Centraal Planbureau

CPB Notitie | 2 april 2014

Ontwerpen voor effectevaluatie

Uitgevoerd op verzoek van het ministerie van Onderwijs, Cultuur en Wetenschap

Centraal Planbureau

CPB Notitie

Centraal Planbureau
Van Stolkweg 14
Postbus 80510
2508 GM Den Haag

T (070) 3383 380
I www.cpb.nl

Contactpersoon

Sander Gerritsen

Aan: Ministerie van OCW

Datum: 2 april 2014
Betreft: Ontwerpen voor effectevaluatie

1 Inleiding

Het CPB heeft op verzoek van het ministerie van OCW vier beleidsinstrumenten beoordeeld op 'geschiktheid voor een effectevaluatie die (quasi-) experimenteel van opzet is'. Deze instrumenten zijn:

1. Differentiatie inspectietoezicht
2. Pilot tweetalig primair onderwijs
3. Bindend studieadvies
4. Studiebijsluiter

In deze notitie geeft het CPB per maatregel aan hoe kansrijk een effectmeting is in het licht van de doelstelling. Dit oordeel kan variëren van kansrijk tot niet kansrijk. We komen tot het oordeel aan de hand van de twee cruciale randvoorwaarden waaraan moet worden voldaan om tot een goede effectevaluatie te komen. Deze randvoorwaarden zijn:

- Een (quasi-) experimentele opzet van het beleidsinstrument. Hieronder wordt verstaan dat er behandelgroepen zijn waar het instrument wordt ingezet en controlegroepen waar het instrument niet wordt ingezet. Deze groepen moeten voor aanvang/implementatie van het beleidsinstrument goed vergelijkbaar zijn. Daarnaast mogen er geen spillovereffecten plaatsvinden tussen controle- en behandelgroepen ten tijde van het experiment. Dat wil zeggen dat controle- en behandelgroepen elkaar niet zodanig mogen beïnvloeden dat effecten onbetrouwbaar gemeten zouden worden.

- De beschikbaarheid van relevante en voldoende data. Hierbij is het van belang dat duidelijk is op welke grootheden en wanneer effecten gemeten gaan worden en dat er genoeg data zijn om statistisch betrouwbare uitspraken te doen over de effecten van het beleidsinstrument.

Om tot een oordeel over deze randvoorwaarden te komen, zijn gesprekken gevoerd met de relevante beleidsdirecties binnen het ministerie van OCW en zijn de officiële stukken geraadpleegd. Dit heeft geresulteerd in een overzicht van

1. het doel van de maatregel;
2. de inhoud van de pilot;
3. wat gemeten kan of zou moeten worden en;
4. een beoordeling kansrijk/niet kansrijk.

De beoordeling kansrijk of niet kansrijk is voor deze maatregel via een vast stramien tot stand gekomen, waarbij de pilot wordt onderzocht op de mogelijkheden om een effectevaluatie te laten plaatsvinden op basis van een 'random design experiment', een 'regression discontinuity analysis' of een 'differences-in-differences analysis'. Alom wordt een 'random design experiment' als de beste methode voor een effectevaluatie beschouwd (goud). De 'regression discontinuity analysis' vormt een goede tweede (zilver) , terwijl een 'differences-in-differences analysis' doorgaans als minst verfijnde methode wordt beschouwd (brons). De eerste twee designs creëren namelijk vergelijkbare behandel- en controlegroepen,[1] terwijl dat bij het laatste design niet het geval hoeft te zijn. Bij 'differences-in-differences analysis' is het van belang dat de behandel- en controlegroep eenzelfde trend hebben in de uitkomstvariabele voorafgaande aan de interventie. Bij een effectevaluatie moet dit eerst worden getest alvorens men verder kan met deze analyse.

Wanneer we aan een maatregel het label kansrijk hechten, is een beschrijving van de opzet van de pilot opgenomen om een effectevaluatie mogelijk te maken. Daarin komen aan de orde de databehoeften en het moment van een mogelijke evaluatie. Of uiteindelijk ook daadwerkelijk een effectevaluatie plaatsvindt, is aan het ministerie van OCW. Merk hierbij op dat de onderzoeksontwerpen zijn gemaakt voor het doel dat het beleidinstrument beoogt. Dit doel hoeft vanuit een maatschappelijk welvaartsperspectief niet per se wenselijk te zijn.

We komen tot de conclusie dat de vier maatregelen kansrijk zijn voor een effectevaluatie. Hierbij biedt de differentiatie inspectietoezicht de meeste mogelijkheden op een effectevaluatie met een 'regression discontinuity analysis'. De overige drie experimenten (tweetalig primair onderwijs, bindend studieadvies,

[1] Bij een 'regression discontinuity analysis' zijn de behandel- en controlegroepen vergelijkbaar rondom de afkapgrens, zie paragraaf 2.

studiebijsluiter) lijken zich te lenen voor een effectevaluatie met een 'differences-in-differences analysis.' Bij deze drie experimenten vallen de andere twee mogelijkheden ('random design experiment' en 'regression discontinuity analysis') af omdat de betrokken scholen/instellingen op basis van vrijwilligheid kunnen deelnemen aan deze pilot. Hierdoor kan 'selection bias' ontstaan. Dat wil zeggen dat door de vrijwillige deelname leerlingen in deelnemende scholen/instellingen kunnen verschillen van leerlingen in niet-deelnemende scholen/instellingen met als gevolg dat behandel- en controlegroepen niet vergelijkbaar zijn.

2 Differentiatie inspectietoezicht

Samenvatting

Het verbeteren van de onderwijskwaliteit bij scholen die de minimumnorm van onderwijskwaliteit al overtreffen, is het doel van gedifferentieerd toezicht zoals geformuleerd in de Kamerbrief Toezicht in transitie van 28-3-2014. Met enige aanpassingen in het beoogde invoeringstraject van deze nieuwe vorm van toezicht is deze beleidswijziging kansrijk voor een effectevaluatie in de vorm van een 'regression discontinuity design'. Aarzelingen hebben we bij het jaar 2016 als beoogd moment van evaluatie. Het is onzeker of al twee jaar na de start van dit nieuwe beleid effecten in de onderwijskwaliteit waarneembaar zijn, terwijl niet uit te sluiten valt dat dat op een langere termijn wel het geval is.

Doel

Op dit moment richt het inspectietoezicht zich vooral op het herstel van de onderwijskwaliteit bij zwakke scholen. Het verbeteren van de onderwijskwaliteit bij scholen die de minimumnorm al overtreffen, krijgt nauwelijks aandacht. Met het gedifferentieerde toezicht wordt beoogd dat ook de onderwijskwaliteit van deze tweede categorie scholen wordt verbeterd. De logisch uit deze doelstelling voortkomende onderzoeksvraag richt zich uiteraard op de effectiviteit van het gedifferentieerde toezicht op de onderwijskwaliteit bij deze tweede groep scholen.

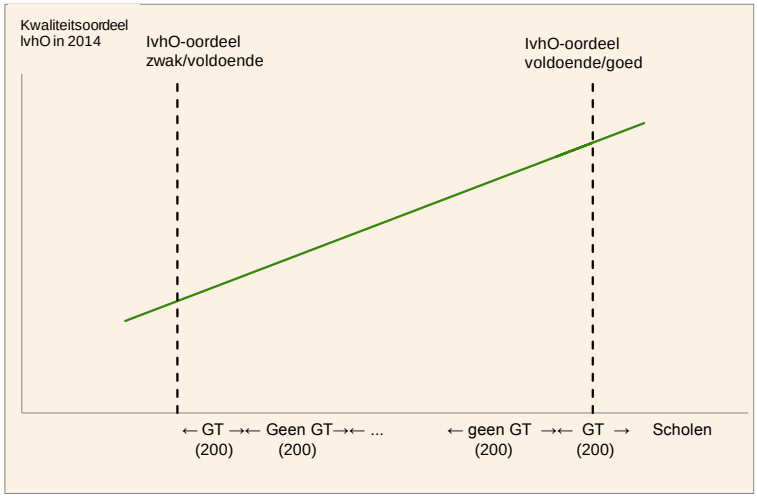
Inhoud pilot

Met ingang van het schooljaar 2014/2015 zal de Inspectie van het Onderwijs met dit gedifferentieerde toezicht gaan starten. Naar verwachting zullen in het eerste schooljaar circa 400 basisscholen onderwerp van toetsing zijn. In het voortgezet onderwijs zal het gaan om 80 scholen.[2] In de daaropvolgende schooljaren zal het gedifferentieerde toezicht verder worden uitgerold. Na circa 6 jaar zijn dan alle scholen met het toetsingsinstrument een keer bezocht. Omdat het hier gaat om een tweejarig experiment is het de bedoeling dat in 2016 een evaluatie van de (eerste) effecten zal plaatsvinden.

[2] Op deze plek zijn de aantallen genoemd die in de Kamerbrief voorkomen. De inspectie werkt op dit moment aan een plan voor monitoring waar andere aantallen in kunnen worden genoemd.

In de nu voorliggende opzet is de Inspectie voor het Onderwijs (IvhO) voornemens de scholen op basis van een nader te bepalen criterium te rangschikken van zwak tot goed. Direct boven de grens tussen zwak en voldoende is de IvhO voornemens 200 scholen te selecteren als pilot groep voor gedifferentieerd toezicht in schooljaar 2014/2015 (stippellijn links in Figuur 1). Het gaat hier dus om scholen die als voldoende worden beschouwd. Ook rond de grens tussen voldoende en goed functionerende scholen (stippellijn rechts in Figuur 1) zullen voor de pilot 200 scholen worden geselecteerd die in het schooljaar 2014/2015) in het kader van het nieuwe toezichtsregime bezocht worden. Op de horizontale as van Figuur 1 is door < GT > en < Geen GT > aangegeven welke scholen wel/niet gedifferentieerd toezicht in 2014/2015 krijgen. Uit de groep scholen die in 2014/2015 geen gedifferentieerd toezicht krijgen, kan bij de latere evaluatie van dit beleid een controlegroep worden samengesteld.

Figuur 1 Ranking scholen van zwak tot goed volgens de Onderwijsinspectie



Wat kun/wil je meten?

De kwaliteit van onderwijs is voor dit effectonderzoek de voor de hand liggende 'doelvariabele'. Bij invoering van gedifferentieerd toezicht zal de Inspectie van het Onderwijs de kwaliteit van het onderwijs vastleggen in een kwaliteitsprofiel waarop vijf aspecten worden gescoord. Het gaat dan om (1) leeropbrengsten, (2) kwaliteit van lesgeven, (3) sociale kwaliteit, (4) kwaliteitszorg van school en bestuur, en (5) belangrijke randvoorwaarden voor het financiële functioneren van de school. Op basis van dit kwaliteitsprofiel zal de Inspectie van het Onderwijs gerichte adviezen voor verbetering gegeven.[3] Bij gedifferentieerd toezicht zullen deze adviezen niet langer alleen aan de zwakke en zeer zwakke scholen worden verstrekt, maar ook aan scholen met een hoger kwaliteitsniveau.

[3] De aard van deze advisering is nog in ontwikkeling.

Voor elk van deze aspecten uit het kwaliteitsprofiel geldt dat deze slechts bij een bezoek van de Onderwijsinspectie wordt gescoord. Dit betekent dat om te kunnen vaststellen of een school na een bezoek van de Onderwijsinspectie voortgang heeft geboekt, een tweede meting vereist is. De interventie is bij deze verandering in het beleid redelijk scherp gedefinieerd, omdat alle onderwijsinspecteurs met dezelfde instrumenten en richtlijnen voor gebruik daarvan op pad gaan. Hoewel verschillende inspecteurs eenzelfde situatie verschillend kunnen beoordelen, mogen we wel aannemen dat zij redelijkerwijs tot consistente beoordelingen zullen komen, zodat er geen systematische vertekening van de eindresultaten zal zijn.

Om het effect van het gedifferentieerde toezicht op de doelvariabele te kunnen vaststellen, is de belangrijkste voorwaarde dat we gegevens hebben van twee groepen. Bij de eerste groep is er sprake van gedifferentieerd toezicht (behandelgroep), terwijl bij de tweede groep nog sprake is van het huidige basistoezicht (controlegroep). Idealiter zijn de twee groepen met uitzondering van dit aspect verder identiek.

