Inquiry-based learning put to the test: Medium-term effects of a science and technology for children programme

Erik Mellander\textsuperscript{1,*} and Joakim Svärdh\textsuperscript{2}

\textsuperscript{1}Institute for Evaluation of Labour Market and Education Policy (IFAU), Uppsala, Sweden, \textsuperscript{2}KTH, Education and Communication in Engineering Science (ECE), Stockholm, Sweden

We evaluate the effects of participation in the Swedish version of the Science and Technology for Children Program on content and process skills in sciences, in grade 9. The Swedish version, called Natural Sciences and Technology for All (NTA), is predominantly employed in grades 1–6. Our outcome measures are scores and grades on nationwide standardised tests, and course grades, in biology, chemistry and physics, for the years 2009 and 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 programme is addressed by propensity score analysis. The matched sample has almost maximum common support and is well behaved in terms of propensity scores. Accounting for selection is shown to be very important. We find significantly positive effects on national test scores (effect size 0.24) and national test grades for physics, but not for biology and chemistry. With respect to course grades, we find no significant effects at all. We consider explanations for the differences in the estimated effects across science subjects and between types of outcome variables, i.e. national standardised tests versus course grades.

Introduction

Inquiry-based learning has its roots in constructivist learning theories, according to which learning is a process requiring active engagement of the student. Specifically, knowledge is constructed through experiments, problem solving and discussion (cf., e.g. James, 2006; Cakir, 2008).

This paper aims to contribute to an empirical strand of the literature on inquiry-based learning, namely the literature on quantitative evaluations of the effects of inquiry-based programmes employing so called science kits. We make five contributions to this literature.

First, to the best of our knowledge, this is the first study conducted on a nationally representative sample of students, namely Swedish students in grade 9, the final year of compulsory school, in which the students generally reach 15 years of age. The value of this contribution is emphasised by the fact that the Science and Technology for Children (STC) programme, whose effects we evaluate, is a very international...
phenomenon. Originally developed in the USA, different versions of the programme today exist in a number of countries such as, e.g. Brazil, Croatia, Germany, Korea, Panama, and Thailand, as well as in Sweden.\(^1\)

Second, despite the internationally widespread use of the STC programme our study appears to be the first large-scale, quantitative, evaluation performed outside the USA.\(^2\)

Third, among the analyses using matching techniques to adjust for pre-intervention differences between treated and non-treated, our analysis handles non-random programme selection more efficiently than the predecessors. It does so by making use of individual-level, multi-dimensional matching in the form of propensity score matching (PSM), cf. Guo and Fraser (2010). Earlier matching approaches have been carried out at the school level and have concerned only a limited number of dimensions.

Fourth, this study is the first to estimate medium-term effects—around three years after treatment—rather than effects immediately after treatment. As the medium-term effects account for potential attenuation of the (initial) effects over time, the distinction between medium-term and short-term effects is an important one.

Fifth, we report effect estimates based on two different outcome measures: results on high-stakes, national standardised tests, on the one hand, and course grades, on the other hand. Accordingly, our analysis is not subject to the critique by Geier et al. (2008) that ‘The lack of student-level distal\(^3\) standardised test data to demonstrate achievement gains from standards-based inquiry science curricula remains a weakness in the literature.’ Furthermore, we are able to investigate the practical importance of this critique, in the Swedish context. We conclude that it does matter. For the standardised test results we find positive and statistically significant effects of the Swedish STC programme, whereas when we employ course grades as outcome variable we find no statistically significant effects at all.

The paper unfolds as follows. In the next section we describe the Science and Technology for Children programme—both the US original and the Swedish version. The third section provides a literature review. In the following two sections the data we use are discussed and methodological issues are considered. Our empirical analyses are reported in the penultimate section, beginning with our propensity score matching estimates, and followed by our estimated effects of participation in the Swedish Science and Technology for Children programme. We also discuss possible explanations to the differences in results when national standardised tests and course grades are employed as outcome measures, respectively, and consider whether our findings are sensitive to the inclusion of control covariates. Finally we summarise our analyses and discuss the heterogeneity in our results, with respect to different science subjects and (choices of) outcome measures, following which, we provide some suggestions for further research.

**The Swedish Science and Technology for Children programme**

The Swedish Science and Technology for Children programme is an inquiry-based programme that contains teaching materials as well as teacher instructions, including teacher training sessions. In Sweden, it is called Natural Science and Technology for All (NTA). It is intended for compulsory school, meaning grades 1–9, with a focus on

grades 1–6. As the normal school starting age is 7 this means that the NTA participants are predominantly 7–12 year olds, but can be up to 15 years old.

*The US original and the Swedish version, the NTA programme*

The NTA programme is a less extensive version of the US STC program. Precursors to the STC program date back to the 1960s. Today’s 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 includes teaching materials in the form of ‘kits’ as well as teacher training and support for school leaders. The program is divided into two parts. The first part concerns teaching from pre-school up to 5th grade. The second part is developed for grades 6–8. The kits are structured by subject and age. The subjects are thematically organised into life sciences, earth sciences, and physical science and technology.

The NTA programme was introduced in 1997. The Swedish Royal Academy of Sciences (KVA) and the Swedish Royal Academy of Engineering Sciences (IVA) translated the STC and adapted it to the Swedish curriculum. The NTA programme primarily addresses grades 1–6, but has also been used in grades 7–9. It does not provide as many kits as the STC—about 20, compared with over 30 in the STC. No school leader support is offered within the NTA programme and the teacher training supplied is somewhat less extensive than in the STC program. However, the utilisation of each kit (theme) is preceded by a teacher training session. The kit is then used for a semester, according to a template: the children formulate a hypothesis, conduct experiments, analyse the results and, finally, document their work in writing. Around the time period that we study, in 2011, NTA was used by around 7000 teachers in a third of Sweden’s 290 municipalities. In total, about 100,000 students were involved in the programme, or 1 out of 8 students in grades 1–9.

*Joining the NTA programme*

In order to appropriately evaluate the effects of the NTA it is crucial to know how students are selected into the programme. Both public and private schools can join the NTA. Almost 11% of the Swedish pupils in grades 1–9 attended private schools in the school year 2009/2010.

The first step in joining the NTA is to become a member of the NTA support organisation. With respect to public schools, this decision is taken by the municipality. For private schools, the decision is taken by the school itself. Membership entails an entrance fee of approximately US$2200. Private schools have to stand the entire fee themselves. For public schools the cost may be lower than US$2200, as the municipality’s entrance cost may be shared among several schools.

The second step is to decide if the school(s) should utilise the NTA programme. Again, the private schools decide themselves whether to do so. For public schools the decision may be taken at the municipality level or locally, by school principals or individual teachers.

When employed, the NTA programme involves a variable cost of about US$250 per kit. Moreover, a US$3.00 fee is charged per student, semester and kit. In public schools these costs are sometimes covered by the municipality, instead of by the schools themselves.

To sum up: the selection into the NTA programme is not decided at the individual level but at the municipality and/or school level. Private as well as public schools can join the programme. Joining the programme is easier for private schools, but also more costly. Nevertheless, for both public and private schools the costs are negligible, compared with teacher salaries.

Quantitative evaluations of inquiry-based learning with science kits

The meta-analysis of the effects of ‘Inquiry-oriented programs with science kits’ among children in grades K–5, in Slavin et al. (2014, pp. 885–890) provides an excellent starting point for this literature review. However, we want to complement their analysis with results concerning children in grades 6–8, for two reasons.

The first reason is that, while the focus of NTA programme is on grades 1–6, it is also available for grades 7–9 (cf. the previous section). The second reason is that there is research indicating that the effects of inquiry-based programmes may be contingent upon in which grades the students attend them. Specifically, Kalyuga et al. (2001) claim that problem solving becomes superior to traditional instruction only once learners have amassed a sufficient amount of experience. Thus, 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. It is conceivable that the cut-off between ‘too early’ and ‘late enough’ may be located after elementary school, i.e. after grade 5.

In the first subsection below we summarise the findings in the literature for children in grades K–5 and in the second we provide a corresponding summary for children in grades 6–8. In third subsection we describe the results in a Swedish study of the NTA programme. Being based on interviews, that study does not constitute a quantitative evaluation. That we, nevertheless, consider it is because its findings indicate that a condition concerning ‘fair evaluation’, discussed by Slavin et al. (2014), is satisfied, namely that participation in the NTA programme implies exposure to novel teaching methods but not to teaching of other topics than the students not participating in the programme were taught. The fourth subsection provides a summary of our findings.

Evaluations concerning children in grades K–5

Slavin et al. (2014) conduct a very thorough and careful ‘best-evidence’ meta-analysis of experimental or quasi-experimental evaluations of science programmes targeting children in grades K–5. In so doing, they distinguish between two categories of inquiry-based programmes: Those that include the provision of science kits and those who come without science kits. Only the former category will be considered here.

The studies examined by Slavin et al. (2014) had several characteristics in common, such as:
• The programme evaluated had to be one that, in principle, could be used in ordinary science classes.
• The studies compared children taught in classes using a given science programme with students in control classes using alternative programmes or standard methods.
• Random assignment or matching with appropriate adjustment for pre-test differences between programme participants and non-participants.
• Quantitative post-test outcome measures of science performance.

