KTH Education and Communication
in Engineering Science

MEASURING LONG-TERM EFFECTS OF A SCHOOL IMPROVEMENT INITIATIVE

JOAKIM SVÄRDH

Licentiate Thesis
Stockholm, Sweden 2013

Papers that this thesis covers are:

I Svärdh, J.
To use or not to use a teacher support program.
Forthcoming in Skogh & de Vries (eds.) *Technology teachers as researchers: Philosophical and empirical technology education studies in Swedish TUFF research school.*
Series: International Technology Education Studies. Sense Publishers.
(Published here with kind permission.)

II Svärdh, J. and Mellander, E.
Inquiry-based learning put to test: Long-term effects of the Swedish Science and Technology for Children program.
Submitted for publication in IFAU's Working paper-series.

Department of Learning
KTH School of Education and Communication in Engineering Science
SE-100 44 Stockholm
Sweden

Typeset by Joakim Svärdh.
Printed by Universitetsservice US-AB, Stockholm.

Trita-ECE 2013:02
ISBN 978-91-7501-928-4

1 Abstract

There is a growing demand for studies applying quantitative methods to large-scale data sets for the purpose of evaluating the effects of educational reforms (UVK, 2010). In this thesis the statistical method, Propensity Score Analysis (PSA), is presented and explored in the evaluating context of an extensive educational initiative within science and technology education; the Science and Technology for All-program (NTA). The research question put forward reads; *under what conditions are PSA-analyses a useful method when measuring the effects from a school improvement initiative in S & T?*

The study considers the use of PSA when looking for long-term effects that could be measured, what to take into consideration to be able to measure this, and how this could be done. The baseline references (outcome variables) used in order to measure/evaluate the long-term effects from the studied program is students' achievements in the national test (score and grades) and their grades in year 9. Some findings revealed regarding the object of study (long-term effects from using NTA) are also presented.

The PSA method is found to be a useful tool that makes it possible to create artificial control groups when experimental studies are impossible or inappropriate; which is often the case in school education research. The method opens up for making use of the rich source of registry data gathered by authorities. PSA proves reliable and relatively insensitive to the effects of covariates and heterogeneous effecter if the number of samples is large enough. The use of PSA (or other statistical methods) also makes it possible to measure outcomes several years after treatment. There are issues of concern when using PSA. One is the obvious demand for organized collection of measurement data. Another issue of concern is the choice of outcome variables. In this study the chosen outcome variables (pupils' score and grading in national tests and grades in year 9) open up for discussions regarding aspects that might not be reflected/measured in national tests and/or teachers' grading. Findings regarding the long-term effects from using NTA) show significantly positive effects in physics on test scores (average increase 16.5%) and test grades, but not in biology and chemistry. In this study no significant effects are found for course grades. PSA approach has proved to be a reliable method. There is however a limitation in terms of the method's ability to capture more subtle aspects of learning. A combination of quantitative and qualitative approach when studying long-term effects from educational intervention is therefore suggested.

Keywords: Propensity Score Analysis (PSA), effect study, quantitative evaluation, educational intervention, compulsory school, Science and Technology for All Initiative, representative sample, non-random selection, multi-level modelling, post-matching multivariate regression.

2 Acknowledgements

I wish to thank everybody for being so engaged in this work.
My supervisor, Professor Inga-Britt Skogh who has helped me through this process. My assistant supervisor, Associate Professor Erik Mellander at The Institute for Evaluation of Labour Market and Education Policy (IFAU), who have made all the statistical work possible with endless advice. My other assistant supervisor, Associate Professor Veronica Bjurulf at The National Agency for Education who has been the fastest reader of my drafts and also filled them with valuable comments. Doctor Per Kornhall, formerly at The National Agency for Education, for supporting my work. Katarina Arkehag and Cecilia Göransson at the City of Stockholm for, together with the Swedish government, funding the program Lärarlyftet (´Boost for teachers') and TUFF (Teknikutbildning för framtiden – Technology education for the future) graduate school. A big thank you to the staff at NTA for providing information, documentation and support. Thanks' also to all my friends' and teachers at TUFF, KTH, SU, LiU, GU, UU and Vällingbyskolan.

A special thanks' to my colleague Eva Hartell, for all the encouragement.

Most of all, I would like to thank my beloved family, who has shown support, and patience with my work.

Joakim Svärdh, Stockholm 31 October 2013

Contents

Part 1

3 Introduction

> The political and economic focus on education means that the requirements for decision support quality increases. The availability of such policy relevant studies of high scientific quality that focuses on the effects of education policy decisions need to increase. … Evaluation should be integrated as a part of every major reform. (Utbildningsdepartementet, 2009)

There is a growing demand for studies applying quantitative methods to large-scale data sets to evaluate the effects of educational reform (Vetenskapsrådet, 2011, 2013). The Swedish government noted this importance in the report cited above. There are, of course, many reasons why the demand for reliable efficacy studies is pointed out today. In Sweden, as in many other countries, an ongoing debate exists about school education and how well it works. The debate is conducted on many levels by various stakeholders who express their views in different ways. Political parties, teacher unions, municipalities, media and representatives from the industry all have different understandings of how school education should be carried out. The focus of their opinions obviously varies, but it mainly revolves around what is wrong in schools and what should be done to resolve these issues (Skolverket, 2013). Results from international tests, such as Trends in International Mathematics and Science Study (TIMSS)[1], Programme for International Student Assessment (PISA)[2] and the Relevance of Science Education (ROSE)[3], have received great attention in the media lately with lists comparing countries to each other. The message concluded from these results is that pupils in the western world (with some exceptions, such as Finland) are doing worse and pupils in Asian countries are doing better. In parallel, a more serious debate regarding strategies for improving teaching and learning in schools has been carried out within the research community. In recent years, assessment has received considerable attention from domestic and international educational researchers (Hartell, 2013; Hattie, 2012; Wiliam, 2009).

In Sweden, evaluations of the effects of educational reforms so far, which are in line with governmental demands, primarily have focused on major reforms in schools. Today, such large-scale evaluations are made by organisations such as the Swedish National Agency for Education (NAE), the Institute for Evalua-

[1] http://timss.bc.edu
[2] http://www.oecd.org/pisa/
[3] http://roseproject.no

tion of Labour Market and Education Policy (IFAU), a number of universities and other organisations (Utbildningsdepartementet, 2009, p. 42). The long-term effects from initiatives undertaken in schools that do not qualify as major reforms have less commonly attracted the attention of national and local stakeholders and policy makers. A number of reasons are clearly behind this lack of follow-up studies. In addition to the financial constraints of most small- and medium-sized educational school initiatives, methodological issues must be resolved. A high-quality evaluation requires reliable measuring methods. In this thesis, one statistical method, the Propensity Score Analysis (PSA), is presented and explored in the evaluative context of an extensive educational initiative within science and technology (S & T) education. PSA is a recognized method (Quigley, 2003) designed to facilitate the establishment of a probabilistically equivalent control group when it is not possible to perform randomized controlled experiments. This design makes it possible to compare groups' performances with each other without the risk of comparing apples and oranges (Blackstone, 2002).

S & T subjects: an educational area with problems

The question of which educational initiative should be chosen for scrutiny of its long-term effects is a challenging one. Each school improvement initiative cannot be measured with statistical methods, but some candidates are appropriate for examination.

Several Swedish and international reports show that pupils' academic performance in S & T has deteriorated through the years (European Commission, 2004; Teknikdelegationen, 2009a, 2010; TIMSS, 2011). The number of students who pursue scientific or technical education in higher education is also considered low. It has been found that "the increased engineer shortage combined with a declining number of applicants to engineering education was deeply disturbing for Sweden" (Teknikdelegationen, 2009b). The importance of having qualified teachers in these subjects has been noted by the (Swedish School Inspectorate, 2009). Today, many teachers lack subject knowledge and instructional training in S & T (Skolverket, 2013; Hartell and Svärdh, 2012).

Several initiatives have been formed over the years to change this development. Some initiatives offer educational materials, teacher support programs and other events or competitions (Rooke, 2013; Teknikdelegationen, 2009c) in support of S & T instruction. A national inventory of different S & T education initiatives resulted in a list of approximately 250 primary and secondary school activities (Teknikdelegationen, 2009d). International programs such as "No Child Left Behind" in the United States indicate that the situation in other western countries is similar.[4]

[4] No Child Left Behind http://www.nea.org/home/NoChildLeftBehindAct.html

In this study, the long-term effects of the Natural Science and Technology for All (NTA) program, which is an educational S & T intervention, are explored. The NTA is a teacher support program aimed at engaging pupils in S & T. The NTA website states that the program "offers and develops methods as well as services and products for improving learning and teaching in science and technology, both at overall municipal level and at the level of individual schools/schools districts" (NTA, 2013). The NTA is well known and is the largest of all S & T teacher support programs in Sweden. In December 2012, the NTA was in use in 110 municipalities by 180,000 students. This particular program originates from the Science and Technology for Children (STC) program developed in the United States.[5] The size and general knowledge of the program makes it suitable for testing with the PSA method. Available project information includes data on school level (e.g. information about participating schools, entrance year and amount of use) that can be combined with registry data.

Outcome variables: national tests and grading (year 9)

Information about students' achievements in different subject areas and at different levels in the school system can be collected and presented in many ways. One way, mentioned above, is the use of internationally organised and nationally processed tests from universities and organizations. Teachers grading students' performance in various subjects are probably the most obvious (and to the public most familiar) way of measuring achievements. In Sweden, as in many other countries, school authorities also provide national tests in several subjects each year (e.g. Swedish, Mathematics and Science). Although the use of national tests is criticised (Lundahl, 2013), they are useful as a nationwide measuring instrument.[6]

The results from national tests and final grades in the science subjects from a nationally representative random sample of approximately 16,000 Year 9 students are used in this study. Unfortunately, there is no national test in the subject of technology, which makes it impossible to measure that subject in this study.

This study draws attention to the use of a specific method (PSA) to evaluate a Swedish teacher support program with an American role model. The research literature examined is primarily limited to Swedish and American literature.

[5] STC http://www.ssec.si.edu/curriculum/about-our-curriculum
[6] http://www.skoloverstyrelsen.se/?p=924

In summary,

- The study explores the effects of a science and technology school improvement initiative (the NTA project) using PSA as a measuring tool.
- The study considers the use of PSA when looking for (1) long-term effects that could be measured; (2) considerations related to these measurements; and (3) how these measurements could be performed.
- The baseline references (outcome variables) used to measure/evaluate the long-term effects of the program studied are the students' achievements on national tests and their grades in Year 9.
- Some of the findings regarding the effects of using the NTA program are also addressed.
- The study does not investigate pedagogical methods or different learning theories. The study is not a complete evaluation of the NTA. It aims to measure the impact of the NTA program in relation to two outcome variables: students' performance on national tests in the science subjects and their grades in Year 9.

4 Purpose and research questions

The main purpose of this study is to explore the effects of one teaching support initiative using a method designed to facilitate the establishment of a probabilistically equivalent control group when it is not possible to perform randomized controlled experiments. Can evaluation with PSA provide useful information for decision makers? This task will be approached using a two-step strategy:

- First, an exploration of considerations that must be taken to create a comparable control group will be performed.
- Second, an exploration of a suitable method for measuring long-term effects (pupils' knowledge) using the chosen improvement initiative will be performed.

The main research question explored in this paper is
"Under what conditions are PSA analyses a useful method for measuring the effects of a school improvement S & T initiative?"

The PSA method requires that multiple choices be made. These are dependent on several conditions. The study explores these choices and conditions based on data collected from the NTA project. The first substudy draws attention to the typical user of the NTA, if the treatment group is representative and

if not, what distinguishes it from the control group. This is primarily addressed in Article 1 and is based on the following research question:

Q1. In what ways does the treatment group differ from a group of average students?

In other words, how do they stand in relation to geographical and socioeconomic-/cultural parameters? Therefore, the question of whether NTA students differ from other students in the country (and if so, in what way) must be answered.

Taking into account the answer to Question 1, the PSA method provides the possibility of creating comparable groups. To test the method's usefulness in measuring the long-term effects of the studied teacher support program (NTA), the following research question is investigated in Article 2:

Q2. What long-term effects could be measured in the treatment group regarding performance in S & T subjects?

In this substudy, the main issue is to explore whether the use of a teacher support initiative provides long-term effects on students' performance. In this case, effects refer to outcomes after four to five semesters of intervention. The term "effect" is interpreted here in a causal sense and is defined as statistically significant differences between students exposed to the NTA and statistically similar students not exposed to the program. The measurement instruments chosen are national test results in science subjects in Year 9 and Year 9 grades in science subjects. This question is addressed primarily in Article 2.

5 Measuring learning outcomes

Section 4 examines international and Swedish studies on the effects of S & T teacher support programs on students' performance. In particular, this section focuses on how the study is done, the instruments and methods used for measuring.

5.1 US research on S & T teacher support programs

In US research, S & T teacher support programs and other similar systems are often called "kits" or "inquiry-based instruction" and are based on constructivist theories. An extensive amount of research has been conducted to evaluate their use and effects[7]. Different studies use several methods to establish their results. Their main findings and methods are described below.

[7] Full Option Science System http://lhsfoss.org/scope/research/search.php

Meta-studies

A meta-study combines results from several other studies to draw conclusions.

A study by (Klentschy et al. 1999) considered instruction using "kits" as providing better outcomes than traditional textbook education and exercises organized by the teacher. These kits were developed in the 1960s and 1970s. Research in the 1980s and 1990s claimed that the kits had positive effects on educational achievement in science, especially among girls, minorities and students of low socioeconomic status. Several large meta-studies of 278 studies showed statistically significant positive results (Bredderman, 1983; Shymansky et al. 1990; Wise, 1996). These early attempts with kits eventually ended due to a lack of support from the government (Shymansky et al. 1990).

Using questionnaires and tests

A pre-test is "a preliminary test administered to determine a student's baseline knowledge". It is usually followed by a post-test, which is "a test given after a lesson or a period of instruction to determine what the students have learned".[8]

(Cuevas et al. 2005) showed how gaps between different social groups were smoothed out through "inquiry-based" teaching and service training. This study had a small number of participants (n = 25). Pre- and post-tests showed significant effects on traditionally low-performing groups. Groups with a low socioeconomic status and a high proportion of immigrants increased their knowledge, while no differences due to gender were detected.

