Comprehensief onderwijs: een bedreiging voor kwaliteit? Een heranalyse van Rindermann en Ceci (2009)

J. Lavrijsen en I. Nicaise

Samenvatting

Uit onderwijskundig onderzoek blijkt over het algemeen dat een vroege opsplitsing van leerlingen tussen algemeen vormend en beroepsgericht secundair onderwijs (vroege tracking) de sociale en etnische ongelijkheden in onderwijsuitkomsten uitvergroot. Bovendien hangt een vroege opsplitsing niet samen met betere gemiddelde prestaties. Rindermann en Ceci (2009) vonden in een landenvergelijkend onderzoek echter wel een positieve samenhang tussen vroege selectie en gemiddelde prestaties. Hoe kan deze tegenspraak worden begrepen? In dit artikel laten we zien dat Rindermann en Ceci niet adequaat gecontroleerd hebben voor de verschillen tussen landen in hun erg heterogene dataset. Ten eerste waren hun drie controlevariabelen niet voldoende om alle relevante verschillen te ondervangen. Ten tweede veronderstelden Rindermann en Ceci dat het effect van de leeftijd van opsplitsing zelf onafhankelijk was van het ontwikkelingsniveau van het land. Door interacties op te nemen laten we zien dat deze aanname niet juist is. Onze heranalyse suggereert integendeel dat in welvarende landen de leeftijd van tracking geen duidelijk effect meer heeft op de gemiddelde prestaties, wat in lijn is met de rest van de literatuur.

1 Inleiding

Eén van de belangrijke kenmerken van een onderwijssysteem is de leeftijd waarop leerlingen worden opgedeeld in aparte onderwijsvormen in functie van hun academische prestaties (de zogenaamde 'tracking' in algemeen vormend versus beroepsgericht secundair onderwijs) (Lavrijsen, 2013). Landen waarin leerlingen al vroeg worden opgesplitst, zoals Duitsland, Nederland en Vlaanderen, worden "vroege trackers" genoemd, terwijl in landen met "comprehensieve" onderwijssystemen de leerlingen veel langer samen blijven (bv. de Scandinavische landen). Onderwijskundig onderzoek heeft vrij eensluidend vastgesteld dat een uitstel van de opsplitsing doorgaans samenhangt met meer gelijke onderwijskansen: de impact van sociale achtergrond op onderwijsprestaties in comprehensieve systemen een stuk minder sterk (zie Van de Werfhorst & Mijs (2010) voor een overzicht). De geplande hervorming van het Vlaamse secundair onderwijs stelde dan ook voor om een brede eerste graad in te voeren, waardoor de effectieve studieoriëntering zou verschuiven van 12 naar 14 jaar (cf. Luyten & Bosker, 2012).

In het debat over deze hervorming was echter de bezorgdheid te horen dat het uitstellen van de opsplitsing negatief zou uitdraaien voor het gemiddelde prestatieniveau in Vlaanderen ("nivellering naar beneden"). Op het eerste zicht lijkt het inderdaad efficiënter om leerlingen al vroeg op te splitsen in functie van hun capaciteiten: in homogene klassen kunnen de leerstof, het lestempo en de onderwijsstijl beter worden afgestemd op het niveau van de leerling. Hierbij moet echter de belangrijke kanttekening worden geplaatst dat studieoriëntering in de praktijk geen perfecte afspiegeling is van de capaciteiten van de leerling, bijvoorbeeld omdat ook sociale achtergrond die oriëntering beïnvloedt. Bovendien blijken de beroepsgerichte tracks, die te vaak gaan fungeren als de "onderkant van de waterval", in de praktijk helemaal niet zo efficiënt te werken. In tegenspraak met het veronderstelde specialisatievoordeel doen de

zwakste leerlingen in Vlaanderen het soms minder goed dan hun tegenhangers in bepaalde comprehensieve systemen. Sommige onderzoekers betogen dan ook dat het voor zwakke leerlingen voordelig kan zijn om de klas te delen met sterkere jaargenoten. Anderen houden dan weer vol dat een leerling beter presteert in het gezelschap van medeleerlingen van een vergelijkbaar niveau. In theorie kan het vroeg sorteren van leerlingen dus zowel positieve als negatieve gevolgen hebben voor de gemiddelde prestaties. Een genuanceerd eindoordeel kan dan ook enkel worden gevormd op basis van empirisch onderzoek.

