Erik Mellander är docent i nationalekonomi vid Uppsala universitet och forskare samt ställföreträdande generaldirektör vid Institutet för arbetsmarknads- och utbildningspolitisk utvärdering (IFAU), i Uppsala. Hans forskning omfattar bland annat frågor kring uppföljning och kvantitativ utvärdering av utbildningsinsatser, effekterna av pedagogik i förskolan samt avvägningen mellan formell utbildning och yrkeserfarenhet i vuxnas kunskapsbildning.

Joakim Svärth är lektor i teknik och ämneslärare i slöjd på grundskolan. Han är också doktorand i Teknikvetenskapens lärande och kommunikation på ECE-skolan (Skolan för teknikvetenskaplig kommunikation och lärande) på KTH, i Stockholm. Arbetet som tekniklärare kombineras på halvtid med statistiska effektstudier av större satsningar på att förbättra resultaten i skolan.

ERIK MELLANDER

Institutet för arbetsmarknads- och utbildningspolitisk utvärdering (IFAU), Sverige
erik.mellander@ifau.uu.se

JOAKIM SVÄRDH

Doktorand på Kungliga Tekniska högskolan, Sverige
jsvardh@kth.se

Tre lärdomar från en effektutvärdering av lärarstödsprogrammet NTA (Naturvetenskap och teknik för alla)

Abstract

This paper discusses three methodological aspects on evaluations of effects of teacher support programmes on students' scholastic achievements. The three aspects are: i) participants and non-participants are, in general, not comparable, which needs to be accounted for, in order to avoid incorrect conclusions regarding the programme's effects, ii) programme effects can differ across subject fields, and iii) to enable precisely estimated effects it is essential to have access to standardized outcome measures. To illustrate the three aspects, we use the results in an earlier evaluation of the effects of the Swedish Science and Technology for Children programme (the NTA programme), on school achievements in grade 9. There, positive effects could be established only after adjusting for systematic differences between participants and non-participants. Moreover, the effects were limited to one subject, Physics, and visible only for results on nationwide tests, not for course grades set by the students' teachers.

INLEDNING

NTA (Naturvetenskap och teknik för alla) är det största lärarstödsprogrammet som finns i Sverige. Det gäller både om man ser till programmets innehåll och till antalet användare. NTA har också funnits under lång tid – det introducerades 1997. Om man vill diskutera olika problem som man kan råka

ut för när man vill utvärdera effekterna av ett lärarstödsprogram på skolprestationer är det därför naturligt att använda NTA som utgångspunkt.

En utvärdering av NTA är också intressant i sig, inte bara i ett svenskt perspektiv utan även internationellt. NTA är nämligen en svensk version av ett amerikanskt program, Science and Technology for Children, STC, som i olika varianter finns i många andra länder: Brasilien, Kroatien, Mexiko, Panama, Sydkorea, Thailand och Tyskland.¹ Det rör sig således inte bara om ett mycket ambitiöst utan även ett mycket spritt initiativ.

Vi har nyligen genomfört en undersökning där vi utvärderar effekterna av NTA på skolprestationer med hjälp av resultat på nationella prov och betyg i årskurs 9 i de naturvetenskapliga ämnena, se Mellander och Svärth (2015). Såvitt vi känner till är detta den första storskaliga undersökningen av orsakssambandet mellan deltagande i NTA och färdigheter i naturvetenskapliga ämnen. Helt nyligen har man också i USA genomfört en stor utvärdering av effekterna av STC på elevresultat i standardiserade tester i årskurserna 5 och 8, se Zoblotsky, Bertz, Gallagher och Alberg (2016).

Vår utvärdering och den amerikanska utvärderingen har genomförts med hjälp av olika metoder. Den amerikanska utvärderingen var upplagd på samma sätt som ett medicinskt experiment. Man valde först ut ett antal jämförbara skolor och bestämde sedan med hjälp av lottning ett antal skolor som skulle få ta del av STC-programmet direkt och ett antal som skulle få göra det först efter studiens genomförande, vilket innebar tre år senare. Med en sådan uppläggning är det förhållandevis enkelt att utvärdera effekterna. Om resultaten bland de (först) deltagande skolorna i genomsnitt är högre än resultaten bland de skolor och om skillnaden är statistiskt säkerställd har STC haft en positiv effekt. Det visade sig att totalt sett, över alla elever, var effekterna förhållandevis små. Däremot fann man starka positiva effekter bland undergrupper studenter med särskilda svårigheter: ekonomiskt efter-satta elever, elever med annat modersmål än engelska och elever med funktionshinder.

Förutsättningarna för vår utvärdering skiljde sig från dem i den amerikanska utvärderingen. I Sverige, liksom i de andra nordiska länderna, är det mycket sällsynt att det ges tillfälle att utvärdera lärarstödsprogram med hjälp av experimentansatser (Randomized Control Trials, RCTs). Ett av de fåtaliga exemplen är Jensen, Holm och Bremberg (2013). Istället är det typiska att deltagandet i ett lärarstödsprogram *inte* är slumpmässigt bestämt utan att det finns systematiska skillnader mellan de skolor som deltar och de som inte gör det. Så är fallet även för NTA. Detta måste man ta hänsyn till när man utvärderar programmets effekter. I vår utvärdering gjorde vi detta med hjälp av en statistisk metod som baseras på information om vilka faktorer som avgör om man en skola använder NTA. Metoden kan för *alla* elever uppskatta hur sannolikt det är att de har deltagit i NTA – oavsett om de faktiskt har deltagit eller inte. Därigenom kan vi jämföra resultaten bland faktiska deltagare och faktiska icke-deltagare som har samma, eller nästan samma, skattade sannolikhet för deltagande. Denna jämförelse kommer att fungera på samma sätt *som om* deltagandet i NTA hade bestämts slumpmässigt, givet att skattningarna av sannolikheterna innefattar samtliga variabler som påverkar deltagandet i NTA.

I likhet med den amerikanska utvärderingen finner vi endast små effekter av NTA när vi tittar på samtliga elever men påtagliga skillnader i effekter bland undergrupper av elever. I vårt fall har undergrupper uppstått genom att eleverna som har gjort nationella proven i de naturvetenskapliga ämnena i årskurs 9 har gjort detta i endast ett av de tre ämnen biologi, fysik eller kemi. Bland dem som gjorde provet i fysik finner vi förhållandevis stora positiva och statistiskt säkerställda effekter medan de skattade effekterna bland dem som gjorde proven i biologi och kemi är mindre och inte statistiskt säkerställda.

¹ Information per e-post 16:e september 2016 från Jean Flanagan, Smithsonian Science Education Center.

En annan beröringspunkt med den amerikanska utvärderingen är att vi får statistiskt säkerställda effekter endast när vi mäter effekterna i termer av resultat på nationella, dvs standardiserade, prov. När elevernas resultat mäts med hjälp av kursbetygen i årskurs 9 finner vi inte några statistiskt säkerställda resultat över huvud taget.

Syftet med denna artikel är diskutera tre metodologiska aspekter på effektutvärderingar av lärarstödsprogram och att därvid utgå från den nordiska kontexten, dvs fokusera på utvärdering av program som inte har genomförts i experimentform. De tre aspekterna är: i) att programdeltagare och icke-deltagare i allmänhet inte är jämförbara, ii) att ett programs effekter kan skilja sig mellan olika tillämpningsområden, samt iii) det är viktigt att använda standardiserade mått på skolprestationer när man ska utvärdera effekter av utbildningsinsatser på elevers lärande. Av dessa tre aspekter är det endast den första som bara är relevant vid utvärdering av program som inte har genomförts i experimentform. De två andra är viktiga att beakta oavsett på vilket sätt ett lärarstödsprogram har genomförts. Vi kommer att använda resultaten i vår effektutvärdering av NTA som *illustrationer* av de tre aspekternas praktiska betydelse.

