly half (51 percent) showed positive and significant effects of inquiry-based instruction on student learning, there was no reason for strong *a priori* beliefs about the results of this study. Moreover, unlike most previous studies, we had access to outcome measures in the form of scores and grades from high-stakes nationwide standardized tests in natural sciences, in addition to course grades. This further added to the uncertainty of our initial expectations. And, indeed, the choice of outcome variable turned out to be important: while we find positive and significant effects of the Swedish STC program (the NTA program) on test results in grade 9, we find no statistically significant effects at all on course grades.

In addition to appropriate outcome measures, a credible effect evaluation of the program needs to properly account for the fact that the participating students constitute a selected group. In our analysis, we have firmly established that ignoring the selectivity issue can lead to very misleading results. We conjecture that this holds true for STC programs in other countries, too. For Sweden, we demonstrate that a direct (unadjusted) comparison of participants and non-participants will lead to the erroneous conclusion that the NTA program has a *negative* impact on test results as well as on course grades.

Our analysis shows that the propensity score method works reasonably well in adjusting for selectivity bias, enabling the formation of a comparable control group. The common support of the matched sample is (almost) maximal and the matched pairs are quite well balanced in terms of the propensity scores incorporating the multiple dimensions covered by our multilevel data. There are significant, but small, differences for some of the variables used to model the selection into the NTA program, however.

According to our results, the effects of NTA are not the same for different science subjects. In fact, we find significant and positive effects for test results in *physics* only. No significant effects (positive or negative) are found for *biology* and *chemistry*. It is a challenge to explain these differences.

Returning to Minner et al. (op.cit.) we note that they only find very small differences in effects across subjects. As this raises the possibility that our result may be due to circumstances that are specific to Sweden, we have discussed the matter with representatives of the NTA program in Stockholm – the largest in Sweden. According to these representatives, the NTA *physics* “boxes” are more popular among teachers and students than the *biology* and *chemistry* facilities. Speculating on the reasons why that is, we venture the guess that the NTA boxes might be better complements to the regular teaching in *physics* than is the case with respect to *biology* and *chemistry*. In *biology*, the boxes do not seem to add that much to the topics already covered by the teaching provided in most schools. Many of the experiments provided in *chemistry* boxes, on the other hand, appear to be perceived as quite abstract, especially among younger pupils. Potentially, the timing of the NTA program is better with respect to *physics* than with respect to *biology* and *chemistry*; cf. the discussion in Klahr & Nigam (2004) about successful discovery learning requiring sufficiently experienced students.

Regarding our finding that the effects are very different with respect to test results and course grades, we first note that there is no reason to expect them to be the same. Course grades should be more wide-ranging in terms of skills covered and pertaining to performance over an entire course rather than merely on a particular test at a specific point in time. Given that, the issue of teaching to the test comes to mind. Could it be that participation in the NTA program primarily fosters high performance at the national tests in science but is less efficient in advancing other relevant science skills? We do not think so. Being an inquiry-based learning support program the NTA should be relevant for a multitude of aspects on natural sciences. Instead, we believe that the discrepancies between (the effects on) test grades and course grades are in line with the phenomenon referred to as grade inflation, manifested in course grades generally being higher than test grades. Whereas the national tests in sciences have been introduced only recently, there have been Swedish national tests in math, Swedish, and English for a long time and in these subjects positive differences between course grades and test grades are clearly visible.[30] To delve deeper into this topic with respect to the natural science subjects is a task for future research, however.[31]

Another interesting topic for future research concerns potentially heterogeneous effects of NTA, across groups of individuals with different characteristics and between similar individuals that participated for shorter and longer periods in the NTA program.

[30] For a discussion between results on standardized tests and grades cf. Klapp Lekholm and Cliffordson (2008).

[31] Comparing Tables 9 and 10 above, we note, however, that the estimated threshold values for *Pass* are lower for course grades and that the distances between different grade levels are smaller for course grades than for test grades. We also note that the student's teacher is the one who marks the student's national test. That is to say, (s)he sets both the test grade and the course grade. Accordingly, it cannot be argued that the test grade is determined by someone who knows the student less well than the student's teacher.

References

Anderhag, P. and P.-O. Wickman (2007), “Utvärdering av NTA hjälper skolorna att nå kursplanemålen för femte skolåret i naturorienterande ämnen” (Evaluation of the NTA program helps schools to reach the learning objectives of the natural sciences curriculum for grade 5), Stockholm: Lärarhögskolan, UKL.

Bredderman, T. (1983). Effects of Activity-based Elementary Science on Student Outcomes: A Quantitative Synthesis. *Review of Educational Research*, *53*(4), 499–518.

Caliendo, M. and S. Kopeinig (2008), “Some practical guidance for the implementation of propensity score matching”, *Journal of Economic Surveys*, Vol. 22(1), 31-72.