The Scaling up Curriculum for Achievement, Learning and Equity Project (SCALE-uP)[9] is an active research project. This five-year project is a large-scale effort to increase "hands-on, inquiry-based" learning in S & T for 85,000 middle school students in Maryland. Study materials are approved and recommended by the American Association for the Advancement of Science (AAAS). The research project has $5.2 million in funding and, the project team is currently writing the final reports. The researchers on this project are trying to identify the conditions under which large-scale implementation of effective training improves student learning.

(Lynch et al. 2005) examined the differences between experience-based learning and traditional textbook learning for 1,500 students in eighth grade. The study included students from ten different schools in Washington who were very diverse, and many had an immigrant background (80%). It used the hands-on Chemistry That Applies (CTA) module over 18 lessons to teach students about "conservation of matter". Schools were matched in pairs, and who partic-

[8] Pre-test and post-test http://www.thefreedictionary.com/posttest
[9] SCALE-uP http://www.gwu.edu/~scale-up/index.html

ipated in the treatment group and the control group was randomly determined. The study used questionnaires to measure attitudes. Pre-tests, post-tests and delayed after-tests that were adapted to meet national standards measured learning outcomes. As expected, pre-tests showed no differences and the groups were considered equally. Post-tests showed that CTA students raised their average test results by 20 points on a 100-point scale, while those with traditional education raised test results by 11 points. Low-performing students with test scores below 23 points were also fewer in the CTA group (22%) compared to the control group (38%). Two of the five measurement scales in the attitude survey (engagement and targeting) also showed significant differences based on the CTA. Differences between subgroups based on socio-economic background and ethnicity were also present. However, no visible differences were seen between genders or among those who participated in language training for immigrants. CTA was found to enhance learning without increasing the knowledge gap between different groups in schools. In the traditional teaching group, the difference in knowledge increased between the strong and weak students.

Multilevel analyses

Multilevel analyses take different levels of data into account, such as individual, class and school levels.

One year later the SCALE-uP researchers involved more modules, schools and students in the study. Using multilevel analyses refined previous results (Rethinam et al. 2008). The positive effects on students' academic performance were more evident when taking into account the factors affecting the classroom. Approximately 15% of the variation between students depended on what class they were in. The effects of interventions were greater when more African-Americans were in the class, and the education gap decreased. More information on class and school levels is necessary, e.g. data on teachers and school resources, so a multilevel analysis with three levels can be implemented.

Matched participants

In these studies, participants are paired according to similar attributes.

A five-year school improvement project in New England sought to enhance students' S & T knowledge using teacher training and inquiry-based educational materials from STC and Full Option Science System (FOSS) (Young & Lee, 2005). The teachers in this study taught an average of 2−3 topics per year, and students received instruction on 12−14 themes over five years. The study compared 226 fifth graders in participating school districts with demographically matched students who received traditional teaching (n = 173). It used a written test designed to match the National Science Education Standards. The larger group of students was divided into two halves; one group had teachers who received many hours of training and the other group's teachers received fewer

hours of training. These groups were formed using randomized sampling. The control group was composed of teachers who volunteered for the study and their students. The tests used were pre- and post-tests. The study showed significant differences between STC/FOSS pupils and those who received traditional teaching. The amount of training teachers possessed did not have any effect on the pupils' test scores. Students in the control group received more hours of instruction on more subjects. Regression analyses showed the difficulty of measuring factors that led to improved results.

Random sampling

In these samples, all participants have an equal and independent chance of being in the treatment group or the control group.

Vanosdall et al. (2007) used rigorous statistical sampling (random selection) to compare the performance of students taught with inquiry-based educational materials. The FOSS "Mixtures and Solutions" module was used in this study. The study comprised several parts comparing different ways of teaching with each other. It was implemented in the Imperial Valley School District in Southern California. This area near the border of Mexico is sparsely populated and poor and has a very high number of immigrants.

The first substudy included 20 teachers who volunteered from four different schools. They were divided randomly into two separate groups. The classes included 563 fifth-grade pupils. Teachers in the treatment group had access to additional training (scaffolding) on how to use the material. In the control group, traditional textbook and instruction were used with "normal" experiments and the exercises that usually occurred. A FOSS test was used as a pre- and post-test for both groups. The California standard test was also used for both groups. Pre-tests showed that the groups were equal and, as expected, the FOSS test showed an advantage in the treatment group. The standard test showed a substantial lead, with an average of 6.03 points in the treatment group compared to 3.41 points in the control group. These differences corresponded to approximately one year of study. Data analyses used a two level hierarchical linear model (HLM).

In the second substudy, 24 teachers who volunteered from 11 different schools were matched on relevant background variables. The pairs were then randomly assigned to the control group or the treatment group. The study also included 762 fifth-grade pupils. These teachers had many years of experience with inquiry-based education. One group of teachers received the same guided teacher training as in the first study. The control group received traditional training from FOSS. The same modules from FOSS were used in this study. The same standard tests were used to measure the effects. The differences between the groups in the standard test were 6.01 points for the treatment group

compared with 3.89 points for the control group. HLM was used for the analyses.

In additional experiments, which used the collected data, comparisons were made between textbook teaching and FOSS. The researchers also studied how guided teacher training affected the pupils of the experienced and inexperienced FOSS teachers. The use of inquiry-based materials without additional guided training led to significant improvements compared to textbook teaching. However, no significant differences based on previous teaching experience were displayed when using scaffolding. According to the authors, the last two substudies did not show high statistical credibility in comparison with the first two since no random assignments between kit-based or textbook-based instruction were made.

Very few studies with negative results are found. One example is Bredderman (1983) meta-study that conducted three follow-up studies of students who used "activity-based program" in elementary and middle school but received traditional science education in high school. No significant differences between the groups were seen.

The studies discussed in this section reported generally positive results for evaluations of different S & T educational initiatives. The studies used different methods to measure the effects of inquiry-based materials and teacher training. The methods varied from questionnaires and meta-studies to using pre-tests and post-tests adapted to meet national standards for measuring learning outcomes. They also used several ways of selecting participants and control groups, including volunteers, matched schools and students and random selection. The most statistically advanced studies used multilevel analyses to detect differences between classrooms. However, these regression analyses showed the difficulties of a lack of information on school and class levels. None of the studies seemed to use any registry data.

For more information on research in this area, FOSS summarizes the research on its website[10] including its own material as well as its competitors' materials.

5.2 Swedish research on S & T teacher support programs

A large amount of research has been conducted on the major S & T teacher support program in Sweden (NTA). A search of literature on this topic found

[10] Full Option Science System http://lhsfoss.org/scope/research/search.php

at least 70 papers[11]. Most of these studies are qualitative and only a few are quantitative to some extent. This section describes five NTA research projects.

Attitudes, subject content and teachers' ability

The early development of the NTA is described in several reports showing that participating teachers were generally satisfied with the material, training and organization. The teachers requested minor adjustments and local adaptations. Many of the teachers who participated in the early stages of the program had some sort of scientific background, and there were hopes to also attract teachers with other backgrounds. The pupils were perceived as being enthusiastic about the work with the boxes. It was also noted that there was a need for dedicated and knowledgeable teachers to take advantage of the development potential of the material. Data collection was made with observations, questionnaires, video recordings, teacher and student notes, participating in meetings and interviews. Four or five schools participated in each study (Gisselberg, 2001; Schoultz et al. 2003; Schoultz and Hultman, 2002).

Anderhag and Wickmans (2006) study evaluated how the teacher's ability to support students conceptual and language development was enhanced by the use of the NTA in their teaching. The pupils demonstrated an increase in scientific linguistics and in the use of scientific concepts, mainly in the oral area. The study showed that the material must evolve and change to support in-depth discussions and increase the pupils' desire to write. Data collection was through observations and interviews. Participants (N = 23 teachers and 96 pupils) came from 21 different schools and 23 classes.

Anderhag and Wickman followed this study with an interview study of 80 Year 6 pupils who used the NTA. Pupils in the treatment group reported improved performance and advanced knowledge in science subjects compared to the control group. This was true for low- and high-performing students. The results differed between the genders, and boys showed better results than girls. Analysis was made using the Statistical Package for the Social Sciences (SPSS), and great care was taken to create equivalent groups (Anderhag and Wickman, 2007).

Ekborg and Lindahl (2006) sent a questionnaire to 700 teachers to evaluate how the NTA works as a tool for teacher training. It revealed how the material could serve as a source of inspiration leading to increased scope and quality of S & T teaching. A difference in how the use of the NTA developed depending on the teacher's educational background was noted. The evaluation also revealed that some teachers were critical of the boxes, stating, "the missions are quite controlled, often but not always a problem is formulated." Ekborg and

[11] http://www.cienciaviva.pt/rede/upload/Swedish_references_on_inquiry_ evaluation.pdf

Lindahl, (2006) examined the NTA as a possible method for school development that depended on how the school management prioritized. They determined that the NTA could be part of a larger school improvement program with the inclusion of prioritized areas such as language development. It could also involve organizational support so teachers could attend courses and NTA-theme gatherings. The questionnaire included 98 questions and was analysed in SPSS.

In general, Swedish studies on the NTA claimed that teachers and pupils were satisfied due to increased knowledge. No follow-up studies have been conducted to evaluate any further long-term effects from teaching using the NTA boxes. None of the studies used registry data.

6 Object of study: the NTA program

The NTA program and the long-term effects of teacher support material is the focus of this study. Subsection 5.1 presents the US Science and Technology for Children program (STC). Subsection 5.2 presents the Swedish version of the STC, which is the NTA.

6.1 The US Science and Technology for Children (STC) program

This description of the STC is drawn from Article 1 and from a Swedish article about the NTA (Svärdh, 2011).

The beginnings of the STC educational program were in the 1960s and came from what has been called Sputnik shock.[12] Americans, who previously saw themselves as the leader in space science, saw themselves overtaken by the Soviet Union in the space race. Great efforts were made to catch up, including launching a series of projects to improve S & T instruction.

Several of these projects used box systems with working materials organised by themes and instructions for experiments; these were similar to those used today in the NTA and STC. Many projects ended on their own, often due to a lack of state funding (Shymansky et al. 1990). In 1983, a commission appointed by President Ronald Reagan published the report *A Nation at Risk: The Imperative for Educational Reform.*[13] The report drew the attention of politicians and the public on the serious shortage of scientists and engineers. The result was that several major institutions became involved in the issue, and various initiatives were undertaken to overcome this deficiency. The National Science and Re-

[12] http://www.theglobalist.com/storyid.aspx?StoryId=2218
[13] http://mathcurriculumcenter.org/PDFS/CCM/summaries/NationAtRisk.pdf

search Center[14] (NSRC) was commissioned to create appropriate educational material that met national standards. The result was the Science and Technology for Children program (STC). The material was ready for use in 1991.

The STC material has evolved over the years with more themes and other materials. NSRC developed STC as part of a system for school development that includes management support for school leaders as well as help implementing training materials and teacher training.[15] The NSRC organisation produces its own material, but it also supports similar teaching material, the Full Option Science System (FOSS)[16].

NSRC's material is organized into two main parts, *Science and Technology for Children (STC) K–6* and *Science and Technology Concepts for Middle Schools (STC/MS) 6–8*. This is complemented with books for deeper understanding intended for grades K–8. Each theme also provides an instructional video for the teacher. The themes are organized by topics with a progression that follows the age of the pupils. They include "physical science, earth science, life science, and technology".[17]

The benefit of using STC is that it is "an inquiry-based learning environment [that] encourages opportunities for children to learn science" (Cuevas et al. 2005). The use of STC has spread to other countries. Chile, China, Thailand, Germany, Mexico, Panama and Sweden have tried to use or are using some variant of STC in schools. The material has been translated into Spanish and Swedish.[18]

6.2 The Swedish Natural Science and Technology for All (NTA) program

This section provides a short description of the NTA and its history. More details can be found in article 1 and in a Swedish description of the NTA (Svärdh, 2011).

The NTA is a teacher support program aimed at engaging pupils in S & T. It was founded in 1997 as a project by the Royal Swedish Academy of ence[19] (KVA) and the Royal Swedish Academy of Engineering Science[20] (IVA) in cooperation with municipalities throughout Sweden.

---

14 NSRC http://www.ssec.si.edu/about/mission

15 Programs and services http://www.ssec.si.edu

16 Full Option Science System http://lhsfoss.org/

17 NSRC Science strands http://www.ssec.si.edu/curriculum/overview

18 FAQ International participation http://www.ssec.si.edu/about/frequently-asked-questions

19 The Royal Swedish Academy of Science http://www.kva.se/en/

20 The Royal Swedish Academy of Engineering Science http://www.iva.se/en/

The NTA program offers and develops methods as well as services and products for improving learning and teaching in science and technology, both at overall municipal level and at the level of individual schools/schools districts. (NTA, 2013)

The NTA is a translated and less-comprehensive version of the STC program. It has been adapted to the Swedish curriculum and developed with more themes. Local NTA coordinators in the municipalities handle administration, distribution and teacher education. The program is mainly used in primary and secondary schools in Years 1–6, but it has been complemented with themes for pre-school and Years 7–9.

The 22 different themes included in the NTA provide boxes with working materials and instructions for the exercises and experiments. Teachers must attend a one-day training session for each theme used. After the training, teachers attend a follow-up meeting to obtain deeper knowledge about the theme. A theme is usually used in a class during a semester.

Despite modest marketing, the NTA has become the most widely used teacher support program in S & T education in Sweden. Since the program's inception, participation increased to 110 municipalities and 27 independent schools, including 180,000 students and 8000 teachers, by December 2012[21]. This represents about 18% of the pupils in Swedish compulsory schools. The number of municipalities that have joined the NTA program has increased by about 10 each year. The NTA is not evenly distributed across the country; it is more common in certain cities and areas. The NTA uses the same pedagogical methods and inquiry-based science education (IBSE) as the STC (NTA, 2013).