Lärare och skolor som funderar på att utnyttja ett lärarstödsprogram kan dra tre lärdomar av denna artikel.

För det första: Väldigt få lärarstödsprogram fungerar lika bra i flera olika ämnen. De kan skilja sig i effektivitet mellan så näraliggande ämnen som biologi, fysik och kemi.

För det andra: Om man vill mäta hur ett lärarstödsprogram påverkar elevernas prestationer bör man använda resultatmått som i möjligaste mån bestäms av andra än de lärare som själva deltagit i programmet. Att använda resultat på nationella prov är därför bättre än att använda ämnesbetyg.

För det tredje: Man kan *inte* mäta ett programs effekter bara genom att jämföra resultaten i de klasser där man har tillämpat det med resultaten i klasser där programmet inte har utnyttjats. Skälet är helt enkelt att i de klasser där man har använt ett lärarstöd har man gjort det av en anledning, exempelvis att de klasserna har ett behov av extra stöd som saknas i de andra klasserna. Detta måste man ta hänsyn till när man mäter programmets effekter, annars kan man dra helt fel slutsatser. Enklast är att vända sig till utbildningsforskare för att få hjälp.

VAD ÄR NTA?

NTA är den svenska versionen av det amerikanska lärarstödsprogrammet Science and Technology for Children (STC).

Föregångaren: Science and Technology for Children (STC)

Tidiga versioner av STC kan spåras ända tillbaka till 1960-talet. I sin nuvarande form lanserades programmet 1995.

Den teoretiska bakgrunden till STC är konstruktivistisk inlärningsteori. Enligt denna teori är inlärning en process som kräver ett aktivt engagemang av den som ska lära. Med utgångspunkt i tidigare kunskaper "konstrueras" nya kunskaper med hjälp av praktisk problemlösning och diskussion. I denna kunskapsuppyggnad ses processförståelse som viktigare än faktakunskaper. Med det konstruktivistiska synsättet innebär kunskap framför allt en förmåga att se strukturer – hur saker hänger ihop (James, 2006; Strandvall, 2000).

STC innehåller tre komponenter: läromaterial – "experimentlådor" – med tillhörande instruktioner, lärarhandledning och -utbildning, samt stöd till skolledare. Programmet är avsett för elever från för-

skolan och upp till årskurs 8. Stoffet är ålders- och ämnesindelad. Det finns ett drygt 30-tal olika "teman" att välja mellan i STC. Temana är tänkta att sträcka sig över en termin och föregås av en heldagsutbildning för den undervisande läraren. Arbetet med temat följer en mall där eleverna under lärarens ledning får formulera hypoteser, utföra experiment, analysera resultaten och skriftligt dokumentera sitt arbete och sina slutsatser.

Skillnader mellan NTA och STC

NTA är på svenska och introducerades 1997. Det var Kungliga Vetenskapsakademien (KVA) som i samarbete med Kungliga Ingenjörsvetenskapsakademien (IVA) översatte STC-programmet och anpassade det till den svenska läroplanen. NTA är i första hand avsett för årskurs1–6 men har också använts i årskurs 7–9.

Antalet teman, och därmed antalet experimentlådor, är något lägre i NTA än i STC – under den period vi studerar uppgick antalet NTA-teman till drygt tjugo stycken. NTA innehåller inte heller lika utförligt stöd för skolledare som STC och lärarutbildningen är något mindre omfattande än inom STC. Lärarnas arbete med NTA stöds av lokala NTA-samordnare som också utgör länkar till regionala NTA-organisationer och till NTA Skolutveckling som skapar och utvecklar det rikstäckande NTA-programmet.

Anslutning till NTA

Både kommunala och privata skolor kan använda sig av NTA. För att skolorna ska kunna göra det måste deras huvudmän vara medlemmar i NTA Skolutveckling. För kommunala skolor är kommunen huvudman. För privata skolor kan skolan själv vara huvudman, eller ett bolag. För medlemskapet betalar huvudmannen en inträdesavgift på 15 000 kr. Skolhuvudmännen beslutar, ofta i samråd med de skolor som omfattas av huvudmannskapet, vilka skolor, eller delar av skolor, som ska delta i NTA. De som deltar betalar en fast kostnad per tema (experimentlåda) på 5 000 kr och terminsavgifter på 20 kr per elev.

År 2010 användes NTA i ungefär en tredjedel av Sveriges 290 kommuner. Antalet deltagande lärare och elever uppgick till ca 6 700 respektive 96 000. Eftersom grundskolan omfattade ungefär 1 miljon elever innebar detta att 10 procent av grundskoleeleverna deltog i NTA. Vid den tiden uppgick den årliga tillväxttakten i verksamheten till omkring 10 kommuner och 10 000 elever.

TIDIGARE ANALYSER AV NTA

Det saknas tidigare effektutvärderingar av NTA, d.v.s. analyser syftande till att fastställa orsakssamband mellan deltagande i NTA och färdigheter i naturvetenskap och teknik. Däremot har ett flertal kvalitativa analyser av NTA genomförts. Dessa visar genomgående på positiva erfarenheter av programmet. Exempelvis finner Gisselberg (2001), Schoultz och Hultman (2002) samt Schoultz, Hultman och Lindkvist (2003) att lärarna i huvudsak är nöjda med den information de har fått om NTA och den undervisning och träning de har erhållit, samt att eleverna är entusiastiska till NTA.

Anderhag och Wickman (2006) fann i en intervjustudie att lärare som använde NTA tyckte att de förbättrade sin förmåga att stödja elevernas förståelse av naturvetenskapliga begrepp. I en efterföljande analys, Anderhag och Wickman (2007), intervjuades 80 elever i årskurs 6, slumpmässigt fördelade på 40 NTA-deltagare och 40 elever som inte deltagit i NTA. Man undersöktes dels hur många av områdena inom naturvetenskap i grundskolan som eleverna ansåg sig ha blivit undervisade i, dels deras kunskaper inom dessa områden. Kunskaper mättes i termer av fem fördefinierade kategorier, tre avseende begrepp och två processkategorier, på så sätt att man räknade antalet begrepp/processer som eleverna nämnde under intervjuerna. Statistiskt signifikanta positiva skillnader mellan NTA-deltagare och icke-NTA-deltagare erhöles för antalet nämnda begrepp i kategorierna "naturvetenskapliga artefakter" och "naturvetenskapliga verksamheter" och för antalet nämnda processer

avseende "relationer mellan naturvetenskapliga och andra verksamheter". Eftersom däremot antalet naturvetenskapliga områden inom vilka eleverna undervisats – "bredden" i kunskaperna – inte skilde sig signifikant mellan grupperna drog man slutsatsen att NTA-deltagarna hade fått en djupare förståelse av naturvetenskapliga begrepp och processer än de elever som inte hade deltagit i NTA.

Studien av Gisselberg (op.cit.) innehåller en intressant observation med avseende på de olika ämnesområden som NTA omfattar. Merparten av de lärare som använde NTA tyckte sig ha tillräckliga sakkunskaper för att tillämpa undervisningsmaterialet i biologi men kände sig mindre säkra på sin kompetens inom fysik, kemi och teknologi.

FRAMTAGANDET AV ANALYSENS DATAUNDERLAG

Det datamaterial som effektutvärderingen baseras på har byggts upp i tre steg. Det första steget utgörs av två stickprov från resultaten i nationella prov i naturvetenskapliga ämnen 2009 och 2010. I det andra steget klassificeras individerna i dessa stickprov med avseende på om de deltagit i NTA eller inte. I det sista steget påförs skol-, kommun- och länsinformation. De tre stegen beskrivs i tur och ordning.