Altogether, seven different studies were evaluated, all of which were conducted in the USA. Aggregate information about these are provided in Table 1, which also includes information about a study published after Slavin et al. (2014), namely Zoblotsky et al. (2016).

It is clear from Table 1 that for children in grades K–5, the effects of inquiry-based learning with science kits are quite modest. Regarding the meta-analysis by Slavin et al. (2014), it should be noted that the upper limit for the effect size range concerns the smallest study in terms of participating students (100).4 Accordingly, it has barely affected the sample-weighted mean effect size. Among the other studies included in the meta-analysis the highest effect size recorded was +0.05. With respect to the study by Zoblotsky et al. (2016) it is noteworthy that the only significant effects concern analytical skills and process skills. This result is consistent with the underlying

Table 1. Experimental or quasi-experimental quantitative evaluations in grades K–5 of inquiry-based programmes with science kits

<table>
<thead>
<tr>
<th>Feature</th>
<th>Slavin et al. (2014): meta-analysis of 7 studies; ranges or categories</th>
<th>Zoblotsky et al. (2016)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year(s) of publication</td>
<td>1992–2012</td>
<td>2016</td>
</tr>
<tr>
<td>Design(s)</td>
<td>Random (1), Cluster random(^1) (5), Matched (1)</td>
<td>Cluster random(^1)</td>
</tr>
<tr>
<td>Duration(s) of intervention(s)</td>
<td>14 weeks–3 years</td>
<td>2.5 years</td>
</tr>
<tr>
<td>Number(s) of students</td>
<td>100–7500</td>
<td>Content skills: 4100 Process skills: 2600</td>
</tr>
<tr>
<td>Grades</td>
<td>(K–3)–5</td>
<td>3–5</td>
</tr>
<tr>
<td>Sample coverage</td>
<td>US school district—US state level</td>
<td>Three regions in three U.S. states</td>
</tr>
<tr>
<td>Effect size(s)</td>
<td>(-0.04 \text{–} +0.48; sample weighted mean = +0.02)</td>
<td>Content skills:(^2) 0.01 Analytical skills:(^3) 0.09** Process skills:(^4) 0.09**</td>
</tr>
</tbody>
</table>

Notes:
1. Cluster random is a two-step design. In the first step, pairs of schools are matched on demographic factors and/or prior performance. In the second step treatment is randomly assigned, within each matched pair.
2. Understanding of scientific facts, concepts, principles, laws, and theories.
3. Ability to think critically, conduct secondary analysis, apply learning, and construct explanations using evidence.
4. Ability to use equipment to make observations, perform investigations, and generate and analyse data.
5. *, **, *** denote the significance of the difference between the treatment and comparison groups at 10%, 5% and 1% levels, respectively.
constructivist theory: inquiry-based learning should be expected to impact more on process and analytical skills than on content knowledge (cf. the very beginning of the introductory section). Still, the effect sizes are not impressive—in terms of the categories suggested by Hattie (2009) they do not even qualify as small.5

Evaluations concerning children in grades 6–8

For evaluations of inquiry-based learning with science kits among children in grades 6–8 (middle grades) there are no meta-analyses available. While meta-analyses exist of studies concerning inquiry-based learning in general, they apply less strict inclusion criteria than Slavin et al. (2014) and also do not specifically identify studies employing science kits (see, e.g. Schroeder et al., 2007; Minner et al., 2010). For example, the study by Marx et al. (2004), included in the meta-analysis of Schroeder et al. (2007), concerns children in grades 6–8 and employs instruction material that potentially could be described as science kits. However, the study does not satisfy the first two of the inclusion conditions in Slavin et al. (2014) that are listed in the previous subsection.

For these reasons, Table 2 below only reports on two evaluations of inquiry-based learning with science kits among children in grades 6–8. The first study, by Lynch et al. (2005) evaluates a quasi-experiment encompassing five pairs of demographically matched schools in a large school district in the state of Maryland. A student-centred, hands-on, ‘guided inquiry’ programme in chemistry was randomly assigned to the 8th grade students in one of the two schools in each pair, resulting in

Table 2. Experimental or quasi-experimental quantitative evaluations in grades 6–8 of inquiry-based programmes with science kits

<table>
<thead>
<tr>
<th>Feature</th>
<th>Lynch et al. (2005)</th>
<th>Zoblotsky et al. (2016)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year of publication</td>
<td>2005</td>
<td>2016</td>
</tr>
<tr>
<td>Design</td>
<td>Cluster random¹</td>
<td>Cluster random¹</td>
</tr>
<tr>
<td>Duration of intervention</td>
<td>18 lessons</td>
<td>2.5 years</td>
</tr>
<tr>
<td>Number of students</td>
<td>1900</td>
<td>Content skills: 2200</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Process skills: 1500</td>
</tr>
<tr>
<td>Grades</td>
<td>8</td>
<td>6–8</td>
</tr>
<tr>
<td>Sample coverage</td>
<td>Large US school district (136,000 students)</td>
<td>Three regions in three US states</td>
</tr>
<tr>
<td>Effect size(s)</td>
<td>+0.5**</td>
<td>Content skills:² −0.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Analytical skills:³ +0.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Process skills:⁴ +0.12***</td>
</tr>
</tbody>
</table>

Notes:
1. Cluster random is a two-step design. In the first step, pairs of schools are matched on demographic factors and/or prior performance. In the second step treatment is randomly assigned, within each matched pair.
2. Understanding of scientific facts, concepts, principles, laws, and theories.
3. Ability to think critically, conduct secondary analysis, apply learning, and construct explanations using evidence.
4. Ability to use equipment to make observations, perform investigations, and generate and analyse data.
5. *, **, *** denote the significance of the difference between the treatment and comparison groups at 10%, 5% and 1% levels, respectively.

approximately 1100 treated students and 800 students in the comparison group. According to non-standardised pre- and post-tests, the treated students clearly outperformed the comparison students, the mean difference being statistically significant, with an effect size of 0.5, i.e. of medium magnitude.

A comparison of Tables 1 and 2 shows that the results in Zoblotsky et al. (2016) for children in grades 6–8 are not very different from the corresponding results for children in grades 3–5. Again, no significant effects are found with respect to science content skills, while significant, but small, effects are found for process skills. However, in contrast to the results for 3–5 graders, no significant impact on analytical skills is estimated for the 6–8 graders.

Essentially no quantitative studies conducted in Sweden

There is no previous quantitative evaluation of the NTA programme. The most quantitatively oriented study to be found is Anderhag and Wickman (2007), which documents the results of an interview study involving 80 students in grade 6. Of these 80 students, 40 were randomly chosen from classes that had participated in the NTA programme and 40 were randomly chosen from classes that had not participated.

No significant differences were found regarding the different aspects on science taught to the two groups of students. However, for several aspects, the students that had participated in the NTA programme could recollect significantly more (science) concepts and processes than the students that had not participated in the NTA programme. The authors interpreted this finding as indicating that the NTA students had gained a ‘deeper’ (but not broader) knowledge of science.

Summary of the literature review

The above discussion gives rise to three conclusions. The first is that, overall, the effects of inquiry-based learning with science kits on student skills appear quite modest. Among the nine studies considered only two had effect sizes above 0.4, i.e. of medium magnitude. Second, we found no distinct effect differences between children in grades K–5 and children in grades 6–8, in spite of arguments in the literature (Kalyuga et al., 2001) indicating that older children ought to benefit more than younger children from inquiry-based learning. Third, the study of Zoblotsky et al. (2016) shows that to the extent that there are positive effects, they concern analytical and process skills, as opposed to content skills. This finding agrees with the constructivist theories underlying inquiry-based learning (James, 2006; Cakir, 2008). It is noteworthy, however, that Anderhag and Wickman (2007) found that the students participating in the NTA programme were better than the comparison students at recollecting both concepts and processes.

The data

The standardised national test in natural sciences for grade 9 students was introduced in 2009. It consists of three subject tests, in biology, chemistry and physics. Each
student only takes one test; which one they take 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. In the years 2009 and 2010, researchers in the Department of Applied Educational Science at Umeå University collected random samples of the test-taking students, intended to cover 10% of the populations. In the first subsection, we consider how our data relate to the population(s) of grade 9 students and the samples drawn by Umeå University. In the following subsection we explain our definition of NTA participation. We then go on to show how the resulting subsamples of NTA and non-NTA participants are distributed across subject tests and how the corresponding test results compare. In the final subsection, we contrast the characteristics of the NTA and non-NTA participants at three levels of aggregation: the individual level, the school level and the municipality level. Our discussion is based on the ‘raw’ data, i.e. before any attempt has been made to control for the non-random selection into the NTA programme. Later, in the Results section, we provide the corresponding comparison of the NTA participants with a matched sample of non-NTA participants, constructed by means of propensity score analysis.

From the population of grade 9 students to our sample

Table 3 provides information about the populations of grade 9 students during the school years 2008/2009 and 2009/2010, the numbers of test-taking students, the random samples drawn from these students and, finally, the samples that we use in our empirical analysis.