7 Method

Assessing the effects of an education initiative

In Swedish school research, there seems to be a lack of balance between quantitative and qualitative methods (Vetenskapsrådet, 2011). Sweden possesses a very large and unique resource in registry data with information about all citizens. Statistics Sweden (SCB) and other official organisations as Swedish National Data Service (SND) and the Institute for Evaluation of Labour Market and Education Policy (IFAU) organize this. These records are useful to this study as they provide information on an individual level about national test results, grades and other personal socioeconomic information.

The Swedish National Agency for Education (NAE) describes their work with national assessments as such:

[21] http://www.nta.kva.se/In-English/

The Agency is responsible for the national system for assessing knowledge. Together with universities and university colleges, we develop national tests and assessment guides for teachers to ensure pupils receive equivalent assessment.[22]

The collected information, including test results, from SCB can be combined with personal data such as family members and socioeconomic situations. Extractions from the databases are available at different levels (country, counties and municipality level). This registry data, combined with interviews, is used in this study.

Qualitative vs. quantitative methods

The NTA has been, with few exceptions, evaluated with qualitative methods. These methods include case studies with questionnaires, interviews and observations. These studies have a short time span and mainly perform real-time evaluations; this study measures effects occurring a few years after implementation. To determine if the use of the NTA provides any long-term effects, other methods with greater samples are needed (Svärdh, 2012).

Co-variation vs. causal effects

The fundamental distinction between correlation and causation requires attention. In particular, what we mean by effects must be examined. In addition, whether the NTA has caused higher results on national tests must be explored. To determine this, it is not enough to see if students who received NTA instruction have good test scores. Whether these results are good in relation to the performance of other students who did not receive NTA instruction must be investigated (the counterfactual group).

The problem of non-random assignments

The differences between the comparison groups must be taken into account (these are presented in detail in Article 1). Since participation in the NTA is not random but is influenced by various decisions at both the municipal and school level, there must be measures to make the groups being compared equal. One method to accomplish this is to use Propensity Score Analysis (PSA) (Guo and Fraser, 2009). In this method, a student who received NTA instruction is matched with a student who has not received NTA instruction but is otherwise as equal as possible. This creates a group of control "twins".

The methodology behind PSA was developed by Rosenbaum and Rubin, (1985; 1983; 1989). PSA has been used in other research areas to compare groups such as studies about life choices (KOMMUT, 2010), economic studies (Dehejia and Wahba, 2002), and choosing a university (Quigley, 2003). The PSA method is considered "the optimal method of establishing a comparison group for an observational study" (Quigley, 2003). Evaluations could be put on

[22] http://www.skolverket.se

a scale measuring the strength of evidence as judged by the value of the study. If randomized controlled trials (RCT) are at the top of the scale, the quasi-experimental PSA is in second place due to "its ability to reconcile variation across many observed factors and still establishing probabilistically equivalent groups" (Quigley, 2003).

PSA combines several variables that may have influenced the choice to participate in the NTA into one shared variable. This is done through a multivariate logistic regression[23] analysis that assigns each student (participating or not participating) a value based on his or her probability of having participated in NTA. The variables used in the regression are tested to provide as high a hit rate of students involved as possible. The variables must be relevant and must have affected municipalities' and school management's decisions to join the NTA. This affect must also have occurred prior to the test date. After each student is assigned a probability value between zero and one, matched pairs of NTA students and non-NTA students with similar values are created.

After matching, a simple T-test[24] and non-parametric tests (Mann-Whitney)[25] are performed on outcome variables to see if significant differences exist between the groups' averages. This is complemented by ordinary linear regression[26] and ordinal regression[27] analysis to determine the impact any remaining differences in control variables and heterogeneity have on the results. This somewhat tedious process is described in more detail in Article 2.

Outcome variables

Data available from the newly introduced national tests (2009) in the science subjects is used in this study. The test results are combined with existing registry data from SCB and NAE. All the Year 9 students (about 100,000 each year) perform written and elaborative tests. The results are collected by SCB and are available for research along with other registry data. National tests are also given in Year 6 and Year 3 (Year 3 is only tested on mathematics and Swedish). The NTA focuses on three science subjects (biology, physics and chemistry) and technology. Unfortunately, no national tests are currently performed for technology, which makes it impossible to measure its impact. More details about the national tests can be found in Article 2.

A randomized anonymous sample of about 8000 individual test scores is collected each year from the Year 9 tests. This sample is performed by Umeå Uni-

[23] http://sph.bu.edu/otlt/MPH-Modules/BS/BS704_Multivariable/BS704_Multivariable8.html
[24] http://www.physics.csbsju.edu/stats/t-test.html
[25] http://www.statisticslectures.com/topics/mannwhitneyu/
[26] http://www.research-training.net/addedfiles/READING/OLSchapter.pdf
[27] http://www.ats.ucla.edu/stat/spss/dae/ologit.htm

versity, which is also the test designer. These samples contain complete test results as well as variables such as test grade, course grade, gender and immigrant status for each test item reported. The randomized test data are combined with registry data from the NAE's SALSA database[28] and the Swedish Association of Local Authorities and Regions (SALAR) "Öppna Jämförelser" (Open Comparison) reports[29] to provide additional information at the school and municipal levels.

Using these combined data together with previously collected categorizations of NTA schools (Article 1), we are able to make statistical comparisons between NTA students and non-NTA students. The classification of schools is as follows: Year 9 students i) participated in the NTA; ii) did not participate in the NTA; or iii) could not be classified according to i) or ii). The results can also be broken down into smaller groups, such as gender and subject.

8 Summary of papers

8.1 Article 1

To use or not to use a teacher support program - A study of what characterizes Swedish schools that apply the inquiry-based teacher support program NTA. In M. de Vries & I.-B. Skogh (Eds.), *Technology teachers as researchers: Philosophical and empirical technology education studies in Swedish TUFF research school*. Sense Publisher.

This study describes the socioeconomic and geographical differences between schools and municipalities using the NTA and those not using the NTA. Through a survey and by personal contact, all NTA schools in Sweden are identified and categorized. Categorization is done at the school level based on how much the school has used the NTA.

Municipality level

Using registry data from SALAR, the article shows that use of the NTA is most common in Stockholm's suburbs and in some other major university cities. In Malmö and Göteborg (Sweden's second- and third-largest cities), the NTA is almost absent and is used only by a few independent schools.

The average NTA municipality has a larger population with a higher average income than non-NTA municipalities. However, no differences between the average percentages of trained teachers or the number of children per employee could be found.

[28] http://salsa.artisan.se/Vad_är_SALSA.htm

[29] http://www.skl.se/vi_arbetar_med/oppnajamforelser/ oppnajamforelser_grundskola

	NTA	Non-NTA
Population in municipality	45,000	27,000
Average yearly income	178.500 SEK	172.700 SEK
Percentage of trained teachers	86%	86%
Pupils per employee	9.5	9.5

At the municipality level, the differences are very small in terms of merit values (sum of all grades) with only a one-point advantage for NTA municipalities. The visible differences are probably associated with the historical disparity between urban and rural areas where differences in income and academic backgrounds are present.

School level

At the school level, the differences are larger. Using 11 years of data from the NEA's SALSA database (approximately 15,000 readings), the differences between NTA schools and other schools in the same municipality can be analysed. The general education level slowly rises each year. The parents of pupils in NTA schools have slightly lower levels of education than average parents (just over 3%, Figure 1). Parents at schools that do not use the NTA have above average levels of education. The scale measures the average of both biological parents' highest level of formal education. The maximum value is three. Value 1 corresponds to graduating from compulsory school, value 2 corresponds to completing upper secondary school and value 3 corresponds to completing at least one semester of university studies. The value is computed as the average from all the pupils' parents at the school.

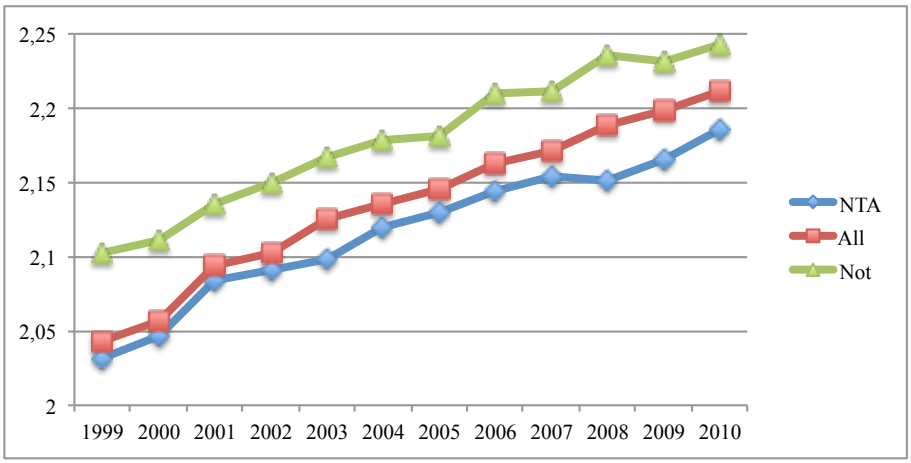

Figure 1. Parent's formal education

The NTA students also have somewhat lower merit values (about 3 points) in Year 9. The difference in merit values also accelerates over the years (Figure 2). The diagram shows the schools' average sum of merit values of the grades. In each of the 16 best subjects, the pupils receive points by grade. Pass (G) is 10 points, pass with distinction (VG) is 15 points and pass with special distinction (MVG) is 20 points. The maximum value is 320 points.

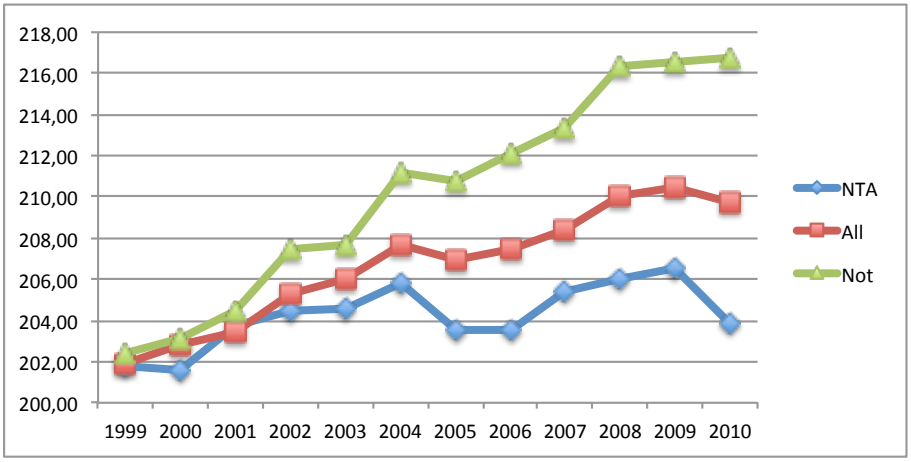

Figure 2. Merit values

When compensating for gender, parental education and immigrant background, the differences become smaller. The schools in NTA municipalities that have chosen not to use the NTA excel the other way around. Their students receive higher merit ratings than the national average and have parents with higher education.

The proportion of immigrants and students with immigrant backgrounds in schools using the NTA compared with those schools who do not use the NTA differ by about 4% in 1999 and then decline. The difference is almost completely absent in 2011.

The focus on supporting low-performing students present in the US studies is not nearly as pronounced in the Swedish studies. Earlier Swedish studies are primarily case studies describing teachers' experiences of working with the NTA.

This study shows differences in merit values that appear to be associated with socioeconomic factors. Over time, it is possible that school choice and increased segregation will affect some variations. Choosing to use the NTA is based on decisions that may be influenced by desires to compensate for low academic performance caused by socioeconomic conditions. Figure 2 presents the results over time and shows a clear trend.

Registry data, such as that found in SALSA and SALAR, could be used as tools by principals and municipalities to make different decisions. Other possible reasons for the use of the NTA that can be traced in the Swedish research (other than the NTA's good reputation) are lack of time, resources and teacher training as well as the low academic performance of certain groups of students. These differences are all adequate reasons for choosing to use the NTA as a tool to transform S & T education.

All of these variables, together with the non-random sample, make it difficult to compare the groups with each other. To determine the long-term effects of using the NTA as a school improvement tool, the schools and the students being compared must have the same conditions as those schools and students who have not used the NTA. The study shows that it is necessary to consider the selection.

8.2 Article 2

Inquiry-based learning put to test: Long-term effects of the Swedish Science and Technology for Children program. Submitted for publication in *IFAU's Working paper-series*. Mellander, E., Svärdh, J.

Selection problems

This study quantitatively evaluates the NTA. To measure if the use of the NTA provides lasting effects, one must consider that students in the schools using the NTA differ systematically from non-participants (see Article 1). Participation in the NTA is not random but is influenced by various decisions at both the municipal and school level. Making a fair comparison requires that the students compared have equal conditions in school. When it is not possible to perform a randomized experiment, an equivalent control group must be obtained by other means.

Method

One method to do this is to use Propensity Score Analysis (PSA), a multivariate method in which a student who participated in the NTA program is matched with a student who did not participate in the NTA but is otherwise as equal as possible (Guo and Fraser, 2009). The advantage of using PSA is that it provides the opportunity to aggregate variables from multiple levels (individual, family, school, etc.) to a common value. Students with similar values are matched, and the two groups are compared to find significant differences.

Data

The ability to perform a quantitative evaluation of the NTA is now possible since standardized national tests in the science subjects are available. The Swedish National Agency for Education introduced national tests in biology, chemistry and physics for Year 9 in the spring of 2009. All students complete the

tests. SCB collects the results and makes them available to researchers. Umeå University (UU), which is the designer of the tests, annually collects a nationally representative 10% sample. The outcome variables available include test scores, test grades and course grades. We use UU's data from 2009 and 2010, including results for approximately 16,000 students out of the approximately 180,000 students who completed the exams.

Using UU's data in combination with previously collected categorizations of the NTA (see Article 1) provides the opportunity to make statistical comparisons in which consideration can be given to gender and subject. In this study, 1000 students participating in the NTA are compared to 1000 non-NTA students. Test data are combined with registry data from SALSA and SALAR to provide additional information at the municipal and school levels. The various data sources are matched in a database and then analysed with statistical software.