Stickprov från nationella prov i biologi, fysik, och kemi

Nationella prov i årskurs 9 i biologi, fysik och kemi introducerades år 2009. Alla elever i årskurs 9 omfattas av proven men en elev i en given skola gör endast provet i ett av de tre ämnena. Ämnesproven fördelas på skolor av Skolverket enligt en modell som i princip är slumpmässig, men som ser till att inga skolor får prov i samma ämne två år i rad.

Forskare vid Institutionen för tillämpad utbildningsvetenskap vid Umeå universitet utformade proven. För att följa upp resultaten beslöt man sig också för att dra ett slumpmässigt stickprov av resultaten bland de elever som skrev provet. Stickprovet omfattade alla elever födda på tre specifika dagar i månaden, vilket innebar att stickprovet skulle motsvara 10 procent av eleverna i årskurs 9. Förutom resultaten på de nationella proven – provpoäng och provbetyg – samlade man också in uppgifter om elevernas kursbetyg, deras kön samt om de deltagit i hemspråksundervisning.

I praktiken kom Umeå universitets stickprov att omfatta en (betydligt) lägre andel än 10 procent av eleverna. Detta framgår av Tabell 1.

Tabell 1: Totala antalet elever i årskurs 9 läsåren 2008/2009 och 2009/2010, antal som skrev nationella prov i biologi, fysik eller kemi, samt antal som ingick i Umeå universitets stickprov

Läsår	Antal elever i årskurs 9	Antal (%) som skrev nationellt prov i bi/fy/ke	Varav antal (%) ingick i Umeå univ.:s stickprov
2008/2009	118 032	88 491 (75,0)	8 028 (9,1)
2009/2010	113 545	98 848 (87,1)	7 839 (7,9)
Summa	231 577	187 339 (80,9)	15 867 (8,5)

Källor: Skolverket; Inst. för tillämpad utbildningsvetenskap, Umeå universitet; egna beräkningar.

Den andra kolumnen i Tabell 1 visar att betydande andelar av eleverna inte skrev det nationella provet i de naturvetenskapliga ämnena – 25 procent 2008/2009 och knappt 13 procent 2009/2010. Detta berodde på att vissa skolor avstod från att genomföra provet, särskilt under läsåret 2008/2009, trots

att deltagande var obligatoriskt. Den tredje kolumnen visar att bland dem som skrev provet var det en lägre andel än 10 procent som kom att ingå i Umeå universitets stickprov. Förklaringen är att vissa skolor inte rapporterade in sina resultat till Umeå universitet – motsvarande 0,9 procent av de elever som skrev provet 2008/2009 och 2,1 procent 2009/2010. Att detta kunde ske berodde på att universitet saknade möjligheter att tvinga de deltagande skolorna att rapportera sina resultat. Sammantaget ledde bortfallet till att stickproven kom att uppgå till knappt 7 procent av eleverna i årskurs 9 båda läsåren (8 028 / 118 032 respektive 7 839 / 113 545).

Hur kan det bortfall som framgår av Tabell 1 ha påverkat våra resultat? Det är rimligt att tro att resultaten i bortfallsgruppen i genomsnitt är lägre än resultaten bland dem som ingår i stickproven – skolor där eleverna har presterat sämre bör vara mindre benägna att rapportera in sina resultat än skolor med goda resultat. Om NTA har haft en positiv effekt på elevernas kunskaper bör bortfallet i så fall vara lägre bland skolor vars elever deltagit i NTA, än i övriga skolor. Detta kommer att medföra en underskattning av effekterna av NTA. Slutsatsen blir att om våra skattningar visar på positiva effekter av NTA kan de ses som en undre gräns för de sanna effekterna.

Uppdelningen av stickproven med avseende på NTA

Utgångspunkten för klassificeringen av enskilda elever med avseende på NTA är uppgifter på skolnivå. Med hjälp av NTA-organisationen har samtliga skolor som hade elever i årskurs 9 år 2009 – drygt 1 850 skolor – hänförs till en av följande tre kategorier:

- i. Skolans samtliga elever hade deltagit i NTA. Dessa skolor utgjorde knappt 100 stycken. I merparten av dem, 73 stycken, hade genomsnittseleven deltagit i NTA under 3–4 terminer upp till och med årskurs 6. I vissa skolor, 23 stycken, hade eleverna också deltagit i NTA under någon eller några av årskurserna 7–9.
- ii. Inte någon av skolans elever hade deltagit i NTA. Dessa skolor uppgick till 1 555 stycken.
- iii. Vissa, men inte alla, av skolans elever hade deltagit i NTA *eller* information saknades om huruvida skolans elever hade deltagit i NTA. Sammanlagt rörde det sig om knappt 200 stycken.

I vår analys bygger klassificeringen av elever direkt på denna klassificering av skolor. Det innebär att elever vars skolor tillhör kategori i. klassificeras som NTA-elever och att elever vars skolor tillhör kategori ii. klassificeras som icke-NTA elever, medan elever vars skolor tillhör kategori iii. inte kan klassificeras varken som NTA- eller icke-NTA-elever.

Genom att matcha denna klassificering med eleverna i den sista kolumnen i Tabell 1 erhåller vi följande uppdelning av eleverna:

Tabell 2: Uppdelningen av Umeå universitets stickprov efter NTA-status

NTA-elever	Icke-NTA elever	Studerade elever: NTA + icke-NTA	Ej klassificerade elever	Summa
1121	11845	12966	2901	15867

Av de 15 867 elever som ingick i Umeå universitets stickprov används således 12 966 stycken i analyserna nedan – 2 901 elever faller bort eftersom de inte kan klassificeras som antingen NTA- eller icke-NTA-elever.

Det ska noteras att även bland de 12 966 elever som ingår i våra analyser föreligger det viss osäkerhet om de har deltagit i NTA eller inte. Denna osäkerhet beror på att elever kan ha bytt skola. På en

skola vars elever har klassificerats som NTA-elever kan det finnas elever som inte har deltagit i NTA, men som har hamnat där efter byte av skola. På motsvarande sätt kan det på skolor vars elever har klassificerats som icke-NTA-elever finnas elever som har deltagit i NTA. Det finns dock ingen anledning att tro att NTA är en viktig faktor vid skolbyten. Felklassificeringarna kan därför betraktas som slumpmässiga mätfel. Dessa medför att om den sanna effekten av NTA är positiv kommer den att underskattas, d.v.s. påverkas på samma sätt som av bortfallet i stickprovsundersökningarna, jfr ovan.

Tillförande av skol-, kommun- och länsdata

Information på skolnivå har inhämtats ifrån Skolverkets databas SALSA (Skolverkets Arbetsverktyg för Lokala SambandsAnalyser). Uppgifterna avser dels skolans elever – andel pojkar, andel utlandsfödda, andel svenskfödda med utlandsfödda föräldrar, andel med minst godkänt betyg i alla ämnen i årskurs 9, dels den genomsnittliga utbildningsnivån hos elevernas föräldrar.

Data på kommunnivå har erhållits från Sveriges kommuner och landstings databas Kolada (Kommun- och landstingsdatabasen). Informationen omfattar demografi – befolkningsstorlek och befolkningstäthet, ekonomi – medianinkomster och utbildningskostnader, lärartäthet, andel behöriga lärare, kostnader per elev i årskurserna 1–9 samt andel elever i årskurs 9 med minst godkänt betyg i alla ämnen.

Länsdata, slutligen, utgörs av uppgifter om i vilka län NTA- och icke-NTA-elevernas skolor är belägna.

DE TRE LÄRDOMARNA FRÅN EFFEKTUTVÄRDERINGEN

I detta avsnitt använder vi resultaten av vår utvärdering av NTA:s effekter på skolresultat i årskurs 9 för att illustrera tre allmänna aspekter på utvärderingar av effekter av utbildningssatsningar: skillnaden mellan en bra och en dålig jämförelsegrupp, skillnaden mellan genomsnittliga effekter och effekter inom olika kunskapsområden, samt skillnaden mellan standardiserade och icke standardiserade mått på skolprestationer.