De verschillende internationale scholierentests die sinds de jaren '90 op regelmatige basis worden uitgevoerd (zoals PISA), bieden hiertoe uitstekend materiaal. Het feit dat de best presterende landen in dergelijke tests vaak comprehensief zijn - met Finland als bekendste voorbeeld - suggereert alvast dat het uitstellen van de opsplitsing tussen onderwijsvormen niet noodzakelijk nadelig hoeft te zijn voor het gemiddelde niveau. Deze eenvoudige vaststelling is natuurlijk nog geen statistisch bewijs: ook andere factoren kunnen de prestaties immers beïnvloed hebben. Onderzoeken met heel verschillende onderzoekdesigns hebben deze stelling echter steviger onderbouwd. Prof. H. van de Werfhorst vatte eerder voor *Pedagogische Studiën* de empirische evidentie dan ook als volgt samen: *"Er is geen enkele aanwijzing dat in vroegselecterende landen de gemiddelde prestaties omhoog zouden gaan"* (Van de Werfhorst, 2011). Op basis van het verzamelde onderzoek stelde ook de OESO (2012) vast dat het uitstellen van de opsplitsing in onderwijsvormen geen nadelige effecten heeft voor de gemiddelde prestaties; een meer gedetailleerde beschrijving van het bestaande onderzoek ter zake kan worden gevonden in Lavrijsen, Nicaise & Wouters (2013).

Toch was er ook een enkel tegengeluid te horen. Rindermann en Ceci (2009) besloten op basis van een landenvergelijkend onderzoek dat een vroege opsplitsing wél positief zou zijn voor het gemiddelde prestatieniveau. Tegenstanders van de hervorming van het Vlaamse secundair onderwijs steunden dan ook uitdrukkelijk op deze studie om hun verzet kracht bij te zetten, zoals in Duyck en Anseel (2012): *"In de wetenschappelijke literatuur is een indrukwekkende cross–nationale vergelijkingsstudie beschikbaar die het effect van early tracking op leerprestaties analyseert. Vreemd genoeg ontbreekt vooralsnog elke verwijzing naar deze studie in het debat. De resultaten tonen eenduidig aan dat early tracking een positief effect heeft op leerprestaties, niet enkel voor de best presterende leerlingen, maar voor het gemiddeld leerniveau, controlerend voor alle andere socio–economische variabelen."*

Hoe kan de tegensprekelijke conclusie van Rindermann en Ceci worden begrepen? In dit artikel zullen we kort uitleggen aan welke problemen elk landenvergelijkend onderzoek het hoofd moet bieden, en welke technieken daarvoor bestaan. Daarna zal het artikel van Rindermann en Ceci meer in detail worden besproken. De oorspronkelijke auteurs stelden hun dataset vriendelijk ter beschikking voor verdere analyse. Op basis van een bijkomende analyse van deze dataset zullen we enkele verbeteringen aan hun analyse voorstellen, in het bijzonder de opname van interactie-effecten tussen context en tracking. Dit zal aantonen dat het oorspronkelijk vastgestelde positieve effect van een vroege opsplitsing toe te schrijven was aan de minder welvarende landen in de dataset: binnen de OESO verdwijnt het significante effect. Voor rijkere regio's zoals Vlaanderen zijn de oorspronkelijke conclusies van Rindermann en Ceci dan ook weinig relevant. Uit de literatuur komt integendeel overheersend de boodschap dat vroege tracking in Westerse landen niet noodzakelijk is om tot een kwaliteitsvol onderwijssysteem te komen.

2 Het probleem in landenvergelijkend onderzoek – en enkele oplossingen

Een belangrijk probleem in elk landenvergelijkend onderzoek is dat de uitkomst (de gemiddelde prestaties) niet alleen beïnvloed wordt door het kenmerk waarin je geïnteresseerd bent (de leeftijd van tracking). Ook heel

wat andere variabelen van binnen en buiten het onderwijssysteem (bv. welvaart, culturele waarden …) kunnen die prestaties beïnvloeden. Om het effect van tracking zuiver te kunnen bepalen, is het nodig om voor deze heterogeniteit te controleren. Hiervoor bestaan er verschillende technieken.

Een heel eenvoudige oplossing is om het onderzoek te beperken tot één enkel land en daarin dan het effect van een onderwijshervorming te bestuderen. Jakubowski (2010) toonde zo bijvoorbeeld aan dat een recente onderwijshervorming in Polen, waarbij de studieoriëntering werd uitgesteld, geleid heeft tot een opvallende verbetering van de resultaten van het land in de PISA-tests.

Een andere mogelijkheid is de "differences-in-differences"-aanpak, die onder meer werd toegepast door Hanushek en Woessmann (2006). Hierbij wordt gesteund op het idee dat verstorende variabelen (bv. welvaart) niet alleen de resultaten op studententests op 15 jaar (bv. PISA) beïnvloeden, maar dat ze een zelfde effect moeten hebben gehad op de resultaten van tests afgenomen op jongere leeftijd (bv. TIMSS, 4de leerjaar), d.w.z. vóór er sprake was van tracking. Het verschil tussen vroege en laat trackende landen wordt dan ook niet rechtstreeks afgeleid uit verschillen in de testresultaten op 15 jaar, maar wel uit verschillen tussen landen in de *toename* van de vaardigheden tussen het meetpunt in het basis- en in het secundair onderwijs (vandaar: "differences-in-differences"). Hanushek en Woessmann toonden zo aan dat er op die manier eigenlijk geen duidelijk verband tussen de leeftijd van tracking en het gemiddelde niveau was vast te stellen.