From the table it is evident that a substantial share of the students supposed to take the test did not do so. Moreover, the share varies over time, from 25% in the school year 2008/2009 to 13% the following school year. The table also shows that the intention to extract samples corresponding to 10% of the test-taking students was not fulfilled, especially not in the school year 2009/2010.

The last column in table provides the number of students included in our analysis. That those numbers are lower than the numbers in the next to last column is due to the fact that about 18% of the students could not be classified as either NTA participants or as individuals that had not participated in the NTA programme. This issue is the topic of the next subsection.

<table>
<thead>
<tr>
<th>School year</th>
<th>Grade 9 students</th>
<th>Students taking the test (%)</th>
<th>Test-taking students sampled (%)</th>
<th>Sampled students in our study (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008/2009</td>
<td>118,032</td>
<td>88,491 (75.0)</td>
<td>8028 (9.1)</td>
<td>6570 (81.8)</td>
</tr>
<tr>
<td>2009/2010</td>
<td>113,545</td>
<td>98,848 (87.1)</td>
<td>7839 (7.9)</td>
<td>6396 (81.6)</td>
</tr>
<tr>
<td>Sum</td>
<td>231,577</td>
<td>187,339 (80.9)</td>
<td>15,867 (8.5)</td>
<td>12,966 (81.7)</td>
</tr>
</tbody>
</table>

Sources: The Swedish National Agency for Education; the Department of Applied Educational Science, Umeå University; own calculations.
Making the definition of NTA participation operational

The definition of NTA and non-NTA participation has been made operational at the school level, by means of a two-step procedure. First, information from the NTA support organisation has been used to classify Sweden’s 290 municipalities as NTA or non-NTA municipalities. The criterion employed was whether at least one of the municipality’s schools or none of its schools, respectively, had joined the NTA programme in the year 2009. This resulted in 87 out of the 290 municipalities being defined as NTA municipalities.

Second, local NTA coordinators, working in the NTA municipalities, were asked to assign the schools in their municipality that had pupils in grade 9 to one of the following four categories:

1. All of the school’s students had participated in the NTA programme at least up to the 6th grade, implying that the average pupil participated 3–4 semesters.
2. Some of the school’s students had participated in the NTA programme.
3. None of the school’s students had participated in the NTA programme.
4. 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 category (1). This results in close to 100 out of slightly more than 1400 schools with grade 9 students, or 7%, being classified as NTA schools. The classification of schools as NTA schools carries over to the classification of students: all students that in grade 9 attended a school classified as an NTA school are defined as NTA participants, yielding 1121 NTA students.

A school is classified as a non-NTA school if it belongs to category (3). In this category there are 11,845 students.

Finally, schools belonging to categories (2) and (4) have to be left out of the analysis because we cannot classify them as NTA or non-NTA schools. This group contains 2901 students.

Table 4 summarises our classifications and shows how our sample relates to the sample collected by Umeå University. It can be seen that the number in the third column in Table 4 equals the last entry in the final column in Table 3, and that the number in the last column of Table 4 equals the last entry in the next to last column in Table 3.

It should be noted that the procedure used to define students as NTA or non-NTA students is associated with a measurement error. As students may move between schools, some of the grade 9 students in a given school may deviate from the typical student in that school. In particular, the deviating students may not have participated

Table 4. The relation between our sample and the Umeå University (UU) sample

<table>
<thead>
<tr>
<th>NTA students</th>
<th>Non-NTA students</th>
<th>Our sample: NTA + non-NTA</th>
<th>Not classified students</th>
<th>NTA + non-NTA + not classified = UU sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>1121</td>
<td>11,845</td>
<td>12,966</td>
<td>2901</td>
<td>15,867</td>
</tr>
</tbody>
</table>

Source: Own calculations.

in NTA although having done so is typical of the students in the school. Similarly, if
the typical student in a school has not participated in the NTA, there may still be
some students that have participated in the programme. However, as long as the
mobility patterns between NTA and non-NTA schools are not systematic, the mea-
surement error can be treated as random. As such, it will bias the estimates of the
effects of NTA towards zero, making it more difficult to establish significant effects
than in the absence of measurement error. For this reason, if we obtain positive esti-
mated effects of NTA these can be regarded as lower limit estimates of the true
effects.

The distributions of NTA and non-NTA students over science subjects

Table 5 shows that for the sample as a whole (i.e. for the school years 2008/2009 and
2009/2010 taken together), the non-NTA participants are evenly spread across sub-
ject tests while the NTA participants are slightly over-represented in the biology and
chemistry tests and, thus, slightly under-represented in the physics test.

We now proceed to look at the test results of the NTA and non-NTA individuals,
respectively. For expositional ease, we only consider the results for the two school
years taken together. However, in the sensitivity analysis in the third subsection of the
Results section we will examine if there is a difference between the two test years.

The outcome variables

The outcome variables available to us are: scores on the nationwide standardised
tests, grades on the standardised tests and course grades. It should be remembered
that the students take the test in only one of the subjects biology, chemistry and
physics.

General information about the standardised science tests. Each test consists of two parts,
conducted separately, at different points in time. One part concerns scientific content
knowledge with a maximum test time of 120 minutes. The other part relates to scien-
tific process skills, for which the test time is limited to 90 minutes.

<table>
<thead>
<tr>
<th>Subject</th>
<th>NTA participants</th>
<th>Non-NTA individuals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>School year</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2008/2009 No. (%)</td>
<td>2009/2010 No. (%)</td>
</tr>
<tr>
<td>Biology</td>
<td>165 (31)</td>
<td>224 (38)</td>
</tr>
<tr>
<td>Chemistry</td>
<td>198 (37)</td>
<td>190 (33)</td>
</tr>
<tr>
<td>Physics</td>
<td>175 (32)</td>
<td>169 (29)</td>
</tr>
<tr>
<td>Sum</td>
<td>538 (100)</td>
<td>583 (100)</td>
</tr>
</tbody>
</table>

|             | School year     |                     |
|             | 2008/2009 No. (%) | 2009/2010 No. (%) | Sum No. (%) |
| Biology     | 2005 (33)       | 1974 (34)           | 3979 (34)   |
| Chemistry   | 2089 (35)       | 1955 (34)           | 4045 (34)   |
| Physics     | 1937 (32)       | 1884 (32)           | 3821 (32)   |
| Sum         | 6031 (100)      | 5813 (100)          | 11,845 (100) |

Sources: The Department of Applied Educational Science, Umeå University; own calculations.

© 2018 The Authors. Review of Education published by John Wiley & Sons Ltd on behalf of
The test items are divided into groups according to subject matter and degree of difficulty. With respect to subject matter, one group of questions may be about, e.g. natural science concepts, models and theories, whereas another group can relate to methods and procedures. Each group is partitioned into subgroups of increasing difficulty and the results of each of the subgroups are assessed separately.

The test scores. The test score is the sum total of the number of points assigned to the student’s answers on the two parts of the test. 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, whereas the corresponding numbers for chemistry were 44 and 43, and for physics 34 and 42. This means that to get the results comparable over time, the scores have to be normalised in some way. We have chosen the common approach of percentile ranked test scores.\footnote{Inquiry-based learning put to the test} 10

Another potential problem with the test scores concerns the assessment of the answers to individual questions. As noted in the previous subsection, the test items are ordered in groups corresponding to different degrees of difficulty. A correct answer to a question of mid-range complexity may yield the same number of points as a correct answer to a question belonging to the most difficult subgroup of questions. However, as long as the students are aware of this structure and act rationally, i.e. begin with the easier problems and consider the more complicated ones if time permits, this should not be a practical problem.

Smoothed frequency distributions of the percentile ranked test scores are provided in Figure 1(a)–(d). The original frequencies have first been aggregated by percentiles, yielding 100 frequencies. Equally weighted moving averages involving six subsequent percentiles have been then computed.\footnote{Inquiry-based learning put to the test} 11

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 a somewhat larger mass on the percentile range 30–50 than the non-NTA distribution, while the latter distribution has a slightly larger mass in the 80–100 percentile range. Nevertheless, the difference between the means of the distributions is significant at the 5\% level, indicating that non-NTA individuals do better at the test. That is to say, Figure 1(a) points to a negative impact of NTA on the results on standardised national tests in natural sciences, albeit not a substantial one. The effect size (Cohen’s $d$) is merely $-0.067$ and, thus, below what Hattie (2009, p. 9) would consider a small effect. Moreover, it should be remembered that we are considering ‘raw’ data—the non-random selection into the NTA programme has not been taken into account. This means that the comparison of the NTA and the non-NTA individuals will be influenced by many other factors beside participation in the NTA programme, and those other factors may dominate the impact of the programme.

Comparing Figure 1(a) and (b), we see that the tendency towards lower test grades among NTA students is more pronounced with respect to biology than for sciences in general. For biology, 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.

Figure 1. Frequency distribution (%) of percentile ranked test scores, by percentile; six percentiles moving averages, raw data, (A) all subjects, (B) biology, (C) chemistry and (D) physics.
For chemistry (Figure 1c), 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 the means are equal. The same is true for physics, too (Figure 1d).