Betydelsen av att ha en jämförbar kontrollgrupp

Ett enkelt sätt att undersöka om deltagande i NTA har haft någon effekt på elevernas skolprestationer är att undersöka om de elever som deltagit i NTA i genomsnitt presterar bättre än de elever som inte har deltagit i NTA. För att testa detta jämför vi poängen på de nationella proven i biologi, fysik och kemi bland NTA-deltagarna med poängen bland dem som inte hade deltagit i NTA.

Det går inte att använda provpoängen direkt, som den är, eftersom det maximala antalet poäng skiljer sig mellan de olika ämnena och även mellan åren. För att göra individernas provpoäng jämförbara mellan ämnen och över år har vi räknat om provpoängen till *percentilrangordningar*. Det är ett vanligt tillvägagångssätt som innebär att vi ersätter elevens provpoäng med hans placering i fördelningen av resultaten för det aktuella provet, där fördelningen går mellan 0 och 100. Den individ som hade det högsta resultatet på provet får percentilrangen 100, den som har precis lika många före sig som efter sig får percentilrangen 50, den som har det lägsta resultat får percentilrangen 0 o.s.v. Formellt ges percentilrangen för det i :te resultatet av $[(c_i + 0,5f_i) / N] \times 100$, där c_i är antalet resultat som är lägre än i :te, f_i är antalet personer som har det i :te resultatet och N är antalet personer som gjorde provet.

Tabell 3 jämför genomsnittliga percentilrangordnade provpoäng bland dem som har deltagit i NTA med dem som inte har deltagit. Denna jämförelse tyder inte på några positiva effekter av NTA. Tvärtom, tabellen visar att den genomsnittliga percentilrangen bland dem som har deltagit i NTA är lägre än bland icke-deltagarna och att skillnaden är statistiskt säkerställd.

Tabell 3: Percentilrangordnade (p-rangordnade) resultat på nationella prov i biologi, fysik och kemi: NTA-deltagare och icke-NTA-deltagare

NTA-deltagare			Icke-NTA-deltagare			NTA – icke-NTA	
Antal individer	Medel p-rang	Standard-avvikelse	Antal individer	Medel p-rang	Standard-avvikelse	Differens ¹	Standard-avvikelse
1121	48,71	27,86	11845	50,56	28,88	-1,85*	0,900

¹ Differens = medelvärde för NTA-deltagare – medelvärde för individer som ej deltagit i NTA. Symbolerna * och ** markerar att ett dubbelsidigt t-test visar att differensen är statistiskt säkerställd på signifikansnivåerna 5 respektive 1 %.

Vad kan ligga bakom resultatet i Tabell 3? Den rimligaste förklaringen har att göra med vad som avgör vilka individer som deltar i NTA och vilka som inte gör det. Testet som redovisas i Tabell 3 gäller endast under ett viktigt underförstått antagande, nämligen att NTA-deltagarna och icke-deltagarna är jämförbara så tillvida att de i genomsnitt är lika i alla avseenden som är relevanta för studieresultaten, och att det enda som skiljer dem åt är att de har, respektive inte har, deltagit i NTA. Detta antagande innebär t.ex. att det inte föreligger några systematiska skillnader när det gäller elevernas socioekonomiska bakgrund, tidigare skolresultat och studiemotivation.

Men eftersom skolor själva väljer att delta i NTA är det rimligt att tro att det uppstår systematiska skillnader mellan NTA-deltagarna och icke-NTA-individerna. Om dessa skillnader inte beaktas vid utvärderingen kommer de skattade effekterna av NTA att bli felaktiga. I utvärderingslitteraturen kallas detta för ”selektionsbias”, jfr. Guo och Fraser (2010, s. 23-24), vilket syftar på att felet beror på att man inte tar hänsyn till (den icke slumpmässiga) process som styr urvalet till ”behandling”.

Vilka skillnader mellan dem som deltagit i NTA och de som inte har gjort det skulle kunna ge upphov till de till synes negativa effekter av NTA som vi just har noterat? En möjlig förklaring är att NTA har använts för att stärka undervisningen i skolor där eleverna presterar sämre än genomsnittet. Det finns vissa tecken i våra data på att det skulle kunna vara så: andelen elever med godkänt i alla ämnen är lägre i NTA-skolor än i andra skolor och i NTA-skolor har elevernas föräldrar lägre utbildning än föräldrarna till barn som går i skolor som inte använder NTA, se Mellander och Svärth (2015). Deltagande i NTA kommer då att vara överrepresenterat bland skolor med låga resultat på nationella prov och låga betyg. Detta måste beaktas när NTA:s effekter ska fastställas, annars kommer effekterna att underskattas.

För att kunna göra en rättvis jämförelse måste vi i våra data välja ut NTA-deltagare och icke-NTA-individer som är så lika som möjligt i alla andra avseenden än just ifråga om NTA. Detta är inte alldeles enkelt eftersom NTA- och icke-NTA-individer kan skilja sig åt på många olika sätt. Det räcker inte med att hitta individer som, exempelvis, har samma kön och föräldrar med samma utbildningsnivå.

Läsare som inte är intresserade av den tekniska beskrivningen av hur vi har valt ut jämförbara NTA-deltagare och individer som inte har deltagit i NTA kan nu hoppa till stycket före Tabell 4.

En metod som samtidigt kan beakta många olika skillnader mellan NTA-deltagarna och de individer som inte har deltagit i NTA är ”propensity score matching”, PSM, se Guo och Fraser (2010). Denna metod innebär att man med statistiska metoder skattar sannolikheten – propensity score – att elever har deltagit i NTA, som en funktion av flera olika variabler. Dessa skattningar (logistiska regressionser) kan göras både för elever som faktiskt har gått NTA och elever som inte har gjort det. Man kan visa att elever med samma propensity score, d.v.s. samma sannolikhet att delta i programmet, kan

betraktas som "statistiskt" lika i alla de dimensioner som variablerna representerar – skillnader kan finnas men är i så fall slumpmässiga.

Idén med PSM är att skapa en kontrollgrupp genom att till varje NTA-deltagare hitta icke-deltagare som har samma, eller nästan samma, propensity score. Jämförelser mellan sådana par av deltagare och icke-deltagare kan tolkas "som om" fördelningen på behandlingsgrupp (NTA) och kontrollgrupp (icke-NTA) har varit slumpmässig och därmed gett upphov till jämförbara grupper. Här finns dock ett viktigt förbehåll: propensity score-skattningen måste inkludera samtliga faktorer som har betydelse för om en individ tar del av behandlingen. Detta krav är i allmänhet svårt att uppfylla eftersom det ofta kan finnas viktiga faktorer som är svåra att observera i data, som t.ex. attityder och ambitioner. I sådana situationer kan det hända att resultaten från PSM inte är tillförlitliga, och att effektutvärderingen därför bör genomföras med hjälp av andra metoder. Vi återkommer till en diskussion om alternativa metoder i det avslutande avsnittet.

Av avsnittet ovan som beskriver NTA framgår dock att valet att delta i NTA sker på kommun- och skolnivå. Detta innebär att icke observerbara individfaktorer utgör ett betydligt mindre problem för utvärderingen än om valet att delta hade fattats av eleverna själva. Men det eliminerar inte problemet helt – kommun- och skollädares beslut kan påverkas av egenskaper hos eleverna som är kända för beslutsfattarna men som inte kan observeras av oss forskare. Resultaten i det följande bygger därför på antagandet att NTA-elever och icke-NTA elever som är jämförbara i termer av propensity scores också är jämförbara i termer av icke-observerade faktorer som t.ex. motivation och studieförmåga.