Een laatste optie is om de verstorende variabelen zelf mee op te nemen in het model. Onder meer Duru-Bellat en Suchaut (2005) toonden zo aan dat een licht positieve samenhang tussen comprehensief onderwijs en gemiddelde resultaten bleef bestaan na controle voor de economische context en de scholingsgraad van het land.

Ook Rindermann en Ceci hanteerden deze laatste aanpak – maar met een tegengesteld resultaat. Hoe is dat te begrijpen? Belangrijk is alvast dat de waarde van deze aanpak volledig afhangt van de adequaatheid van de controle voor de mogelijke verstorende variabelen: als we niet "alle" relevante verstorende variabelen in het model opnemen, dan blijven we met een verstoorde schatting zitten. Een moeilijkheid daarbij is dat we nooit zeker kunnen weten of we inderdaad alle verstoorders hebben opgenomen. Dit is vooral een probleem bij erg heterogene datasets: hoe meer de bestudeerde landen van elkaar verschillen op zaken die niets met onderwijs te maken hebben, hoe groter de kans dat we relevante verschillen over het hoofd zien.

Onderwijskundig onderzoek heeft zich traditioneel vooral gefocust op de ontwikkelde (Westerse) landen. Rindermann en Ceci hebben dit studiegebied uitgebreid met landen van over de hele wereld, waarbij ze ook "exotischer" landen zoals India, Iran en Zuid-Afrika in beschouwing namen. Op zich is dit zeker een verdienstelijke poging om meer "mondiale" uitspraken te kunnen doen. Tegelijkertijd is het echter de achilleshiel van de studie. Een zo heterogene dataset vraagt immers om een zeer doorgedreven controle voor contextvariabelen, omdat de kans dat een relevant verschil vergeten wordt erg groot is. De waarde van de resultaten van Rindermann en Ceci zal dan ook volledig afhangen van de adequaatheid waarmee ze gecontroleerd hebben voor verschillen in de context.

3 Opzet van het oorspronkelijke artikel

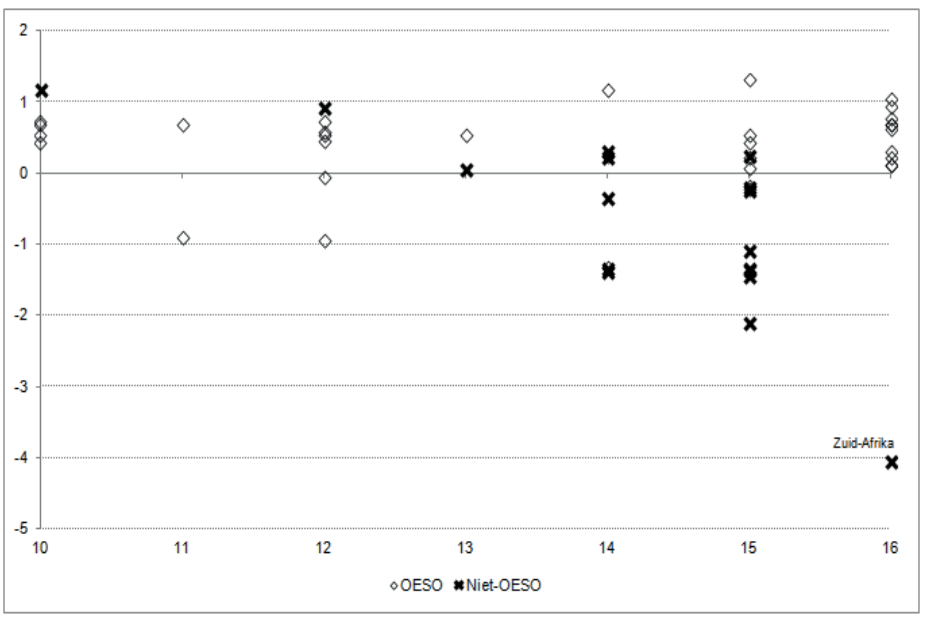
Rindermann en Ceci gebruikten als uitkomstvariabelen een aantal aggregaten1, d.w.z. de gemiddelde score van een land over een hele reeks scholierentests uit verschillende jaren (PIRLS: 4de leerjaar, TIMSS: 4de leerjaar en 2de middelbaar, IEA: 9- en 14-jarigen, PISA: 15-jarigen).