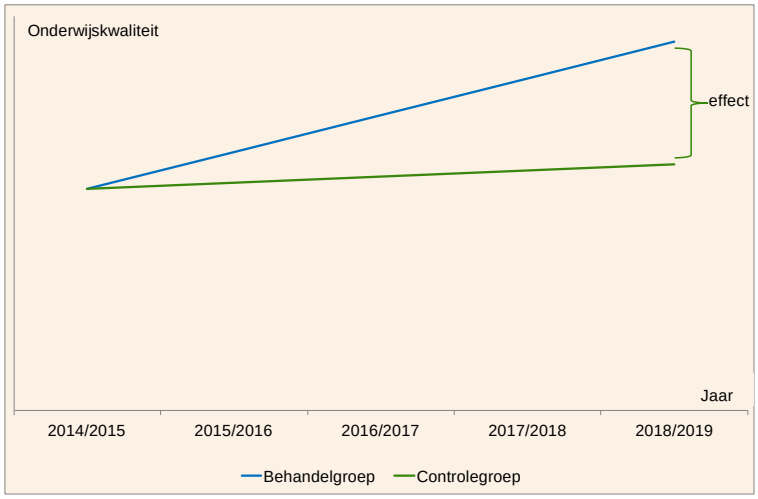
Beoordeling

In theorie zijn er drie manieren om binnen deze pilot tot een behandel- en controlegroep te komen. De eerste manier is door de scholen willekeurig te verdelen over een groep scholen die gedifferentieerd toezicht ondergaat en een groep die dit toezicht niet ondergaat (en waar het basistoezicht gewoon blijft gelden).[4] De pilot krijgt dan het karakter van een 'random design' experiment. De scholen in beide groepen worden vervolgens in het schooljaar 2014/2015 bezocht. Voor de scholen uit de behandelgroep wordt geadviseerd conform het gedifferentieerde toezicht. Voor de scholen uit de controlegroep blijft het huidige toezicht ongewijzigd van kracht. In het beoogde jaar van evaluatie (2016) worden de scholen opnieuw bezocht en wordt voor de scholen uit beide groepen een kwaliteitsprofiel opgemaakt. Het effect van het gedifferentieerde toezicht kan dan geschat worden door het kwaliteitsprofiel (doelvariabele) in het evaluatiejaar in de behandelgroep te vergelijken met dat van de controlegroep. Figuur 2 schetst een dergelijk 'random design' experiment waarbij op de verticale as de onderwijskwaliteit, of een van de vijf determinanten daarvan, is opgenomen. Het verschil tussen beide groepen op dezelfde doelvariabele in het evaluatiejaar (schooljaar 2015/2016) geeft dan het effect van het gedifferentieerde toezicht weer. Een voordeel van een dergelijk experiment is dat een voormeting in principe niet nodig is. Dit wil zeggen dat de doelvariabele niet aan het begin van het experiment gemeten hoeft te worden.[5] De loting zorgt er namelijk voor dat uitkomsten van het kwaliteitsprofiel in de behandel- en controlegroep voorafgaand aan het experiment vergelijkbaar zijn. In de figuur kan dit worden afgelezen aan het

[4] Loting wordt geprefereerd omdat dit garandeert dat willekeurig ook echt willekeurig is. Andere vormen van toewijzing, zoals vrijwilligheid, kunnen tot selectie-effecten leiden.
[5] In dit geval vindt er wel een voormeting plaats, aangezien deze meting een integraal onderdeel is van het gedifferentieerde toezicht. In principe is dit voor een evaluatie niet nodig.

feit dat de lijntjes van de controle- en behandelgroep op elkaar liggen vóór invoering van het gedifferentieerde toezicht (schooljaar 2014/2015).

Figuur 2 ‘Random design experiment’ bij gedifferentieerd toezicht



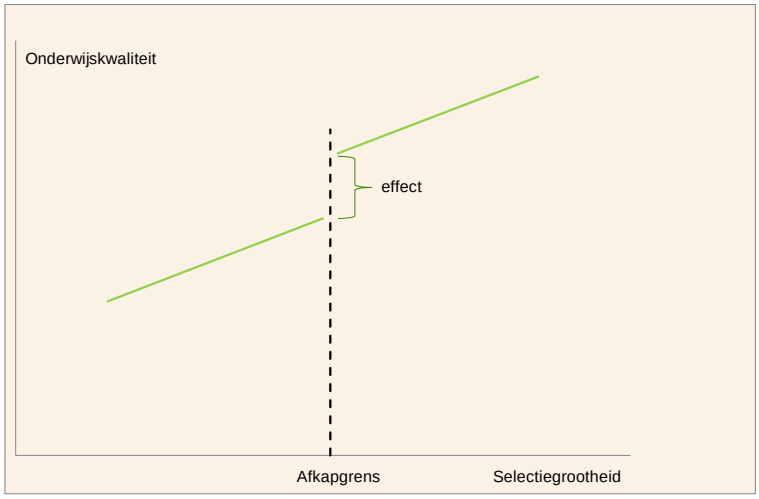
De hiervoor beschreven verdeling in een behandel- en controlegroep wordt in de nu voorliggende plannen voor invoering van het gedifferentieerde toezicht niet gevolgd. Omdat daarin de scholen op grond van een nader te bepalen criterium geselecteerd worden, vervalt de mogelijkheid van een ‘random design experiment’ voor een effectevaluatie.

Een alternatieve, tweede, manier om twee groepen te krijgen, is de instelling van een afkapcriterium op grond waarvan scholen worden verdeeld over een behandel- en een controlegroep. Zo’n afkapcriterium vereist dan een ranking van scholen op basis van een selectievariabele, waarbij de scholen onbekend zijn met de hoogte van de afkapgrens. Deze variabele kan van alles zijn, bijvoorbeeld het aantal leerlingen op een school, of een maat voor kwaliteit van scholen. Scholen met een waarde onder het afkapcriterium krijgen dan geen gedifferentieerd toezicht, terwijl scholen met een waarde boven dit criterium dit in schooljaar 2014/2015 wel krijgen. Door een dergelijk selectiecriterium toe te passen, zullen de scholen rondom de afkapgrens vergelijkbaar zijn. Het effect van het gedifferentieerde toezicht kan dan in 2016 gemeten worden door de *doel*variabele van de scholen rondom deze afkapgrens te vergelijken. Dit betekent dat op het evaluatiemoment voor zowel scholen uit de behandelgroep, als de scholen uit de controlegroep het kwaliteitsprofiel (de doelvariabele) moet worden bepaald door een bezoek in schooljaar 2015/2016.

Deze methode staat bekend als ‘regression discontinuity analysis’. Figuur 3 schetst dit type analyse. Het effect van het gedifferentieerde toezicht wordt dan weergegeven door de discontinuïteit in de score op het kwaliteitsprofiel die is ontstaan op de grens

tussen behandel- en controlegroep na afloop van de pilot.[6] Als het gedifferentieerde toezicht een positief effect heeft op de kwaliteit van scholen, dan zullen de scholen rechts van het criterium (behandelgroep) een hogere score op het kwaliteitsprofiel hebben dan de scholen links hiervan (controlegroep). Het voordeel van een regressiediscontinuïteitsanalyse is dat een voormeting in principe niet nodig is.[7]

Figuur 3 ‘Regression discontinuity analysis’ bij gedifferentieerd toezicht



Dit type analyse lijkt voor een effectevaluatie van gedifferentieerd toezicht mogelijk, aangezien de opzet van de pilot voorziet in groepsindelingen op basis van een selectievariabele. Hieronder zal de opzet van een dergelijke evaluatie verder besproken worden.

De derde methode is een ‘differences-in-differences’ methode (‘diff-in-diff’ analyse) waarbij de verandering in de ontwikkeling van het kwaliteitsprofiel tussen controle- en behandelgroep wordt beschouwd. Deze controle- en behandelgroep hoeven niet op elkaar te lijken. Wat wel belangrijk is, is dat zij voor aanvang van het experiment dezelfde trend in de doelvariabele (het kwaliteitsprofiel) hebben. Onder de aanname dat de ontwikkeling van de behandelgroep in afwezigheid van het gedifferentieerde toezicht dezelfde trend zou hebben gevolgd als de ontwikkeling van de controlegroep, kan het causaal effect van het gedifferentieerde toezicht geschat worden. Figuur 4 illustreert de ‘differences-in-differences’ methode. Het verschil tussen de gestippelde blauwe lijn en de doorgetrokken blauwe lijn geeft het causale effect van het gedifferentieerde toezicht weer. Het kan geschat worden door het verschil tussen de

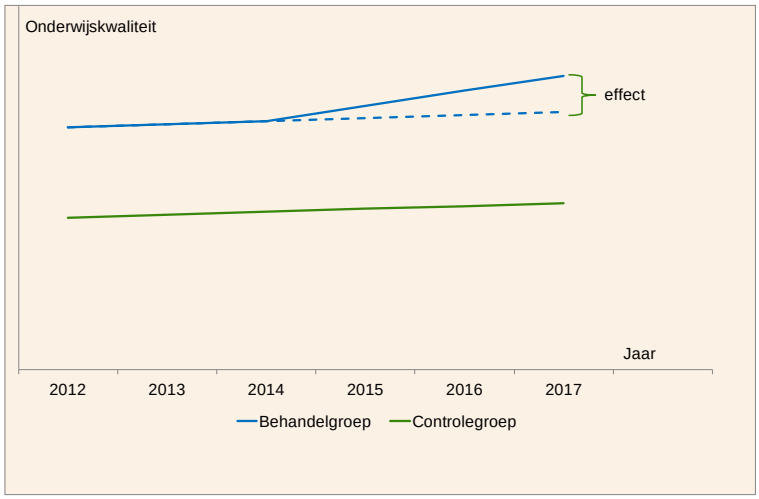
[6] Bij de opzet van de evaluatie zullen andere afkapgrenzen worden benut dan in Figuur 1. Figuur 5 schetst waar de afkapgrenzen voor het onderzoek zullen liggen. Zoals te zien is, vallen zij niet samen met de afkapgrenzen van de Inspectie.

[7] Een voormeting zal wel de geloofwaardigheid van de analyse kunnen vergroten. Een voormeting kan namelijk de veronderstelling testen dat scholen rondom de afkapgrens vóór invoering van het gedifferentieerde toezicht vergelijkbaar zijn. De groene lijnen in figuur 3 zouden bij een voormeting samenvallen tot 1 lijn die continu doorloopt rondom de afkapgrens. De nameting moet dan de discontinuïteit te zien geven, zoals in figuur 3.

ontwikkeling van de blauwe lijn en de ontwikkeling van de groene lijn te nemen. Belangrijk voor deze methode is dus dat er voor zowel behandel- als controlegroep voor- en nametingen van de doelvariabele plaatsvinden. De metingen voorafgaand aan de invoering van het gedifferentieerde toezicht moeten immers de veronderstelling dat behandel- en controlegroep in die periode gelijk op gingen, bevestigen. Concreet betekent dit dat voor alle scholen historische gegevens voor het kwaliteitsprofiel (de doelvariabele) beschikbaar moeten zijn.[8]

Deze laatste eis maakt een 'diff-in-diff' analyse lastig voor het effect van het differentieerde toezicht op de vijf aspecten van het kwaliteitsprofiel. Er zullen immers weinig voormetingen voorhanden zijn. Het profiel is namelijk pas net in gebruik. Er zullen dus geen metingen beschikbaar zijn van een aantal jaar voor invoering van gedifferentieerd toezicht. Hierdoor ontbreken de observaties die nodig zijn voor een 'diff-in-diff' analyse.

Figuur 4 'Diff-in-Diff' analyse bij gedifferentieerd toezicht



Eindoordeel

De overstap op een gedifferentieerd toezicht biedt potentieel goede mogelijkheden voor een adequate effectiviteitsstudie op de onderwijskwaliteit. Wanneer als maat voor onderwijskwaliteit het kwaliteitsprofiel wordt genomen, lijken er mogelijkheden te liggen voor een 'regression-discontinuity analysis' om de effecten van het gedifferentieerde toezicht op de vijf items van het kwaliteitsprofiel te schatten. Een 'diff-in-diff' analyse lijkt echter niet mogelijk met deze uitkomstmaat, aangezien er onvoldoende informatie is over de vijf items van het kwaliteitsprofiel voor de jaren voorafgaand aan de invoering van het gedifferentieerd toezicht (voormetingen).

[8] Dus gegevens over het kwaliteitsprofiel van scholen van de jaren vóór invoering van het gedifferentieerd toezicht.

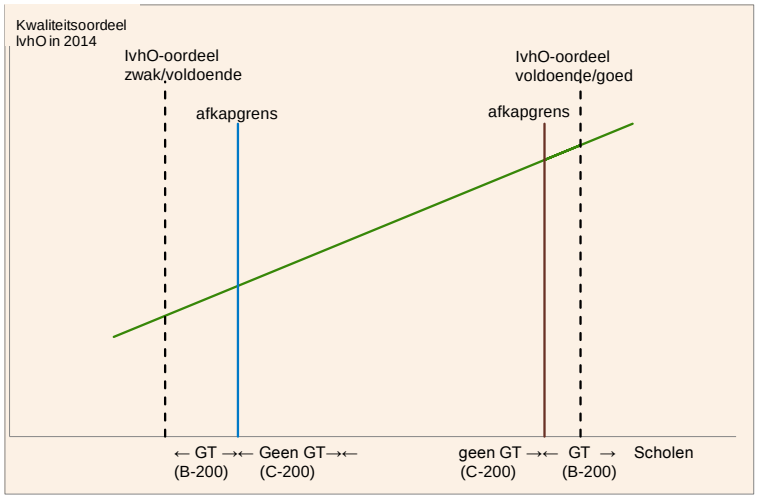
Een kanttekening is wel op zijn plaats bij 2016 als het gewenste moment van evaluatie. De keuze voor dit evaluatiemoment betekent namelijk dat scholen slechts maximaal twee jaar aan de slag zijn met de gesignaleerde verbeterpunten uit het kwaliteitsprofiel. Gelet op de lengte van het traject vanaf het bezoek van de inspectie tot en met implementatie van veranderingen op scholen, mag betwijfeld worden of al zo snel resultaten zullen optreden. We werpen dan ook de vraag op of het niet beter zou zijn het moment voor een effectevaluatie enige jaren uit te stellen.