Resultatet av den propensity score-regression som vi har skattat redovisas i Tabell A1 i Appendix. Av tabellen framgår det att alla de variabler på skol- och kommunnivå som har inkluderats i regressionen har ett statistiskt säkerställt samband med huruvida en elev har gått i en NTA-skola eller inte.

På basis av propensity score-skattningen har vi valt ut 1 000 par av NTA- och matchande icke-NTA-individer. Individerna i dessa par är jämförbara i den meningen att deras propensity scores ligger nära varandra. Figur 1 visar de relativa frekvenserna av olika propensity scores för de 1 000 NTA-individerna och de 1 000 icke-NTA-individerna. Av figuren framgår att de relativa frekvenserna av olika

Not: Korrelationen mellan propensity scores för NTA och icke-NTA är 0,991.

Figur 1: Relativa frekvenser av propensity scores – skattade sannolikheter för deltagande i NTA för faktiska deltagare och matchade icke-deltagare

propensity scores är mycket lika i NTA-gruppen och den matchande gruppen av icke-NTA-individer. Bortsett från enstaka avvikelser, t.ex. i intervallet 0,25 – 0,29, ligger de två kurvorna påfallande nära varandra. Som framgår av noten till figuren är korrelationen mellan de två gruppernas propensity scores 0,991.

När vi undersöker skillnaderna i resultat mellan NTA- och icke-NTA-individer bland de 1 000 jämförbara paren blir slutsatsen en helt annan än i Tabell 3. Av Tabell 4 framgår att skillnaden i genomsnittlig percentilrang mellan jämförbara NTA-deltagare och icke-NTA-deltagare är positiv och statistiskt säkerställd på signifikansnivån 1 procent.

Tabell 4: Percentilrangordnade (p-rangordnade) resultat på nationella prov i biologi, fysik och kemi för NTA-deltagare och icke-NTA-deltagare som parats ihop (matchats) med hjälp av skattade propensity scores

NTA-deltagare			Icke-NTA-deltagare			NTA – icke-NTA	
Antal individer	Medel p-rang	Standard-avvikelse	Antal individer	Medel p-rang	Standard-avvikelse	Differens ¹	Standard-avvikelse
1 000	48,42	28,07	1 000	44,38	28,44	4,04**	1,263

¹ Differens = medelvärde för NTA-deltagare – medelvärde för individer som ej deltagit i NTA. Symbolerna * och ** markerar att ett dubbelsidigt t-test visar att differensen är statistiskt säkerställd på signifikansnivåerna 5 respektive 1 %.

De fullständigt olika slutsatser som Tabell 3 och Tabell 4 medför visar tydligt på betydelsen av att ha en jämförbar kontrollgrupp när man ska utvärdera effekter av utbildningsinsatser.

Skiljer sig effekten av NTA mellan biologi, fysik och kemi?

Den positiva NTA-effekt som redovisas i Tabell 4 är en aggregerad effekt som inte gör någon skillnad mellan de tre olika naturvetenskapliga ämnena. Det kan dock vara så att NTA har olika effekter på biologi, fysik och kemi. För att undersöka detta bryter vi ner NTA-effekten på de olika ämnena, jfr. Tabell 5. Som jämförelse återger den sista raden i Tabell 5 resultatet i Tabell 4, för alla de tre naturvetenskapliga ämnena sammantagna.

Av Tabell 5 framgår att den positiva NTA-effekten hänförs till fysik. Skillnaderna i genomsnittlig percentilrang mellan NTA- och icke-NTA-deltagare är positiva även för biologi och kemi men för dessa ämnen är skillnaderna inte statistiskt säkerställda. Tabellen visar också att den aggregerade effekten för samtliga ämnen (på sista raden i tabellen), naturligt nog, är betydligt mindre än effekten i fysik. Effekten i fysik, knappt sju percentiler, är betydande. Den innebär exempelvis att om en elev som inte har gått NTA har bättre resultat än en 1/4 av de övriga eleverna (25:e percentilen) så har den jämförbara NTA-eleven bättre resultat än knappt en 1/3 av de andra eleverna (32:a percentilen).

I litteraturöversikten framkom ett resultat som är konsistent med Tabell 5: Gisselberg (2001) noterade att bland lärare som använde NTA tyckte sig de flesta ha bättre förkunskaper i biologi än i fysik och kemi. Följaktligen bör NTA ha haft större potential att förbättra undervisningen i fysik och kemi än i biologi. Denna hypotes stöds av Tabell 5 genom att den näst sista kolumnen (Differens) visar på större positiva differenser för fysik och kemi än för biologi. Mera allmänt finns det i litteraturen en diskussion om att ett framgångsrikt lärande kräver att pedagogiska strategier – som, t.ex., NTA – anpassas till ämnesinnehållet och kunskapsnivån inom de undervisningsområden där strategin tillämpas, se Shulman (1987) och temanumret av *Journal of Curriculum Studies* 2004, volym 36, nummer 2.

Tabell 5: Percentilrangordnade (p-rangordnade) resultat på nationella prov i biologi, fysik och kemi för NTA-deltagare och icke-NTA-deltagare som parats ihop (matchats) med hjälp av skattade propensity scores

Ämne	NTA-deltagare			Icke-NTA-deltagare			NTA – icke-NTA	
	Antal individer	Medel p-rang	Standard-avvikelse	Antal individer	Medel p-rang	Standard-avvikelse	Differens ¹	Standard-avvikelse
Biologi	339	45,38	28,40	339	43,88	26,85	1,50	2,120
Fysik	322	49,75	27,45	313	42,47	29,14	7,28**	2,242
Kemi	339	50,18	28,15	348	46,60	29,23	3,58	2,187
Samtliga	1 000	48,42	28,07	1 000	44,38	28,44	4,04**	1,263

¹ Differens = medelvärdet för NTA-deltagare – medelvärdet för individer som ej deltagit i NTA. Symbolerna * och ** markerar att ett dubbelsidigt t-test visar att differensen är statistiskt säkerställd på signifikansnivåerna 5 respektive 1 %.

Vikten av standardiserade utfallsmått

I den internationella vetenskapliga litteraturen framhålls fördelarna med standardiserade utfallsmått när utbildningsinsatser ska utvärderas. Argumentet är att standardiserade utfallsmått innebär att skolprestationerna blir mer jämförbara än om de mäts i termer av resultat som (delvis) kan vara olika definierade och operationaliserade i olika skolor, se t.ex. Geier m.fl. (2008). I denna studie har vi dragit nytta av att införandet av nationella prov i de naturvetenskapliga ämnena år 2009 innebar att standardiserade utfallsmått för dessa ämnen blev tillgängliga, för första gången – de nationella proven är lika för alla i årskurs 9 och ska rättas efter enhetliga instruktioner.

Men hade det spelat någon roll om vi istället hade utvärderat effekterna av NTA genom deras inverkan på eleverna kursbetyg i årskurs 9? Detta är den fråga vi besvarar i detta avsnitt.

Tabell 6: Kursbetyg i årskurs 9 i biologi, fysik och kemi för NTA-deltagare och icke-NTA-deltagare som parats ihop (matchats) med hjälp av skattade propensity scores

Ämne	NTA-deltagare			Icke-NTA-deltagare			NTA – icke-NTA	
	Antal individer	Medel-betyg ¹	Standard-avvikelse	Antal individer	Medel-betyg ¹	Standard-avvikelse	Differens ³	Standard-avvikelse
Biologi	339	12,17	5,053	335	11,87	4,646	0,30	0,373
Fysik	321	11,95	4,351	312	11,94	5,120	0,01	0,377
Kemi	335	12,04	4,886	343	11,94	4,937	0,10	0,377
Samtliga	995 ²	12,06	4,776	990 ²	11,91	4,895	0,15	0,217

¹ Vid beräkningen av betygsmedelvärden har betygen värderats enligt: IG = 0, G = 10, VG = 15, MVG = 20.