Een eerste, heel belangrijke opmerking daarbij is dat de uitkomstvariabelen dus gebaseerd waren op een mix van scores uit het basis- en secundair onderwijs. Een dergelijke mix is nochtans zeker niet de beste manier om de effecten van de structuur van het *secundair* onderwijs op de uitkomsten te meten. Ter verdediging van de auteurs merken we hier op dat hun artikel niet specifiek over het effect

van tracking ging, maar dat ze op basis van één serie uitkomstvariabelen het effect van verschillende kenmerken van het onderwijssysteem probeerden te schatten (zij het in afzonderlijke analyses, dus zonder bij de studie van één kenmerk te controleren voor de effecten van een ander). Rindermann en Ceci gaan bv. ook na hoe de gemiddelde prestaties afhangen van de mate waarin een land in onderwijs investeert. Voor dit soort verbanden is het beter te verdedigen om "prestaties" te definiëren in termen van een gemiddelde score over het basis- en secundair onderwijs. Voor de studie van het effect van tracking is deze maat echter minder geschikt.

Omdat Rindermann en Ceci enkel de geaggregeerde data ter beschikking stelden, zijn we genoodzaakt verder te werken met deze beperking. In die zin kan ook onze heranalyse *op zichzelf* geen voldoende bewijs leveren voor het ontbreken van een effect van tracking op de gemiddelde prestaties. We zullen wél laten zien dat de door Rindermann en Ceci gevonden effecten foutief waren, en dat hun artikel dus niet kan worden ingeroepen om de globale consensus uit de literatuur in vraag te stellen. Eerder onderzoek waarin slechts één - in het secundair - afgenomen test werd bestudeerd (zie bv. Horn, 2009; Dupriez, Dumay & Vause, 2008; Duru-Bellat & Suchaut, 2005), gaf immers eensluidend aan dat een vroege tracking geen positieve effecten had op de gemiddelde prestaties.

De belangrijkste reden waarom Rindermann en Ceci met geaggregeerde data werkten, was dat ze hierdoor een zeer ruime dataset konden samen stellen (met in totaal 58 landen die aan minstens 1 van de tests uit hun dataverzameling deelnamen). De op die manier geconstrueerde dataset is echter erg heterogeen. Dit kan eenvoudig worden geïllustreerd door de ruwe scores te plotten in functie van de leeftijd van tracking en daarbij aan te duiden of het land behoort tot de elite van meest welvarende landen of niet (hier op basis van OESO-lidmaatschap). Figuur 1 maakt duidelijk hoe belangrijk de context is: OESO-leden scoren beduidend beter dan niet-leden. Een tweede vaststelling is dat de niet-OESO-leden over het algemeen ook erg late trackers zijn. Zowel de afhankelijke als de onafhankelijke variabele correleren dus met de context: dit is een typevoorbeeld van een situatie waarbij de statistische alarmlampen op rood gaan staan. Het betekent dat de context, wanneer ze niet goed onder controle



Figuur 1. Ruwe scores (gestandaardiseerd) in functie van de leeftijd van tracking

wordt gehouden, de schatting van het effect van tracking danig zal verstoren. Een goede controle is dus uitermate belangrijk.
Rindermann en Ceci erkenden uitdrukkelijk dat niet-geobserveerde verschillen tussen landen de schatting van het effect van tracking kunnen vertroebelen. Ze gaven daarbij ook een omstandig overzicht van allerlei mogelijk relevante verschillen. Uiteindelijk controleerden ze echter slechts voor drie contextvariabelen: welvaart (BBP/capita), het onderwijsniveau van de samenleving (op basis van het percentage ongeletterden, het gemiddeld aantal scholingsjaren en het aantal mensen zonder een diploma secundair onderwijs) en moderniteit (op basis van het persoonlijke oordeel van vier onderzoekers m.b.t. criteria zoals respect voor de mensenrechten, een gelijke behandeling van vrouwen, en een democratische staatsordening).

Controlerend voor deze drie contextvariabelen berekenden ze vervolgens de correlatiecoëfficiënten tussen de gemiddelde scores en de leeftijd van tracking. Op basis hiervan concludeerden ze dat een vroege tracking een positief effect had op de kwaliteit van het onderwijs. We repliceren deze vaststelling in het regressiemodel in Tabel 1, waarbij de prestaties worden verklaard op basis van de drie contextvariabelen en de leeftijd van tracking (gestandaardiseerde coëfficiënten). Een late tracking hangt in dit model inderdaad significant negatief samen met de prestaties.

Tabel 1
Replicatie oorspronkelijk model
(afh. variabele: gemiddelde prestaties)

Parameter	Est.	Sign.
(Intercept)	0.00	
Scholingsgraad	0.54	***
BBP/capita	-0.08	
Moderniteit	0.36	**
Leeftijd van tracking	-0.27	***

*** $p \leq 0.01$ ** $p \leq 0.05$ * $p \leq 0.10$

4 Hoe adequaat is de controle voor de context?

Maar hoe goed houden de contextvariabelen de verschillen tussen de landen nu onder controle? Tabel 1 wijst alvast op een belangrijk probleem: het gemodelleerde effect van BBP op de prestaties blijkt negatief te zijn. Dat is vervelend, want er is geen enkele reden denkbaar waarom welvaart de prestaties negatief zou beïnvloeden. In de praktijk betekent het dat van de drie contextcontroles er maximaal twee echt werken zoals verondersteld.