Opzet van de evaluatie

De nu voorliggende pilot biedt twee mogelijkheden voor een 'regression discontinuity analysis'. Ten eerste bij de groep scholen die zich direct boven de zwakke scholen bevinden (rechts van de linker stippellijn in Figuur 1). Ten tweede bij de scholen die zich rond de grens van voldoende/goed bevinden (nabijheid van de rechterstippellijn uit Figuur 1).

Voor beide analyses geldt dat in vergelijking met Figuur 1 de afkapgrenzen tussen behandel- en controlegroep voor een 'regression discontinuity analysis' anders liggen. Figuur 5 illustreert dit.

Figuur 5 Afkapgrenzen bij regression discontinuity analysis bij gedifferentieerd toezicht



De linker blauwe lijn is het afkapcriterium waarbij de scholen links van het afkapcriterium gedifferentieerd toezicht krijgen (behandelgroep), terwijl de scholen rechts van dit afkapcriterium dit type toezicht niet krijgen (controlegroep). De rechter bruine lijn is de afkapgrens waarbij de scholen links van afkapcriterium geen gedifferentieerd toezicht krijgen (controlegroep), terwijl de groep scholen rechts van dit afkapcriterium dit wel krijgen (behandelgroep). Merk op dat hier de behandelgroep een mix is van scholen die voldoende en goed zijn.

De effectevaluatie van gedifferentieerd toezicht bestaat dus uit twee 'regression discontinuity analyses', een bij de blauwe grens, en een bij de bruine grens. Het voordeel hiervan is dat het effect van het gedifferentieerde toezicht wordt bepaald voor zowel 'voldoende' scholen (bij blauwe lijn) als 'goede' scholen (bruine lijn). Om voor deze analyses voldoende statistische power te hebben, zijn wel voor elke afkapgrens, 200 scholen in de controlegroep en 200 in de behandelgroep nodig. Dat betekent dus dat in totaal 800 scholen meedoen in de analyse (400 bij de blauwe afkapgrens en 400 bij de bruine afkapgrens).

De twee 'regression discontinuity analyses' kunnen op verschillende tijdstippen worden uitgevoerd. Belangrijk is dat ten tijde van de evaluatie de doelvariabele, bijvoorbeeld scores op een kwaliteitsprofiel of Cito-toetsscores bij de scholen uit zowel de controle- als behandelgroep, gemeten wordt.[9] Afhankelijk van het moment van evaluatie kan de set van doelvariabelen kleiner of groter zijn. Voorwaarde voor de scholen uit de controlegroep is dat zij voorafgaand aan het evaluatiemoment nog niet bezocht zijn in het kader van het gedifferentieerde toezicht. Daarnaast is informatie over de selectievariabele uit 2014 noodzakelijk.

3 Pilot tweetalig primair onderwijs

Samenvatting

Doel van de pilot tweetalig primair onderwijs is dat de Engelse, Duitse of Franse taalvaardigheden van de kinderen in het primair onderwijs verbeteren, zonder dat dit ten koste gaat van hun Nederlandse taalvaardigheden. Deze pilot achten wij kansrijk voor een evaluatie van de effecten op laatstgenoemde vaardigheden. Omdat de scholen op basis van vrijwilligheid kunnen deelnemen aan deze pilot, kan deze beleidswijziging slechts met een zogenaamde 'diff-in-diff' analyse worden geëvalueerd. Zo'n type analyse vereist een meting van de taalvaardigheden voorafgaand aan de pilot. Voor de Nederlandse taalvaardigheden is deze in ruime mate voorhanden in de vorm van bijvoorbeeld Cito-scores. Voor de vaardigheden in de vreemde taal zijn deze gegevens echter niet aanwezig. Dat betekent dat een effectevaluatie van deze pilot noodgedwongen beperkt zal blijven tot de gevolgen voor de Nederlandse taalvaardigheid.

[9] Bij de afkapgrens voldoende/goed wordt waarschijnlijk geen kwaliteitsprofiel in 2016 afgenomen, en zullen de Cito-scores benut worden.

Doel

De pilot tweetalig primair onderwijs staat twintig scholen toe 30 tot 50% van de lestijd in het Engels, Duits of Frans te geven. Doel van de pilot is dat de Engelse, Duitse of Franse taalvaardigheden van de kinderen op die scholen te verbeteren, zonder dat dit ten koste gaat van hun Nederlandse taalvaardigheden.[10]

Inhoud pilot

In deze pilot kunnen basisscholen op vrijwillige basis deelnemen. Er zijn echter wel strenge selectiecriteria gesteld. Maximaal twintig scholen kunnen aan de pilot meedoen. In augustus 2014 mogen twaalf scholen met de pilot beginnen. In augustus 2015 komen daar nog eens acht scholen bij. De scholen krijgen de gelegenheid 30 tot 50% van de onderwijstijd in het Engels, Duits of Frans te geven. De looptijd van de pilot is ongeveer vijf jaar; hij loopt in 2019 af.

Wat kun/wil je meten?

Hoewel de pilot tweetalig primair onderwijs ruimte biedt voor het lesgeven in de Franse of Duitse taal, zullen de meeste scholen voor Engels kiezen als tweede taal. In het verdere verloop van onze beschrijving van mogelijkheden voor een effectevaluatie zullen we ons dan ook beperken tot Engels als tweede taal.

Gelet op de doelstelling van het tweetalig primair onderwijs zijn de Nederlandse en Engelse taalvaardigheden de meest voor de handliggende doelvariabelen om het effect op te meten. De doelstelling is dan dat Engelse taalvaardigheden erop vooruitgaan, terwijl de Nederlandse en andere (taal)vaardigheden er niet op achteruit mogen gaan. De Nederlandse taalvaardigheden zouden gemeten kunnen worden met de Cito-eindtoets of andere toetsen uit de leerlingvolgsystemen van scholen. Echter, om het Engelse niveau te meten zullen specifieke toetsen moeten worden ingezet voor de pilot, aangezien die niet standaard voorhanden zijn. In een eerdere studie naar de effecten van tweetalig onderwijs, uitgegeven door de Rijksuniversiteit Groningen en de Universiteit Utrecht, zijn hiervoor verschillende meetinstrumenten benut.[11] Deze zouden ook kunnen worden ingezet bij deze pilot.

Om het causale effect van het tweetalig onderwijs op de Nederlandse en Engelse taalvaardigheden te kunnen vaststellen, is de belangrijkste voorwaarde dat er gegevens zijn van twee groepen scholen. Bij de eerste groep is sprake van tweetalig onderwijs (de behandelgroep), terwijl dat bij de tweede groep niet het geval is (de controlegroep). Idealiter zijn de twee groepen met uitzondering van dit aspect verder identiek.

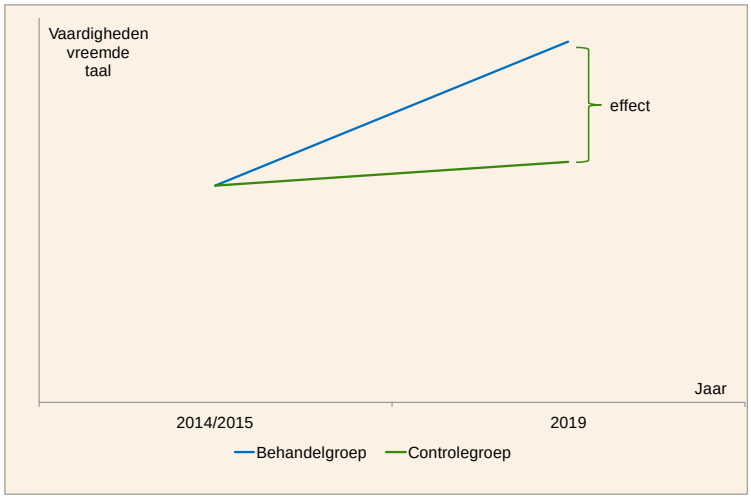
---

[10] Zie blz. 5 van OCW-nota “Startnota - Pilot Tweetalig Primair Onderwijs” van 6 september 2013.
[11] Zie http://www.europeesplatform.nl/sf.mcgi?3856&cat=769

Beoordeling

In theorie zijn er drie manieren om binnen deze pilot tot een behandel- en controlegroep te komen. De eerste manier is door scholen willekeurig te verdelen over scholen met en zonder tweetalig onderwijs. Dit betekent dat door loting wordt bepaald of een school in de controle- of behandelgroep terechtkomt.[12] De pilot krijgt dan het karakter van een 'random design' experiment. Het effect van het tweetalig onderwijs kan dan geschat worden door de doelvariabele (taalvaardigheid) van de leerlingen in de scholen van de behandelgroep te vergelijken met de taalvaardigheid van leerlingen van de scholen in de controlegroep. Figuur 6 schetst een dergelijk 'random design' experiment. Een voordeel van een dergelijk experiment is dat een voormeting niet nodig is. Dit wil zeggen dat de taalvaardigheden niet aan het begin van het experiment gemeten hoeven te worden. De loting zorgt er namelijk voor dat de behandel- en controlegroep vooraf aan het experiment vergelijkbaar zijn. In de figuur kan dit worden afgelezen aan het feit dat de lijntjes van de controle- en behandelgroep op elkaar liggen bij invoering van het tweetalig onderwijs (schooljaar 2014/2015). Het verschil tussen de taalvaardigheden van beide groepen op het moment van evalueren, bijvoorbeeld na afloop van de pilot in het schooljaar 2018/2019, geeft dan het effect van het tweetalig onderwijs weer. Gelet op de doelstelling van de pilot zou men, in het geval dat Engelse taalvaardigheden gemeten worden, een positief effect willen zien zoals weergegeven in Figuur 6. In het geval Nederlandse taalvaardigheden gemeten worden, wil men geen negatief effect zien. Met andere woorden, er zou geen verschil tussen controle- en behandelgroep moeten ontstaan en in Figuur 6 zouden de blauwe en groene lijn gedurende de gehele duur van de pilot op elkaar moeten liggen.

Figuur 6 'Random design experiment' bij tweetalig primair onderwijs



[12] Loting garandeert dat willekeurig ook echt willekeurig is. Andere vormen van toewijzing aan behandel- of controlegroep, zoals vrijwilligheid, kunnen tot selectie-effecten leiden.

Helaas is het in de praktijk onmogelijk scholen door middel van loting aan een controle- en behandelgroep toe te wijzen. Scholen kunnen zichzelf aanmelden voor de pilot 'tweetalig onderwijs'. Hierdoor vervalt het 'random design' karakter en daarmee deze wijze van effectevaluatie voor tweetalig onderwijs. Een na afloop van de pilot geobserveerd verschil in taalvaardigheden tussen de leerlingen van beide groepen scholen wordt dan immers veroorzaakt door mogelijke selectie-effecten. Scholen mochten immers zelf bepalen in welke groep zij kwamen, waardoor de leerlingenpopulaties tussen controle- en behandelscholen niet vergelijkbaar zijn. Daarmee wordt de effectevaluatie van tweetalig primair onderwijs onzuiver.

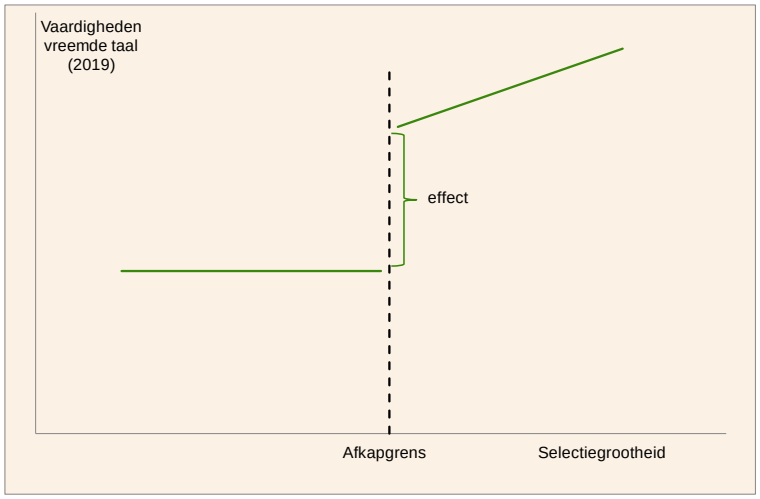
Een alternatieve, tweede, manier om twee groepen te krijgen, is de instelling van een afkapcriterium op grond waarvan scholen worden verdeeld over een behandel- en een controlegroep. Zo'n afkapcriterium vereist dan een ranking van scholen op basis van een selectievariabele, waarbij de scholen onbekend zijn met de hoogte van de afkapgrens. Deze variabele kan van alles zijn, bijvoorbeeld het aantal ingeschrevenen op een school of het financieringsbudget voor de school. Scholen met een waarde onder het afkapcriterium krijgen dan geen tweetalig onderwijs, terwijl scholen met een waarde boven dit criterium dit wel krijgen. Door een dergelijk selectiecriterium toe te passen, zullen de scholen rondom de afkapgrens vergelijkbaar zijn. Het effect van tweetalig onderwijs op de taalvaardigheid van de studenten kan dan gemeten worden door de scholen rondom deze afkapgrens te vergelijken. Deze methode staat bekend als 'regression discontinuity analysis'. Figuur 7 schetst dit type analyse. Het effect van het tweetalig onderwijs wordt weergegeven door de discontinuïteit in het taalniveau die is ontstaan op de grens tussen behandel- en controlegroep na afloop van de pilot. Uitgaande van Engelse taalvaardigheden als doelvariabele hebben de scholen rechts van het criterium (behandelgroep) een hogere taalvaardigheid dan de scholen links hiervan (controlegroep). Als de Nederlandse vaardigheden als doelvariabele worden gekozen, is het gewenste effect van de pilot dat er geen negatief effect zichtbaar is. In dat geval zouden op het evaluatiemoment in Figuur 7 de twee groene lijnen rond de afkapgrens nog steeds op elkaar moeten aansluiten en zou er geen discontinuïteit (dat wil zeggen geen effect) te zien moeten zijn. Het voordeel van een regressie-discontinuïteitsanalyse is dat een voormeting niet nodig is.[13]