The test grades. There are four test grades: Fail ($F$), Pass ($P$), Pass with Distinction ($PD$), and Pass with Special Distinction ($PSD$). These are assigned points according to: $F = 0 \, p$, $P = 10 \, p$, $PS = 15 \, p$ and $PSD = 20 \, p$. Figure 2(a)–(d) show the test grade distributions for NTA and non-NTA individuals, for all subjects taken together and by subject. When comparing the NTA and non-NTA distributions we use a non-parametric test, the Mann–Whitney U test, which compares the medians of the two distributions, rather than the means. 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).

For all subjects taken together (Figure 2a) the difference between NTA and non-NTA distributions is small, as with respect to the test scores. Similar to 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 ($= 20$). Furthermore, like the $t$-test in Figure 1(a), the Mann–Whitney U test rejects the null hypothesis that the distributions for the NTA and non-NTA individuals are equal, instead indicating that the median of the non-NTA distribution is larger than the median of the NTA distribution.

Again, in biology there is a distinct difference between the NTA pupils and the non-NTA pupils, to the advantage of the non-NTA pupils, cf. Figure 2(b). The null hypothesis that the two distributions are equal is rejected at the 1% level. Finally, also consistent with the test score distributions in Figures 1(c)–(d), there are no significant differences between the test grade distributions in chemistry and physics.

The course grades. Figures 3(a)–(d) show that the tendency towards better results for non-NTA students just observed with respect to the test scores and test grades is even more marked with respect to course grades. Compared with the 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 diagram, Figure 3(c), which shows the course grades for chemistry. The Mann–Whitney tests support the visual impression: except for chemistry, the hypothesis that the NTA and non-NTA course grade distributions are equal is decisively rejected.

Background characteristics of the NTA and non-NTA individuals

In Table 6, the NTA and non-NTA participants in the original sample are characterized in terms of individual-level, school-level, and municipality-level data. The individual-level data come from the sample collected by Umeå University and the school- and municipality-level data from the Swedish Association of Local Authorities and Regions.

Figure 2. Test grade distributions for NTA (dark) and non-NTA (light) individuals: raw data, (A) all subjects, (B) biology, (C) chemistry and (D) physics.
Mann-Whitney U test rejects equality of distributions at 1% level of significance, \( p = 0.006 \)

Mann-Whitney U test rejects equality of distributions at 5% level of significance, \( p = 0.043 \)

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

Mann-Whitney U test rejects equality of distributions at 1% level of significance, \( p = 0.004 \)

Figure 3. Course grade distributions for NTA (dark) and non-NTA (light) individuals: (A) all subjects, (B) biology, (C) chemistry and (D) physics.
### Table 6. Background characteristics of the NTA and non-NTA individuals, raw data

<table>
<thead>
<tr>
<th>Variable</th>
<th>NTA participants (N = 1121)</th>
<th>Non-NTA individuals (N = 11,845)</th>
<th>Mean differences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Std. dev.</td>
<td>Mean</td>
</tr>
<tr>
<td>A. Individual level</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex; female = 1</td>
<td>0.470</td>
<td>0.499</td>
<td>0.500</td>
</tr>
<tr>
<td>Taking first language classes(^2)</td>
<td>0.110</td>
<td>0.313</td>
<td>0.110</td>
</tr>
<tr>
<td>B. School level(^3)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Share of boys</td>
<td>0.518</td>
<td>0.040</td>
<td>0.509</td>
</tr>
<tr>
<td>Parents’ educational level(^4)</td>
<td>2.134</td>
<td>0.148</td>
<td>2.176</td>
</tr>
<tr>
<td>Share of foreign-born students</td>
<td>0.065</td>
<td>0.061</td>
<td>0.074</td>
</tr>
<tr>
<td>Share of students with foreign-born parents</td>
<td>0.062</td>
<td>0.096</td>
<td>0.061</td>
</tr>
<tr>
<td>Share of grade 9 students with at least Pass in all subjects</td>
<td>0.740</td>
<td>0.100</td>
<td>0.775</td>
</tr>
<tr>
<td>C. Municipality level(^5)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Population in millions</td>
<td>0.081</td>
<td>0.135</td>
<td>0.140</td>
</tr>
<tr>
<td>Share of population in densely populated areas(^6)</td>
<td>0.844</td>
<td>0.113</td>
<td>0.836</td>
</tr>
<tr>
<td>Median net earnings and incomes, millions of SEK(^7)</td>
<td>0.180</td>
<td>0.023</td>
<td>0.177</td>
</tr>
<tr>
<td>Share of teachers with university pedagogical exam(^8)</td>
<td>0.842</td>
<td>0.055</td>
<td>0.842</td>
</tr>
<tr>
<td>Students per 100 teachers(^9)</td>
<td>0.124</td>
<td>0.008</td>
<td>0.124</td>
</tr>
<tr>
<td>Total costs for grades 1–9, millions of SEK, over resident grade 1–9 students(^10)</td>
<td>0.072</td>
<td>0.006</td>
<td>0.071</td>
</tr>
<tr>
<td>Share of grade 9 students with at least Pass in all subjects</td>
<td>0.745</td>
<td>0.055</td>
<td>0.764</td>
</tr>
</tbody>
</table>

Notes:
1. The number of observations differs slightly across variables owing to missing observations. For NTA participants numbers range between 1112 and 1121, while the range for the non-NTA participants is 11,756–11,845.
2. Binary indicator = 1 for students that do not have 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 students’ 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 university 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+ (SEK = Swedish Krona).
8. The pedagogical exams considered here include pre-school teacher exams and exams for recreational pedagogues.
9. The numbers of teachers are measured in terms of full-time equivalents.
10. Municipality costs are measured net of central government grants.
11. *, **, *** denote the significance levels at 10%, 5% and 1%, respectively.
12. Underlined differences in the penultimate column denote differences with effect sizes (Cohen’s d) ≥ 0.20.

The last two columns of Table 6 show the mean differences between the NTA and the non-NTA participants, with the corresponding standard deviations.

At the individual level (Panel A in Table 6), no differences between the NTA and the non-NTA participants are significant at the 5% level of significance. This is consistent with our discussion in the second section about the selection into the NTA programme not taking place at the individual level.

At the school and municipality level (Panels B and C, respectively), in contrast, where the selection into NTA should occur (as mentioned in the ‘Joining the NTA programme’ subsection), several significant differences between the NTA and the non-NTA participants can be noted. For instance, with respect to the school level, differences in average educational level of the students’ parents, the share of foreign-born students, and the share of grade 9 students with at least pass grade in all subjects are significant at the 1% level. At the municipality level, differences in, for example, population, median net earnings and incomes, and average costs spent on grade 1–9 students are also significant at the 1% level.

It should be noted, however, that while many of the school-level and municipality-level differences in Table 6 are strongly significant, they are still quite small. This is demonstrated by only four of the differences having effect sizes of at least 0.2 (cf. note 12 to Table 6).

Methodological issues

The most important methodological consideration in our empirical analysis is to account for the non-random selection into the NTA programme. We do this by estimating propensity scores, cf. the first subsection below. The procedure creating the balanced (matched) sample by means of the propensity scores is described in the second subsection below. In the third subsection below we explain how we allow for differences between NTA participants and NTA individuals that are unrelated to the selection into the programme. Finally, the fourth subsection below concerns the fact that some of our outcome measures—test grades and course grades—are discrete variables.

Sample selection and propensity scores

Participation in the NTA programme is voluntary. This creates a problem for the evaluation of the programme’s effects in so far as the characteristics of the programme participants differ from the characteristics that would have been found in a group of participants randomly assigned to the NTA. Earlier analyses suggest that this is indeed the case. Svärdh’s (2013) descriptive analysis indicates that NTA is over-represented among low-performing students with poor family background. NTA might thus have been used as a compensatory device. That would be in line with the findings in Cuevas et al. (2005). Accordingly, there is reason to believe that the sample of NTA participants can be viewed as selected in terms of certain properties, properties that differ from those of 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 the previous section might simply be due to NTA and non-NTA individuals being intrinsically different.

Ideally, to infer the 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 if they had not joined the programme. Of course, that counterfactual outcome cannot be observed. 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 programme. Propensity score matching is a method for identifying such non-participants.

The main advantage of propensity score matching is that it makes it possible to account for the fact that meaningful characterisations of individuals generally involve many dimensions. The dimensions may involve several levels of aggregation: the individual level, the school level, the community level and the regional level. The choice of variables (dimensions) should be guided by what is known about the enrolment in the programme. The ‘Joining the NTA programme’ subsection of the second section tells us that participation in NTA is not determined by individual students or parents, but by municipalities and/or schools.

The first step in the construction of the propensity score is a multivariate logistic regression analysis. 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 the estimated regression parameters. The propensity score thus aggregates the multiple dimensions captured by the regressors into a scalar value. Being derived from a logistic regression, this scalar will, by construction, belong to the (0,1) interval, implying that it can be interpreted as the probability of participating in the NTA programme. As the computation of the propensity score does not involve information about whether the individual actually participated in the NTA programme, it can be estimated for participants and non-participants alike. It has been shown that if the propensity score regression contains all the variables that matter for the selection into treatment, then a treated (NTA) individual and a control (non-NTA) individual with equal propensity scores have the same distributions of the observed selection variables. This means that 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 are due to chance (Guo & Fraser, 2010, pp. 132–133).