Een blik op de residuen - wat er overblijft van de ruwe scores na controle voor de context, m.a.w. wat nog verklaard moet worden door de leeftijd van tracking - insinueert bovendien dat de controle voor de context onvolledig was. Het duidelijkst is dit voor Zuid-Afrika. In figuur 1 was te zien dat dit land erg slecht scoorde. Slagen de drie contextvariabelen er nu in om hier iets van te verklaren? De residuen suggereren van niet: terwijl de standaarddeviatie op de residuen 0.67 punten bedraagt, bedraagt het negatieve residu van Zuid-Afrika nog steeds -3.04 punten. Zuid-Afrika blijft dus een extreme outlier, ook na controle voor de context. De slechte onderwijsprestaties in dit land zijn inderdaad al veel langer bekend, en de belangrijkste oorzaak ook: het is een erfenis van de Apartheid. De drie contextvariabelen zijn duidelijk niet adequaat genoeg om hier voor te corrigeren. Merk op dat Zuid-Afrika bij de zeer late trackers hoort (16 jaar). De – door de contextvariabelen onverklaarde - negatieve prestaties van dit land worden door het model dus onterecht aan die late leeftijd van tracking toegeschreven.

Zuid-Afrika is ongetwijfeld een apart geval, gezien zijn specifieke geschiedenis. Wat dit voorbeeld echter wel duidelijk maakt, is dat het te eenvoudig is om te veronderstellen dat verschillen in een zo heterogene dataset voldoende worden opgevangen door te controleren voor slechts een zeer beperkt aantal contextvariabelen.

5 Heeft tracking overal hetzelfde effect?

Los van de vorige vaststelling is het belangrijk dat Rindermann en Ceci enkel de uitkomsten zélf controleerden voor verschillen in de context. Het *effect* van tracking op die uitkomsten werd verondersteld overal hetzelfde te zijn, onafhankelijk van de context: er werden geen interacties gemodelleerd tussen context en tracking. Dit betekent dat ze aannamen dat de invloed van de context volledig kon gescheiden worden van het effect van tracking: ze maten het effect van het niveau van economische of culturele ontwikkeling op de gemiddelde prestaties, ze maten het effect van de leeftijd van tracking op de prestaties, maar ze gingen niet na hoe dit laatste effect mogelijk zelf afhing van het niveau van ontwikkeling.

Dit lijkt om verschillende redenen een discutabele aanname. Zoals we hoger zagen, kan een latere tracking zowel positieve als negatieve effecten hebben. Welke van deze effecten de bovenhand neemt, hangt mee af van de omstandigheden waaronder het comprehensief systeem moet werken. Een voorbeeld: comprehensief onderwijs veronderstelt een meer gedifferentieerde manier van lesgeven. Om zoiets goed te kunnen laten werken moeten leerkrachten goed zijn opgeleid, mogen de klassen niet overdreven groot zijn, moet het belang van schools onderwijs door iedereen wordt erkend, enz. In ontwikkelingslanden is aan deze voorwaarden veel minder voldaan dan bij ons. Klassen zijn er bijvoorbeeld vaak dubbel zo groot als in het Westen. Als er eerst aan een aantal randvoorwaarden moet worden voldaan vooraleer comprehensief onderwijs succesvol kan zijn, dan lijkt het moeilijk verdedigbaar dat de leeftijd van tracking overal ter wereld gelijke effecten zou hebben. Merk trouwens op dat eerder onderzoek al heeft aangetoond dat ook het effect van tracking op gelijke kansen een andere vorm aanneemt in niet-Westerse landen, zie bv. Broaded (1997) of Buchmann en Hannum (2001).

Rindermann en Ceci gaven zelf inderdaad aan dat het effect van tracking niet los kan worden gezien van de context: *"Systems with later tracking (…) can also be very successful under favorable conditions. (…) Tracking is not necessary to achieve high competence in countries in which at risk students avoid the development of problematic school careers through support by special teachers, additional instruction (…) The success of educational systems is not independent of the larger matrix of social, political, and cultural conditions, and there appears to be several paths to achieve successful outcomes."* Helaas werden dit soort interactie-effecten in het oorspronkelijke artikel niet expliciet gemodelleerd, waardoor de impact ervan op de vastgestelde verbanden niet duidelijk werd gemaakt.

6 Verbeteringen aan het oorspronkelijke artikel

Tabel 2 geeft een overzicht van de resultaten van onze eigen regressiemodellen, waarin we wel interactie-effecten opnamen.