Deze manier van effectevaluatie zou in theorie haalbaar kunnen zijn. In het selectieproces van scholen worden een selectiegrootheid en een afkapcriterium benut op grond waarvan scholen wel/niet 30-50% van de lestijd in de vreemde taal mogen geven.[14] Echter, omdat slechts een klein aantal scholen afvalt door dit

[13] Een voormeting zal wel de geloofwaardigheid van de analyse kunnen vergroten. Een voormeting kan namelijk de veronderstelling testen dat scholen rondom de afkapgrens vóór invoering van tweetalig primair onderwijs vergelijkbaar zijn. De groene lijnen in figuur 7 zouden bij een voormeting samenvallen tot 1 lijn die continu doorloopt rondom de afkapgrens. De nameting moet dan de discontinuïteit te zien geven, zoals in figuur 7.
[14] Scholen worden beoordeeld op hun basiskwaliteit en op aspecten die van belang zijn voor het succesvol kunnen uitvoeren van de pilot. Deze aspecten bestaan uit haalbaarheid, invoering, startsituatie, motivatie, internationalisering, investering van school en bestuur in competenties van leerkrachten, en investering van

selectiecriterium, zullen er onvoldoende scholen zitten in de controlegroep om een serieuze effectevaluatie te kunnen uitvoeren.[15] Dat betekent dat een evaluatie met de hiervoor beschreven 'regression discontinuity analysis' in de praktijk waarschijnlijk niet mogelijk zal zijn.

Figuur 7 `Regression discontinuity analysis' bij tweetalig primair onderwijs



De derde methode is de 'differences-in-differences' methode ('diff-in-diff'analyse) waarbij de verandering in de ontwikkeling van de taalvaardigheden tussen controle- en behandelgroep wordt beschouwd. Deze controle- en behandelgroep hoeven niet op elkaar te lijken. Wat wel belangrijk is, is dat zij voor aanvang van het experiment dezelfde trend in de doelvariabele (taalvaardigheden) hebben. Onder de aanname dat de ontwikkeling van de behandelgroep met tweetalig onderwijs dezelfde trend zou hebben gevolgd als de ontwikkeling van de controlegroep, kan het causaal effect van tweetalig onderwijs geschat worden. Figuur 8 illustreert de 'differences-in-differences' methode. Het verschil tussen de gestippelde blauwe lijn en de doorgetrokken blauwe lijn illustreert het causale effect van het tweetalig onderwijs op Engelse taalvaardigheden. Het kan geschat worden door het verschil tussen de ontwikkeling van de groene lijn en de ontwikkeling van de blauwe lijn te nemen.

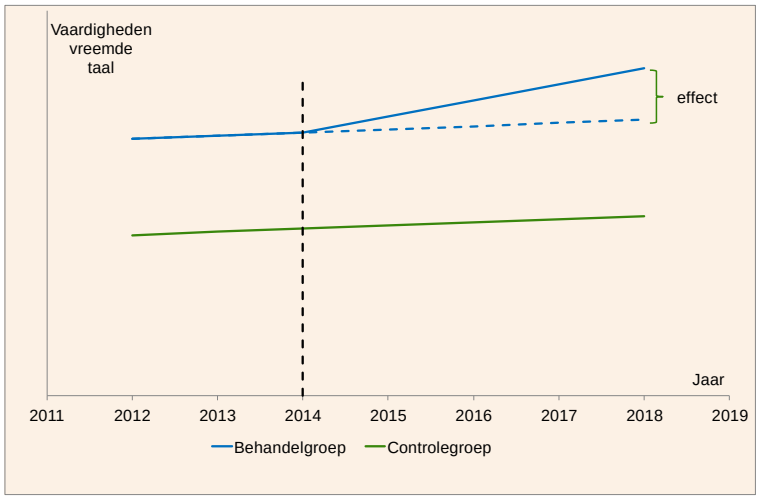
Voor de Nederlandse taalvaardigheden lijkt de ideale uitkomst van de pilot dat er geen negatieve effecten optreden. In dat geval moet de ontwikkeling van de behandelgroep zich voortzetten via de gestippelde blauwe lijn.

---

school en bestuur in de pilot zelf. Op deze aspecten zijn de scholen gescoord op een schaal van 1 (laag) tot 5 (hoog). Uit de scores van deze afzonderlijke aspecten is een totaalscore afgeleid. Scholen met een totaalscore lager dan een 3 zijn afgevallen. Dit betekent dat er een selectiegrootheid is (de totaalscore) en een afkapcriterium (3) op basis waarvan scholen wel/niet tweetalig onderwijs mogen invoeren.

[15] In de eerste ronde zijn slechts drie scholen afgevallen. Vijftien scholen hadden zich aangemeld, waarvan er twaalf met tweetalig onderwijs mogen starten in het schooljaar 2014/2015.

Figuur 8 ‘Diff-in-Diff’ analyse bij tweetalig primair onderwijs



Belangrijk voor deze methode is dat er voor zowel de behandel- als de controlegroep voor- en nametingen van de doelvariabele plaatsvinden. De meting voorafgaand aan de invoering van het tweetalig onderwijs moet immers de veronderstelling dat behandel- en controlegroep in die periode gelijk op gingen, bevestigen. Omdat historische gegevens ontbreken voor de Engelse taalvaardigheid van kinderen zal een ‘diff-in-diff’ analyse zich daardoor moeten beperken tot de beheersing van de Nederlandse taal. Als een ‘diff-in-diff’analyse zich hiertoe beperkt, zal de voormeting bestaan uit het meten van Nederlandse taalvaardigheden van de cohorten leerlingen die in de controle- en behandelgroep zitten voorafgaand aan de invoering van het tweetalig onderwijs. Zij dienen dan dus nog niet met het tweetalig onderwijs in aanraking te zijn gekomen. Het is hierbij van belang om een aantal cohorten te nemen, aangezien dan kan worden nagegaan of controle- en behandelgroep dezelfde trend hebben voor interventie. De nameting zal bestaan uit het meten van de Nederlandse taalvaardigheden van leerlingen in beide groepen na invoering van het tweetalig onderwijs in de behandelgroep.

Een risico bij deze ‘diff-in-diff’ analyse is dat ouders bij de schoolkeuze voor hun kind anticiperen op de invoering van het tweetalig onderwijs. Hierdoor kunnen selectie-effecten optreden. Ouders van leerlingen die belang hechten aan tweetalig onderwijs, kunnen ervoor kiezen hun kinderen naar scholen te brengen waar dit type onderwijs wordt aangeboden. Als deze kinderen over het algemeen de betere leerlingen zijn, zullen de vaardigheden in de behandelgroep toenemen ten opzichte van de controlegroep. De toename van de vaardigheden in de behandelgroep is dan mede het resultaat van selectie, en niet van tweetalig onderwijs. Dit mogelijke probleem kan in een ‘diff-in-diff’ analyse ondervangen worden door te bestuderen of een ander type leerlingen naar de behandelscholen gaat na invoering van het tweetalig onderwijs.

Eindoordeel

De uitvoering van een effectevaluatie van de invoering van tweetalig opnderwijs lijkt kansrijk met een 'diff-in-diff' analyse als we deze beperken tot een analyse van de effecten op de beheersing van de Nederlandse taal. Cijfers over deze taalvaardigheid zijn voorhanden en ook de interventie (tweetalig onderwijs) is scherp gedefinieerd. Aangezien de looptijd van de pilot ongeveer vijf jaar is - hij loopt in 2019 af - kan een vijfjaarseffect in kaart worden gebracht. Hierbij dient wel te worden opgemerkt dat bij een evaluatie in 2019 dan niet het effect van de 'full size' implementatie wordt gemeten. Het cohort dat in 2014 instroomt en te maken krijgt met tweetalig primair onderwijs in groep 1 zal in 2019 pas in groep 5 zitten. Ze hebben dan slechts vijf jaar tweetalig primair onderwijs gehad, en niet de volledige 8 jaar. Voor het meten van het 'full size' implementatie-effect zal dus als evaluatiemoment het jaar 2022 moeten worden gekozen, aangezien in dat jaar het cohort dat in 2014 is ingestroomd acht jaar tweetalig primair onderwijs erop heeft zitten.[16]

Opzet van de evaluatie

Voor het schatten van de effecten van tweetalig onderwijs op de Nederlandse taalvaardigheid wordt een 'diff-in-diff' analyse benut. Hierbij wordt de ontwikkeling van dit type vaardigheden van de groep scholen die tweetalig onderwijs invoert (behandelgroep), vergeleken met die van een groep scholen die dit niet doet (controlegroep). Deze controlegroep kan zodanig gekozen worden dat ze voor aanvang van de pilot veel lijkt op de scholen in de behandelgroep. Een manier om dat te doen is bijvoorbeeld via matching. Stel dat bijvoorbeeld op de behandelscholen veel kinderen van hoogopgeleide 'expats' zitten, dan kunnen er controlescholen worden bij gezocht waar ongeveer een even grote populatie van dit type kinderen op zit. Naast deze expatvariable als matching variabele zou ook op meerdere variabelen gematcht kunnen worden, bijvoorbeeld op geslacht, thuistaal, en het leerlinggewicht.

De eerste twaalf scholen beginnen met tweetalig onderwijs in schooljaar 2014/2015. Voor een controlegroep zouden bijvoorbeeld twaalf of meer scholen kunnen worden gezocht die op deze scholen lijken of dezelfde trend in de doelvariabele (Nederlandse taalvaardigheden) hebben. In 2015/2016 starten nog eens acht scholen met tweetalig onderwijs. Ook voor deze scholen kan op eenzelfde wijze een controlegroep worden bijgezocht. Uiteindelijk zullen er dan twintig scholen in de behandelgroep zitten en twintig (of meer) in de controlegroep. Voor het meten van de Nederlandse

[16] Een kanttekening bij een analyse die zich alleen richt op de taalvaardigheden op de basisschool, is dat deze effecten mogelijkerwijs wegebben in het voortgezet onderwijs. Dit zou kunnen gebeuren als het taalonderwijs in het voortgezet onderwijs niet goed aansluit op het taalniveau van de leerlingen die het tweetalig primair onderwijs hebben doorlopen. Immers, het taalcurriculum in het voortgezet onderwijs is doorgaans afgestemd op leerlingen die dit type onderwijs niet hebben gevolgd. De kinderen afkomstig uit het tweetalig primair onderwijs leren dan in het voortgezet onderwijs minder bij als het gaat om de vreemde taal, terwijl kinderen die niet uit tweetalig primair onderwijs afkomstig zijn meer bijleren. Gevolg is dat beide groepen aan het eind van het voortgezet onderwijs ongeveer hetzelfde taalniveau hebben, en dat dit niveau waarschijnlijk vergelijkbaar is met de vaardigheden in de huidige situatie. Per saldo heeft tweetalig primair onderwijs dan weinig toegevoegd aan de taalvaardigheden.

taalvaardigheden kunnen Cito-toetsen of vergelijkbare toetsen uit de leerlingvolgsystemen worden benut.

4 Bindend studieadvies

Samenvatting

Met de uitbreiding bindend studieadvies wordt beoogd te stimuleren dat studenten hun opleiding binnen de nominale studieduur afronden door middel van een samenhangend pakket van maatregelen ter bevordering van de studievoortgang waardoor de kwaliteit en de doelmatigheid van het hoger onderwijs worden verbeterd. Dit experiment is naar onze mening kansrijk voor een evaluatie van de effecten op de studieduur. De evaluatie zal uit een 'diff-in-diff' analyse bestaan waarbij bijvoorbeeld de studieduur van cohorten studenten wordt vergeleken in opleidingen met en zonder een bindend studieadvies. Om de kans op selectie-effecten te verkleinen, lijkt het verstandig de effectevaluatie te beperken tot het eerste cohort studenten dat bij een bepaalde opleiding geconfronteerd wordt met de uitbreiding van het bindend studieadvies. Voor enkele opleidingen, zoals opleidingen aan Rijksuniversiteit Leiden, zal dit het cohort eerstejaarsstudenten uit studiejaar 2013/2014 betreffen, bij andere opleidingen de eerstejaars van collegejaar 2014/2015. Een evaluatie zal evenwel pas aan het eind van dit decennium plaats vinden, omdat dan vrijwel alle studenten uit deze cohorten hun studie zullen hebben afgerond.

Doel

Met het experiment uitbreiding bindend studieadvies wordt beoogd te stimuleren dat studenten hun opleiding binnen de nominale studieduur afronden door middel van een samenhangend pakket van maatregelen ter bevordering van de studievoortgang waardoor de kwaliteit en de doelmatigheid van het hoger onderwijs worden verbeterd.[17] Het is de bedoeling dat medio 2018 een evaluatie plaats vindt aan de hand van een viertal criteria.[18] Ten eerste, het aantal studenten dat de opleiding binnen de nominale studieduur afrondt, het aantal studenten dat is uitgevallen en geen andere opleiding is gestart en het aantal studenten dat een andere opleiding is gestart. Ten tweede de mate waarin de doelmatigheid binnen de instelling is verbeterd. Ten derde de mate waarin de kwaliteit van het onderwijs is verbeterd en ten vierde de mate waarin de experimentele uitbreiding bindend studieadvies van invloed is geweest op de verbetering van de kwaliteit en doelmatigheid.