Given the propensity scores, the possible effect(s) of NTA can be evaluated 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. There are several ways to form these pairs of NTA and non-NTA participants, i.e. several matching procedures, see the following subsection.

The propensity score matching method rests on two important assumptions. The first is the so called ‘Conditional Independence Assumption’ (CIA). This assumption states that conditional on a vector of observable covariates \( \mathbf{x} \), enrolment in the NTA programme is independent of the outcome of not participating and the outcome of participating. The implication of the CIA 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 programme was determined through random assignment. The key here is observability—the CIA rules out unobservable factors that matter for the participation in the NTA.

In some contexts, the CIA assumption can be seriously questioned. For instance, if, hypothetically, participation in the NTA programme had been determined by the individual student, factors like ambitions and attitudes would most likely have mattered for enrolment. Since ambitions and attitudes often are very hard to measure they are, for practical purposes, usually unobservable. Fortunately, the decision to participate is, however, taken at the school or municipality level, as explained in earlier.

The second assumption underlying the propensity score approach is the stable unit treatment value assumption (SUTVA). SUTVA requires that if a student participates in the NTA programme his/her test results will not depend either on (i) how her/his participation in the NTA was decided or on (ii) whether her/his fellow students participated. The first condition is essentially equivalent to the CIA assumption. Regarding (ii), one interpretation is that it rules out social interactions (Heckman, 2005). In our context, this requirement appears hard to satisfy, as 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, because the students that we study have been randomly sampled (see the previous section). We can thus safely assume that the probability is very close to zero that any of the NTA participating students have attended the same class.\textsuperscript{13}

The matching procedure

As discussed by Guo and Fraser (2010, ch. 5.4), many different matching procedures have been suggested. We have used the most simple and intuitive one, namely ‘nearest neighbour matching, without replacement’, which can be described in the following way.

For each NTA participant, we identify the non-NTA participant whose propensity score is closest to that of the NTA participant. When this non-NTA individual, the neighbour, is found, the pair is set aside. The process is repeated until all of the NTA participants have been assigned non-NTA neighbours. Eventually, there be will as many pairs as there are NTA individuals.

An obvious issue here is how to define ‘close enough’. Sometimes a constraint (a caliper) is imposed on the maximum distance between the treated individuals (i.e. the NTA participants) and the corresponding controls (i.e. the non-NTA individuals)—see Caliendo and Kopeinig (2008). We have chosen a different, but equivalent, procedure that amounts to reducing the subsample of NTA participants (cf. ‘The logistic regression’ in the next section).

Post-matching multivariate analysis

Even when the sample has been balanced with respect to participation in the NTA programme, outcomes may differ across groups like, e.g. male and female students. Regression analysis makes it possible to control for such differences when estimating the effect of participation in the NTA programme. For example, to control for gender...
differences, a linear regression can be run with test scores as the dependent variable and a gender dummy variable alongside with an NTA dummy variable as right-hand side variables. Additional control variables may also be included in the regression.

Post-matching multivariate analysis can be viewed as a kind of sensitivity analysis, in the following sense. If the propensity score matching analysis has been carried out correctly then the (point) estimates of the NTA effects should be (almost) unaffected by the inclusion of the control variables. However, the standard errors of the effect estimates may decrease, i.e. the control variables may result in the effects being more precisely estimated.

Discrete outcome measures

Two of the outcome variables, test grades and course grades, are essentially discrete outcome measures. Although the grades have been assigned values—Fail = 0, Pass = 10, Pass with Distinction = 15, and Pass with Special Distinction = 20—and these values often are added together it is in general not clear whether they correspond to an interval scale or merely an ordinal scale, the latter implying that the grades can be ordered but that the numbers assigned to them are not informative about the distance between the grades.

Of course, the discussion in the previous subsection 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 statistical considerations. The first is the choice of functional form or, more correctly, the transformation applied to the cumulative probability corresponding to the ordinal outcome variable. When the probability distributions look like the distributions in Figures 2(a)–(d) and Figures 3(a)–(d), with most of the mass on the low values of the outcome variable, a so called ‘negative log–log’ form is appropriate. The second consideration concerns the modelling of the different values of the outcome variable, i.e. the different grades. The simplest alternative assumes that the different values of the outcome variable are all related in the same way to the treatment indicator and the control variables. In other words, the slope coefficients of the regression model do not vary by grade. This means that differences across grades will be captured by three threshold parameters that separate the grades fail and pass, pass and pass with distinction, and pass with distinction and pass with special distinction, respectively. We will adopt this so called ‘parallel lines’ assumption. The reason is that it allows the ordinal regression to be interpreted similarly to a standard multivariate regression. Furthermore, it makes it possible to construct estimates of the magnitudes of the NTA effects (if any). Specifically, we will suggest estimates defined by the ‘distances’ between the different grades, as measured by differences between the threshold parameters.

Results

The first subsection below documents the logistic regression used to generate the propensity scores and the properties of the matched sample. The second subsection
reports the estimated effects of participating in the NTA programme, for all natural science subjects taken together and by subject, while the third subsection contains the results from the sensitivity analysis, i.e. the post-matching regression analysis controlling for differences between the NTA participants and the non-participants, over and above the differences accounted for by the matching procedure.

The propensity scores and the properties of the matched sample

In this subsection we first discuss the parameter estimates from the logistic regression. The construction of the matched sample of NTA and non-NTA individuals by means of the propensity scores from the logistic regression are then described. Finally, we compare the properties of the NTA individuals and the matched non-NTA nearest neighbors. The information is structured in the same way as in the earlier ‘Background characteristics of the NTA and non-NTA individuals’ subsection, making it easy to contrast the matched data with the corresponding raw data in Table 6.

The logistic regression. From the earlier ‘Joining the NTA programme’ subsection we know that the decision to join the NTA programme is determined at the municipality and/or school level. Accordingly, we have modelled the selection into the NTA programme by means of school-level and municipality-level variables. We have also controlled for regional fixed effects, by means of (county) dummy variables.

Table 7 provides the parameter estimates of our preferred logistic regression. Among the alternative regressions that we considered this one yielded the largest common support, 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 7 correspond to the following partial derivative:

\[ \frac{\partial \ln[P/(1 - P)]}{\partial x_i} \]

where \( P \) denotes the unknown probability of participating in the NTA programme, \( P/(1 - P) \) is the odds ratio, and \( x_i \) is the variable associated with the parameter estimate. Since the natural logarithm is a monotonic transformation, the impact of \( x_i \) on the logarithm of the odds ratio is qualitatively the same as its impact on the odds ratio itself.

The parameter estimates associated with the school-level variables are all in agreement with the findings in Svärdh’s (2013) descriptive analysis, where NTA schools are compared with non-NTA schools within the same municipality, i.e. controlling for municipality characteristics. Svärdh’s (2013) interpretation that the NTA programme may be used as a device to compensate for disadvantageous learning conditions, seems to apply here, too. 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. The compensatory interpretation needs some explanation with respect to the variables Share of boys and Share of foreign-born students, however.
Concerning the result that a larger share of boys increases the likelihood that the school participates in the NTA programme, what we have in mind is that classes with many boys tend to give rise to different kinds of behavioural issues to a larger extent than classes with a majority of girls. Thus, we are speculating that science experiments might catch the boys’ attention and, thereby, reduce behavioural problems.

That a larger share of foreign-born students instead decreases the probability that the school participates in NTA seems reasonable— with many foreign-born students, improving the skills in natural sciences should be a second-order priority compared with 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 participating in the NTA programme, just as at the school level, but that the influence is much weaker at the municipality level. This is to be expected—if there is a negative impact it should be weaker at the municipality level, provided that there is variation across schools.

Further, we find that NTA participants are more likely to live in municipalities with high median income and in municipalities that spend more on schooling, per student. This indicates that, ceteris paribus, the larger the municipality’s financial resources the more likely its schools are to participate in the NTA programme, which stands to

---

**Table 7. Logistic regression; dependent variable 1 for NTA and 0 for non-NTA**

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

Notes:
1. The number of observations is 12,952.
2. For details on the variable definitions (cf. the notes to Table 6).
3. The regression’s intercept has been suppressed, to increase the precision in the estimated slope coefficients.
4. 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.

Regarding the result that a larger share of boys increases the likelihood that the school participates in the NTA programme, what we have in mind is that classes with many boys tend to give rise to different kinds of behavioural issues to a larger extent than classes with a majority of girls. Thus, we are speculating that science experiments might catch the boys’ attention and, thereby, reduce behavioural problems.

That a larger share of foreign-born students instead decreases the probability that the school participates in NTA seems reasonable—with many foreign-born students, improving the skills in natural sciences should be a second-order priority compared with 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 participating in the NTA programme, just as at the school level, but that the influence is much weaker at the municipality level. This is to be expected—if there is a negative impact it should be weaker at the municipality level, provided that there is variation across schools.