Een eerste belangrijke conclusie is dat het negatieve effect van een late tracking steeds afneemt wanneer de scholingsgraad, het BBP of de moderniteit toeneemt (model 1-3). Dit is volledig in lijn met wat we verwachtten. Voor de scholingsgraad van de volwassen bevolking, de contextvariabele die het belangrijkst was bij het begrijpen van prestaties (grootste hoofdeffect), is dit interactie-effect ook significant. Merk op dat het om gestandaardiseerde coëfficiënten gaat, waarbij de standaardisering werd uitgevoerd t.o.v. het gemiddelde over de mondiale dataset. De scholingsgraad in de meest ontwikkelde Westerse landen ligt typisch in de orde van 1 standaarddeviatie hoger dan het mondiale gemiddelde. Bij een dergelijk niveau van ontwikkeling kan het effect van tracking als praktisch onbestaande worden beschouwd (-0.32+0.23=-0.09).

Hoger suggereerden we een aantal mogelijke verklaringen voor een negatief effect van late tracking in ontwikkelingslanden. Eén ervan was dat de klassen in die landen vaak erg groot zijn, wat lesgeven aan een heterogeen publiek fel bemoeilijkt. Zonder hiermee een definitieve uitspraak te willen doen over het precieze kanaal achter de vastgestelde interactie-effecten, namen we in model 4 een

Tabel 2
Modellen met interacties tussen leeftijd van tracking en een contextvariabele (afh. variabele; gemiddelde prestaties)

	Model 1: Scholings-graad		Model 2: BBP/capita		Model 3: Moderniteit		Model 4: Klasgrootte		Model 5: Lidm. OESO	
(Intercept)	-0.01		0.01		0.00		-0.01		0.09	
Scholingsgraad	0.49	***	0.52	***	0.52	***	0.50	***	0.46	***
BBP/capita	-0.13		-0.03		-0.06		-0.11		-0.15	
Moderniteit	0.40	**	0.34	*	0.35	**	0.34	**	0.33	**
Klasgrootte							-0.13			
Lidm. OESO (ref. = leden)									-0.13	
Leeftijd van tracking	-0.32	***	-0.29	***	-0.33	***	-0.30	***	-0.10	
Tracking * Scholingsgraad	0.23	*								
Tracking * BBP/capita			0.09							
Tracking *Moderniteit					0.16					
Tracking * Klasgrootte							-0.22	**		
Tracking * Lidm. OESO									-0.94	***

*** $p \leq 0.01$ ** $p \leq 0.05$ * $p \leq 0.10$

interactie met klasgrootte op. De significant negatieve interactieterm in model 4 suggereert inderdaad dat late tracking vooral negatief is voor prestaties wanneer de klasgrootte hoog is. Merk opnieuw op dat de gecentreerde waarden verwijzen naar de afwijking van een mondiaal gemiddelde klasgrootte, d.w.z. in de welvarende Westerse landen is de gecentreerde waarde sterk negatief (bv. België: -1.3). Met de in Westen courante klasgroottes is het effect van de leeftijd van tracking dus opnieuw verwaarloosbaar.

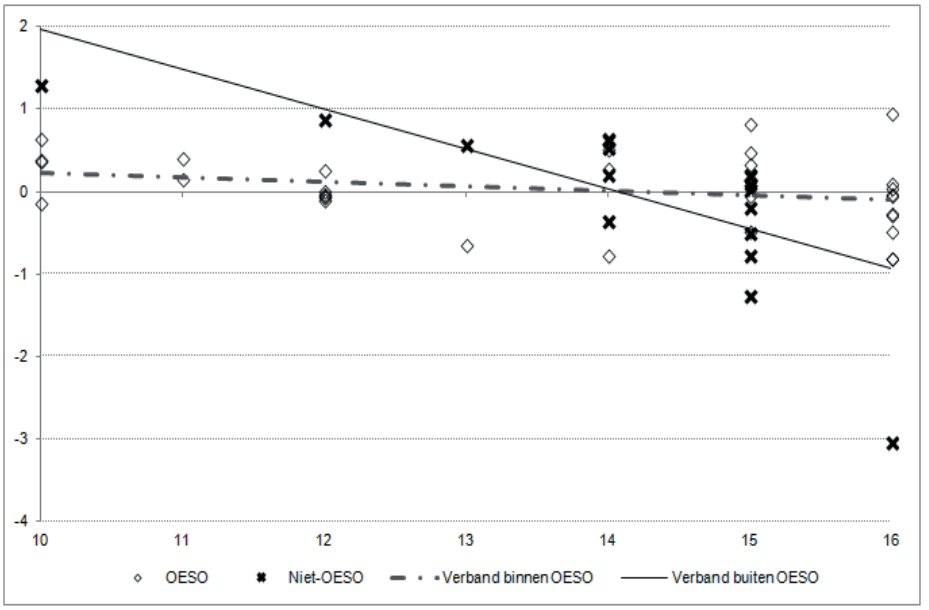
Er zijn evenwel nog twee problemen met dit soort specificaties. Ten eerste werd telkens maar één interactie-effect tegelijkertijd gemodelleerd, waardoor de impact van de context op het effect van tracking mogelijk niet volledig wordt doorzien. Verschillende interacties samen in één model opnemen is in principe mogelijk, maar minder robuust wegens een gebrek aan vrijheidsgraden. Ten tweede veronderstellen de interactie-effecten dat de impact van de context op het effect van tracking lineair is. Een lineair effect is misschien onrealistisch: de redenering uit de vorige paragraaf suggereert bv. dat comprehensief onderwijs goed kan werken wanneer aan bepaalde minimumvoorwaarden is voldaan, d.w.z. vanaf dat een bepaald ontwikkelingsniveau is overschreden. De impact van de context op het effect van tracking hoeft dan niet lineair te zijn.