Results

The results for natural sciences in general show a positive effect on standardized test scores. T-tests show that the difference in test scores is significantly different at the 1% level. The mean of the percentile-ranked test scores for NTA students is 48.3 while the mean for non-NTA students is 44.5. When the results are divided into subjects, biology and chemistry do not show any significant differences.

The difference of 7% comes from the physics results, as shown in Figure 3. Means: mNTA = 49.5, mnon-NTA = 42.5; t-test for equality rejected at 1 % level. Effect size (Cohen's d): 0.247

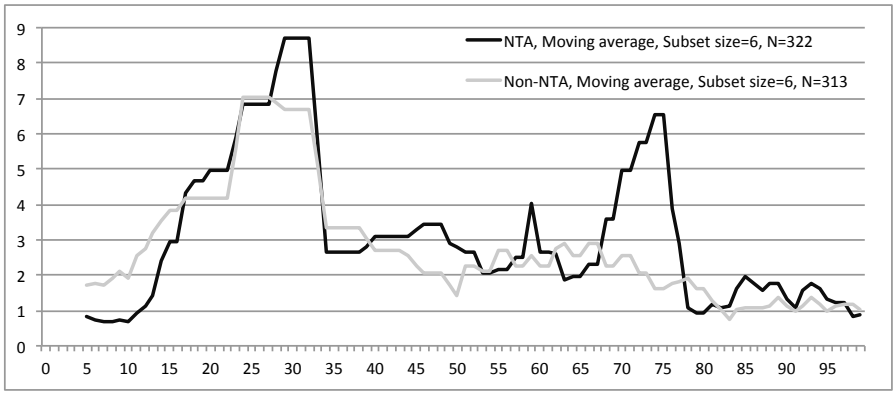

Figure 3. Frequency distribution (%) of percentile-ranked test scores, by percentile; 6 percentiles moving averages for NTA participants and non-NTA individuals, matched data, ***physics***

When comparing test grades, the results are similar. The non-parametric Mann-Whitney U tests show that the NTA students' performance differs positively from non-NTA pupils. The entire difference comes from the subject of physics (Figure 4). Mann-Whitney U test rejects equality of distributions at 1% level of significance.

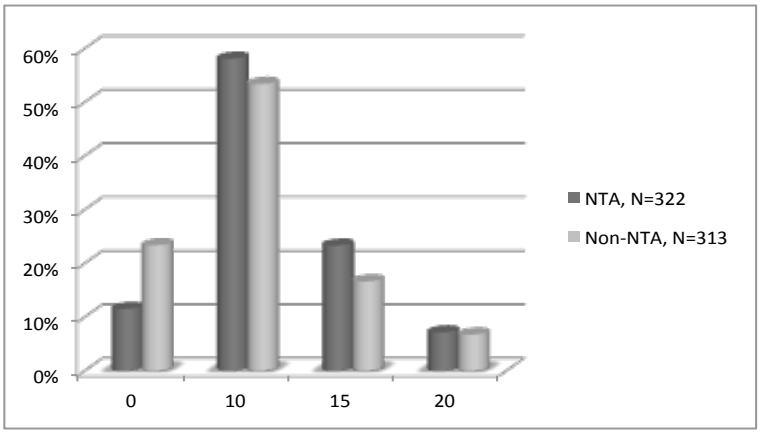

Figure 4. Test grade distributions for NTA and matched non-NTA individuals, ***physics***. Note: 0 = Fail (F), 10 = Pass (P), 15 = Pass with Distinction (PD), 20 = Pass with Special Distinction (PSD)

The course grade distribution shows the very different result that no significant differences exist between the groups. This demonstrates the benefit of having access to measurement instruments with higher precision than grades.

Control covariates and heterogeneous participation effects

The results using multivariate regression are similar to those from previous unconditional tests. The regressions take into account the control variables (parental characteristics, foreign background, student/teacher ratio, etc.) and heterogeneity (differences between the years 2009 and 2010, gender and participating in mother tongue). The first case, with only control variables, shows a significant positive effect from participation in the NTA for all natural science subjects together and for physics. The differences are slightly smaller than the unconditional tests; the effect is reduced to 6.4%. No significant effects are present for course grades. The conclusion is that the basic estimates are not sensitive to control covariates. Only one difference, for test grades in biology, shows a weakly significant negative effect.

Differences related to time, gender and immigrant background do not affect the results significantly, except in one case. Those students participating in mother-tongue education would benefit from using the NTA. However, the number of individuals in the sample is very small, which makes these conclusions unreliable. This would then make the general NTA effect insignificant. From these results we draw the conclusion that our samples are not large enough to provide precise estimates of both general and heterogeneous effects of participation in the NTA program.

The results can be summarized as follows. Participation in the NTA provides a general positive effect on the results of national tests in science carried out at the end of Year 9. The effect is limited to physics, but the average score on the physics test for a non-NTA pupil would have increased by 16.5% had (s)he participated in the NTA program, which we consider to be a large effect. This affects the results and provides a positive effect even when all subjects are weighted together. Attempts to account for heterogeneous effects show that the number of individuals in the study is insufficient to divide into subgroups.

9 Results

9.1 Under what conditions are PSA analyses a useful method for measuring the effects from a school improvement S & T initiative?

The aim of this paper is to compare the performance of two groups that do not have the same opportunities. PSA offers the opportunity to solve this problem. However, several factors must be considered for this to be feasible in a more systematic manner. Technical details are covered in-depth in the two articles

summarized above. General considerations are presented below and are further discussed in section 9.

Adequate data and control

First, there must be adequate control over when, where, by whom and how often an initiative is used. A lack of this information may have led to many schools being unnecessarily classified as unable to participate in the NTA or to schools choosing not to participate in the NTA. These schools have been placed in category V: "No information was available about whether the school's students had participated or not". The presence of this group reduces the possibility of matching participants and thus reduces precision.

It is also important to keep track of the level at which decisions are made. This information is necessary in the process of determining which variables should be included in PSA calculations. This information includes data on the level at which decisions are made about participation, on what grounds the decisions are made, and so forth. The variables available to this study could have been more numerous. The availability of more variables on individual and class level also improves accuracy.

Measuring similar initiatives in the future will require sufficient participants to provide necessary data. The system of collecting school results at the national level provides great opportunities to retrospectively match records, but it is important that all necessary steps involved are well organised.

Identifying group characteristics

Article 1 addresses the first question: "Does the treatment group differ from a group of average students?" The article shows the necessity of considering the differences that exist between different groups of pupils in schools. The use of the NTA in schools is not randomly distributed across the country geographically or socioeconomically. Choosing to participate in the NTA is primarily decided on the municipal and school levels. The study shows differences between NTA students and non-NTA students and suggests that the NTA is used for compensatory purposes.

Geographically, major differences are present in where the NTA is used the most. It is more commonly used in larger cities with more financial resources. It is used more often in suburban schools with lower performing students. The spread between the cities is also very uneven; usage is high in the capital but the program is virtually absent in Sweden's two other major cities. Measuring which of the two groups received the best test scores without taking into account these differences would yield unreliable results. Our investigation concludes that NTA students differ from non-NTA students.

Creating comparable groups

The differences highlighted in Article 1 must be addressed in a statistically appropriate way to answer the next research question, "What long-term effects could be measured in the treatment group regarding performance in S & T subjects?" This question is answered in Article 2.

The article describes a possible method for making the comparison as fair as possible. If the previously described conditions for handling necessary information are met, PSA provides the opportunity to make reliable comparisons between the groups. The article carefully addresses all the steps involved in PSA that must be taken to obtain comparable groups and to measure any long-term effects from participating in a school development program.

9.2 Findings from the object of study

The effect measures of the NTA by means of PSA show some significant positive results. Large differences are present between the three different subjects, but large differences also exist depending on the outcome variable used. As expected, the results of the test scores and test grades are clearly linked.

- Statistically significant results in favour of the NTA are shown for the subject of physics. The effect size (Cohen's d) is 0.247, which is considered small by Hattie (2009). However, a non-NTA student could have increased his/her test results by 16.5% had he/she participated in the NTA program.
- In biology and chemistry, no significant differences can be observed.
- When the final grades from Year 9 are used as an outcome variable, differences between the groups disappear. None of the subjects shows any significance.
- Attempts to measure the heterogeneous effects prove unreliable when the numbers of observations are too few.

10 Discussion

10.1 Reasons for using PSA

Measuring effects using the PSA method is one way to perform statistical effect studies. This section examines why this particular method is appropriate in this case and if it is a reliable method.

Rosenbaum[30] and Rubin[31] developed the PSA method (Quigley, 2003). The mathematics behind the method is well described in the literature (Guo and Fraser, 2009); this information is too voluminous to include in this paper. Some of the method's characteristics are, however, worth highlighting in this context.

- The method makes it possible to create artificial control groups when experimental studies are impossible or inappropriate, which is often the case in school education research. In this study of the NTA, randomized trials in which participation is not influenced by decisions' at the municipal and school levels cannot be organized. It seems common that initiatives to improve education are launched but that quantitative measurability has not been considered. The PSA could be used in other school studies. (Quigley, 2003) argued that the PSA is the optimal method when randomized experiments are not available. The different variants for random selection (Vanosdall et al. 2007) or demographically matched students (Young and Lee, 2005) is not as thoroughly instruments as PSA combined with registry data of the type available in Sweden.
- This study includes observational data that works well with PSA. If other data were available, another statistical method may have been more appropriate. The registry data gathered by Swedish authorities should be used more often. The information is more or less readily available (with some time delay) and can be obtained inexpensively. The information is not limited to any particular school subject or initiative. With appropriate outcome variables and enough participants, similar effect studies are possible. It should be noted that merely having a large number of participants is insufficient to qualify as an effect study. For example, (Ekborg and Lindahl, 2006) used many participants in their study (N = 700), but they measured/collected data regarding teachers' opinions and attitudes rather than measuring the long-term effects of a school improvement initiative.
- In this study, PSA proves reliable and relatively insensitive to covariate and heterogeneous effects when the number of samples is large enough. This suggests that future studies using registry data in terms of number of participants should be able to increase measurement precision.
- Using PSA or other statistical methods in a more systematic way creates opportunities to find interesting issues; these issues can be investigated more thoroughly with qualitative studies. This is fully in line with the wishes of more quantitative measurements using large-scale data sets as desired (Vetenskapsrådet, 2013). One example of an interesting issue is the indication that low-performing students seem to benefit from the use of the NTA. This is also suggested in US research (Lynch et al. 2005;

[30] https://statistics.wharton.upenn.edu/profile/580/
[31] http://www.iza.org/en/webcontent/personnel/photos/index_html?key=6721

Rethinam et al. 2008) and in Swedish research (Anderhag and Wickman, 2007). The large difference between test grades and course grades found in this study is an example of findings that would benefit from further qualitative research.

- The ability to measure outcomes several years after treatment, described by (Bredderman, 1983), is possible using the PSA or other statistical methods. PSA requires some sort of "matching" data. The delay in obtaining official records from SCB (from reporting data to accessing data) is around one year, which is a factor to consider when planning a study.

Need for organized collection of measurement data

This study demonstrates that a better method for recording data could have improved and eased the evaluation of effects. The NTA organisation would have less of a burden if there were working practices and a better way to record data. Performing impact studies on a large scale requires a more organized collection of measurement data at the outset of a project. This area is where major problems seem to occur, this cost time and money.

Authorities have implemented legal barriers to recording data systematically using databases. Instead of introducing obstacles, authorities and research organizations should offer opportunities to collect data that enable advanced studies. The collected data could then more easily be linked with existing registry data at Statistics Sweden. This would be even more useful now that national tests are present for other subjects and grades (Years 3 and 6) that provide new opportunities for efficacy studies. It would also be helpful with any practice that identifies variables that need to be collected and could offer practical advice for this collection. Another question is how many individuals' data must be sampled from to get reliable results. The 1000 participants in this study were insufficient to divide the results into subgroups to study heterogeneous effects.

Measurability as criteria for founding

Whether measurability is a criterion for obtaining various grants may be called into question. In many instances, various observations and surveys suffice as proof that something has succeeded. Participating students, teachers and other stakeholders involved express their appreciation for the material. Requirements should be set before money is granted that call for measuring whether the initiative actually works. If no organised data collection exists from the outset, it may be impossible, or at least unnecessarily difficult and expensive, to measure the impact of school initiatives.

Today, researchers often depend on natural experiments. These experiments are based on coincidences that make it possible to measure an initiative. Early involvement of statisticians in the initiative planning and development process should result in improved measurability.

The missing level

Data collection at SCB should also be extended with data linking individual students to the teacher and his/her certification. If this information were available as registry data, it would provide better opportunities for using PSA and other statistical methods to study classroom effects. Classroom variables are currently missing from the registries.

This proposal may be controversial, but the variation between different classes is a factor that shows one of the largest variations in school research. It is important to investigate classroom variation and teachers' authorisation. This is clearly shown in (Rethinam et al., 2008). Ambiguous messages about the benefits of qualified teachers are available from (Andersson et al. 2011; Wiliam, 2009; Young and Lee, 2005) and from the widely cited meta-study by (Hattie, 2009).

Various professional development programs and in-service training, such as the extensive guided teacher training or scaffolding that (Vanosdall et al. 2007) describes, should be statistically evaluated. It should then be compared with the relatively short training the NTA provides for teachers. Does the difference in training matter? Is it possible that additional training would provide more of an effect on the NTA students' knowledge?

Within this data, it should not be possible to identify individual teachers or groups. This should instead be seen as an opportunity to obtain a deeper understanding of the factors related to long-term results in school. Measurements and quantitative methods should be used to find interesting anomalies that are then explored with qualitative methods.

Higher precision without more work

The data collected from national tests should include test scores instead of only test grades. This is demonstrated in this study, where the results from test grades differed somewhat from test scores and excessively from course grades. Standardized tests provide increased precision through their scale of measure, which has about 40 steps instead of the 7 steps provided by test grades. The increase in work for teachers to register a numerical value and a letter is negligible. As shown in this study, evaluating effects on course grades only is insufficient because the difference from the test grades is very large.