² Antalet elever med kursbetyg är något lägre än de 1 000 elever vardera som utgör de matchade NTA- och icke-NTA-deltagarna, p.g.a. några elever inte uppfyllde kraven för att kunna få kursbetyg.

³ Differens = medelvärdet för NTA-deltagare – medelvärdet för individer som ej deltagit i NTA. Symbolerna * och ** markerar att ett dubbelsidigt t-test visar att differensen är statistiskt säkerställd på signifikansnivåerna 5 respektive 1 %.

Av Tabell 6 framgår att vi inte hittar några statistiskt säkerställda effekter av NTA på kursbetyg, varken för de tre naturvetenskapliga ämnena sammantagna eller var för sig. Det spelar alltså stor roll om man utvärderar NTA-effekten i termer av resultat på nationella prov eller i termer av kursbetyg.

En möjlig invändning mot jämförelsen mellan Tabell 6 och Tabell 5 är att det inte bara är typen av utfallsmått som skiljer utan också sätten att mäta. Medan percentilrang, som används i Tabell 5, ligger nära en kontinuerlig variabel kan betygspoäng, som används i Tabell 6, endast anta fyra olika värden, nämligen 0, 10, 15 och 20. För att undersöka hur viktig denna invändning är kan vi utnyttja att resultaten på de nationella proven inte bara redovisas som en provpoäng utan också i form av ett provbetyg, mått i samma enheter som kursbetyget.

Det visar sig att kvalitativt sett blir vår slutsats när vi mäter resultaten på det nationella provet i termer av provbetyg identisk med när vi mäter resultaten i termer av (percentilrangordnade) provpoäng. Detta framgår av Tabell 7 som dels innehåller den information som finns i Tabell 5 och Tabell 6, dels motsvarande information när utfallsmåttet utgörs av betyg på det nationella provet. För fysik motsvarar effekten i det senare fallet en knapp tredjedel av avståndet mellan betygsstegen G, VG och MVG (1,56 / 5, där nämnaren är lika med poängskillnaden VG – G resp. MVG – VG).

Slutsatsen blir alltså att den skillnad i skattade effekter av NTA som vi finner när vi använder (percentilrangordnade) provpoäng respektive kursbetyg *inte* beror på att de mäts utefter olika skalor. Vad orsaken till skillnaden är kan vi bara försöka gissa oss till. En sak att notera är att kursbetyg är ett ”bredare” utfallsmått än provpoäng; resultatet på det nationella provet ska vara av betydelse för kursbetyget men kursbetyget ska även avspegla en mängd andra förhållanden, t.ex. resultat på andra prov, aktivitet på lektioner o.s.v. Ju högre vikt dessa andra förhållanden tilldelas desto svagare kommer sambandet mellan provpoäng och kursbetyg att bli. Och så länge andra faktorer än resultatet på det nationella provet viktas på olika sätt av olika lärare så kvarstår det grundläggande problemet med kursbetyg i detta sammanhang: de är inte lika jämförbara som resultaten på nationella prov. Att olika lärare faktiskt tillmäter de nationella proven olika betydelse vid betygsättningen är högst troligt. Enligt Skolverket (2011, s.1): ”... i hur hög grad ett nationellt prov ska påverka betygsättningen finns inte reglerat. Det är läraren själv som bestämmer hur hon eller han ska tolka att det nationella provets resultat ska utgöra ett stöd vid betygsättningen.”

I Skolverket (2007) diskuteras ett antal tänkbara förklaringar till avvikelser mellan provbetyg och kursbetyg. Där nämns bland annat att läraren har att ta hänsyn till flera mål än de som ingår i det nationella provet och även har ett mer omfattande underlag för sin bedömning än elevens provresultat (jfr. ovan), att särskilda insatser kan ha satts in för elever som inte klarat proven bra samt att eleverna kan ha arbetat med ett annat stoff än det provet behandlar.²

Den sista förklaringen, att eleverna kan ha arbetat med ett annat stoff än det provet behandlar, är av speciellt intresse i detta sammanhang. Skälet är att de nationella proven i de naturvetenskapliga ämnena testar två olika färdigheter: dels teori- och faktakunskaper, dels förmågan att planera och genomföra en vetenskaplig undersökning. Det är framför allt den senare färdigheten som deltagande i NTA kan förbättra. I enlighet med den konstruktivistiska teori som NTA baseras på bör elever som

² En intressant fråga är i vilken utsträckning Skolverkets syn på relationen mellan nationella prov och betyg, så som den kommer till uttryck i Skolverket (2007) och Skolverket (2011), är representativ även för de andra nordiska länderna. Efter diskussioner med nordiska kollegor kan vi notera följande. I Finland har man inte nationella för alla elever i grundskolan utan bara för ett urval och då för att se på överensstämmelsen mellan den faktiska undervisningens och målen i läroplanen. Danmark, Norge och Island har nationella prov för alla elever. I Danmark och Norge har man inte nationella prov i sista årskursen i grundskolan (årskurs 9 i Danmark och årskurs 10 i Norge) men däremot i näst sista årskursen. Lärarna kan låta resultaten på dessa nationella prov påverka betygen och liksom i Sverige väljer lärarna själva i vilken utsträckning detta ska ske. Motsvarande gäller för Island där man dock har nationella prov i sista årskursen, årskurs 10, vilket torde innebära att Island är det land som är mest likt Sverige ifråga om förhållandet mellan resultat på nationella prov och betyg.

Tabell 7: Resultat på nationella prov – percentilrangordnade poäng och provbetyg – samt kursbetyg i biologi, fysik och kemi för NTA-deltagare och icke-NTA-deltagare som parats ihop (matchats) med hjälp av skattade propensity scores

	NTA-deltagare			Icke-NTA-deltagare			NTA – icke-NTA	
Percentil-rang provpoäng	Antal individer	Medel ¹	Standard-avvikelse	Antal individer	Medel	Standard-avvikelse	Differens	Standard-avvikelse
Biologi	339	45,38	28,40	339	43,88	26,85	1,50	2,120
Fysik	322	49,75	27,45	313	42,47	29,14	7,28**	2,242
Kemi	339	50,18	28,15	348	46,60	29,23	3,58	2,187
Samtliga	1 000	48,42	28,07	1 000	44,38	28,44	4,04**	1,263
Provbetyg								
Biologi	339	10,15	5,729	339	10,13	5,293	0,02	0,423
Fysik	322	10,73	4,899	313	9,17	5,797	1,56**	0,425
Kemi	339	10,10	5,486	348	9,38	5,879	0,72	0,433
Samtliga	1 000	10,32	5,392	1 000	9,57	5,670	0,75**	0,247
Kursbetyg								
Biologi	339	12,17	5,053	335	11,87	4,646	0,30	0,373
Fysik	321	11,95	4,351	312	11,94	5,120	0,01	0,377
Kemi	335	12,04	4,886	343	11,94	4,937	0,10	0,377
Samtliga	9952	12,06	4,776	9902	11,91	4,895	0,15	0,217

¹ Vid beräkningen av betygsmedelvärden har betygen värderats enligt: IG = 0, G = 10, VG = 15, MVG = 20.

² Antalet elever med kursbetyg är något lägre än de 1 000 elever vardera som utgör de matchade NTA- och icke-NTA-deltagarna, p.g.a. några elever inte uppfyllde kraven för att kunna få kursbetyg.