Om die redenen werd ook een interactie gemodelleerd tussen de leeftijd van tracking en het lidmaatschap van de OESO (model 5, referentiegroep: OESO-leden). OESO-lidmaatschap dient hier als een algemene indicatie voor het horen bij de welvarende Westerse elite; we veronderstellen dat de randvoorwaarden voor succesvol comprehensief onderwijs in landen met een dergelijk hoog ontwikkelingsniveau vervuld zijn. De laatste kolom in tabel 2 laat zien dat er binnen de OESO inderdaad geen betekenisvol effect van tracking meer vast te stellen is. Dit effect is trouwens nog duidelijker in de andere globale maat die Rindermann en Ceci gebruikten, CCS (definitie: zie noot). Met deze maat als uitkomstvariabele is het effect van tracking op de cognitieve prestaties binnen de OESO gelijk aan nul (0.03, $p = 0.84$).

Verder stellen we opnieuw een significante interactie tussen OESO-lidmaatschap en tracking vast. Dit bewijst opnieuw dat er zeer

belangrijke verschillen bestaan in het effect van tracking afhankelijk van de context.

Deze conclusies worden tot slot grafisch geïllustreerd in Figuur 2. Hierin zijn de scores - ná controle voor de context - uitgezet t.o.v. de leeftijd van tracking. Het niet opnemen van interactie-effecten, zoals Rindermann en Ceci nalieten te doen, komt neer op de veronderstelling dat er één verband bestaat dat van toepassing is op alle punten, onafhankelijk van OESO-lidmaatschap. De figuur illustreert echter dat de werkelijke verbanden duidelijk verschillend zijn voor de groep van de OESO- resp. niet-OESO-leden, en dat de opname van interactie-effecten dus nodig is. Binnen de OESO is er bovendien geen effect van de leeftijd van tracking op de prestaties.



Figuur 2. Verbanden tussen residuele scores (na controle voor de context) en de leeftijd van tracking

7 Conclusie

De verdienste van Rindermann en Ceci (2009) is dat ze geprobeerd hebben om het blikveld uit te breiden tot de minder ontwikkelde landen. Cruciaal bij een dergelijke aanpak is echter dat eventuele verstorende verschillen accuraat onder controle worden gehouden. Onze analyse laat zien dat dit helaas niet voldoende gebeurd is. Ten eerste lijken de drie controlevariabelen niet alle relevante verschillen te hebben gevat. Belangrijker nog is dat onze analyse laat zien dat het belangrijk is om ook interacties tussen de context en het effect van tracking te modelleren.

In het bijzonder demonstreert onze analyse dat binnen de OESO de leeftijd van tracking geen duidelijk effect had op de gemiddelde prestaties. Het in het oorspronkelijke artikel gerapporteerde negatieve effect van een late opsplitsing van leerlingen werd volledig veroorzaakt door het bestaan van een dergelijk effect buiten de OESO.

Een afzonderlijk aandachtspunt is dat Rindermann en Ceci een aggregaat van scores uit het basis – en het secundair onderwijs als uitkomstvariabelen hebben gebruikt. Dit soort aggregaten is eigenlijk niet zo geschikt voor de studie van het effect van de structuur van het secundair onderwijs op de prestaties. Ook onze heranalyse blijft uiteraard onderhevig

aan deze beperking. Daarom is het belangrijk om onze conclusie te kaderen in de volledige literatuur rond de gevolgen van tracking op gemiddelde prestaties. Analyses van scholierentests afgenomen in het secundair onderwijs in welvarende landen laten immers keer op keer zien dat een vroege opsplitsing geen duidelijke voordelen heeft voor de gemiddelde prestaties (zie Van de Werfhorst, 2011 voor een overzicht). Dat de vaststellingen van Rindermann en Ceci (2009) van deze consensus afweken, werd dus veroorzaakt door de onvoldoende controle voor de heterogeniteit en door het samennemen van tests in basis- en secundaire scholen.