Outcome variables

The previously mentioned criticism of national tests must be taken seriously, and what is really measured by the tests should be examined. NAE claims that they "assess knowledge". Since no other larger or more appropriate measurements are available, we have to accept what is available as good enough. The national tests are very thoroughly tested and evaluated by the creators, who have extensive experience at different universities. Other data at the municipal

level, such as local tests, would also be applicable for efficacy studies. Such data are available from several municipalities in Sweden.

Size matters

The number of potential participants in studies makes a difference between quantitative measurements and qualitative measurements. Provided information linking the identity of the participants to the test results is present, the opportunity to combine this information with SCB registry data is a strength. This information includes a great deal of the socioeconomic information found to have a significant impact on academic success (see SALSA). The registry data eliminates much of the burden of providing data from participating students and teachers. All the available variables are presented by SCB in a 89-page PDF file (Statistiska centralbyrån, 2013). The data is made available yearly and provides further possibilities for studying trends, thereby improving the accuracy of studies. This is unique to a few countries; however, it should be provided more often. None of the Swedish or US studies found in searches for this study used registry data.

Believe(r)s

The scarcity of quantitative follow-up research on small school improvement initiatives is surprising. More evaluations should be performed on initiatives, including smaller educational initiatives. Increased quantitative evaluation could improve decisions made about schools in the future. While information on school issues is often provided based on stakeholders' thoughts or beliefs, information provided based on measurements should be used more often to improve the quality of school initiatives in the future.

Methodological issues

The usefulness of a PSA with only one example could be seen as problematic. It would be preferable to evaluate other school improvement initiatives as well. The problem with that approach is that the study would be too complex and time consuming. This study should be seen as the first step in developing a working practise for evaluating smaller initiatives in schools.

Another issue is that the outcome variables could be seen as kind of a blunt instrument that cannot measure students' different abilities. IBSE, the method used by the NTA, might improve skills that cannot be measured with the variables available from the national tests.

10.2 Concluding comments regarding findings in relation to the object of study

Some forward-looking possibilities for further research that have been observed in this study are as follows.

Physics in the NTA stands out

Anderhag and Wickman (2006) showed increases in scientific linguistics and in the use of scientific concepts in general. This study shows that the effects are limited to physics. Significant differences are present in the subject of physics between the treatment group and the control group in test scores and test grades; no differences are present in the course grades. The test grades are more in line with course grades in NTA schools. The reason for this discrepancy in grading can only be speculated on. Is it possible that NTA schools provide more honest grading than non-NTA schools, which in that case would put to high grades? (Klapp Lekholm, 2008) writes about the issue of how schools' grading varies according to sociocultural conditions.

Low performing students seem to gain the most from the NTA

It also appears that those who benefit the most from this initiative are the low-performing students; the number of students with the lowest grades is small compared to the control group. Even if the treatment is performed in the lower years and the measures are made in year 9. However, this difference is not significant as the number of observations is too small. If the low-performing students benefit in this case, this is a positive effect of the program.

Biology and Chemistry?

Another question is why the study does not show any differences in biology and chemistry. (Bredderman, 1983) showed how the effect of treatment fades after a few years. This is a possibility in this study. It could also be due to the teachers' subject knowledge and which NTA boxes they choose to use in their instruction. Information provided by NTA representatives suggests this. Some of the physics themes are used most often. NTA has implemented more comprehensive data collection to determine how the material is used in schools. This information would provide the possibility for precise evaluations in the future. This is another example of an opportunity to combine a quantitative and a qualitative approach. Interviewing teachers or observing classrooms could provide information that is not visible in registry data.

10.3 Contribution

Since STC and similar systems are international phenomena, this study may be of interest to other countries that do not have access to registry data. The study shows that the PSA method is a useful way to measure the impact of initiatives on school improvements. Thus, it is possible to provide additional information

to policymakers and local school administrators that qualitative studies cannot offer. The rich source of registry data in Sweden makes this method very suitable for evaluating initiatives in school. The method is useful if the requirements for the organization of the collection are fulfilled.

However, the lack of a simple way to register and manage data is apparent. This makes it unnecessarily difficult to perform efficacy studies. In the future, it would be helpful if those who fund school initiatives and school research could also provide the opportunity to record measurement data. This could also solve problems with data legislation if this is handled in a professional manner. Connecting collected data with registry data at the SCB should also be handled more efficiently. As the situation stands now, data registration must be implemented locally by each individual project, which is very tedious.

A next step could be to place demands on measurability as a prerequisite for funding. Not all projects can be measured with quantitative methods, but it should be possible to evaluate effects in more cases than are currently evaluated. Bredderman´s, (1983) study showing that the effects of initiatives in schools subside after a few years raises the question of whether some education is wasted if students who received treatment did not differ from those who did not receive treatment after only three years.

The NTA has other positive effects, such as how teachers' work is affected when the work with tools and supplies in the institution are facilitated.

During the time I have worked on this study, I have more than once heard the phrase "either you have NTA or you have nothing". School improvement initiatives should be measured with quantitative methods if possible. It is irresponsible not to attempt obtaining measurements. Future researchers should determine interesting issues using quantitative methods and investigate these issues more thoroughly with qualitative studies. This research area would benefit if this were done more often.

Another area left to investigate is whether the use of the NTA influences students' choices of upper secondary schools. Will Sweden wake up from its own "Sputnik shock", and will more students choose scientific or technical programs? This question will be investigated in a forthcoming study, in which a survey of two years of national test data in science subjects from SCB will be utilized. The number of observations in this survey will be approximately 200,000.

The randomized data from 15,000 participants used in this study also includes test results on an item level, which could be used in the future to investigate if the use of NTA affects different abilities or knowledge. In a few years, it will also be possible to investigate students' choices for higher education.

Evaluate thoroughly if possible instead of believe!

Stockholm 2013

Joakim Svärdh

Part 2

Summary in Swedish

SVENSK SAMMANFATTNING

ATT MÄTA LÅNGVARIGA EFFEKTER AV SKOLSATSNINGAR

11 Introduktion

> De politiska besluten behöver följas upp och utvärderas. Som en konsekvens av detta har regeringen uttryckt en önskan om att få tillgång till beslutsunderlag av hög vetenskaplig kvalitet. Enligt regeringen behöver antalet studier med sikte på effekter av utbildningspolitiska beslut öka. …Utvärdering ska ingå som en del i varje större reform (Utbildningsdepartementet, 2009).

Det uttrycks ett ökande intresse för statistiska effektstudier med hjälp av registerdata. Både Vetenskapsrådet och Regeringen pekar på vikten av att fler pålitliga effektmätningar görs (Vetenskapsrådet, 2011, 2013). I Sverige liksom i många andra länder finns en debatt om hur väl skolan fungerar. Stora internationella mätningar som TIMSS[32], PISA[33] och ROSE[34] uppmärksammas i media, där listor jämför länders olika skolframgångar. Frågan är vad som orsakar vissa länders framgångar och andra länders försämrade skolresultat engagerar många. I forskarvärlden bedrivs en mer seriös debatt över strategier för förbättra utbildning och lärande. Bland annat så har bedömning uppmärksammats mycket på senare tid (Hartell, 2013; Hattie, 2012; Wiliam, 2009).

Effektutvärderingar i Sverige har koncentrerats till storskaliga skolreformer (i linje med regeringens önskemål). Mindre skolsatsningar har inte uppmärksammats lika mycket av nationella och lokala beslutsfattare. Förutom ekonomiska svårigheter att effektutvärdera mindre skolsatsningar så finns det dessutom metodologiska svårigheter som måste lösas. Högkvalitativa effektstudier kräver pålitliga metoder. *Huvudsyftet i denna studie* är därför att utforskas möjligheterna hos *en* statistisk metod, Propensity Score Analysis (PSA). Metoden är väletablerad och är utformad för att skapa statistiskt jämförbara grupper i de fall då det inte är möjligt att genomföra randomiserade experiment (Quigley, 2003).

---

[32] http://timss.bc.edu
[33] http://www.oecd.org/pisa/
[34] http://roseproject.no

Naturvetenskap och Teknik – ett utbildningsområde med problem

Frågan är då vilken av alla mindre skolsatsningar som skall väljas ut för denna effektstudie. Alla skolsatsningar går kanske inte att mäta med kvantitativa metoder men det finns lämpliga kandidater.

Elevers kunskaper i naturvetenskap och teknik (N & T) visar på sjunkande resultat. Antalet studenter som söker sig till högre utbildning i ämnena anses också vara för få (European Commission, 2004; Teknikdelegationen, 2009a, 2010; TIMSS, 2011). Det finns också en brist på behöriga Naturvetenskaps- och tekniklärare som bekymrar (Skolverket, 2013; Hartell & Svärdh, 2012).

Nationella inventeringar inom N & T listar minst 250 olika satsningar som försöker förbättra denna situation. Dessa omfattar allt från mindre tävlingar till större engagemang som sträcker sig över flera år (Rooke, 2013; Teknikdelegationen, 2009c). NTA är ett mycket välkänt och välanvänt material för N & T undervisning i grundskolan. Det är därför lämpligt att använda som exempel för att utvärdera dess effekter med hjälp av PSA. I december 2012 användes NTA i 110 kommuner av 180 000 elever. NTA är en vidareutvecklad och översatt version av det Amerikanska materialet Science and Technology for Children (STC).[35] För att testa användbarheten hos PSA har skolsatsningen Naturvetenskap och Teknik för Alla (NTA) valts ut.

> Syftet med NTA-programmet är att stödja lärarna i deras ansträngningar för att stimulera elevernas nyfikenhet, intresse och kunskaper inom naturvetenskap och teknik. NTA-programmet erbjuder och utvecklar metoder samt tjänster och produkter för att förbättra lärande och undervisning i naturvetenskap och teknik, både på övergripande kommunal nivå och för enskilda skolor (NTA, 2013)

Nationella prov i år 9 som utfallsvariabel

Det finns många sätt att mäta elevers studieframgångar. I Sverige skriver eleverna nationella prov i flera ämnen. Trots att det finns kritik mot användningen av dessa prov så är de användbara för storskaliga nationella effektstudier.[36] I denna studie används provresultaten, provbetygen och elevernas slutbetyg (år 9) i biologi, kemi och fysik som utfallsvariabler. Ett nationell representativt urval om nästan 16 000 elever från 2009 och 2010 års prov utnyttjas. Tyvärr finns det inga nationella prov i ämnet teknik vilket gör det omöjligt att mäta ämnet i denna studie.

Valet av forskningslitteratur är i huvudsak begränsat till svenska och amerikanska studier.

---

[35] STC http://www.ssec.si.edu/curriculum/about-our-curriculum

[36] http://www.skoloverstyrelsen.se/?p=924

Sammanfattningsvis

- Studien undersöker användningen av PSA som ett verktyg för att mäta effekter av ett skolutvecklingsinitiativ i N & T (NTA-projektet).
- Studien behandlar användning av PSA för att leta efter (1) långsiktiga effekter som kan mätas, (2) vad man behöver ta hänsyn till för att kunna mäta detta, och (3) hur detta skulle kunna göras.
- Utfallsvariabler som används för att mäta/utvärdera de långsiktiga effekterna från det studerade programmet är elevernas prestationer på det nationella provet och elevernas slutbetyg i årskurs 9.
- Några av resultaten om effekterna från användning av NTA-programmet tas också upp.
- Studien undersöker inte pedagogiska metoder eller olika teorier om lärande. Studien är inte en fullständig utvärdering av NTA, utan mäter endast en aspekt av NTA: s påverkan.

12 Syfte och forskningsfrågor

Den övergripande forskningsfrågan är: **Under vilka förhållanden är PSA en användbar metod för att mäta effekter från skolsatsningar i N & T?**

Första delstudien uppmärksammar hur den typiska NTA-skolan ser ut, dvs. är behandlingsgruppen representativ, och om inte, vad skiljer den från kontrollgruppen? Denna fråga behandlas i huvudsak i artikel 1.

Fråga 1. Skiljer sig behandlingsgruppens elever från genomsnittseleven? Här avses geografiska och socioekonomiska parametrar. När hänsyn tas till hur svaret på fråga 1 utfaller, ger PSA möjligheter att skapa jämförbara grupper.

För att testa PSA som ett sätt att mäta långsiktiga effekter från att använda NTA ställs nästa forskningsfråga.

Fråga 2. Vilka långsiktiga effekter är det möjligt att mäta hos behandlingsgruppen avseende deras prestationer i naturvetenskap?
Denna delstudie undersöker om användningen av NTA under 4-5 terminer i grundskolan ger några långsiktigt mätbara effekter på elevernas nationella provresultat i år 9. Med *effekt* avses här en statistiskt signifikant skillnad mellan elever som har deltagit i NTA-undervisning jämfört med statistiskt likvärdiga elever som inte har deltagit i NTA-undervisning. Frågan behandlas i huvudsak i artikel 2.

13 Mäta kunskaper

Här följer en översikt av svensk och amerikansk forskning som behandlar försök att mäta effekter av naturvetenskaplig utbildning. Hur gör man och vilka mätinstrument används?

Amerikanska studier om skolutvecklingsinitiativ i N & T

Det finns en omfattande forskning som försöker etablera effekter av skolsatsningar som ofta benämns ”kits” eller ”inquiry-based instructions”. Satsningarna är baserade på konstruktivistiska teorier. Studierna använder flera olika metoder och dessa tillsammans med huvudsakliga resultat beskrivs nedan.

En metastudie kombinerar resultat från många andra studier.

Att använda ”kits” i undervisningen anses ge bättre resultat än traditionell skolboksundervisning (Klentschy m.fl. 1999) Metoderna är inte nya och föregångare fanns redan på 60- och 70-talet där de ansågs ge goda effekter på bl.a. lågpresterande elever. Flera stora metastudier med 278 ingående studier visar på signifikant positiva resultat (Bredderman m.fl. 1996).

Ett förtest används för att etablera på vilken nivå eleverna ligger på före en behandling. Detta följs sedan upp med ett eftertest för att se hur mycket eleverna har lärt sig.

Cuevas m.fl. (2005) visar i en liten studie (n=25) med hjälp av för- och eftertest hur skillnader mellan olika socioekonomiska grupper jämnas ut med hjälp av ”inquiry-based” undervisning och fortbildning för lärare.