³ Differens = medelvärdet för NTA-deltagare – medelvärdet för individer som ej deltagit i NTA. Symbolerna * och ** markerar att ett dubbelsidigt t-test visar att differensen är statistiskt säkerställd på signifikansnivåerna 5 respektive 1 %.

har deltagit i NTA ha större förmåga att strukturera och behandla sitt naturvetenskapliga kunnande än andra elever (Strandvall, 2000, s. 12). Vid betygssättningen är det därför rimligt att lärare vars elever har deltagit i NTA fäster större vikt vid de resultat i de nationella proven som avser planering och genomförande av vetenskapliga undersökningar än vad andra lärare gör. Denna hypotetiska skillnad i betygssättningen har två konsekvenser som vi bör kunna observera i data.

Den första konsekvensen är att effekterna av NTA ska vara mera påtagliga när elevernas färdigheter mäts i termer av nationella prov än när de mäts i termer av kursbetyg. Detta har vi redan konstaterat.

Den andra konsekvensen är att skillnaderna mellan provbetyg och kursbetyg ska vara mindre bland de elever som har deltagit i NTA än de elever som inte har deltagit. Om vi beräknar den genomsnittliga skillnaden mellan kursbetyg och provbetyg för icke-deltagare i NTA och minskar den med den genomsnittliga skillnaden mellan kursbetyg och provbetyg bland NTA-deltagare bör vi således få ett positivt tal. Sådana beräkningar redovisas i Tabell 8. Det visar sig att i tre fall av fyra är skillnaden positiv, som förväntat, nämligen för fysik, kemi och för de tre ämnena sammantagna, men inte för biologi. De tre positiva skillnaderna är också signifikant större än noll. Tabell 8 ger alltså också stöd

för hypotesen att lärare vars elever har deltagit i NTA i sin betygssättning fäster större vikt vid de resultat i de nationella proven som avser planering och genomförande av vetenskapliga undersökningar än vad andra lärare gör.

Tabell 8: Skillnader mellan kursbetyg och provbetyg i årskurs 9 i biologi, fysik och kemi för matchade icke-NTA-deltagare och NTA-deltagare samt skillnaderna mellan dessa skillnader

	Icke-NTA-deltagare			NTA-deltagare			Icke-NTA – NTA	
Ämne	Antal individer ²	Medeldiff (kursbetyg – provbetyg)		Antal individer	Medeldiff (kursbetyg – provbetyg)		Differens i medeldiff ³	Standard-avvikelse
Biologi	335	1,74	3,906	339	2,02	4,067	-0,29	0,307
Fysik	312	2,77	4,384	321	1,25	3,331	1,52**	0,309
Kemi	343	2,48	4,273	335	1,93	3,939	0,55*	0,316
Samtliga	990	2,32	4,207	995	1,74	3,812	0,58**	0,180

¹ Vid beräkningen av betygsmedelvärden har betygen värderats enligt: IG = 0, G = 10, VG = 15, MVG = 20.

² Antalet elever med kursbetyg är något lägre än de 1 000 elever vardera som utgör de matchade NTA- och icke-NTA-deltagarna, p.g.a. några elever inte uppfyllde kraven för att kunna få kursbetyg.

³ Differens i medeldiff = medeldiff (kursbetyg – provbetyg) för icke-NTA-deltagare – medeldiff (kursbetyg – provbetyg) för NTA-deltagare. Symbolerna * och ** markerar att ett *enkelsidigt t-test* visar att differensen är statistiskt säkerställd på signifikansnivåerna 5 respektive 1 %.

SAMMANFATTNING OCH DISKUSSION

I denna artikel har vi gått från det partikulära till det generella. Vi har visat hur resultaten från en utvärdering av effekterna av en speciell utbildningsinsats – NTA – kan användas för att diskutera tre olika aspekter som är relevanta för effektutvärderingar av utbildningsinsatser i största allmänhet.

Den första aspekten avser vikten av att konstruera en rättvisande jämförelsegrupp. Vi har visat att om man bortser från denna aspekt kan analysens slutsatser bli felaktiga inte bara ifråga om effekternas storlek utan även ifråga om deras riktning. Effekten av NTA på nationella prov i naturvetenskap i årskurs 9 framstod som signifikant negativ när systematiska skillnader mellan NTA-deltagare och icke-deltagare inte beaktades. Vid kontroll för skillnaderna mellan grupperna blev effekten istället signifikant positiv.

Den andra aspekten har att göra med det välkända faktumet att en och samma medicin kan vara olika verksamt beroende på mot vilka åkommor den används. Vår analys visade att NTA har en statistiskt säkerställd positiv inverkan på resultaten i fysik men inte i biologi och kemi.

Den tredje aspekten visar på betydelsen av att effekterna mäts i förhållande till ett utfallsmått som är standardiserat och därmed direkt jämförbart mellan olika elever och olika skolor. Resultaten på de nationella proven uppfyller detta villkor – frågorna är desamma för alla elever och svaren rättas efter en enhetlig mall. Kursbetygen uppfyller inte villkoret – de baseras på flera andra prestationer än resultaten på de nationella proven och vilka dessa andra faktorer är och hur de vägs ihop skiljer sig mellan olika lärare och skolor, och möjligen även mellan elever. Liksom ifråga om den första aspekten utgör vår analys ett exempel på att det verkligen spelar roll hur man gör. När resultat på

nationella prov används som utfallsvariabel finner vi signifikanta effekter av NTA. Med kursbetyg är skillnaderna mellan NTA-deltagare och elever som inte deltagit i NTA mycket små och inte i närheten av att vara statistiskt säkerställda. Vi finner stöd för att lärare vars elever har deltagit i NTA tenderar att sätta kursbetyg som ligger närmare resultaten på de nationella proven än vad kursbetygen gör för de elever som inte har deltagit i NTA. Vår hypotetiska förklaring till detta resultat är att lärare vars elever inte har deltagit i NTA i sin betygssättning fäster lägre avseende vi den del av de nationella proven som avser planering och genomförande av vetenskapliga undersökningar än vad lärare med NTA-deltagare som elever gör. Vi finner också stöd även för denna förklaring.

Utöver dessa tre aspekter på effektutvärderingar finns det naturligtvis ytterligare distinktioner som är av allmänt intresse. En som vi indirekt har berört är skillnaden mellan kvalitativa och kvantitativa analyser. Vår litteraturgenomgång kan ge vid handen att de kvalitativa analyserna har resulterat i en mera positiv bedömning av NTA än vår kvantitativa studie har gjort. Men resultaten behöver inte stå i motsättning till varandra. De kvalitativa studierna kan ha avsett lärare och/eller elever vars erfarenheter av NTA av olika skäl har varit ovanligt positiva. Det skulle till exempel kunna bero på att det är lättare att höra talas om skolor där man framgångsrikt har tillämpat NTA än skolor där användningen av NTA har fungerat mindre bra. På motsvarande sätt kan det vara lättare att få samtycke till intervjuer och frågor om NTA i skolor där lärare och elever är positiva till NTA. I så fall kan studiernas representativitet ifrågasättas men inte nödvändigtvis deras resultat. Annorlunda uttryckt: analysernas *externa validitet* kan ifrågasättas men deras *interna validitet* kan vara god. Dessutom har vi i diskussionen av dataunderlaget till utvärderingen av NTA-programmet pekat på två skäl till att vår analys tenderar att underskatta de positiva effekterna av NTA: både bortfall i stickprovsundersökningen av elever som skrev de nationella proven och fel i klassificeringen av elever som deltagare respektive icke deltagare i NTA har denna konsekvens. Skillnaden i resultat mellan de tidigare kvalitativa studierna och vår kvantitativa studie är alltså troligtvis i realiteten mindre än de i förstone verkar vara.