Noot

1 De in het oorspronkelijke artikel gebruikte uitkomstvariabelen waren: “PISA” (de optelsom van de scores in PISA2000 and PISA2003), “Grade” (de optelsom van TIMSS1994, TIMSS1999, TIMSS2003 en PIRLS2001 – deze variabele ontbrak in de door Rindermann en Ceci ter beschikking gestelde dataset), “SASS” (de optelsom van alle hoger genoemde tests, plus IEA1991) en “CCS”: de optelsom van alle hoger genoemde tests, plus een “nationale IQ score” op basis van een databank uit de literatuur. We gebruiken verder in deze tekst “SASS” als uitkomstvariabele, omdat in het oorspronkelijke artikel het effect van tracking het duidelijkst was voor deze variabele. De analyses voor de andere uitkomstvariabelen zijn vergelijkbaar.

Literatuur

Broaded, C. M. (1997). The limits and possibilities of tracking: Some evidence from Taiwan. *Sociology of Education, 70*(1), 36-53.

Buchmann, C. & Hannum, E. (2001). Education and stratification in developing countries: A review of theories and research. *Annual review of sociology*, 27, 77 - 102.

Dupriez, V., Dumay, X., & Vause, A. (2008). How Do School Systems Manage Pupils’ Heterogeneity?, *Comparative Education Review, 52*(2), 245 - 273.

Duru-Bellat, M. & Suchaut, B. (2005). Organisation and Context, Efficiency and Equity of Educational Systems: what PISA tells us. *European Educational Research Journal, 4*(3), 181-194.

Duyck, W. & Anseel, F. (2012). Gelijke Kansen, Gelijke Kinderen, Gelijke Klassen? Early Tracking in het Onderwijs. *Itinera Institute Discussion Papers*, 2012/4.

Hanushek, E. A. & Woessmann, L. (2006). Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *Economic Journal*, 116, C63 - C76.

Horn, D. (2009), Age of selection counts: a cross-country analysis of educational institutions. *Educational research and evaluation, 15* (4), 343 - 366.

Jakubowski, M. (2010). The Impact of the 1999 Education Reform in Poland. *Policy Research Working Paper Series - The World Bank,* 5236.

Lavrijsen, J. (2013). Characteristics of educational systems. How they influence outcomes in the short and the long run. Steunpunt SSL, publicatienr. SSL/2012.04/1.1.1.

Lavrijsen, J., Nicaise I. & Wouters T. (2013). Vroege tracking, kwaliteit en rechtvaardigheid. Wat het wetenschappelijk onderzoek ons leert over de hervorming van het secundair onderwijs. *HIVA Working Paper*, KU Leuven, 2031.

Luyten, H. & Bosker, R. (2012). Naar een hervorming van het Vlaams secundair onderwijs: Evaluatieve bemerkingen ex ante vanuit Nederlands perspectief. *Pedagogische Studiën, 89*(5), 317-326.

OESO (2012). Equity and Quality in Education. Paris, 2012.

Rindermann, H. & Ceci, S. J. (2009). Educational Policy and Country Outcomes in International Cognitive Competence Studies. *Perspectives on Psychological Science, 4*(6), 551 - 568.

Van de Werfhorst, H. (2011). Selectie en differentiatie in het Nederlandse onderwijsbestel. Gelijkheid, burgerschap en onderwijsexpansie in vergelijkend perspectief. *Pedagogische Studiën*, 2011 (88), 283-297

Van de Werhorst, H. & Mijs, J. (2010). Achievement inequality and the institutional structure of educational systems: A comparative perspective, *Annual Review of Sociology*, 36, p. 407 - 428.

Manuscript aanvaard op: 17 februari 2014

Auteurs

Jeroen Lavrijsen en **Ides Nicaise** zijn verbonden aan het HIVA van de KU Leuven.

Correspondentieadres:
jeroen.lavrijsen@kuleuven.be

Abstract

Comprehensive education: a threat to quality? A reanalysis of Rindermann and Ceci (2009)

In general, educational research has concluded that early tracking does not lead to a better average performance. However, a cross-country comparison by Rindermann and Ceci (2009) did find a positive effect of early tracking on performance. How can we understand this contradiction? In this paper we show that Rindermann and Ceci did not adequately control for confounding differences between countries in their very heterogeneous dataset. First, their three background variables prove to have been insufficient to control all relevant differences. Secondly, Rindermann and Ceci assumed that tracking had the same effect in all countries, independently from other national system characteristics. By explicitly accounting for interactions, we show that this assumption is not valid. Our reanalysis demonstrates that in developed countries early tracking does not have a significant effect on performance. This untangles the contradiction and confirms the general message from the literature.