Scaling up Curriculum for Achievement, Learning and Equity Project (SCALE-uP) [37] är ett stort 5-årigt projekt som försöker implementera "hands-on, inquiry-based learning " i N & T för 85 000 mellanstadieelever i delstaten Maryland. Studiematerialet är sådant som är godkänt och rekommenderat av American Association for the Advancement of Science AAAS. Forskningsprojektet har förfogat över 5,2 millioner dollar och är nu inne i slutfasen. Forskarna försöker identifiera vilka villkor och metoder som fungerar för storskalig implementering.

En första delstudie tittar på skillnader mellan att använda traditionella skolböcker jämfört med ett “kitt” i kemiundervisning (Lynch m.fl. 2005). Skolorna i studien är slumpvis matchade till behandlingsgruppen eller kontrollgruppen. 1500 elever i år 8 ingår i studien som utnyttjar enkäter för att mäta attityder samt för- och eftertester för att mäta kunskapsförändringar. Testerna är anpassade efter nationella riktlinjer. På Förtesterna syns inga skillnader mellan grup-

[37] http://www.gwu.edu/~scale-up/index.html

perna däremot visar Eftertesterna på ett signifikant högre resultat i behandlingsgruppen. Lågpresterande elever verkar gynnas i högre grad.

Flernivåanalys tillhör en grupp med olika statistiska metoder som hanterar data från flera nivåer som kommun, skola, klassrum etc.

I samma studie som ovan används också flernivåanalys för att förfina resultaten. Skillnaden mellan grupperna blir ännu tydligare och det visar sig också att 15 % av variationen kan härledas till skillnader mellan klasser (Rethinam m.fl. 2008). Utbildningsgapet mellan låg och högpresterande elever minskar, effekten blir större ju fler färgade elever som finns i klassen. Studien visar också att mer information på klassnivå om t.ex. lärare behövs.

Deltagare matchade på olika egenskaper

I en studie med demografiskt matchade elever i år 5 jämförs hur användningen av STC och liknande material påverkar kunskaperna jämfört med traditionell undervisning (Young & Lee, 2005). Skriftliga test anpassade efter nationella standarder används som mätinstrument. Signifikant positiva resultat visas för behandlingsgruppen. Lärares olika utbildning jämförs också men visar inte på några signifikanta skillnader på elevernas resultat.

Alla deltagare har en lika stor och oberoende chans att hamna i behandlings- eller kontrollgruppen.

20 frivilliga lärare (563 elever i år 5) från 4 skolor delas slumpvis in i två olika grupper (Vanosdall m.fl. 2007). Studien använder undervisningsmaterial från FOSS, Full Option Science System[38], i kemi. Lärarna i kontrollgruppen får fortbildning i hur man använder materialet. I kontrollgruppen används traditionell undervisning. Standardiserade för- och eftertest används som mätinstrument. Studien visar signifikant positiva resultat. Data analyseras med hjälp av en hierarkisk linjär modell (HLM) i två nivåer.

I en andra delstudie matchas 24 lärare från 11 skolor (762 elever i år 5) för att därefter slumpas in i behandlings- eller kontrollgrupp. Lärarna med lång erfarenhet får nu antingen ta del av en utökad fortbildning (scaffolding) eller den fortbildning som ingår i FOSS. Även denna studie visar på signifikant positiva resultat, analyser gjordes med HLM.

Det finns mycket få studier som påvisar negativa resultat. Ett undantag är Breddermans (1983) metastudie som visar att eventuella effekter av att använda ”activity-based programs” i låg och mellanstadiet inte finns kvar i slutet av högstadiet.

Dessa studier får exemplifiera den amerikanska forskningen. I huvudsak rapporteras positiva resultat. Flera olika metoder för att försöka åstadkomma jäm-

[38] Full Option Science System http://lhsfoss.org/scope/research/search.php

förbara grupper används. De mest avancerade studierna använder sig av flernivåanalyser, ingen studie använder sig av registerdata.

Svenska studier om skolutvecklingsinitiativ i N & T

Det finns mycket forskning om NTA, minst 70 olika studier har listats. De flesta studier är kvalitativa och bara några få kan till viss del beskrivas som kvantitativa.

Attityder, innehåll och lärares förmågor

Den tidiga forskningen beskriver lärare som nöjda med materialet. Mindre ändringar föreslås. Eleverna beskrivs som entusiastiska inför arbetet med NTA. Det noteras att för att utnyttja materialet till fullo krävs kunniga lärare. Metoder har varierat från observationer och enkäter till lärarnotiser och intervjuer. 4-5 skolor deltog i de flesta studier (Gisselberg, 2001; Schoultz et al., 2003; Schoultz & Hultman, 2002).

Anderhag och Wickman (2006) visar hur elevernas naturvetenskapliga ordförråd och begreppsanvändning ökar när NTA används i undervisningen. I studien ingick 23 lärare och 96 elever.

Studien följs upp med en intervjustudie av 80 elever i år 6. NTA-eleverna visar ökade kunskaper i naturvetenskap jämfört med kontrollgruppen. Effekten är störst hos låg- och högpresterande elever, även pojkar gynnas. SPSS har använts i analyserna och man har ansträngt sig noga för att skapa jämförbara grupper (Anderhag & Wickman, 2007).

Ekborg och Lindahl (2006) använder enkäter för att mäta hur lärares utbildning påverkar användningen av NTA (n=700). Studien visar att lärare med lite utbildning i NO följer anvisningarna i materialet mer noggrant än lärare med längre NO-utbildning. Enkäten innehåller 98 frågor och analyseras i SPSS.

De svenska studierna i allmänhet hävdar att lärare är nöjda med NTA och att elever ökar sina kunskaper. Det finns inga uppföljningsstudier där långsiktiga effekter av att undervisa med hjälp av NTA utvärderas. *Ingen* av studierna använder registerdata.

14 Studieobjekt – NTA programmet

14.1 Den amerikanska förlagan

NTA härstammar från det amerikanska undervisningsmaterialet Science and Technology for Children STC. Det fanns tidigt lådsystem i USA, redan på 60-talet gjordes ansträngningar för att förbättra undervisningen i naturvetenskap.

På 80-talet publicerades rapporten *A nation at Risk*[39] om bristen på kvalificerade vetenskapsmän och ingenjörer. Flera stora forskningsorganisationer engagerades i att förbättra undervisningen och 1991 var STC klart för att börja användas. Materialet har utvecklats över åren och även konkurrerande system finns numera[40]. STC har organiserats i två spår, STC K-6 för låg och mellanstadiet samt STC/MC 6-8 för högstadiet. Materialet har översatts till svenska och spanska och används i flera länder.

14.2 Naturvetenskap och Teknik för Alla

NTA är ett lärarstödsmaterial som har som mål att engagera elever i naturvetenskap och teknik. Det startade 1997 som ett projekt av Kungliga Vetenskapsakademien och Kungliga Ingenjörsvetenskapsakademien i samarbete med olika kommuner. NTA är en översatt och utvecklad version av STC. Materialet används främst i år 1-6 men teman finns numera även för år 7-9. 2012 användes NTA i över 110 kommuner av över 180 000 elever. Detta motsvarar ca 18 % av eleverna i grundskolan. Antalet kommuner som deltar i NTA har ökat med ca 10 st. varje år. NTA är inte jämnt fördelat över landet.

15 Metod

Mäta effekter av utbildningsinitiativ

Det verkar finnas en obalans mellan kvantitativ och kvalitativ skolforskning (Vetenskapsrådet, 2011). Sverige har en unik och mycket stor resurs i form av registerdata insamlade av Statistiska Centralbyrån (SCB). Dessa registerdata är mycket betydelsefulla för denna studie eftersom de bidrar med socioekonomisk bakgrundsinformation och provresultat från nationella prov i år 9.

Kvantitativa metoder jämfört med kvalitativa metoder

De i huvudsak kvalitativa mätningar som har gjorts av NTA har ett kort tidsspann till skillnad från denna studie som försöker mäta effekter flera år efter behandlingen. För att kunna säga något om långsiktiga effekter behövs andra metoder med fler deltagare. Registerdata har den stora fördelen att den ger möjligheter till stora behandlings- och kontrollgrupper.

Samvariation eller orsakssamverkan

Vad menas här med att mäta effekter? Frågan som ska besvaras är om NTA har orsakat högre resultat på nationella prov i NO-ämnen. För att kontrollera detta är det inte tillräckligt att studera om de som har deltagit i NTA har höga

[39] http://mathcurriculumcenter.org/PDFS/CCM/summaries/NationAtRisk.pdf
[40] Full Option Science System http://lhsfoss.org/

testresultat (i allmänhet), utan man måste se om dessa resultat är höga jämfört med elever som inte har deltagit i NTA *men som skulle kunna ha gjort det*.

Problemet med icke slumpvisa urval

Deltagandet i NTA är inte styrt av slumpen utan är påverkat av flera olika beslut på kommun- och skolnivå. Skillnaderna mellan behandlingsgruppen och kontrollgruppen som behandlas i artikel 1 måste hanteras på lämpligt sätt för att göra det möjligt att jämföra provresultaten. Ett sätt att göra detta på är att använda Propensity Score Analysis (PSA) (Guo & Fraser, 2009). Elever som har deltagit i NTA-undervisning matchas med elever som inte har deltagit i NTA-undervisning men som har haft liknande förutsättningar för att få delta. PSA skapar konstgjorda kontrollgrupper bestående av ”tvillingar” med liknande förutsättningar. Metoden anses som ”optimal” (Quigley, 2003) för att konstruera kontrollgrupper i observationsstudier där inte randomiserade experiment går att utföra. PSA summerar flera olika variabler, som har påverkat valet att delta i en behandling (NTA), till en gemensam variabel. Detta görs genom en multvariat logistisk regression[41] som ger alla elever ett värde i form av en statistisk sannolikhet att ha deltagit i NTA-undervisning. Elever som *har* deltagit i NTA-undervisning matchas med elever som har så lika sannolikhetsvärden (propencity score) som möjligt men *inte har* deltagit i NTA-undervisning. Efter Matchningen används T-test[42] och icke-parametriska test (Mann-Whitney)[43] för att se om signifikanta skillnader finns mellan gruppernas provresultat. Resultaten kontrolleras för kvarvarande skillnader med hjälp av regressionsanalyser. Denna process beskrivs utförligt i artikel 2.

Utfallsvariabler

2009 introducerade skolverket nationella prov i NO-ämnen för år 9. Resultaten från dessa prov kompletteras med registerdata från SCB. Ett statistiskt slumpvis urval med provresultat och betyg från ca 16 000 elever för åren 2009 och 2010 har samlats in av provkonstruktören, Umeå Universitet. Dessa provresultat tillsammans med de kategoriseringar av NTA-skolor som gjorts i artikel 1, gör det möjligt att jämföra provresultaten för NTA-elever och icke NTA-elever.

[41] http://sph.bu.edu/otlt/MPH-Modules/BS/BS704_Multivariable/BS704_Multivariable8.html

[42] http://www.physics.csbsju.edu/stats/t-test.html

[43] http://www.statisticslectures.com/topics/mannwhitneyu/

16 Sammanfattningar av ingående artiklar

16.1 Artikel 1

To use or not to use a teacher support program - A study of what characterizes Swedish schools that apply the inquiry-based teacher support program NTA. In M. de Vries & I.-B. Skogh (Eds.), *Technology teachers as researchers: Philosophical and empirical technology education studies in Swedish TUFF research school*. Sense Publisher.

Denna studie beskriver den socioekonomiska och geografiska skillnad som finns mellan skolor och kommuner som använder NTA jämfört med de som inte använder NTA. Genom en enkätundersökning och personliga samtal, har alla Sveriges skolor identifierats och kategoriserats. Kategorierna baseras på hur mycket en skola använder NTA.

Kommunnivå

Genom att använda registerdata från Sveriges kommuner och landsting (SKL) kan man se hur NTA är vanligast i Stockholms förorter och i en del andra större universitetsstäder. I Malmö och Göteborg används det inte alls med undantag för någon enstaka friskola. Den genomsnittliga NTA-kommunen har en större befolkning med högre inkomster än icke NTA-kommuner. Däremot så skiljer det inget mellan andelen behöriga lärare eller hur många anställda som finns på skolorna per elev.

	NTA	Icke-NTA
Befolkning i kommunen	45 000	27 000
Medelinkomst	178 500 SEK	172 700 SEK
Andel behöriga lärare	86 %	86 %
Elever per anställd	9.5	9.5

På kommunnivå är skillnaderna i meritvärden mycket små med endast en poängs övervikt för NTA-kommunerna. De synliga skillnaderna är i stället troligen förknippade med de historiska skillnader som har funnits mellan stad och landsbygd där inkomstskillnader och skillnader i akademisk bakgrund finns.

Skolnivå

På skolnivå så är skillnaderna större. Genom att använda 11 år med data från Skolverkets databas SALSA (ca 15 000 mätvärden) kan man se hur NTA-skolor skiljer sig från icke NTA-skolor i samma kommun. Den generella utbildningsnivån hos föräldrar ökar något varje år. Föräldrarna på NTA-skolor har något lägre utbildningsnivå (ca 3 %, figur 1) än genomsnittsföräldern. De skolor inom kommunen som inte använder NTA har istället 3 % högre utbildningsnivå än medelföräldern.

NTA-elever har också något lägre meritpoäng (ca 3 p) i år 9 än medeleleven. Skillnaden verkar dessutom accelerera (Figur 2). Även här utmärker sig icke NTA-eleverna med att ha högre meritvärden än medeleleven (ca 3 p).

När man använder SALSA-värden som är justerade för kön, invandrarbakgrund och föräldrars utbildning sjunker skillnaderna något.

Andelen elever med invandrarbakgrund i skolor som använder NTA var 4 % högre för år 1999 men har minskat till när 0 för år 2011.

Det fokus som finns i de Amerikanska studierna på att stödja lågpresterande elever är inte alls uttalat i den svenska kontexten. Tidigare svenska studier beskriver i huvudsak lärares erfarenheter av att arbeta med NTA. Denna studie visar på skillnader i meritvärden som verkar vara förknippade med socioekonomiska faktorer. Över tid kan man misstänka att skolval och ökad segregation har påverkat resultaten. Valet att delta i NTA är inte styrt av slumpen, utan bakom beslutet kan ligga en önskan att kompensera låga skolprestationer orsakade av socioekonomiska faktorer. Diagrammen som presenterar data över tid (figur 2) visar på en tydlig trend med ökande skillnader.