Om te bereiken dat meer studenten hun bachelordiploma behalen in de nominale studieduur moeten de studenten in hun tweede leerjaar een minimaal aantal studiepunten behalen. Doelmatigheids- en kwaliteitswinst moeten worden bereikt

[17] Zie artikel 2 Besluit experiment bindend studieadvies, gepubliceerd op 21 januari 2014.
[18] Zie artikel 13 Besluit experiment bindend studieadvies, gepubliceerd op 21 januari 2014.

doordat gedurende de pilot de deelnemende onderwijsinstellingen een deel van de prestatieafspraken naar verwachting versneld implementeren.

Inhoud pilot

In deze pilot kunnen onderwijsinstellingen voor hoger onderwijs op vrijwillige basis deelnemen. Er zijn slechts twee restricties gesteld. Ten eerste mogen opleidingen die niet elders in Nederland gevolgd kunnen worden (de unica), niet onderdeel zijn van de pilot. Ten tweede mag het aantal studenten dat in het experiment meedoet, niet meer dan 10% van de totale studentenpopulatie beslaan.

Wat kun/wil je meten?

In het artikel 13 van het Besluit experiment bindend studieadvies (publicatiedatum 21 januari 2014) wordt gesproken over een evaluatie gericht op aantallen studenten die de opleiding binnen de nominale duur afronden, die uitvallen en geen andere opleiding starten en die uitwijken naar een andere opleiding (zie subkopje Doel). Dit heeft een paar beperkingen. Ten eerste hebben absolute aantallen weinig zeggingskracht. Deze aantallen hangen immers ook samen met de populariteit van de opleiding en het aantal beschikbare plaatsen. Ten tweede is het lastig uit enkele van deze maatstaven een oordeel over het effect van dit experiment te destilleren. Bijvoorbeeld: hoe moet een groter aantal studenten dat na een negatief bindend studieadvies een andere opleiding start, gewogen worden? Is dit nu positief of negatief?

Als alternatief stellen we daarom voor de effecten op de studieduur van een instroomcohort studenten bij een opleiding met/zonder uitgebreid bindend studieadvies als maatstaf te hanteren. Het gaat daarbij om de studieduur waarmee een student zijn bachelor voltooit. Het nadeel van deze variabele is dat het vrij lang duurt voordat deze beschikbaar komt in de pilot. De studenten die in het studiejaar 2013/2014 zijn ingestroomd, krijgen te maken met het bindend studieadvies aan het eind van hun tweede jaar, dat wil zeggen rond de zomer van 2015.[19] Vervolgens duurt het nog minimaal één jaar voordat de eerste studenten van dit cohort hun bachelor afronden. Omdat voor een goede evaluatie de studieduur van alle bachelorstudenten die in 2013/2014 zijn ingestroomd nodig is, kan het wellicht wel vijf jaar of langer duren voordat deze variabele volledig beschikbaar zal zijn. Dat betekent dat een effectevaluatie waarschijnlijk niet in 2018 kan plaats vinden.

Een kanttekening bij de focus op de studieduureffecten is dat we mogelijk niet de effecten op de investeringen in menselijk kapitaal meten. Door uitbreiding van het bindend studieadvies naar latere jaren kunnen sommige leerlingen besluiten na

[19] De pilot 'uitbreiding bindend studieadvies' biedt instellingen de ruimte om ook voor latere studiejaren een bindend studieadvies uit te brengen. Omdat op dit moment geen enkele instelling voornemens is hiervan gebruik te maken, laten we een expliciete uitwerking hiervan in deze notitie achterwege. Onze verkenning van de mogelijkheden voor een effectevaluatie is echter ook van toepassing op een bindend studieadvies in latere leerjaren.

afronding van het voortgezet onderwijs niet te gaan doorleren, terwijl deze zonder bindend studieadvies wellicht wel daartoe hadden besloten met eventueel een langere studieduur als gevolg. Het feit dat sommige kinderen besluiten niet door te leren, betekent dat zij minder investeren in hun menselijk kapitaal. Het bindend studieadvies veroorzaakt dan mogelijk een welvaartsverlies. Het kan dus zijn dat uitbreiding van het bindend studieadvies effectief is om de studieduur te beperken, maar tegelijkertijd tot welvaartsverliezen leidt door teruglopende investeringen in menselijk kapitaal.

Een doelvariabele die wellicht beter in staat is welvaartseffecten in kaart te brengen, is het aantal behaalde studiepunten (ECTS) dat een student binnen een bepaalde periode (bijvoorbeeld 2 jaar) haalt. Immers, als iemand besluit niet meer verder te studeren door uitbreiding van het studieadvies, dan haalt hij/zij minder ECTS. Dat zie je dan terug in de analyse. Een voordeel is ook dat deze variabele sneller beschikbaar komt dan de studieduur. Een nadeel is dat deze informatie bij de instellingen zelf moet worden opgehaald, aangezien zij niet door de Dienst Uitvoering Onderwijs (DUO) centraal wordt bijgehouden. Een ander nadeel is dat dit slechts een indicator is van de studieduur: van een student die binnen twee jaar veel studiepunten haalt, maar daarna op zijn lauweren rust, wordt niet meegenomen dat hij/zij wellicht toch nog lang over zijn studie doet.

Omdat de pilotuitbreiding bindend studieadvies ook gericht is op het behalen van doelmatigheidswinst en kwaliteitsverbetering, zou een effectevaluatie zich ook hierop moeten richten. Probleem is wel dat in tegenstelling tot de studieduur, de indicatoren voor doelmatigheid en kwaliteit niet scherp gedefinieerd zijn en ook niet voor alle instellingen en opleidingen op een uniforme wijze verzameld worden.

Om het causale verband tussen de introductie van het bindend studieadvies in het tweede studiejaar op de studieduur te kunnen vaststellen, is de belangrijkste voorwaarde dat we uiteindelijk gegevens hebben van twee groepen.[20] Bij de eerste groep is sprake van bindend studieadvies (de behandelgroep), terwijl dat bij de tweede groep niet het geval is (de controlegroep). Idealiter zijn de twee groepen met uitzondering van dit aspect verder identiek. Op dit moment, maart 2014, lijken de opleidingen aan de Universiteit Leiden hiervoor het meeste potentieel te bieden.[21] Binnen deze universiteit zijn namelijk twee groepen studenten te onderscheiden: studenten die met het bindend studieadvies te maken krijgen en studenten die hier niet mee te maken krijgen. De studenten in de zogeheten unica-opleidingen worden

[20] Dit geldt ook voor een effectevaluatie gericht op aantallen studenten, doelmatigheid en kwaliteit. Omwille van de leesbaarheid zullen we waar nodig deze andere doelstellingen van de pilot in de voetnoten aan de orde stellen.

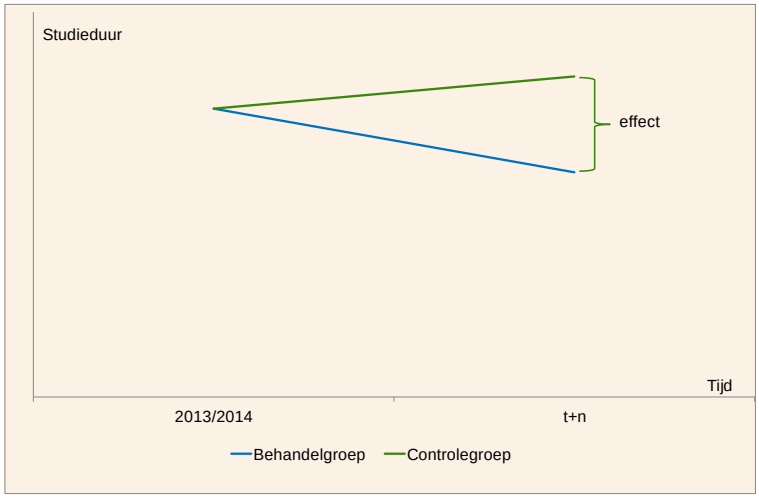
[21] Bij aanvang van het collegejaar 2013/2014 hebben ook twee opleidingen van de Gerrit Rietveld Academie en een opleiding van de combinatie UVA/VU een uitgebreid studieadvies. Ook de resultaten van deze opleidingen zouden in aanmerking kunnen komen voor een effectevaluatie als er een bijpassende controlegroep kan worden gedefinieerd.

immers vrijgesteld van het bindend studieadvies in het tweede studiejaar, terwijl de rest van de studenten wel is gebonden aan dit studieadvies. Bij de start van het collegejaar 2014/2015 komen daar vier nieuwe instellingen bij.

Beoordeling

In theorie zijn er drie manieren om binnen deze pilot tot een behandel- en controlegroep te komen. De eerste manier is door opleidingen bij aanvang willekeurig te verdelen over opleidingen met en zonder bindend studieadvies. Dit betekent dat door loting wordt bepaald of een opleiding in de controle- of behandelgroep terechtkomt.[22] De pilot krijgt dan het karakter van een 'random design' experiment. Nadat alle studenten hun studie hebben afgerond kan dan het effect van het bindend studieadvies eenvoudig geschat worden door de doelvariabele (studieduur) van de studenten in de behandelgroep te vergelijken met die van de studenten in de controlegroep. Figuur 9 schetst een dergelijk 'random design' experiment. Een voordeel van een dergelijk experiment is dat een voormeting niet nodig is. Dit wil zeggen dat de studieduur (doelvariabele) niet aan het begin van het experiment gemeten hoeft te worden. De loting zorgt er namelijk voor dat de behandel- en controlegroep vooraf aan het experiment vergelijkbaar zijn. In de figuur kan dit worden afgelezen aan het feit dat de lijntjes van de controle- en behandelgroep op elkaar liggen bij invoering van het bindend studieadvies (collegejaar 2013/2014). Het verschil op tijdstip t+n tussen de studieduur van beide groepen geeft het effect van het bindend studieadvies weer.[23]

Figuur 9 'Random design experiment' bij uitbreiding bindend studieadvies



[22] Loting garandeert dat willekeurig ook echt willekeurig is. Andere vormen van toewijzing aan behandel- of controlegroep, zoals vrijwilligheid, kunnen tot selectie-effecten leiden.
[23] Als aantallen studenten die binnen de nominale studieduur hun bachelordiploma behalen, doelmatigheid en/of kwaliteit als doelvariabele in de effectevaluatie worden gebruikt, liggen de lijnen uit figuur 9 waarschijnlijk andersom. De instellingen uit de behandelgroep kennen na afloop een sterkere toename van het aantal studenten dat binnen de nominale studieduur afstudeert, een hogere doelmatigheid en/of kwaliteit waardoor de blauwe lijn boven de groene lijn uitkomt.

Helaas is het in de praktijk onmogelijk opleidingen door middel van loting aan opleidingen toe te wijzen. Instellingen en opleidingen kunnen zich onder enkele voorwaarden zelf aanmelden voor de pilot 'uitbreiding bindend studieadvies'. Hierdoor vervalt het 'random design' karakter en daarmee deze wijze van effectevaluatie voor uitbreiding bindend studieadvies. Een na afloop van de pilot geobserveerd verschil in studieduur tussen studenten van beide groepen opleidingen wordt dan immers veroorzaakt door de selectie-effecten, omdat opleidingen en studenten bepalen in welke groep zij terechtkomen. Daarmee wordt de effectevaluatie van deze maatregel onzuiver.