Further, we find that NTA participants are more likely to live in municipalities with high median income and in municipalities that spend more on schooling, per student. This indicates that, ceteris paribus, the larger the municipality’s financial resources the more likely its schools are to participate in the NTA programme, which stands to
reason. As explained in the ‘Joining the NTA programme’ subsection, participation entails some, albeit small, costs.

Regarding the result that municipalities with large populations are less likely to employ the NTA programme than municipalities with smaller populations, it should be noted that this result does not apply to the capital, Stockholm. This is because in the estimation of the county dummies, the county of Stockholm has been used as the reference county and the estimated county parameters (not shown in Table 7) are all negative.

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 university exam increases the likelihood of NTA participation seems to go in the opposite direction. It should be noted, though, that, a high share of teachers with university exams does not necessarily mean that science teachers are highly qualified in science—it merely says that teachers in general are highly qualified in some subject.

Concerning the overall properties of the regression, all conventional goodness-of-fit measures for logistic regressions compare the estimated model with a model where the only non-zero parameter is the intercept. As we have suppressed the intercept, to increase the precision of the slope estimates, this comparison is not informative. However, the capability of the model to predict the observed outcomes is of interest. How the model fares in this respect is shown in Table 8.

A first thing to observe about Table 8 is that there is a simple way to obtain a share of correct predictions quite close to the 92.4% given in the table: predict that none of the individuals participated in the NTA programme—that would yield 91.4% correct predictions (11,834/12,952). However, such a model would obviously not be very useful.

Instead, the model’s capability to predict participation in the NTA programme, i.e. to generate propensity scores > 0.5, is of particular interest. Table 8 shows that there are 264 such cases. Of these, 200 are correct predictions, i.e. predictions corresponding to individuals that actually participated in the NTA programme. There are thus only 64 propensity scores larger than 0.5 that correspond to individuals who did not participate in the NTA programme. This means that it will be rather difficult to find neighbours to NTA participants with propensity scores above 0.5.

We choose to address this problem by slightly reducing the subsample 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, less precise effect estimates.

The results reported below have been obtained by means of a matched sample containing 1000 NTA participants, i.e. we have reduced the NTA subsample by eliminating 112 observations, or 10%, 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 subsample. Qualitatively, the effect estimates reported below are the same as the ones resulting from the sample containing all of the 1112 NTA participants. However, with the
smaller sample the effects are more precisely estimated (the standard errors are smaller). Regarding the trade-off noted in the previous paragraph, it thus appears that the improved matching dominated the loss of degrees of freedom.

**The matching of NTA and non-NTA individuals by propensity scores.** The nearest neighbour matching (without replacement) of the NTA and non-NTA individuals was implemented by hand. Neighbours to the NTA participants with the highest propensity scores were located first, and then neighbours to NTA participants with successively lower propensity scores were identified. The outcome of this process is illustrated in Figure 4.18

Figure 4 shows that both the NTA participants and the matched non-NTA individuals effectively cover the entire (0,1) interval. Accordingly, the common support region is maximal. Moreover, the propensity scores of the NTA participants and the non-NTA individuals closely follow each other. This is evident both from visual inspection of Figure 4 and by the fact that the coefficient of correlation between the two sets of propensity scores is 0.991 (cf. the note to Figure 4).

**Properties of the matched sample.** A first observation to be made from Table 9 below is 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. This is what we should expect—in the earlier ‘Joining the NTA programme’ subsection we have argued that the selection into the NTA programme is not affected by individual-level variables. Thus, 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, the majority of the differences are still statistically significant. This means that the matched sample is not balanced with respect to the corresponding variables. However, the crucial issue is whether the common support is large and the sample is well matched in terms of propensity scores. That this is the case has been shown in the previous subsection.

---

**Table 8. Numbers of observed versus predicted NTA and non-NTA individuals**

<table>
<thead>
<tr>
<th>Category</th>
<th>Indicator</th>
<th>Predicted outcomes</th>
<th>Percentage correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-NTA</td>
<td>0</td>
<td>11,770</td>
<td>64</td>
</tr>
<tr>
<td>NTA</td>
<td>1</td>
<td>918</td>
<td>200</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>12,688</td>
<td>264</td>
</tr>
</tbody>
</table>

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.

---

It should also be noted that while they are statistically significant, the mean differences in Table 9 are very small. For nine of the twelve variables in Panels B and C they are smaller than the corresponding raw data differences in Table 6—in spite of the fact that the differences between the unmatched variables were quite small to begin with, as noted in connection with Table 6. Moreover, in contrast to Table 6, all of the differences in Table 9 have effect sizes below 0.2. This is seen by that fact that, in contrast to Table 6, none of the differences in the next to last column are underlined (cf. note 11 to Table 9).

A scalar measure of the decrease in the distances between the mean values of the NTA and non-NTA variables that has been achieved by the matching procedure is provided by the Euclidean norm—the square root of the sum of the squared differences. When applied to the raw data mean differences in panels B and C in Table 6, the Euclidean norm is 0.0833. In contrast, the Euclidean norm corresponding to the matched data in Table 9 is only 0.0428, implying a reduction by almost 50%.

The effect estimates

We compute the effects of the NTA programme by comparing the test score, test grade and course grade distributions of the NTA participants and non-participants,

Note: The correlation between the propensity scores of the NTA and non-NTA individuals is 0.991.
Table 9. Background characteristics of the NTA and non-NTA individuals, matched data

<table>
<thead>
<tr>
<th>Variable</th>
<th>NTA participants (N = 1000)</th>
<th>Non-NTA individuals (N = 1000)</th>
<th>Mean differences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Std. dev.</td>
<td>Mean</td>
</tr>
<tr>
<td>A. Individual level</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex; female = 1</td>
<td>0.470</td>
<td>0.499</td>
<td>0.470</td>
</tr>
<tr>
<td>Taking first language classes(^1)</td>
<td>0.120</td>
<td>0.324</td>
<td>0.130</td>
</tr>
<tr>
<td>B. School level(^2)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Share of boys</td>
<td>0.517</td>
<td>0.041</td>
<td>0.516</td>
</tr>
<tr>
<td>Parents’ educational level(^3)</td>
<td>2.137</td>
<td>0.145</td>
<td>2.110</td>
</tr>
<tr>
<td>Share of foreign-born students</td>
<td>0.066</td>
<td>0.062</td>
<td>0.063</td>
</tr>
<tr>
<td>Share of students with foreign-born parents</td>
<td>0.063</td>
<td>0.099</td>
<td>0.052</td>
</tr>
<tr>
<td>Share of grade 9 students with at least pass in all subjects</td>
<td>0.742</td>
<td>0.102</td>
<td>0.738</td>
</tr>
<tr>
<td>C. Municipality level(^4)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Population in millions</td>
<td>0.082</td>
<td>0.137</td>
<td>0.061</td>
</tr>
<tr>
<td>Share of population in densely populated areas(^5)</td>
<td>0.841</td>
<td>0.112</td>
<td>0.820</td>
</tr>
<tr>
<td>Median net earnings and incomes, millions of SEK(^6)</td>
<td>0.179</td>
<td>0.022</td>
<td>0.175</td>
</tr>
<tr>
<td>Share of teachers with university pedagogical exam(^7)</td>
<td>0.843</td>
<td>0.057</td>
<td>0.849</td>
</tr>
<tr>
<td>Students per 100 teachers(^8)</td>
<td>0.123</td>
<td>0.008</td>
<td>0.123</td>
</tr>
<tr>
<td>Total costs for grades 1–9, millions of SEK, over resident grade 1–9 students(^9)</td>
<td>0.071</td>
<td>0.006</td>
<td>0.071</td>
</tr>
<tr>
<td>Share of grade 9 students with at least pass in all subjects</td>
<td>0.746</td>
<td>0.056</td>
<td>0.738</td>
</tr>
</tbody>
</table>

Notes:
1. Binary indicator = 1 for students that do not have Swedish as a 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 students’ 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 university 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. *, **, *** denote the significance levels at 10%, 5% and 1%, respectively.
11. Underlined differences in the penultimate column denote differences with effect sizes (Cohen’s $d \geq 0.20$, of which there are none in this table, in contrast to Table 6.

respectively, in the matched sample. The comparisons are carried out in the same way as the corresponding raw data comparisons in the earlier ‘The outcome variables’ subsections. Accordingly, $t$-tests and effect sizes, and Mann–Whitney tests for equality of distributions are provided, for all subjects taken together and by individual subjects.

Figure 5(a)–(d) show how the percentile ranked test score distributions compare in the matched sample. The note to Figure 5(a) tells 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 mean percentile ranked test score of the NTA participants is statistically significantly higher than the corresponding mean of the matched non-NTA individuals. The size of the effect is not impressive, however. Cohen’s $d$ statistic is 0.13, implying less than a small effect by Hattie’s (2009, p. 9) standard according to which a small effect requires $d \geq 0.20$.

Still, it is interesting to compare Figure 5(a) with Figure 1(a), which provides the corresponding result for the raw data. In Figure 1(a), the hypothesis of equality of means was rejected, too, but for a very different reason, namely that the mean NTA test score was smaller than the mean test score of the non-NTA individuals. Accordingly, after the matching the direction of the estimated effect is reversed. This goes to show that the matching procedure really matters.