Cuevas, P., O. Lee, J. Hart, and R. Deaktor (2005), “Improving science inquiry with elementary students of diverse backgrounds”, *Journal of Research in Science Teaching*, 42(3), 337-357.

Geier, R., P.C. Blumenfeld, R.W. Marx, J.S. Krajcik, B. Fishman, E. Soloway, and J. Clay-Chambers (2008), “Standardized test outcomes for students engaged in inquiry-based science curricula in the context of urban reform”, *Journal of Research in Science Teaching*, 45(8), 922-939.

Gou, S. and M.W. Fraser (2010), *Propensity Score Analysis: Statistical methods and applications*, Advanced Quantitative Techniques in the Social Sciences 11, Los Angeles, Sage Publications.

Hattie, J. (2009), Visible learning. A synthesis of over 800 meta-analyses relating to achievement, New York, Routledge.

Heckman, J.J. (2005), “The scientific model of causality”, *Sociological Methodology*, 35, 1-97.

Hmelo-Silver, C., R.G. Duncan, and C.A. Chinn (2007), “Scaffolding and Achievement in Problem-Based and Inquiry Learning: A Response to Kirschner, Sweller, and Clrak, *Educational Psychologist*, 42(2), 99-107.

Kalyuga, S., P. Chandler, J. Tuovinen, and J. Sweller (2001), ”When problem solving is superior to studying worked examples”, *Journal of Educational Psychology*, 93, 579-588.

Kenny, D.A. (1987), Statistics for the social and behavioral sciences, Boston, Little, Brown.

Klapp Lekholm, A. and C. Cliffordson (2008), ”Discrepancies between school grades and test scores at individual and school level: effects of gender and family background”, *Educational Research and Evaluation*, 14, 181-199.

Kirschner, P.A., J. Swellet, and R.E. Clark (2006) “Why minimal guidance during instruction does not work: an analysis of failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching”, *Educational Psychologist*, 41(2), 75-86.

Klahr, D. and M. Nigam (2004), “The Equivalence of Learning Paths in Early Science Instruction: Effects of Direct Instruction and Discovery Learning”, *Psychological Science,* 15(10), 661-667.

Lynch, S., J. Kuipers, C. Pyke, and M. Szesze (2005), “Examining the effects of a highly rated science curriculum unit on diverse students: Results from a planning grant”, *Journal of Research in Science Teaching*, 42, 921-946.

Mayer, R. (2004), “Should there be a three-strike rule against pure discovery learning? The case for guided methods of instruction”, *American Psychologist*, 59, 14-19.

Minner, D.D., A. Jurist Levy, J. Century (2010), ”Inquiry-Based Science Instruction – What Is It and Foes It Matter? Results from a Research Synthesis Years 1984 to 2002”, *Journal of Research in Science Teaching*, 47(4), 474-496.

Smith, J. and P. Todd (2005), “Does Matching Overcome LaLonde’s Critique of Nonexperimental Estimators?”, *Journal of Econometrics,* 125(1-2), 305-353.

Svärdh, J. (2013). To use or not to use a teacher support program - A study of what characterizes Swedish schools that apply the inquiry-based teacher support program NTA, in Skogh, I.-B. & de Vries, M. (eds.): *Technology Teachers as Researchers: Philosophical and Empirical Technology Education Studies in theSwedish TUFF Research School*. Rotterdam, Sense Publishers.

Young, B. J., & Lee, S. K. (2005). The Effects of a Kit-Based Science Curriculum and Intensive Science Professional Development on Elementary Student Science Achievement. *Journal of Science Education and Technology*, *14*(5), 471–481.

Publication series published by IFAU – latest issues

Rapporter/Reports

2015:1 Albrecht James, Peter Skogman Thoursie and Susan Vroman ”Glastaket och föräldraförsäkringen i Sverige”

2015:2 Persson Petra ” Socialförsäkringar och äktenskapsbeslut”

2015:3 Frostenson Magnus ”Organisatoriska åtgärder på skolnivå till följd av lärarlegitimationsreformen”

2015:4 Grönqvist Erik and Erik Lindqvist ”Kan man lära sig ledarskap? Befälsutbildning under värnplikten och utfall på arbetsmarknaden”

2015:5 Böhlmark Anders, Helena Holmlund and Mikael Lindahl ”Skolsegregation och skolval”

2015:6 Håkanson Christina, Erik Lindqvist and Jonas Vlachos ”Sortering av arbetskraftens förmågor i Sverige 1986–2008”

2015:7 Wahlström Ninni and Daniel Sundberg ” En teoribaserad utvärdering av läroplanen Lgr 11”

2015:8 Björvang Carl and Katarina Galic´ ”Kommunernas styrning av skolan – skolplaner under 20 år”

2015:9 Nybom Martin and Jan Stuhler ”Att skatta intergenerationella inkomstsamband: en jämförelse av de vanligaste måtten”

2015:10 Eriksson Stefan and Karolina Stadin ”Hur påverkar förändringar i produktefterfrågan, arbetsutbud och lönekostnader antalet nyanställningar?”