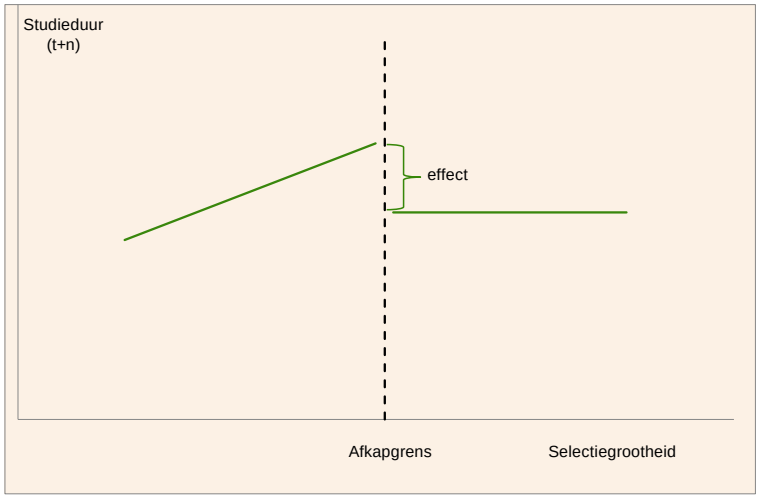
Een alternatieve, tweede, manier om twee groepen te krijgen, is de instelling van een afkapcriterium op grond waarvan opleidingen verdeeld worden over een behandel- en een controlegroep. Zo'n afkapcriterium vereist dan een ranking van opleidingen op basis van een selectievariabele, waarbij de opleidingen onbekend zijn met de afkapgrens. Deze variabele kan van alles zijn, bijvoorbeeld het aantal ingeschrevenen in een opleiding of de gemiddelde studieduur van de opleiding. Opleidingen met een waarde onder het afkapcriterium krijgen dan geen bindend studieadvies, terwijl opleidingen met een waarde boven dit criterium dit wel krijgen. Door een dergelijk selectiecriterium toe te passen, zullen de opleidingen rondom de afkapgrens vergelijkbaar zijn. Het effect van het bindend studieadvies op de studieduur kan dan gemeten worden door de opleidingen rondom deze afkapgrens te vergelijken. Deze methode staat bekend als 'regression discontinuity analysis'. Figuur 10 schetst dit. Het effect van het bindend studieadvies wordt weergegeven door de discontinuïteit in de studieduur die is ontstaan op de grens tussen behandel- en controlegroep na afloop van de pilot: de studenten in opleidingen rechts van het criterium (behandelgroep) hebben een kortere studieduur dan de studenten in opleidingen links hiervan (controlegroep).[24] Het voordeel van een regressie-discontinuïteits-analyse is dat een voormeting niet nodig.[25]

Helaas is ook deze manier van effectevaluatie voor deze pilot niet haalbaar. Omdat opleidingen zelf de keuze maken omtrent wel/niet bindend studieadvies in het tweede leerjaar, ontbreekt voor het toewijzen van opleidingen aan de groepen met wel/niet een bindend studieadvies een selectievariabele. Dat betekent dat de uiteindelijk gevonden effecten het gevolg zijn van de uitbreiding bindend studieadvies en selectie-effecten door de samenstelling van de behandel- en controlegroep. Daarmee levert een evaluatie met de hiervoor beschreven 'regression

[24] Als aantallen studenten die binnen de nominale studieduur hun bachelordiploma behalen, doelmatigheid en/of kwaliteit als doelvariabele in de effectevaluatie worden gebruikt, ligt de lijn rechts van de afkapgrens uit figuur 10 waarschijnlijk hoger dan de lijn links van deze grens.
[25] Een voormeting zal wel de geloofwaardigheid van de analyse kunnen vergroten. Een voormeting kan namelijk de veronderstelling testen dat opleidingen rondom de afkapgrens vóór invoering van het bindend studieadvies in het tweede jaar vergelijkbaar zijn. De groene lijnen in figuur 10 zouden bij een voormeting samenvallen tot 1 lijn die continu doorloopt rondom de afkapgrens. De nameting moet dan de discontinuïteit te zien geven, zoals in figuur 10.

discontinuity analysis' niet het causale effect van uitbreiding bindend studieadvies op.

Figuur 10 'Regression discontinuity analysis' bij uitbreiding bindend studieadvies



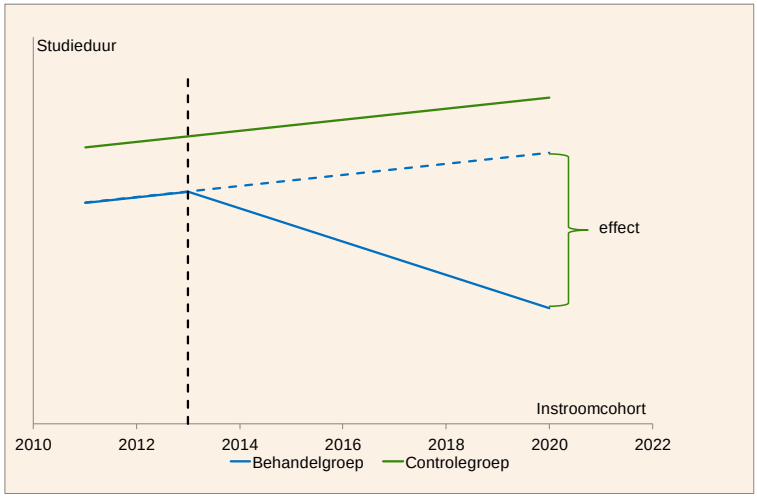
De derde methode is de 'differences-in-differences' methode ('diff-in-diff' analyse), waarbij de verandering in de ontwikkeling van de studieduur tussen controle- en behandelgroep wordt beschouwd. Deze controle- en behandelgroep hoeven niet op elkaar te lijken. Wat wel belangrijk is, is dat zij voor aanvang van het experiment dezelfde trend in de doelvariabele (studieduur) hebben. Onder de aanname dat de ontwikkeling van de behandelgroep zonder uitbreiding bindend studieadvies tot het tweede studiejaar dezelfde trend zou hebben gevolgd als de ontwikkeling van de controlegroep, kan het causaal effect van het bindend studieadvies worden geschat. Figuur 11 illustreert de 'differences-in-differences' methode. Het verschil tussen de gestippelde blauwe lijn en de doorgetrokken blauwe lijn illustreert het effect van het bindend studieadvies. Het kan geschat worden door het verschil tussen de ontwikkeling van de groene lijn en de ontwikkeling van de blauwe lijn te nemen.[26] Belangrijk voor deze methode is dus dat er voor zowel behandel- als controlegroep voor- en nametingen van de studieduur (doelvariabele) plaatsvinden. De meting voorafgaand aan de uitbreiding van het studieadvies moet immers de veronderstelling dat behandel- en controlegroep in die periode gelijk op gingen, bevestigen.

Voor een 'diff-in-diff'analyse van de effecten van het bindend studieadvies zou bijvoorbeeld de studieduur van de cohorten studenten die deelnemen in opleidingen die in augustus 2013 starten met een bindend studieadvies op de Universiteit Leiden (blauwe doorgetrokken lijn) vergeleken kunnen worden met de studieduur van de

[26] Als doelmatigheid en/of kwaliteit als doelvariabele in de effectevaluatie worden gebruikt, ligt het meer voor de hand dat de blauwe lijn zal stijgen in plaats van dalen.

cohorten studenten aan de andere opleidingen (in dit geval de unica) van deze universiteit (groene lijn). Gelet op de beschikbaarheid van de toegankelijkheid van cijfers over studie-uitval lijkt een 'diff-in-diff' analyse daarmee mogelijk. De voormeting zal bestaan uit het meten van de studieduur van de studenten die in de studiejaren voorafgaande aan het studiejaar 2013/2014 als eerstejaars zijn ingestroomd. Hierbij is het van belang om een aantal cohorten te nemen. Zo kan worden nagegaan of controle- en behandelgroep dezelfde trend hebben voor interventie. [27]

Figuur 11 'Diff-in-Diff' analyse bij uitbreiding bindend studieadvies



Een risico bij deze 'diff-in-diff' analyse is dat studenten hun keuze voor een opleiding mede laten afhangen van de aanwezigheid van een bindend studieadvies aan het eind van het tweede leerjaar. Er ontstaat dan een zogenaamde 'selection bias' in de resultaten. Studenten die niet vrezen voor het bindend studieadvies in het tweede leerjaar ('goede' studenten) laten zich namelijk niet afschrikken door uitbreiding van het bindend studieadvies, maar de mindere studenten die zo'n advies vrezen, kiezen wellicht voor een soortgelijke opleiding aan een andere instelling zonder zo'n advies. Als gevolg hiervan leidt uitbreiding van het bindend studieadvies ook tot een andere verdeling van de studenten over de instellingen. Als voorbeeld nemen we de opleiding politicologie aan de Universiteit Leiden. Als 'slechte' studenten weten dat dit een opleiding is met een bindend studieadvies, dan kunnen zij ervoor kiezen om uit te wijken naar andere universiteiten waar deze opleiding ook wordt aangeboden, maar waar een dergelijk advies niet van kracht is, bijvoorbeeld bij de Rijksuniversiteit Groningen. Een eindevaluatie binnen de Universiteit Leiden waarbij opleidingen met en zonder bindend studieadvies met elkaar vergeleken worden, zal dan laten zien dat de studieduur van de studenten politicologie afneemt ten opzichte

[27] Voor opleidingen die in augustus 2014 starten met de uitbreiding bindend studieadvies verschuiven de genoemde jaartallen uiteraard een jaar.

van de studieduur van de studenten aan de unica-opleidingen van deze universiteit. Maar wat niet wordt waargenomen is dat de studieduur van de studenten politicologie aan de Rijksuniversiteit Groningen is toegenomen. Een eindevaluatie geeft dan dus een partieel beeld, aangezien er met de totale studieduur onder alle studenten niets hoeft te zijn gebeurd. Dit mogelijke probleem zou ondervangen kunnen worden als ook de opleidingen waarnaar de 'mindere' studenten uitwijken een bindend studieadvies zouden hebben. Maar dat is in deze pilot niet het geval, omdat bepaald is dat maximaal 10% van de studenten te maken krijgt met een bindend studieadvies in het tweede studiejaar. In de pilot blijven dus uitwijkmogelijkheden bestaan. Dit mogelijke probleem kan in een 'diff-in-diff' analyse ondervangen worden door te bestuderen of een ander type studenten zich bij de opleidingen met een uitgebreid bindend studieadvies gaat inschrijven na invoering van dit advies in het tweede collegejaar.

Een tweede ontsnappingsroute aan de 'selection bias' is het overvallen van de studenten met informatie over de uitbreiding van het bindend studieadvies nadat de studiekeuze heeft plaatsgevonden. Uitbreiding van het bindend studieadvies heeft dan immers geen effect op de studiekeuze gehad. Van een zuivere 'overval' is echter bij de invoering van het uitgebreide studieadvies geen sprake. De informatie hiervoor was bij inschrijving voor de studenten beschikbaar. In hoeverre deze informatie al een rol heeft gespeeld bij de studiekeuze van deze studenten en dus tot selectie-effecten heeft geleid, is moeilijk na te gaan. Het lijkt ons waarschijnlijk dat de kans daarop het kleinst is bij het eerste cohort dat met de maatregel geconfronteerd wordt. Een effectevaluatie die gebruik maakt van de resultaten van dit cohort, lijkt dan ook de beste bescherming te bieden tegen een vertekening door selectie-effecten.

Eindoordeel

De uitvoering van een effectevaluatie met behulp van een 'diff-in-diff' analyse, waarbij de interventie gedefinieerd is als de aanwezigheid van een bindend studieadvies in het tweede leerjaar, lijkt kansrijk. Met name cijfers over studieduur zijn goed voorhanden en ook de interventie (uitbreiding bindend studieadvies) is scherp gedefinieerd. Beperking is dat de effecten op studieduur niet dezelfde zijn als effecten voor economische welvaart/welzijn. Verder zal een evaluatie van deze interventie nog enige jaren op zich moeten laten wachten. Immers, tussen het moment van het bindend studieadvies in het tweede leerjaar en afronding van de bachelorstudie zullen nog enige jaren verstrijken voordat van alle studenten hun studieduur is vastgesteld.

Een evaluatie gericht op de behaalde doelmatigheids- of kwaliteitswinst zal lastiger zijn. Allereerst ontbreekt een scherpe definitie van deze begrippen. Ten tweede ontbreekt momenteel een georganiseerde en geüniformeerde dataverzameling van indicatoren voor kwaliteit en doelmatigheid bij de opleidingen voor hoger en wetenschappelijk onderwijs. Dat betekent dat de noodzakelijke gegevens voor de behandel- en controlegroep ontbreken voor heden en verleden. Met name de voor een 'diff-in-diff'analyse noodzakelijke voormeting wordt daarmee problematisch.

Opzet van de evaluatie

Een 'diff-in-diff' analyse zal zich, gelet op de risico's van een 'selection bias', het beste kunnen beperken tot de resultaten van de studieduur van het eerste cohort studenten dat met een uitbreiding van het bindend studieadvies wordt geconfronteerd. Studenten van opleidingen zonder een uitgebreid studieadvies, zoals de unica, vormen in de evaluatie dan de controlegroep. Naast de resultaten van de studieduur van de cohorten studenten die zijn ingestroomd ten tijde van de uitbreiding van het bindend studieadvies, zullen ook gegevens beschikbaar moeten zijn over de studieduur van de cohorten studenten die zijn ingestroomd in eerdere studiejaren. Dit is noodzakelijk om te testen of de ontwikkeling van de studieduur in de behandel- en controlegroep gelijk opgingen.

Eenvoudshalve uitgaand van het cohort 2013/2014 zullen zij in de zomer van 2015 aan het eind van het tweede leerjaar een bindend studieadvies ontvangen. In beginsel zullen alle studenten met een positief studieadvies aan het eind van het tweede leerjaar in de periode 2016-2018 hun bachelordiploma behalen.[28] De studenten met een negatief studieadvies zullen voor een deel hun opleiding elders voortzetten. In veel gevallen zal zo'n overstap tot verdere vertraging leiden. Omdat de studieduur van deze groep studenten niet mag ontbreken in de evaluatie zal het daarom tot

[28] Voor een positief studieadvies is het niet nodig dat alle studiepunten van het tweede leerjaar behaald zijn. Dat betekent dat een deel van het cohort 2013/2014 al enige vertraging in zijn studie zal hebben na het tweede leerjaar. Het is dan ook niet te verwachten dat alle studenten van cohort 2013/2014 in de zomer van 2016 hun bachelor hebben. Voor opleidingen die later starten met een uitgebreid studieadvies schuift de tijdlijn navenant op.

2019-2020 duren voordat de studieduur van alle studenten uit cohort 2013/2014 bekend is.[29] Dat is dan ook het eerste moment dat een evaluatie van deze pilot zou kunnen plaatsvinden.