I detta sammanhang vill vi påpeka att det faktum att den kvantitativa metod som ligger till grund för denna artikel är Propensity Score Matching (PSM) inte innebär att vi menar att PSM är en bra utvärderingsmetod i största allmänhet. När så är möjligt, är det bästa att planera och genomföra en utbildningsåtgärd så att åtgärden kan utvärderas som om den genomförts på samma sätt som ett medicinskt experiment. Det behöver inte innebära att man lottar om att vissa ska få delta och andra inte. Det kan vara så att alla får delta men att man lottar om *när* olika individer ska påbörja sitt deltagande, se Mellander (2015).

Skälet till att ett experimentliknande upplägg är att föredra är att det medför mycket lägre krav på den information som krävs för utvärderingen än vad PSA gör. Med experimentupplägget kan man som mått på åtgärdens effekt helt enkelt beräkna skillnaden mellan det genomsnittliga utfallet (resultatet) i behandlingsgruppen (deltagarna) med det genomsnittliga utfallet i kontrollgruppen (icke-deltagarna, alternativt de som deltagit senare). Det beror på att genom lottningen av deltagare kan individerna i behandlings- och kontrollgruppen förväntas vara jämförbara både vad gäller observerbara och icke-observerbara egenskaper. Som nämnts i diskussionen av PSM kräver denna metod däremot att utvärderaren också kan definiera och mäta samtliga faktorer som har påverkat huruvida en individ har deltagit. Detta stränga villkor beror på att utvärderaren inte har haft kontroll över hur deltagandet i åtgärden har bestämts.

Vad gäller framtida forskning bör det finnas goda möjligheter att kombinera kvantitativa och kvalitativa ansatser. En studie där man först genomför en kvantitativ analys och sedan följer upp denna med kvalitativa undersökningar av har tre viktiga fördelar. Den första är att den kvantitativa analysen kan användas som urvalsinstrument för den kvalitativa analysen. Man vet då hur de enheter man väljer att studera närmare förhåller sig till den representativa enheten, den som beskrivs av genomsnittsresultaten. Den andra fördelen är att man genom att närstudera enheter vars resultat överensstämmer med den representativa enhetens kan få information om vilka mekanismer som har genererat

genomsnittresultaten. Den tredje fördelen är att fallstudier av enheter vars resultat tydligt avviker från den representativa enhetens kan ge information om enheter som ligger långt ifrån genomsnittet.

TACK

Many thanks to Mary James for encouragement and very constructive comments. Tack även till Helena Holmlund för kommentarer på en tidigare version av denna artikel samt till tidskriftens två granskare.

REFERENSER

- 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: Lärarhögskolan, UKL.
- Anderhag, P. & Wickman, P. – O. (2007). Utvärdering av hur NTA hjälper skolorna att nå kursplanemålen för femte skolåret i naturorienterande ämnen. Stockholm: Lärarhögskolan, UKL.
- Geier, R. m. fl., (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*. Umeå: Umeå universitet, Lärarutbildningen, IMN.
- Guo, S. & Fraser, M.W. (2010). *Propensity Score Analysis: Statistical methods and applications*, Advanced Quantitative Techniques in the Social Sciences 11. Los Angeles: Sage Publications.
- Mellander, E. (2015). Möjligheterna att bedöma, följa upp och effektvärdera utbildnings- och fortbildningspolitiken. *Nordic Studies in Education*, No. 3-4, 299-312.
- Mellander, E. & Svärth, J. (2015). Inquiry-based learning put to test: long-term effects of the Swedish Science and Technology for Children program. IFAU Working Paper 2015:23. Uppsala: IFAU.
- James, M. (2006). Assessment, Teaching and Theories of Learning. Chapter 3 in J. Gardner (ed.) *Assessment and Learning*, 1st ed., London: Sage Publications.
- Jensen, B., Holm, A. & Bremberg, S. (2013). Effectiveness of a Danish early year pre-school program: A randomized trial. *International Journal of Educational Research*, 62, 115-128.
- Journal of Curriculum Studies* (2004). Disciplinary perspectives on fostering a community of teachers as learners, 36(2).
- Schoultz, J. & Hultman, G. (2002). *Det är bra med NTA. Vi gör inte saker för att tråka ut oss utan för att lära oss. Utvärdering av elevers och lärares lärande och utveckling inom NTA-projektet*. Linköping: Linköpings universitet: ITUF, IUUV.
- 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*. Linköping: Linköpings universitet, IUUV.
- Shulman, L. S. (1987). Knowledge and teaching. *Harvard Educational Review*, 57(1), 1-22.
- Skolverket (2007). *Provbetyg – Slutbetyg – Likvärdig bedömning? En statistisk analys av sambandet mellan nationella prov och slutbetyg i grundskolans årskurs 9, 1998-2006*. Rapport 300. Stockholm: Skolverket.
- Skolverket (2011). *Skillnaden mellan betygsresultat på nationella prov och ämnesbetyg i årskurs 9, läsåret 2009/10*. PM 71-2011-14. Stockholm: Skolverket.
- Strandvall, T. (2000). *Internet som stöd för inläring och distansutbildning*. (Avhandling för pedagogie magisterexamen.) Vasa: Åbo Akademi, Pedagogiska fakulteten, Institutionen för lärarutbildning.
- Zoblotsky, T., Bertz, C., Gallagher, B. & Alberg, M. (2016). *The LASER Model: A Systemic and Sustainable Approach for Achieving High Standards in Science Education*. (SSEC i3 Validation Final Report of Confirmatory and Exploratory Analyses.) University of Memphis: The Center for Research in Educational Policy (CREP).

APPENDIX

Tabell A1: Logistisk regression för beräkning av propensity scores. Beroende variabel 1 för NTA-individer och 0 för icke-NTA-individer

Variabel	Parameter-skattning	Standard-fel	Signifikans-nivå
Skolnivå:			
Andel pojkar	4,565** ⁹	0,731	0,000
Föräldrarnas utbildningsnivå ²	- 0,833**	0,310	0,007
Andel utlandsfödda elever	- 3,269**	0,711	0,000
Andel elever med utlandsfödda föräldrar	3,641**	0,616	0,000
Andel elever i åk 9 med minst Godkänt i alla ämnen	- 1,946**	0,516	0,000
Kommunnivå:			
Befolkning, i miljoner	- 4,921**	0,342	0,000
Andel av befolkningen i tätbebyggda ⁴ områden	2,865**	0,447	0,000
Medianinkomst ⁵ , miljoner SEK	52,223**	3,105	0,000
Andel lärare med pedagogisk universitetsutbildning ⁶	9,253	0,869	0,000
Antal elever per 100 lärare	19,742	6,515	0,002
Totala kostnader ⁷ för åk 1-9 per elev i åk 1-9 som bor i kommunen	114,351**	8,318	0,000
Andel elever i åk 9 med minst Godkänt i alla ämnen	- 0,112**	0,010	0,000
Länsnivå:			
Länsdummy-variabler ⁸	Ja		

1. Skolnivåvariablerna utgör genomsnitt för åren 2004-06.

2. Föräldrarnas utbildning är kodad som 1 för grundskola, 2 för gymnasium och 3 om utbildningen innehåller minst en termins universitetsstudier. Variabeln är beräknad som genomsnittet av utbildningskoderna bland föräldrarna till eleverna på skolan.

3. Kommunivåvariablerna avser år 2006.

4. Områden med minst 200 invånare där avståndet mellan bostäderna är högst 200 m.

5. Medianinkomsten, efter skatter och transfereringar, bland 20+ år gamla invånare.

6. I pedagogiska universitetsutbildningar ingår förskollärare och fritidspedagoger. Antalet är räknat i heltidsekvivalenter.

7. Kostnaderna är räknade netto, efter statsbidrag.

8. För att spara på utrymme rapporteras inte de skattade koefficienterna för länsdummy-variablerna.

9. Symbolerna * och ** markerar att ett t-test visar att parameterskattningen är signifikant skild ifrån noll på signifikansnivåerna 5 respektive 1 %.