16.2 Artikel 2

Inquiry-based learning put to test: Long-term effects of the Swedish Science and Technology for Children program. Submitted for publication in *IFAU's Working paper-series*. Mellander, E., Svärdh, J.

Urvalsproblem

Denna studie gör en kvantitativ utvärdering av Naturvetenskap och Teknik för Alla (NTA). För att kunna mäta om deltagande i NTA-undervisning ger några långsiktiga effekter måste man ta hänsyn till den systematiska skillnad mellan NTA-elever och icke NTA-elever som framkommer i artikel 1. Eftersom det inte är möjligt att utföra randomiserade experiment för att mäta eventuella effekter, så måsta man åstadkomma en jämförbar kontrollgrupp på något annat sätt.

Metod och data

Genom att använda Propensity Score Analysis (PSA), matchas en NTA-elev med en icke NTA-elev som har så statistiskt lika förutsättningar som möjligt (Guo & Fraser, 2009). Bakgrundsvariabler från SCB används för att få fram ytterligare information på kommun- och skolnivå. Ett slumpat urval om ca 16 000 elever, med provresultat och betyg från Nationella proven för år 2009 och 2010 i NO-ämnen, används som mätdata. Ur dessa matchas 1000 NTA-elever med 1000 icke NTA-elever för att därefter göra jämförelser av provresultat, provbetyg och kursbetyg.

Resultat

De standardiserade *provresultaten* i allmänhet visar på positiva skillnader. Ett T-test ger en signifikansnivå på 1 %, NTA-eleverna har ett medelvärde på 48,3 % medan icke NTA-eleverna har ett medelvärde på 44,5 %. Om man delar upp resultaten i respektive ämne visar det sig att hela skillnaden kan härledas till ämnet fysik. Ämnena biologi och kemi visar inte upp några signifikanta skillnader mellan grupperna. I fysik skiljer det ca 7 % mellan behandlingsgruppen och kontrollgruppen (figur 1).

Testbetygen ger liknande resultat, Mann-Whitney test ger också en signifikansnivå på 1 %. Även här står ämnet fysik för hela skillnaden (figur 2).

Kursbetygen visar på något helt annat, det finns *inga* signifikanta skillnader mellan grupperna i något ämne.

Regressionsanalyser används för att kontrollera för kvarvarande skillnader mellan grupperna (t.ex. föräldrars utbildning, invandrarbakgrund, lärare/elev etc.). Effekten blir något mindre och sjunker till 6,4 % för fysikämnet, för övriga ämnen består de uteblivna skillnaderna. Försök att dela in resultaten efter heterogena grupper som kön och år för provet visar sig inte fungera eftersom antalet deltagare blir för få.

Resultatet kan sammanfattas som att en icke NTA-elevs provresultat i fysik skulle ha ökat med 16,5 % i medelvärde om denne hade deltagit i NTA-undervisning. Vi anser detta som en stor effekt.

17 Resultat

Under vilka förhållanden är då PSA en användbar metod för att göra effektutvärderingar av satsningar i skolan? Det man vill åstadkomma är en jämförelse av två grupper som inte har samma förutsättningar. Detta innebär att det är många faktorer att ta hänsyn till för att man ska kunna använda PSA i mer systematiska mätningar.

Kontroll på data

Satsningar som ska utvärderas måste ha god kontroll på sina mätdata: vilka har använt, när, hur mycket etc. Om inte denna noggranna kontroll sker finns stora risker att mycket mätdata inte går att klassificera som deltagit eller icke deltagit, utan hamnar i en gråzon med bortfall. Information måste också finnas om beslutsvägar, vem har fattat beslut om att delta och på vilken nivå. Fler variabler på individnivå skulle också behövas för att beräkningarna som används i PSA ska fungera så bra som möjligt.

Framtida mätningar kommer också att kräva många deltagare för att bli så pålitliga som möjligt. De nationella provresultaten och registerdata hos SCB erbjuder goda möjligheter till mätningar om det bara finns en välorganiserad insamling av övriga data.

Identifiera vad som karaktäriserar en grupp elever

Artikel 1 behandlar frågan: *Skiljer sig behandlingsgruppen från genomsnittseleven?* Studien visar att stora systematiska skillnader finns mellan grupperna. De skiljer sig åt med avseende på socioekonomiska faktorer både på kommun-, skol- och individnivå. Skolorna i behandlingsgruppen är inte heller jämt fördelade varken över landet eller inom kommunerna. Det ser ut som att NTA i viss mån har använts i kompensatoriskt syfte. Skulle man försöka jämföra grupperna utan att ta hänsyn till dessa skillnader skulle det var som att försöka jämföra äpplen och päron (Blackstone, 2002).

Skapa jämförbara grupper

De skillnader mellan grupperna som framkommer i artikel 1 måste hanteras. Hur detta går till beskrivs utförligt i artikel 2.

PSA erbjuder möjligheten att skapa jämförbara grupper för att kunna mäta eventuella långsiktiga effekter. Artikeln beskriver utförligt hur detta hanteras.

Resultat avseende NTA

Effektstudien visar på signifikant positiva provresultat och provbetyg för fysikämnet. Biologi och kemi visar inte på några signifikanta resultat. Resultaten med kursbetygen som mätinstrument skiljer sig mycket från provresultaten och inga signifikanta skillnader finns mellan grupperna. Försöka att mäta heterogena effekter, från t.ex. invandrarbakgrund, visar sig opålitliga då antalet observationer är mycket få.

18 Diskussion

Att använda PSA

Det finns många olika statistiska metoder som går att använda i effektutvärderingar. Här beskrivs varför PSA är lämplig i detta fall. Matematiken bakom

PSA finns väl beskriven i litteraturen (S.Y. Guo & Fraser, 2009) men är väl omfattande för att rymmas i denna studie.

- Metoden gör det möjligt att skapa konstgjorda kontrollgrupper i de fall då inte experimentella studier är lämpliga eller möjliga att göra. Detta är ofta fallet i skolforskning. Quigley (2003) anser att PSA är den optimala metoden när slumpade experiment inte är möjliga. PSA skulle gå att använda även i andra fall där inte statistisk utvärdering ursprungligen var planerad. De varianter av slumpade urval som används i övriga refererade studier är inte lika precisa mätinstrument som PSA kombinerat med svenska registerdata.
- I detta fall finns det observationsdata lämpliga för PSA men även andra statistiska metoder skulle kunna ha varit lämpliga om det fanns andra typer av mätdata. Registerdata borde användas oftare, den är relativt billig och lättillgänglig. Att använda PSA är inte begränsat till studier i N & T utan bör gå bra att använda i andra ämnen också.
- I denna studie visar sig PSA vara relativt okänsligt för samvariation och heterogena effekter bara antalet observationer är tillräckligt stort. Framtida studier med större antal observationer bör betyda ökad precision i mätningarna.
- Att använda PSA på ett mer systematiskt sätt ger möjlighet att upptäcka andra intressanta frågeställningar. Dessa kan sedan undersökas mer noggrant med kvalitativa metoder. Ett exempel på sådan frågeställning är om lågpresterande elever gynnas av att använda NTA. Den stora skillnaden mellan provbetyg och kursbetyg är också mycket intressant att undersöka.
- Möjligheten att mäta effekter flera år efter behandling bör också lyftas fram. Att det finns en viss fördröjning på ca 1 år för leverans av data från SCB måste tas med i planeringen.

Nödvändigt att organisera insamling av data

I denna studie visas på svårigheter med datainsamling. Om utvärdering ska kunna ske på ett mer systematiskt sätt krävs att datainsamling kan hanteras på ett mer effektivt sätt på central nivå. Myndigheter och lagstiftning försvårar möjligheten att samla in data. Detta borde i stället underlättas med centraliserad insamling och förbättrade kopplingar mot SCBs registerdata. Ytterligare möjligheter till effektstudier finns numera i och med att nationella prov införs i fler ämnen och i fler årskurser.

Mätbarhet som ett kriterium för att få finansiering

Denna fråga kan anses kontroversiell men ofta används olika typer av observationsstudier och enkäter som ett kvitto på att en skolsatsning verkar fungera. I stället bör statistisk mätbarhet ingå som en del av planeringen. Görs inte detta i tid kan mätning försvåras eller fördyras. I dag är forskning ofta bero-

ende av så kallade naturliga experiment där något av en slump råkar bli mätbart. Använd statistiker som rådgivning tidigt i planeringen för att undvika detta problem.

Den saknade nivån

Datainsamling hos SCB bör utökas med klassrumsnivån, i dag saknas möjligheten att koppla ihop undervisande lärare med dennes elever. Om klassinformation fanns tillgänglig skulle olika typer av klassrumseffekter kunna studeras mer effektivt Forskning visar på mycket stora skillnader mellan olika klassrum.

Högre precision utan ökad arbetsinsats

De insamlade provresultaten från nationella proven bör utökas till att omfatta även provpoängen. Att mata in ett betygsvärde *och* ett provresultat kan inte anses som allt för betungande. Precisionen i mätningar kommer att öka om denna mätskala med ca 40 nivåer kan användas i stället för de få steg som finns i betygsskalan i dag. Denna studie visar att slutbetyg är ett trubbigt mätinstrument.

Utfallsvariabler

Kritiken som finns mot nationella prov som mätinstrument måste tas på allvar. Men forskarna bör ändå använda dessa mätdata i större studier. De nationella proven är ändå utprovade av experter på sina områden. Det finns även en del lokala och kommunala prov tillgängliga som bör gå att använda i PSA studier.

Storleken har betydelse

Möjligheten att koppla på registerdata ger möjligt till att göra mycket stora studier. Detta är en stor fördel gentemot kvalitativa studier. Registerdata innehåller också mycket socioekonomisk information som man slipper samla in på nytt genom t.ex. enkäter. Registerdata är tillgänglig för varje år vilket gör att trender kan studeras och därmed ökas precisionen i mätningarna.

Inga tidigare refererade studier kring detta fenomen, att utvärdera mindre skolutvecklingsinitiativ, har så vitt jag vet använt någon form av registerdata.

Att tro

Det är förvånande att inte fler kvantitativa studier finns för mindre skolsatsningar. Att öka den typen av studier kan ge ett bättre beslutsunderlag i framtiden. Ofta baseras information om skolsatsningars effekter på vad beslutsfattare tror och tycker. Information baserad på mätningar *kombinerade* med kvalitativa metoder skulle förbättra kvalitén på skolsatsningar i framtiden.

Metodologiska frågor

Att utforska användbarheten hos PSA med endast ett exempel (NTA) kan ses som problematiskt. Det skulle så klart vara en fördel om detta testades med andra skolsatsningar också men detta ryms inte inom denna studie. Denna

studie bör ses som ett första steg för att få till en fungerande praxis hur man kan utföra framtida mätningar av mindre skolsatsningar.

En annan invändning mot att använda nationella provresultat som mätinstrument är att det kan finnas andra förmågor som förbättras av att använda NTA. Dessa förmågor kanske inte syns i provresultaten.

Avslutande kommentarer om resultaten av effekter av att använda NTA

Fysik sticker ut som det ämne där mätbara effekter finns. Detta trots att effektmätningen sker i år 9, och den mesta användningen sker på låg och mellanstadiet men. Detta måste anses som ett gott resultat.
Provbetygen och kursbetygen stämmer bättre överens på NTA-skolor än på icke NTA-skolor. Vad detta kan bero på kan man bara spekulera i. Är det möjligt att lärare på NTA-skolor sätter mer rättvisa betyg än lärare på icke NTA-skolor?

Lågpresterande elever verkar gynnas av att använda NTA. Andelen elever med de lägsta betygen verkar, utifrån resultatet här, vara lägre i NTA-skolorna. Skillnaden är dock inte signifikant då antalet observationer är för få. Om detta ändå är fallet måste det anses som ett positivt resultat.

Biologi och kemi? Varför syns inga signifikanta effekter i dessa ämnen, möjligen kan eventuella effekter ha klingat av liksom i Breddermans (1983) studie. Det skulle också kunna bero på vad lärarna redan har för ämneskunskaper och vilka teman som lånas ut mest. Enligt representanter för NTA är några fysikteman de mest populära. NTA har påbörjat en insamling av data för att kunna göra fördjupade studier i framtiden.

Bidrag

Då STC och liknande system är en internationell företeelse bör denna studie även vara intressant i andra länder. Studien visar att PSA är en användbar metod för att mäta effekter från satsningar i skolan. Statistiska effektutvärderingar är ett värdefullt bidrag för att skapa förbättrade beslutsunderlag i skolfrågor. Den registerdata som finns i Sverige är mycket värdefull och erbjuder goda möjligheter till vidare studier. Det krävs *bättre insamlingsrutiner* för att PSA-metoden ska bli rationell att använda systematiskt.

Det finns tyvärr brister i hur datainsamling för mindre skolsatsningar ska gå till, detta gör det onödigt besvärligt att mäta effekter. En central insamling vore att föredra då även juridiska besvärligheter gentemot SCB kunde underlättas.

En annan god sak vore om mer energi lades på att se till att satsningar faktiskt blev *möjliga* att utvärdera statistiskt, eventuellt skulle det kunna ställas krav redan vid finansieringen. En intressant fråga är om det är värt att satsa pengar på något som har klingat av redan tre år efter behandlingen? Kanske det vore mer verkansfullt med fortbildning av lärare i stället?

NTA har andra mycket positiva effekter så som att t.ex. underlätta lärares arbetssituation. Jag har mer än en gång i arbetet med denna studie hört uttrycket ”antingen så har man NTA eller så har man inget”.

Skolsatsningar bör mätas med kvantitativa metoder om det är möjligt. Det är oansvarigt att inte ens försöka! De intressanta frågeställningar som då dyker upp kan då utforskas ytterligare med kvalitativa metoder.