Als een evaluatie van uitbreiding bindend studieadvies na 2020 ongewenst is, zou overwogen kunnen worden de studieduur als doelvariabele te vervangen door 'percentage leerlingen dat na het tweede leerjaar voldoet aan de norm bindend studieadvies'. Voor de populatie 2013/2014 zijn cijfers voor deze grootheid al rond de zomer van 2015 voorhanden.[30] Echter, de vraag is of het 'percentage leerlingen dat na het tweede leerjaar voldoet aan de norm bindend studieadvies' een goede voorspeller is van de verandering in de studieduur. Te verwachten valt wel dat een stijging van dit percentage zich ook vertaalt in een kortere studieduur van de studenten met een positief advies bij de desbetreffende opleiding. Maar het is onduidelijk wat er zal gebeuren met de studieduur van die studenten die als gevolg van een negatief studieadvies hun opleiding elders moeten vervolgen. Deze doelvariabele zal naar alle waarschijnlijkheid dan ook een geflatteerd beeld geven van de effecten van uitbreiding van het bindend studieadvies.

5 Studiebijsluiter

Samenvatting

Het verminderen van de studie-uitval door het geven van realistische voorlichting aan potentiële studenten in het middelbaar beroepsonderwijs (mbo) en hoger onderwijs is het doel van de studiebijsluiter. In het verleden is al op basis van vrijwilligheid door verschillende instellingen geëxperimenteerd met deze bijsluiter en met ingang van het studiejaar 2014/2015 zullen alle opleidingen in het hoger onderwijs een studiebijsluiter hebben. Als gebruik wordt gemaakt van de cijfers betreffende studie-uitval van de studentencohorten tot en met augustus 2013, lijkt een effectevaluatie van deze maatregel kansrijk. Deze evaluatie zal dan uit een differences-in-differences analyse bestaan waarbij de verandering in de ontwikkeling van de studie-uitval onder eerstejaarsstudenten tussen instellingen met en zonder een studiebijsluiter wordt beschouwd.

Doel

Doel van de studiebijsluiter is het geven van realistische voorlichting aan potentiële studenten in het middelbaar beroepsonderwijs (mbo) en hoger onderwijs. Een realistische voorlichting die betrekking heeft op het arbeidsmarktperspectief na

[29] Er zullen ook studenten zijn die na een negatief studieadvies besluiten niet verder te leren. Zij zorgen daarmee voor een verkorting van de gemiddelde studieduur, maar ook tot een daling in de investeringen in menselijk kapitaal. Zoals eerder vermeld wordt dit effect in de evaluatie niet meegenomen.
[30] Alternatief is het aantal behaalde studiepunten per student. Deze grootheid heeft een meer continue karakter dan het discrete wel/niet voldaan aan de norm gesteld door de uitbreiding bindend studieadvies.

voltooiing van de opleiding en de inhoud en kwaliteit van de aangeboden opleiding.[31] Met het beschikbaar komen van deze betere informatie kunnen potentiële studenten dan een betere keuze maken en dat zou moeten leiden tot minder studie-uitval.

Inhoud pilot

Bij de ontwikkeling van de studiebijsluiter loopt het hoger onderwijs iets voor op het mbo. Sinds 2012 ondersteunt Studiekeuze123 het project Studie in Cijfers en eind 2013 doen daaraan 27 van de 38 hogescholen mee. Ultimo 2013 participeert vanuit het wetenschappelijk onderwijs alleen de TU Delft. Verwacht wordt dat in mei 2014 de overige hogescholen en opleidingen voor wetenschappelijk onderwijs zullen deelnemen.[32] Vanaf dat moment kunnen dus alle potentiële studenten deze bron raadplegen. In navolgende verkenning van de mogelijkheden voor een effectevaluatie van de studiebijsluiter zullen we ons eerst beperken tot het hoger onderwijs, alwaar de bijsluiter al is ingevoerd. Tot slot zullen we kort ingaan op de mogelijkheden voor het mbo, alwaar de implementatie van de studiebijsluiter nog minder duidelijk is.

Wat kun/wil je meten?

Op grond van bovenstaande beschrijving lijken in beginsel drie doelvariabelen in aanmerking te komen om het effect van de studiebijsluiter te meten. Ten eerste de behaalde kwaliteitswinst ten aanzien van de verstrekte informatie. Ten tweede de effecten op de studie-uitval. Ten derde is het mogelijk de effecten van de studiebijsluiter op het switchgedrag van studenten te bestuderen. Op dit moment is er geen adequate maatstaf voorhanden om de kwaliteit van de verstrekte informatie te meten. Dit beperkt de mogelijkheden om de effecten van de studiebijsluiter op de kwaliteit van de informatie vast te stellen. Cijfers over studie-uitval en switchen tussen opleidingen zijn wat dit betreft beter voorhanden.

Om het causale verband tussen de introductie van de bijsluiter en de studie-uitval te kunnen vaststellen, is de belangrijkste voorwaarde dat we uiteindelijk gegevens hebben van twee groepen.[33] Bij de eerste groep is sprake van een studiebijsluiter (de behandelgroep), terwijl dat bij de tweede groep niet het geval is (controlegroep). Idealiter zijn de twee groepen met uitzondering van dit aspect verder identiek.

Beoordeling

In theorie zijn er drie manieren om voor een evaluatie van de effecten van de studiebijsluiter tot een behandel- en controlegroep te komen. De eerste manier is door instellingen willekeurig te verdelen over instellingen die een studiebijsluiter hebben en instellingen die dit niet hebben. Dit betekent dat door loting wordt

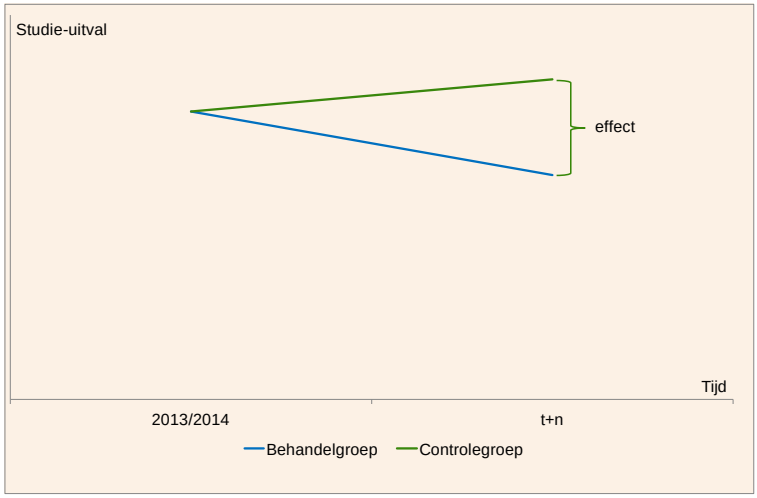
---

[31] Zie blz. 1, 2 en 5 Brief Tweede Kamer van 19 december 2013 over “Voortgang van de invoering van Studie in Cijfers (“studiebijsluiter”)”.
[32] Zie blz. 2 Brief Tweede Kamer van 19 december 2013 over “Voortgang van de invoering van Studie in Cijfers (“studiebijsluiter”)”.
[33] In de verdere beschrijving van de mogelijkheden voor een effectevaluatie wordt steeds gesproken over de studie-uitval als doelvariabele. Maar hiervoor kan ook de studieswitch gelezen worden .

bepaald of een instelling in de controle- of behandelgroep terechtkomt.[34] De evaluatiestudie krijgt dan het karakter van een 'random design' experiment. Het effect van de studiebijsluiter kan dan worden vastgesteld door doelvariabele (de studie-uitval in het eerste leerjaar) van studenten bij instellingen met de studiebijsluiter (de behandelgroep) te vergelijken met instellingen zonder studiebijsluiter (de controlegroep). Figuur 12 schetst een dergelijk 'random design' experiment. Een voordeel van een dergelijk experiment is dat een voormeting niet nodig is. Dit wil zeggen dat de doelvariabele niet aan het begin van het experiment gemeten hoeft te worden. De loting zorgt er namelijk voor dat de behandel- en controlegroep vooraf aan het experiment vergelijkbaar zijn. In de figuur kan dit worden afgelezen aan het feit dat de lijntjes van de controle- en behandelgroep op elkaar liggen bij invoering van de studiebijsluiter (tijdstip t). Nadat alle studenten hun studie hebben afgerond (tijdstip t+n), geeft het verschil tussen de studie-uitval van beide groepen het effect van de studiebijsluiter weer.

Helaas is het bij de evaluatie van de effecten van de studiebijsluiter onmogelijk instellingen door middel van loting toe te wijzen aan een behandelgroep en een controlegroep. Een deel van de instellingen heeft zich namelijk inmiddels al aangemeld voor de studiebijsluiter en deze ook al ingevoerd. Hierdoor vervalt het 'random design' karakter en daarmee deze wijze van effectevaluatie voor de studiebijsluiter bij hbo-opleidingen.

Figuur 12 'Random design experiment' bij effectevaluatie studiebijsluiter

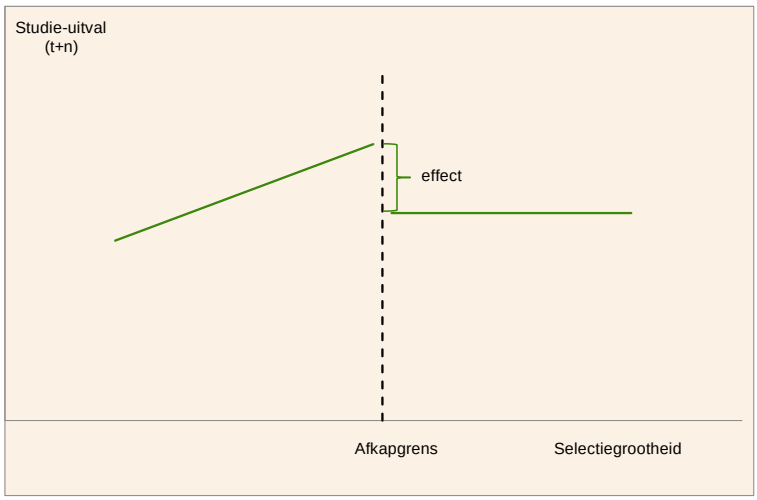


Een alternatieve, tweede, manier om twee groepen te krijgen, is de instelling van een afkapcriterium op grond waarvan instellingen verdeeld worden over een behandel- en een controlegroep. Zo'n afkapcriterium vereist dan een ranking van instellingen op basis van een selectievariabele, waarbij de instellingen onbekend zijn met de

[34] Loting garandeert dat willekeurig ook echt willekeurig is. Andere vormen van toewijzing aan behandel- of controlegroep, zoals vrijwilligheid, kunnen tot selectie-effecten leiden.

afkapgrens. Deze variabele kan van alles zijn, bijvoorbeeld het aantal ingeschrevenen bij een instelling, of het aantal opleidingen bij een instelling. Instellingen met een waarde onder het afkapcriterium krijgen dan geen studiebijsluiter, terwijl instellingen met een waarde boven dit criterium deze wel krijgen. Door een dergelijk selectiecriterium toe te passen, zullen de instellingen rondom de afkapgrens vergelijkbaar zijn. Het effect van de studiebijsluiter op de studie-uitval in het eerste leerjaar kan dan gemeten worden door de scholen rondom deze afkapgrens te vergelijken. Deze methode staat bekend als 'regression discontinuity analysis'. Figuur 13 schetst een dergelijke analyse. Het effect van de studiebijsluiter wordt weergegeven door de discontinuïteit in de studie-uitval die is ontstaan op de grens tussen behandel- en controlegroep na afloop van de pilot: de instellingen rechts van het criterium (behandelgroep) hebben een lagere studie-uitval dan de opleidingen links hiervan (controlegroep). Het voordeel van een regressie-discontinuïteitsanalyse is dat een voormeting niet nodig is.[35]

Figuur 13 'Regression discontinuity analysis' bij effectevaluatie studiebijsluiter



Ook deze manier van effectevaluatie is voor evaluatie van de effecten van de studiebijsluiter bij de hbo-instellingen niet haalbaar. Omdat instellingen zelf al wel/niet een studiebijsluiter hebben ingevoerd, ontbreekt voor het toewijzen van instellingen aan de groepen wel/niet studiebijsluiter een selectievariabele. Eventuele effecten van de studiebijsluiter op de studieduur zijn dan het ook gevolg van verschillen tussen behandel- en controlegroep.