2015:11 Grönqvist Hans, Caroline Hall, Jonas Vlachos and Olof Åslund ”Utbildning och brottslighet – vad hände när man förlängde yrkesutbildningarna på gymnasiet?”

2015:12 Lind Patrik and Alexander Westerberg ”Yrkeshögskolan – vilka söker, vem tar examen och hur går det sedan?

2015:13 Mörk Eva, Anna Sjögren and Helena Svaleryd ”Hellre rik och frisk – om familjebakgrund och barns hälsa”

2015:14 Eliason Marcus and Martin Nilsson ”Inlåsningseffekter och differentierade ersättningsnivåer i sjukförsäkringen”

2015:15 Boye Katarina ”Mer vab, lägre lön? Uttag av tillfällig föräldrapenning för vård av barn och lön bland svenska föräldrar”

2015:16 Öhman Mattias ”Smarta och sociala lever längre: sambanden mellan intelligens, social förmåga och mortalitet”

2015:17 Mellander Erik and Joakim Svärdh ”Tre lärdomar från en effektutvärdering av lärarstödsprogrammet NTA”

Working papers

2015:1 Avdic Daniel “A matter of life and death? Hospital distance and quality of care: evidence from emergency hospital closures and myocardial infarctions”

2015:2 Eliason Marcus “Alcohol-related morbidity and mortality following involuntary job loss”

2015:3 Pingel Ronnie and Ingeborg Waernbaum “Correlation and efficiency of propensity score-based estimators for average causal effects”

2015:4 Albrecht James, Peter Skogman Thoursie and Susan Vroman “Parental leave and the glass ceiling in Sweden”

2015:5 Vikström Johan “Evaluation of sequences of treatments with application to active labor market policies”

2015:6 Persson Petra “Social insurance and the marriage market”

2015:7 Grönqvist Erik and Erik Lindqvist "The making of a manager: evidence from military officer training"

2015:8 Böhlmark Anders, Helena Holmlund and Mikael Lindahl "School choice and segregation: evidence from Sweden"

2015:9 Håkanson Christina, Erik Lindqvist and Jonas Vlachos "Firms and skills: the evolution of worker sorting"

2015:10 van den Berg Gerard J., Antoine Bozio and Mónica Costa Dias "Policy discontinuity and duration outcomes"

2015:11 Wahlström Ninni and Daniel Sundberg "Theory-based evaluation of the curriculum Lgr 11"

2015:12 Frölich Markus and Martin Huber "Direct and indirect treatment effects: causal chains and mediation analysis with instrumental variables"

2015:13 Nybom Martin and Jan Stuhler "Biases in standard measures of intergenerational income dependence"

2015:14 Eriksson Stefan and Karolina Stadin "What are the determinants of hiring? The role of demand and supply factors"

2015:15 Åslund Olof, Hans Grönqvist, Caroline Hall and Jonas Vlachos "Education and criminal behaviour: insights from an expansion of upper secondary school"

2015:16 van den Berg Gerard J. and Bas van der Klaauw "Structural empirical evaluation of job search monitoring"

2015:17 Nilsson Martin "Economic incentives and long-term sickness absence: the indirect effect of replacement rates on absence behavior"

2015:18 Boye Katarina "Care more, earn less? The association between care leave for sick children and wage among Swedish parents"

2015:19 Assadi Anahita and Martin Lundin "Tenure and street level bureaucrats: how assessment tools are used at the frontline of the public sector"

2015:20 Stadin Karolina "Firms' employment dynamics and the state of the labor market"

2015:21 Öhman Mattias "Be smart, live long: the relationship between cognitive and non-cognitive abilities and mortality"

2015:22 Hägglund Pathric, Per Johansson and Lisa Laun "Rehabilitation of mental illness and chronic pain – the impact on sick leave and health"

2015:23 Mellander Erik and Joakim Svärdh "Inquiry-based learning put to test: long-term effects of the Swedish science and technology for children program"

Dissertation series

2014:1 Avdic Daniel "Microeconometric analyses of individual behaviour in public welfare systems"

2014:2 Karimi Arizo "Impacts of policies, peers and parenthood on labor market outcomes"

2014:3 Eliasson Tove "Empirical essays on wage setting and immigrant labor market opportunities"

2014:4 Nilsson Martin "Essays on health shocks and social insurance"

2014:5 Pingel Ronnie "Some aspects of propensity score-based estimators for causal inference"

2014:6 Karbownik Krzysztof "Essays in education and family economics"