En av frågorna som är kvar att utreda är om NTA påverkar elevers val till gymnasiet. Blir det fler som söker naturvetenskapligt eller tekniskt program? Detta är något jag ska försöka utreda i nästa studie som omfattar 200 000 elever.

Det material om 15 000 observationer som har använts i denna studie innehåller outnyttjad information på frågenivå. Detta ger möjlighet att titta på om enskilda förmågor påverkas av att delta i NTA-undervisning.

I framtiden kommer även elevers val av högre studier att kunna studeras.

Utvärdera noga om möjligt i stället för att tro!

19 Bibliography

Anderhag, P., & Wickman, P.-O. (2006). *NTA som kompetensutveckling för lärare. Utvärdering av hur lärares deltagande i NTA utvecklar deras kompetens att stödja elevernas begrepps- och språkutveckling*. Stockholm.

Anderhag, P., & Wickman, P.-O. (2007). *Utvärdering av hur NTA hjälper skolorna att nå kursplanemålen för femte skolåret i naturorienterande ämnen*. Stockholm.

Andersson, C., Johansson, P., & Waldenström, N. (2011). Do you want your child to have a certified teacher? *Economics of Education Review*, 30(1), 65–78. doi:10.1016/j.econedurev.2010.07.003

Biriell, F., & Josefsson, F. (1998). *Utprovningen av STC 1997/1998. Science and Technology for Children. Erfarenheter från de två första terminerna*. Linköping.

Blackstone, E. H. (2002). Comparing apples and oranges. *The Journal of Thoracic and Cardiovascular Surgery*, 123(1), 8–15. doi:10.1067/ mtc.2002.120329

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

Bybee, R. W. (2003). *Science curriculum reform in the United States*. Washington, DC.

Böhlmark, A., & Lindahl, M. (2012). *SNS Analys nr 7. Friskolereformens långsiktiga effekter på utbildningsresultat.*

Cuevas, P., Lee, O., Hart, J., & Deaktor, R. . (2005). Improving Science Inquiry with Elementary Students of Diverse Backgrounds. ", *Journal of Research in Science Teaching*, 42(3), 337–357.

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental casual studies. *The Review of Economics and Statistics*, 84(1), 151–161.

Ekborg, M., & Lindahl, B. (2006). *NTA som skolutvecklingsprogram. utvärdering av effekten av kompetensutveckling på lärarna och deras värderingar samt effekten på kommun- och rektorsnivå.*

European Commission. (2004). *Europe needs more scientists: EU blueprint for action.* Luxembourg.

FileMaker Pro. (2013). http://www.filemaker.com/products/filemaker-pro.

Full Option Science System - FOSS. (2013). http://lhsfoss.org/index.html

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

Gisselberg, K. (2001). *NTA-projektets tre första år - en positionsbestämning.*

Guo, Shenyang Y., & Fraser, M. W. (2009). *Propensity Score Analysis: Statistical Methods and Applications (Advanced Quantitative Techniques in the Social Sciences)*. SAGE Publications.

Hansen, M., & Lander, R. (2009). Om statens verktyg för skoljämförelser: Vem vill dansa SALSA? *Pedagogisk Forskning i Sverige*, 14(1).

Hartell, E. (2013). Exploring the (un-) usefulness of mandatory assessment documents in primary technology. *International Journal of Technology and Design Education*. doi:10.1007/s10798-013-9250-z

Hartell, E., & Svärdh, J. (2012). Unboxing technology education part I – Starting point. In *Technology Education in the 21st Century* (pp. 211–222). Stockholm: Linköping

Hartman, L., Anell, A., Mörk, E., Vlachos, J., Hanspers, K., Lundin, M., … Wiklund, S. (2011). *Konkurrensens konsekvenser. Vad händer med svensk välfärd?* http://www.sns.se/forlag/konkurrensens-konsekvenser-vad-hander-med-svensk-valfard

Hattie, J. (2009). *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*. Routledge.

Hattie, J. (2012). *Visible learning for teachers Maximizing impact on student learning*. Exeter, Devon: Routledge.

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

IFAU Institute for Evaluation of Labour Market and Education Policy. (2013). http://www.ifau.se/en/

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

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why Minimal Guidance During Instruction Does Not Work: An Analysis of the Failure of Constructivist, Discovery, Problem-Based, Experiential, and Inquiry-Based Teaching. *Educational Psychologist*, 41(2), 75–86. doi:10.1207/ s15326985ep4102_1

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

Klapp Lekholm, A. (2008). *Grades and grade assignment: effects of student and school characteristics*.

Klentschy, M., Garrison, L., & Amaral, O. M. (1999). *Valle Imperial Project in Science (VIPS) Four -Year Comparison of Student Achievement Data 1995-1999. Educational Research*. Calexico, CA.

Kommunal utvärdering av ungas livsval - KOMMUT. (2010). https://sites.google.com/site/kommutinfo/

Lee, O., & Luykx, A. (2006). *Science Education and Student Diversity: Synthesis and Research Agenda*. Cambridge University Press.

Lekholm, A. K., & Cliffordson, C. (2008). Discrepancies between school grades and test scores at individual and school level: effects of gender and family background. *Educational Research and Evaluation*, 14(2), 181–199. doi:10.1080/13803610801956663

Lundahl, C. (2013). Vilket är elevens rätta resultat? Retrieved from http://www.skoloverstyrelsen.se/?p=924

Lynch, S. J. (2000). *Equity and Science Education Reform* (p. 320). Routledge. Retrieved from http://www.amazon.com/Equity-Science-Education-Reform-Sharon/dp/0805832491

Lynch, S., Kuipers, J., Pyke, C., & Szesze, M. (2005). Examining the effects of a highly rated science curriculum unit on diverse students: Results from a planning grant. *Journal of Research in Science Teaching*, 42(8), 912–946. doi:10.1002/tea.20080

Marx, R. W., Blumenfeld, P. C., Krajcik, J. S., Fishman, B., Soloway, E., Geier, R., & Tal, R. T. (2004). Inquiry-based science in the middle grades: Assessment of learning in urban systemic reform. *Journal of Research in Science Teaching*, 41(10), 1063–1080. doi:10.1002/tea.20039

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

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

Moore, D., McCabe, G., & Craig, B. (2007). *Introduction to the Practice of Statistics*. W. H. Freeman.

National Education Goals Panel. (1995). *Data volume for the national education goals report (vol. 1)*. Retrieved from http://govinfo.library.unt.edu

National Science Education Standards. (1996). Washington, D.C.: The National Academies Press. Retrieved from http://www.nap.edu

National Science Resources Center - NSRC. (2013). Retrieved 2013, from http://www.nsrconline.org/about_the_nsrc/index.html

Naturvetenskap och teknik för alla - NTA. (2013). Retrieved 2013, from http://www.nta.kva.se/

NRC National Research Council - NRC. (2013). Retrieved 2013, from http://www.nationalacademies.org/nrc/index.html

NSRC Annual Reports. (2013). Retrieved 2013, from http://www.nsrconline.org/publications/annual_reports.html

NTA. (2013). NTA in English.

Retrieved from http://www.ntaskolutveckling.se/In-English/

Quigley, D. D. (2003). Using Multivariate Matched Sampling That Incorporates the Propensity Score to Establish a Comparison Group. *CSE Technical Report*, 1522(596).

Research on STC. (2013). Retrieved 2013, from http://www.nsrconline.org

Rethinam, V., Pyke, C., & Lynch, S. (2008). Using Multilevel Analyses to Study the Effectiveness of Science Curriculum Materials. *Evaluation & Research in Education*, 21(1), 18–42. doi:10.2167/eri418.0

Rooke, G. (2013). *In Search for Gender awareness in Technology Education*. Retrieved from http://www.diva-portal.org

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(33-38).

Rosenbaum, Paul R. (1989). Optimal Matching for Observational Studies. *Journal of the American Statistical Association*, 84(408), 1024–1032. doi:10.1080/01621459.1989.10478868

Rosenbaum, Paul R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55. doi:10.1093/biomet/70.1.41

S-TEAM Science Education. (2013). Retrieved 2013, from http://www.s-teamproject.eu/

SCB. (2013). Statistics Sweden - SCB. Retrieved 2013, from http://www.scb.se

Schmidt, W. H., McKnight, C. C., & Raizen, S. (1997). *A Splintered Vision: An Investigation of U.S. Science and Mathematics Education.*

Schoultz, J., & Hultman, G. (2002). *Det är bra med NTA. Vi gör inte saker för att tråka ut oss utanför att lära oss. Utvärdering av elevers och lärares lärande och utveckling inom NTA-projektet.*

Schoultz, J., Hultman, G., & Lindkvist, M. (2003). *I början fick vi använda vår fantasi. Utvärdering av elevers och lärares lärande och utveckling inom NTA–projektet.*

Schwartz, Daniel, L., Lindgren, R., & Lewis, S. (2009). Constructivism in an Age of Non-Constructivist Assessments. In S. Tobias & M. Duffy, Thomas (Eds.), *Constructivist Theory Applied to Instruction: Success or Failure?* New York: Routledge.

Science and Technology for Children Concepts Program - STC. (2013). Retrieved 2013, from http://www.nsrconline.org

Science, T. N. C. on M. and. (2000). *Before It's Too Late: A Report to the Nation from the National Commission on Mathematics and Science Teaching for the 21st Century.*

Segregation, skolval och skolresultat - Ekonomistas. (2013). Retrieved from http://ekonomistas.se/2011/03/21/segregation-skolval-och-skolresultat/

Shymansky, J. A., Hedges, L. V., & Woodworth, G. (1990). A Reassessment of the Effects of Inquiry-Based Science Curricula of the 60's on Student Performance. *Journal of Research in Science Teaching*, 27(2), 127–44.

Skatteverket. (2013). Folkbokföringen i går och i dag. Retrieved 2013, from http://www.skatteverket.se/privat/folkbokforing

Skatteverket. (2013). SALSA Betygsresultat för kommuner och skolor år 9. Retrieved from http://salsa.artisan.se

Skolverket. (2005). *Undervisningen per ämne i grundskolan hösten 2002 – Resultat av en undersökning om lärares undervisning och utbildning i undervisningsämnet*. Stockholm.

Skolverket. (2009). *Vad påverkar resultaten i svensk grundskola? Kunskapsöversikt om betydelsen av olika faktorer*. Retrieved from http://www.skolverket.se

Skolverket. (2011). *Läroplan för grundskolan, förskoleklassen och fritidshemmet 2011 lgr 11*. Retrieved from http://www.skolverket.se

Skolverket. (2013). *Skolverkets lägesbedömning 2013, Rapport 387*.

Smithsonian. (2013). Retrieved 2013, from http://www.si.edu/

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

SND Swedish National Data Service. (2013). SND Swedish National Data Service. Retrieved from http://snd.gu.se/en

SNS Analys nr 7. *Friskolereformens långsiktiga effekter på utbildningsresultat* | SNS - Studieförbundet Näringsliv och Samhälle. (2013). Retrieved from http://www.sns.se/forlag

Statistical Package for the Social Sciences - SPSS. (2013). IBM Corporation. Retrieved 2013, from http://www-01.ibm.com/software/analytics/spss/

Statistiska centralbyrån. (2013). SCB:s data för forskning. Retrieved from http://www.scb.se

Statskontoret. (2007). *Lärares utbildning och undervisning, Rapport 2007:8*.

Sveriges Kommuner och Landsting - SALAR. (2013). Öppna jämförelser. Sveriges kommuner och landsting. Retrieved June 04, 2013, from http://www.skl.se

Svärdh, J. (2013). To use or not to use a teacher support program - A study of what characterizes Swedish schools that apply the inquiry-based teacher support program NTA. In M. de Vries & I.-B. Skogh (Eds.), *Technology teachers as researchers: Philosophical and empirical technology education studies in Swedish TUFF research school*. Sense Publisher.

Swedish School Inspectorate. (2009). Skolinspektionsmyndigheten (Swedish School Inspectorate). Retrieved from http://www.skolinspektionen.se

Teknikdelegationen. (2009a). *Finns teknik och är matte svårt? Rapport 2009;2*.

Teknikdelegationen. (2009b). *Samverkan mellan skola och arbetsliv – flaskhalsar och framgångsfaktorer, Rapport 2009:3*. Stockholm.

Teknikdelegationen. (2010). *Vändpunkt Sverige – ett ökat intresse för matematik, naturvetenskap, teknik och IKT,* SOU 2010:28. Stockholm.

Teknikföretagen. (2005). *Alla barn har rätt till teknikundervisning*. Retrieved from http://www.teknikforetagen.se

The American Association for the Advancement of Science - AAAS. (2013). Retrieved 2013, from http://www.aaas.org/

The Fibonacci-Project. (2013). Retrieved from http://fibonacci.uni-bayreuth.de/

The Royal Swedish Academy of Sciences - KVA. (2013). Retrieved 2013, from http://www.kva.se/en/

TIMSS. (2011). *International Science Report: TIMSS 2011 International Results in Science*. Retrieved from http://timssandpirls.bc.edu/timss2011

Utbildningsdepartementet. (2009). *Att nå ut och nå ända fram, SOU 2009:94*. Stockholm.

Vanosdall, R., Klentschy, M., Hedges, L. V., & Weisbaum Sloane, K. (2007). *A Randomized Study of the Effects of Scaffolded Guided-Inquiry Instruction on Student Achievement in Science*. Chicago, Illinois: Annual Meeting of the American Education Research Association.

Vetenskapsrådet. (2013). *CODEX - regler och riktlinjer för forskning*. Retrieved 2013, from http://codex.vr.se/

Vetenskapsrådet. (2011). *Inventering av svensk utbildningsvetenskaplig forskning*.

Vetenskapsrådet. (2013). Forskningsinriktning för utbildningsvetenskap 2013. Vetenskapsrådet. Retrieved 2013, from http://www.vr.se

Wenglinsky, H. (2000). *How Teaching Matters: Bringing the Classroom Back Into Discussions of Teacher Quality*.

Wiliam, D. (2009). *Assessment for learning: why, what and how? An inaugural professorial lecture by Dylan Wiliam*. Institute of Education University of London.

Wise, K. C. (1996). Strategies for Teaching Science: What Works?. *Clearing House*, 69(6), 337–38.

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