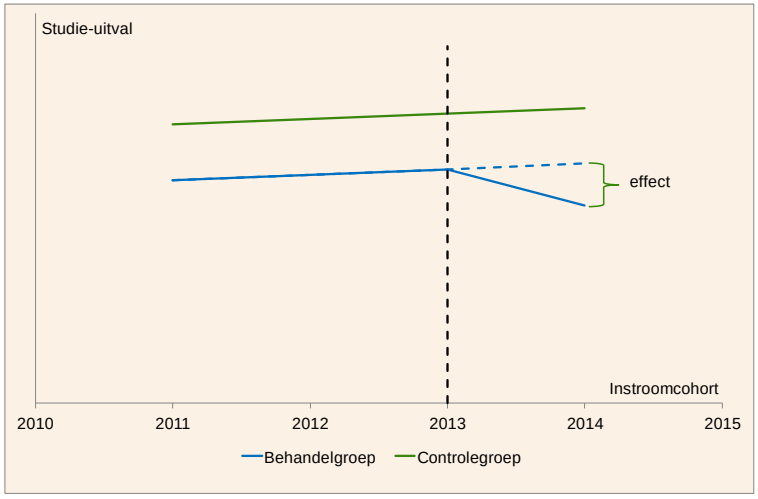
De derde methode is de 'differences-in-differences' methode ('diff-in-diff' analyse) waarbij de verandering in de ontwikkeling van de studie-uitval onder

[35] Een voormeting zal wel de geloofwaardigheid van de analyse kunnen vergroten. Een voormeting kan namelijk de veronderstelling testen dat scholen rondom de afkapgrens vóór invoering van de studiebijsluiter vergelijkbaar zijn. De groene lijnen in figuur 13 zouden bij een voormeting samenvallen tot 1 lijn die continu doorloopt rondom de afkapgrens. De nameting moet dan de discontinuïteit te zien geven, zoals in figuur 13.

eerstejaarsstudenten tussen een controle- en behandelgroep wordt beschouwd. Deze controle- en behandelgroep hoeven niet op elkaar te lijken. Wat wel belangrijk is, is dat zij voor aanvang van het experiment dezelfde trend in de doelvariabele (studie-uitval in het eerste leerjaar) hebben. Onder de aanname dat de ontwikkeling van de behandelgroep zonder studiebijsluiter dezelfde trend zou hebben gevolgd als de ontwikkeling van de controlegroep, kan het causale effect van de studiebijsluiter worden geschat. Figuur 14 illustreert de 'differences-in-differences' methode. Hierbij is aangenomen dat bij aanvang van het studiejaar 2013/2014 de studiebijsluiter is ingevoerd bij de groep hbo-instellingen waarvan de blauwe lijn de studie-uitval in het eerste jaar beschrijft. De groene lijn illustreert de studie-uitval bij de instellingen zonder studiebijsluiter. De blauwe stippellijn beschrijft het veronderstelde verloop van de studie-uitval bij instellingen met een studiebijsluiter als deze niet was ingevoerd bij aanvang studiejaar 2013/2014. Het verschil tussen de gestippelde blauwe lijn en de doorgetrokken blauwe lijn illustreert vervolgens het causale effect van de studiebijsluiter. Het kan geschat worden door het verschil tussen de ontwikkeling van de groene lijn en de ontwikkeling van de blauwe lijn te nemen. Belangrijk voor deze methode is dus dat er voor zowel de behandel- als de controlegroep (resp. blauwe en groene lijn) voor- en nametingen van de studie-uitval plaatsvinden. De meting voorafgaand aan de invoering van de studiebijsluiter moet immers de veronderstelling dat de studie-uitval in de behandelgroep en die in de controlegroep in die periode gelijk opgingen, bevestigen.

Gelet op de beschikbaarheid van cijfers over studie-uitval lijkt een 'diff-in-diff' analyse daarmee mogelijk. De voormeting zal dan bestaan uit het meten van de studie-uitval van de cohorten eerstejaarsstudenten die zijn ingestroomd vóór invoering van de studiebijsluiter (bijvoorbeeld vóór studiejaar 2013/2014).

Figuur 14 'Diff-in-Diff' analyse bij effectevaluatie studiebijsluiter



Er zijn twee risico's bij deze analyse. Ten eerste wordt de studiebijsluiter op instellingsniveau geïntroduceerd en kunnen instellingen tegelijk met deze studiebijsluiter ook nog een ander pakket aan maatregelen nemen om de studiekeuze te bevorderen. Hierdoor wordt de interventie 'studiebijsluiter' minder scherp gedefinieerd. Dit zou kunnen betekenen dat bij een effectevaluatie niet alleen het effect van de studiebijsluiter geschat wordt, maar ook dat van de andere maatregelen. Het effect van de studiebijsluiter wordt dan dus niet geïsoleerd van deze overige maatregelen.

Een tweede risico is dat studenten door gebruik van de studiebijsluiter besluiten naar een andere instelling te gaan, of helemaal niet te gaan studeren. Er ontstaat dan een mogelijke 'selection bias' in de resultaten. Studenten die zich laten afschrikken door de informatie uit de studiebijsluiter kunnen de minder goede studenten zijn. Als gevolg hiervan leidt de studiebijsluiter tot een andere verdeling van de studenten over de behandel- en controlegroep. Een eindevaluatie zal dan laten zien dat bij de instellingen met een studiebijsluiter de studie-uitval in het eerste leerjaar is teruggelopen, terwijl bij de instellingen zonder zo'n bijsluiter de studie-uitval juist is toegenomen. In een 'diff-in-diff' analyse lijkt het er dan op dat de studiebijsluiter op een succesvolle wijze de studie-uitval heeft verminderd, terwijl dat eigenlijk niet het geval is; de bijsluiter heeft slechts geleid tot een andere verdeling van studenten over de instellingen, maar de totale studie-uitval onder *alle* studenten is feitelijk niet afgenomen.

De aanwezigheid van dit mogelijke probleem kan in de analyse worden bestudeerd door als doelvariabele achtergrondkenmerken van de studenten te nemen en te kijken of over de tijd een ander type studenten zich bij de onderwijsinstellingen met studiebijsluiter inschrijft dan bij de onderwijsinstellingen zonder studiebijsluiter.

Eindoordeel

De uitvoering van een effectevaluatie van de studiebijsluiter bij hbo-instellingen met behulp van een 'diff-in-diff' analyse is technisch gesproken mogelijk, aangezien cijfers over studie-uitval en switchgedrag in principe goed voorhanden zijn. Er kleven echter wel risico's aan deze analyse. Ten eerste wordt de studiebijsluiter op instellingsniveau geïntroduceerd en kunnen instellingen tegelijk met deze studiebijsluiter ook nog een ander pakket aan maatregelen nemen om een betere studiekeuze te bevorderen. Hierdoor wordt de interventie 'studiebijsluiter' minder scherp gedefinieerd. Dit zou kunnen betekenen dat bij een effectevaluatie niet alleen het effect van de studiebijsluiter geschat wordt, maar ook dat van de andere maatregelen. Het effect van de studiebijsluiter wordt dan niet geïsoleerd van deze overige maatregelen. Ten tweede bestaat er een mogelijkheid van selectie-effecten: studenten kunnen door gebruik van de studiebijsluiter bij de instellingen in de behandelgroep besluiten naar instellingen in de controlegroep te gaan of helemaal niet te gaan studeren. In een 'diff-in-diff' analyse lijkt het er dan op dat de studiebijsluiter op een succesvolle wijze de studie-uitval heeft verminderd, terwijl dat

eigenlijk niet het geval is; de bijsluiter heeft slechts geleid tot een andere verdeling van studenten over de instellingen, terwijl de totale studie-uitval onder alle studenten feitelijk niet is afgenomen. Een ander risico is dat het aantal beschikbare instellingen (thans 38) aan de kleine kant is. Wellicht dat een gefaseerde invoering van de studiebijsluiter in het mbo hier nog verandering in kan brengen.

Opzet van de evaluatie

Een 'diff-in-diff' analyse kan worden uitgevoerd door de studie-uitval in het eerste jaar van de studenten in instellingen met een studiebijsluiter te vergelijken met de studenten in instellingen zonder studiebijsluiter. Voor het hbo gaat het dan in het studiejaar 2013/2014 om 27 instellingen die vergeleken kunnen worden met 11 instellingen die geen studiebijsluiter kennen.[36]

Belangrijk bij een dergelijke analyse is dat er voor- en nametingen plaatsvinden in zowel behandel- als controlegroep. Dat betekent dat voor de voormeting bij alle 38 hbo-instellingen de studie-uitval gemeten wordt van de cohorten eerstejaarsstudenten die in de studiejaren vóór 2013/2014 zijn ingestroomd. Voor deze cohorten geldt immers dat zij allen niet de beschikking hadden over een studiebijsluiter. Wat betreft de nameting moet het cohort eerstejaarsstudenten worden gebruikt dat in het studiejaar 2013/2014 is ingestroomd. Voor dit cohort geldt immers dat het zich deels in de behandelgroep (in een van de 27 instellingen) en deels in de controlegroep (in een van de overige 11) bevindt. De cohorten die na 2013/2014 instromen in het hbo kunnen niet voor het onderzoek benut worden omdat elke instelling dan een studiebijsluiter kent, waardoor alle studenten zich met deze bijsluiter geconfronteerd zien. Er is dan geen controlegroep meer.

Mogelijkheden tot een effectevaluatie van de studiebijsluiter voor het mbo

Op dit moment van schrijven hebben drie van de zestig mbo-instellingen aangegeven een studiebijsluiter te publiceren die voor studenten voor het collegejaar 2014/2015 een rol kan spelen bij de studiekeuze.

Het is de bedoeling dat in april 2014 door het ministerie een definitief besluit over de vormgeving van de studiebijsluiter voor het mbo wordt vastgesteld en dat per 1 januari 2015 alle mbo-instellingen verplicht worden tot publicatie van die bijsluiter. Dat laatste betekent dat alle eerstejaars van het collegejaar 2015/2016 bij hun keuze gebruik kunnen maken van die bijsluiter (voor zover zij niet voor de

---

[36] We beschrijven de opzet van een evaluatiestudie alsof de 27 instellingen in het studiejaar 2013/2014 voor het eerst de studiebijsluiter hebben gepubliceerd. Dat is niet conform de realiteit waarin al een aantal instellingen eerder de studiebijsluiter beschikbaar hebben gesteld. Bij deze instellingen zijn de effecten op de studie-uitval dus ook al bij eerdere cohorten zichtbaar. In termen van figuur 14 betekent dit een soortgelijke figuur voor eerdere cohorten. Om te voorkomen dat voor de analyse het aantal instellingen dat wel/niet een studiebijsluiter heeft te beperkt wordt, ligt een gepoolde 'diff-in-diff' analyse dan voor de hand. Verder is belangrijk voor deze evaluatie dat bij deze 27 instellingen tegelijkertijd geen andere specifieke wijzigingen hebben plaatsgevonden. Als dat wel het geval is, zal een effectevaluatie het gecombineerde effect van meerdere wijzigingen te zien geven,

inwerkingtreding van de wettelijke verplichting al een keuze hebben gemaakt). Daarmee ontbreekt voor dat cohort dus de controlegroep. Indien de inwerkingtreding van het wetsvoorstel later dan 1 januari 2015 zal zijn, dan is er wellicht wel een mogelijkheid om te komen tot een controlegroep van mbo-instellingen die geen gebruik maken van de bijsluiter.

Een eventuele effectevaluatie voor de mbo-studiebijsluiter zal dus moeten plaatsvinden op de resultaten van de eerstejaars uit collegejaar 2014/2015 (of collegejaar 2015/2016 indien de inwerkingtreding later zal zijn). Deze evaluatie zal net als bij de evaluatie van de studiebijsluiter bij het hoger onderwijs uit een 'diff-in-diff' analyse bestaan. Hierbij zal de studieduur (of het switchgedrag) van de cohorten studenten die zijn ingestroomd in een van de drie instellingen die een studiebijsluiter hebben gepubliceerd, in 2014 worden vergeleken met die van de cohorten studenten die zijn ingestroomd in de overige 57 instellingen. Een aandachtspunt is wel dat de eerstejaars in het studiejaar 2014/2015 mogelijk weinig beïnvloed zijn door de studiebijsluiter gezien het late moment waarop deze beschikbaar kwam.

Alternatieve indelingen om te komen tot een behandel- en een controlegroep, bijvoorbeeld naar leeftijd of naar type mbo (voltijds, beroepsopleidende leerweg, beroepsbegeleidende leerweg) lijken ons van beperkte toegevoegde waarde. Belangrijk is dat bij een van de beide groepen een studiebijsluiter wordt ingezet, terwijl dat bij de andere niet gebeurt. Met de huidige opzet van de regeling lijkt dit niet het geval te zijn. Zodra de studiebijsluiter bij een instelling wordt ingevoerd, geldt zij voor alle groepen studenten. Het is echter ook denkbaar om te onderzoeken of het gebruik van de bijsluiter verschilt tussen groepen. Een van de groepen wordt dan gedefinieerd als controlegroep en de effectevaluatie geeft dan inzicht in de mate waarin de behandelgroep anders reageert op de studiebijsluiter dan deze controlegroep. Omdat deze laatste ook kan reageren op de studiebijsluiter, is dit echter niet hetzelfde als een effectevaluatie van de introductie van de studiebijsluiter. Dat zou slechts het geval zijn als we vooraf met zekerheid zouden weten dat de controlegroep niet reageert op de studiebijsluiter. Maar dat is nu juist onbekend.

Tot slot is een mogelijkheid te onderzoeken of meerdere mbo-instellingen, vooruitlopend op inwerkingtreding van de verplichte informatievoorziening, bereid zijn om in een zuiver experimenteel onderzoek de meerwaarde van de bijsluiter te onderzoeken. Dat vergt echter uitstel van de inwerkingtreding van het wetsvoorstel tot 1 januari 2016 of later, zodat een deel van de mbo-instellingen in het collegejaar 2015/2016 wel een bijsluiter publiceert (behandelgroep) en een deel nog niet (controlegroep).

Dit is een uitgave van:

Centraal Planbureau
Van Stolkweg 14
Postbus 80510 | 2508 GM Den Haag
T (070) 3383 380

info@cpb.nl | www.cpb.nl

April 2014