Figure 5(b)–(d) show the effect estimates by subject. It can be then seen that the aggregate effect stems from a positive effect from NTA on the test scores in physics. According to Figure 5(d), the physics test score frequencies of the NTA students are mostly lower than the corresponding frequencies of the non-NTA students up to around the 40th percentile, while the relation is reversed for the 40+ percentiles, and the effect size is non-negligible, too ($d = 0.24$).

Positive but statistically insignificant effects are obtained with respect to biology and chemistry, cf. Figure 5(b) and Figure 5(c). Nevertheless, the results differ markedly from the corresponding findings based on the raw data, which indicated a significantly negative effect of NTA on biology (cf. Figure 1b), again emphasising the importance of accounting for the non-random selection into the NTA programme.

In Figure 6(a)–(d) the NTA individuals’ test grade distributions are compared with the test grade distributions of the non-NTA individuals in the matched sample. Qualitatively, these results are precisely the same as the results in Figures 5(a)–(d). The significance levels at which equality of the NTA and non-NTA distributions are rejected are the same, too, in spite of the fact that the nature of the outcome variable as well as the nature of the statistical test are different. Thus, the results concerning the test scores and the test grades are consistent and corroborate one another.

When course grades are used as outcome variables, the results are quite different from the results obtained for the test scores and the test grades, cf. Figures 7(a)–(d). No significant differences are found between the NTA distributions and the matched non-NTA distributions, but, again, this is still an ‘improvement’ as the raw data pointed to significantly negative effects of NTA on course grades for all subjects but chemistry (see the fourth subsection of ‘The outcome variables’).
Figure 5. Frequency distribution (%) of percentile ranked test scores, by percentile; six percentiles moving averages, matched data, (A) all subjects, (B) biology, (C) chemistry and (D) physics.
Why can it be that our results for tests grades and course grades are different? According to the Swedish National Agency for Education, one possible explanation might be that the teaching of the students has not (fully) covered all of the topics included in the national test (Skolverket, 2007). Moreover, the extent to which the results on the national tests shall impact on the grades is not regulated, but left to the teacher to determine (Skolverket, 2011). These considerations are of interest in view of the fact that the national tests assess both science content knowledge and science process skills, cf. ‘General information about the standardised science tests’ above. Being a constructivist theory-based programme, the NTA should primarily affect science process skills. It is conceivable that teachers whose students have participated in the NTA programme will attach more importance to the national test results that concern science process skills when they set their grades than teachers whose students have not participated in the NTA programme. *Ceteris paribus*, this hypothetical difference should have two empirical implications.

First, the effects of the NTA programme should be more visible when national test grades are used as outcome variables, than when course grades are used. This empirical implication is supported by the differences in results reported in Figure 6(a)–(d) and Figure 7(a)–(d). Second, the difference between course grades and test grades should be smaller among NTA students than among non-NTA students. This implication is tested in Table 10. It should be noted that the test in Table 10 is a difference-in-differences test. The differences between course grades and test grades are first computed separately for non-NTA and NTA participants. Then the difference between these differences is computed. The null hypothesis is that the difference-in-differences is zero. The alternative hypothesis is that the difference for the non-NTA students is larger than the difference for the NTA students, making a one-sided test appropriate. Table 10 shows that the null hypothesis is decisively rejected in favour of the alternative hypothesis for all science subjects taken together and for chemistry and physics. The null is retained only for biology. Accordingly, our conjecture that the difference between course grades and test grades should be smaller among NTA students than among non-NTA students is supported.

*The sensitivity analysis—adding control covariates*

The sensitivity analyses, i.e. the post-matching multivariate analyses, consists of ordinary least squares (OLS) regressions for percentile ranked test scores and ordinal regressions for test grades and course grades. With respect to the choice of covariates, the following two considerations have been made.

First, individual-level factors should not affect the selection into the NTA programme, but it is likely that they matter for the results on the standardised tests and for the course grades. If so, they should increase the precision in the estimated NTA effects. Therefore, we include the two individual-level variables to which we have access, the student’s gender and if (s)he has a foreign background. The latter is made operational by means of a dummy variable equal to 1 if the student has attended first language classes in another language than Swedish.

Second, a time dummy has been included to account for possible differences between the test years 2009 and 2010 owing to, e.g., changes in the schools taking the
Mann-Whitney U test rejects equality of distributions at 1% level of significance.

Mann-Whitney U test does not reject equality of distributions at 5% level of significance.

Mann-Whitney U test does not reject equality of distributions at 5% level of significance.

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

Figure 6. Test grade distributions for NTA (dark) and matched non-NTA (light) individuals: (A) all subjects, (B) biology, (C) chemistry and (D) physics.
Figure 7. Course grade distributions for NTA (dark) and matched non-NTA (light) individuals: (A) all subjects, (B) biology, (C) chemistry and (D) physics.
Table 10. Differences between course grades and national test grades for non-NTA and NTA participants, respectively, and differences in the non-NTA and NTA differences

<table>
<thead>
<tr>
<th>Subject</th>
<th>Non-NTA participants</th>
<th>NTA participants</th>
<th>Non-NTA–NTA difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of students$^2$</td>
<td>Mean difference (course grade – test grade)$^1$</td>
<td>Standard error</td>
</tr>
<tr>
<td>All</td>
<td>990</td>
<td>2.32</td>
<td>4.207</td>
</tr>
<tr>
<td>Biology</td>
<td>335</td>
<td>1.74</td>
<td>3.906</td>
</tr>
<tr>
<td>Chemistry</td>
<td>343</td>
<td>2.48</td>
<td>4.273</td>
</tr>
<tr>
<td>Physics</td>
<td>312</td>
<td>2.77</td>
<td>4.384</td>
</tr>
</tbody>
</table>

Notes:
1. The grades are expressed in terms of points according to: Fail = 0, Pass = 10, Pass with Distinction = 15, and Pass with Special Distinction = 20.
2. The numbers of non-NTA and NTA students are slightly below 1000 because not all students fulfilled the requirements for achieving course grades.
3. *, **, *** denote the significance levels at 10%, 5% and 1%, respectively, for a one-sided t-test.
national tests and/or changes in the extent to which the test-taking schools reported their results to Umeå University (cf. ‘From the population of grade 9 students to our sample’ above).

The dummy variable indicating foreign background ($D_{\text{First language class}}$) is significantly negatively related to all of the outcome variables, and across all subjects (cf. Tables 11–13). The results with respect to gender, in contrast, show a mixed pattern. In the relatively few instances where 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. Finally, regarding the issue of systematic differences for the test years 2009 and 2010, our results are inconclusive. Table 11 shows that there are no significant differences between the years with respect to test scores, whereas Tables 12 and 13 indicate that the results for test grades and course grades partly differ between the two test 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 12 and Table 13.

Still, the estimated effects of the NTA programme are very close to the corresponding results reported in the previous subsection. Especially with respect to the test score regressions, reported in Table 11, the results are very similar to the results reported in Figure 5(a)–(d) above. The estimated effects are only slightly larger in Table 11. Specifically, the effect estimates in Table 11 are 3.90 and 7.46 percentiles for all subjects and for physics, respectively. The corresponding estimates in Figure 5(a) and Figure 5(d) are $47.91 - 44.21 = 3.69$ and $48.25 - 44.41 = 6.81$ (cf. the previous subsection) 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 with 1.28 in Figure 5(a), and for physics the standard error of the OLS estimate is 2.24 percentiles while the corresponding estimate in Figure 5(d) is 2.30. In summary, we conclude that controlling for individual-level covariates and time heterogeneity barely changes our effect estimates but make them marginally more precise.

Regarding the ordinal regressions we first note that in these the constants are replaced by threshold values. The thresholds act 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 explained earlier, the grades have been assigned points ($p$): Fail ($F$) = 0 $p$, Pass ($P$) = 10 $p$, Pass with Distinction ($PD$) = 15 $p$, Pass with Special Distinction ($PSD$) = 20 $p$. If the 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 equal, as the corresponding points always differ by 5. This property can be checked by the estimated thresholds. Furthermore, the NTA effects can be assessed by the differences between the thresholds, answering the question: ‘How much does the NTA contribute to increasing the student’s grade by, say, one level?’

In line with the relationship between Figure 5(a)–(d) and Figure 6(a)–(d), the test grade regressions in Table 12 are qualitatively equivalent to the test score regressions in Table 11. The only notable difference is that, except for biology, test grades differ between 2009 and 2010, whereas the test scores were time invariant.

Regarding the threshold estimates in Table 12, the differences between adjacent thresholds are roughly equal, ranging between 1.5 and 1.8. Accordingly, they are approximately consistent with the equidistant differences between the grade levels $P$, $PD$ and $PSD$ in terms of the numbers of points. Assessing the NTA effects in terms of
thresholds, we can say that for all subjects the effect corresponds to about slightly more than 8% of the distance between two grade levels, while for physics the effect amounts to 23% of the distance between $P$ and $PD$ and almost 28% of the distance between $PD$ and $PSD$.

Turning to the ordinal course grade regressions, reported in Table 13, we find that these are entirely in line with the results in Figure 7(a)–(d). That is to say, none of the estimated effects of participation in the NTA programme are anywhere near significant.

### Summary and discussion

Given the earlier literature review, showing quite modest effects of inquiry-based learning with science kits, we had no strong *a priori* beliefs about the results of this study. Moreover, in contrast to most previous studies, we had access to outcome measures from high-stakes nationwide standardised tests in natural sciences, in addition to course grades. This seemed to add further uncertainty to our expectations. Rightly so, it turned out: while we found positive and significant effects of the Swedish STC programme (the NTA programme) on test results in grade 9, we did not find any statistically significant effects on course grades.

In addition to use appropriate outcome measures, to be credible, the effect evaluation needs to properly account for the fact that the participating students are likely to constitute a selected group. We have firmly established that ignoring the selectivity...
issue can yield very misleading results. A reasonable conjecture is that this holds true for STC programmes in other countries, too. For Sweden, we demonstrate that a direct, i.e. unadjusted, comparison of participants and non-participants will lead to the erroneous conclusion that the NTA programme has negative impacts on both test results and course grades.

It appears that the propensity score method works quite 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 well balanced in terms of the propensity scores, which embody information provided by a large number of variables in our rich data set. While there are significant differences between the NTA participants and the non-participants with respect to some of the school and municipality variables used to model the selection into the NTA programme, these differences are small—for all of them the effect sizes are below 0.2.

That we have been successful in accounting for the selection into the NTA programme is further confirmed by our sensitivity analyses. These show that controlling for student properties—gender and foreign background—results in negligible changes in the estimated effects, which is consistent with our hypothesis that the selection is not determined at the student level, but at the school and municipality level.

According to our results, the effects of NTA are not the same for different science subjects. We find significant and positive effects only for test results in physics. No significant effects (positive or negative) are found for biology and chemistry. Possibly, the difference may be due to circumstances that are specific to Sweden. One indication that is consistent with this conjecture is Gisselberg’s (2001) finding that among Swedish science teachers that made use of NTA a larger number considered themselves to have adequate knowledge of biology than there were teachers who felt confident about their knowledge of physics and chemistry. Accordingly, the potential benefits of the NTA programme should be larger for physics and chemistry than for biology, which is in line with our results.

Concerning our finding that the effects are very different with respect to test results and course grades, a first remark is that course grades are more wide-ranging both in terms of skills covered and in the sense that they reflect performance over an entire course rather than just on one particular test. This raises the issue of teaching to the test. Perhaps participation in the NTA programme fosters high performance at the national tests in science but is less efficient in advancing other relevant science skills? We do not believe that to be the case. As an inquiry-based learning support programme, the NTA should be relevant for many aspects on natural sciences. Instead, we believe that the discrepancy between the effects on test grades and course grades is consistent with grade inflation, manifested in course grades generally being higher than test grades. In line with this conjecture, our analysis shows that the differences between course grades and grades on the national tests are invariably positive, for all of the natural science subjects and for NTA and non-NTA students alike. However, our results also show that the differences for the NTA students are smaller than the differences for the non-NTA students and significantly so, except with respect the subject biology. Thus, if anything, the NTA programme appears to have a dampening effect on grade inflation. As discussed in the previous section, this may be due to the NTA students’ teachers knowing that their students partly are better prepared for the
national standardised test, namely for the part examining science process skills, than the non-NTA students.

Relating our results to those of Zoblotsky et al. (2016), we conclude the following. While both studies establish effects for the aggregates of all students that are statistically significant, neither finds results that are substantively important, in terms of effect size. However, both analyses studies report statistically significant and substantively important results for subgroups of students. For Zoblotsky et al. (2016) the subgroups are students whose mother tongue is not English and students with disabilities, while in our case the subgroup is the students taking standardised tests in physics, as opposed to biology and chemistry. Interestingly, these similarities are observed despite the fact that Zoblotsky et al. (2016) evaluate the programme’s (very) short-term effects while we, in general, estimate the effects three years after the intervention.

For future research, it appears to be important to delve deeper into heterogeneous effects of the STC programme, both across groups of individuals with different characteristics and between similar individuals that participated for shorter and longer periods in the programme.

Acknowledgements

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 and Monika Rosén, as well as from Christina Wikström and the reviewers of the Review of Education. Many thanks also to Mary James for very helpful suggestions along the way. An earlier version of this paper has been circulated as IFAU Working Paper 2015:23, ‘Inquiry-based learning put to test: long-term effects of the Swedish Science and Technology for Children program’.

NOTES

1 Email information from Jean Flanagan at the Smithsonian Education Center in September 2016.
2 It should be noted that we are here considering evaluations of inquiry-based programmes employing science kits. Inquiry-based programmes not employing science kits have been evaluated in several countries outside the USA, e.g. in England, Israel, Kuwait and Taiwan (cf. Slavin et al., 2014).
3 The term ‘distal’ refers to tests that are applied all across US states, as opposed to tests employed locally.
4 The study is also smallest in terms of the duration of the intervention (14 weeks). Although we have not had access to the study itself, it being an unpublished doctoral dissertation (Leach, 1992), it appears to have been quite carefully conducted, however. In particular, it is the only one in the meta-analysis with a random design at the student level.
5 Hattie (2009, p. 9) suggests that, in terms of absolute values, the effect size (Cohen’s $d$) should be categorised as follows: $d = 0.20$ constitutes a small effect, and $d = 0.4$ and $d = 0.6$ medium-sized and large effects, respectively. It should be noted that Hattie’s (2009) threshold values concern applications in educational contexts and are somewhat less conservative than the values originally suggested by Cohen (1988) for behavioural sciences in general ($d = 0.2$ for small, $d = 0.5$ for medium, and $d = 0.8$ for large).
6 From 2010 and onwards the assignment of tests by subject is not entirely random since the Swedish National Agency for Education then imposed a constraint on the assignment: a given school shall not be assigned tests within the same subject in two consecutive years.
7 The Swedish National Agency for Education reports that despite the fact that the test was compulsory, a number of schools choose not to take it, more so in the school year 2008/2009 than in 2009/2010.
8 This was not related to the sample frame(s). Rather, it was due to Umea University having no power to force the schools to submit their test results and that some schools refused to do the test, cf. the previous footnote.
Not all Swedish schools offer all the grades 1–9. The local NTA coordinators therefore had to keep track both of which schools that participated in the NTA programme 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.

The test score’s percentile rank is computed according to: \[ \left( \frac{f_r + 0.5}{N} \right) \times 100 \] where \( f_r \) is the number of scores less than the score of interest, \( f_r \) is the frequency of the score of interest and \( N \) is the total number of examiners.

When we had completed the empirical analysis, it came to our knowledge that recent versions of SPSS include a missing category. The researchers at Umeå University have told us that the most likely reason is that some teachers forgot to fill in grades for some students. The students with missing course grades—which happen to be non-NTA individuals only—are not included in the empirical analysis reported in the penultimate section.

However, the probability that some students have attended the same school is 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 less intensive than the social interaction among classmates and, hence, less of a problem.

It is important to note that this regression is run on matched data. If run on raw data, the effect estimate will be biased. Because of selection bias, the NTA dummy will not be uncorrelated with the model’s stochastic residual term.

In contrast, the standard logit form amounts to assuming that the probabilities of the different outcome values are equal.

The variability in several of the regressors is quite small, especially among the municipality-level variables. This means that these variables will be correlated with the intercept term, if that term is included, resulting in multicollinearity and, hence, imprecisely estimated parameters.

Another option is to match with replacement. We tried that, without success. As noted by Smith and Todd (2005), matching with replacement has the disadvantage that it increases the variance of the effect estimator, through the creation of duplicate non-participants.

When we had completed the empirical analysis, it came to our knowledge that recent versions of SPSS include an algorithm for random matching, which is advocated by Caliendo and Kopeinig (2008) because 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 matching procedure with the result of the random matching procedure, accompanied by a constraint on the maximum distance between treated and controls formulated such that the resulting sample included almost exactly the same number of matched pairs as our sample, namely 1002 pairs. The two matched samples are surprisingly similar. In particular, disregarding the two pairs with the smallest propensity scores among the 1002 pairs, we find that the correlation between the remaining 1000 neighbours and the 1000 neighbours from our matching procedure is 0.990. The randomly matched pairs yield a diagram that is very hard to distinguish from Figure 4. We thus feel confident that our effect estimates do not hinge upon our chosen matching procedure.

References


Gisselberg, K. (2001) *NTA-projektets tre första år—en positionsbestämning (A characterisation of the first three years with the NTA programme in terms of its development and position)* (Umeå, Umeå universitet, Lärarutbildningen).


Skolverket. (2011) *Skillnaden mellan betygresultat på nationella prov och ämnesbetyg i årskurs 9, läsåret 2009/2010 (Differences between grades on national test and course grades in the last, 9th, year of compulsory school)*. PM 71-2011-14 (Stockholm, Skolverket).


Svärdh, J. (2013) *To use or not to use a teacher support program—a study of what characterizes Swedish schools that apply the teacher support program NTA*, in: I.-B. Skogh & M. de Vries (Eds) *Technology teachers as researchers: Philosophical and empirical technology education studies in the Swedish TUFF Research School* (Rotterdam, Sense), 119–142.