Bruikbaarheid van internationaal vergelijkende surveys naar leerprestaties

H. Luyten, R. Janssen en S. Karsten

De laatste jaren nemen de resultaten van internationaal vergelijkende surveys een steeds grotere plaats in het debat over de staat van ons onderwijs. Systematische vergelijking van de prestaties van leerlingen begon al in de jaren zestig van de vorige eeuw. Sinds de jaren negentig zijn die internationale prestatiemetingen geïnstitutionaliseerd in twee verschillende organisaties. De eerste is de International Association for the Evaluation of Educational Achievement (IEA), een organisatie opgezet door vertegenwoordigers van nationale regeringen en onderzoekers. Deze organisatie is bekend door zijn, om de vier jaar herhaalde, studies naar wiskunde en exacte vakken (Trends in International Mathematics and Science Study of kortweg TIMMS) en de, om de vijf jaar herhaalde, toetsen van leesvaardigheid (Progress in International Reading Literacy Study; PIRLS). De toetsen voor TIMMS en PIRLS worden afgenomen in groep 6 van de basisschool (internationaal 4th grade) en in geval van TIMMS ook in de tweede klas van het voortgezet onderwijs (internationaal 8th grade).

De tweede organisatie is de Organisatie voor Economische Samenwerking en Ontwikkeling (OESO). Dit is een samenwerkingsverband van de rijke industriële landen die onder meer om de drie jaar de taalvaardigheid, rekenkundige vaardigheid en vaardigheid op het gebied van exacte vakken (Programme for International Student Assessment, PISA) van leerlingen in de deelnemende landen toetst. PISA gebruikt een leeftijdsgebonden steekproeftrekking, namelijk vijftienjarigen. De filosofie achter en reikwijdte van toetsen van deze organisaties verschillen; TIMMS en PIRLS proberen het werkelijk onderwezen curriculum te toetsen, terwijl PISA zoiets als competenties probeert te meten. Desalniettemin vertonen de resultaten vrij grote samenhang. De politieke impact echter is verschillend. De uitkomsten van PISA kunnen in alle deelnemende landen op grote politieke en mediabelangstelling rekenen en hebben in enkele gevallen een ware schokgolf veroorzaakt. Alle reden om een balans op te maken.

Wat kunnen deze internationaal vergelijkende surveys ons vertellen over de kwaliteit van het onderwijs in Nederland of Vlaanderen en in de rest van de wereld? In de hierna volgende discussiebijdragen wordt nader ingegaan op deze vraag. In elke bijdrage worden verschillende accenten gelegd, maar een terugkerend thema is steeds het verschil tussen de manier waarop in de politiek-maatschappelijke arena wordt omgegaan met de onderzoeksresultaten tegenover de discussie in wetenschappelijke kring.

Zodra nieuwe bevindingen van internationaal vergelijkende studies bekend worden, kunnen beleidsvoerders en journalisten zelden de neiging onderdrukken om er direct vergaande conclusies en beleidsimplicaties aan te verbinden. In wetenschappelijke kring is men doorgaans veel terughoudender. Uit de verschillende bijdragen blijkt steeds weer dat wetenschappers veel meer voorzichtigheid in acht nemen voordat men verstrekkende conclusies verbindt aan de uitkomsten.

Politici en journalisten willen een daling op de internationale ranglijsten al snel opvatten als “bewijs” voor de noodzaak tot ingrijpende maatregelen. Vanuit wetenschappelijke kring wordt juist gewezen op het gegeven dat lang niet alle veranderingen en variaties in leerprestaties zonder meer kunnen worden toegeschreven aan de kwaliteit van het onderwijs op school. Ook buitenschoolse factoren zijn immers van invloed op het behaalde niveau. Daarnaast is het goed mogelijk om op de internationale ranglijst te dalen zelfs als de leerprestaties zijn verbeterd. Als andere landen nog meer winst weten te boeken kan een verbetering van het nationaal gemiddelde immers toch samengaan met een daling op de internationale ranglijst. Verder spelen ook allerlei andere factoren een rol die de verge-

lijkbaarheid van toetsscores tussen diverse landen bemoeilijken. De bijdrage van Van Rijn, Kordes en Gille gaat nader in op een aantal meer methodologische punten in dit verband (populatiedefinitie, meetinvariantie en vergelijkbaarheid van contextvariabelen).

Het concept *gemotiveerd scepticisme*, dat in de bijdrage van Karsten naar voren wordt gebracht, geeft een treffende karakterisering van de manier waarop in Nederland wordt omgegaan met de uitkomsten van studies als TIMSS, PISA en PIRLS. Negatieve bevindingen worden zonder veel kritiek voor waar aangenomen, maar positieve resultaten worden genegeerd of in twijfel getrokken. Informatie die de indruk lijkt te bevestigen dat het bijzonder slecht gesteld is met de kwaliteit van het Nederlandse onderwijs, past in het bestaande beeld en wordt gemakkelijk geaccepteerd. Een merkwaardige consequentie is dat het ene jaar uitkomsten die wijzen op een hoge internationale positie van Nederland niet serieus genomen worden, maar een aantal jaren later toch zonder meer geaccepteerd worden, zij het op impliciete wijze. Wanneer de scores in een bepaald jaar lager uitvallen dan voorheen, wordt door niemand betwijfeld dat er sprake is van een dalende trend. Deze conclusie kan natuurlijk alleen correct zijn als de hoge scores op de eerdere metingen wel degelijk klopten.

Dronkers besteedt in zijn bijdrage veel aandacht aan het gevaar van politieke beïnvloeding bij onderzoek waarvan de uitkomsten grote politieke en maatschappelijke consequenties kunnen hebben. Zowel bij de dataverzameling als bij analyse en rapportage ligt dit gevaar op de loer. Bij de dataverzameling kan politieke beïnvloeding ertoe leiden dat besloten wordt bepaalde informatie (zoals land van herkomst van de leerlingen) niet te verzamelen. Ook bij het analyseren en rapporteren is het mogelijk dat bepaalde gevoelige uitkomsten onderbelicht blijven. De risico's van politieke beïnvloeding worden echter in het geval PISA, TIMSS en PIRLS in belangrijke mate geneutraliseerd door het feit dat de datasets kort na het verschijnen van de eerste rapportages via het internet beschikbaar worden gesteld. Zodoende is het ook voor kritische geesten die er weinig moeite mee hebben om politiek minder correcte uitkomsten te rapporteren, mogelijk om hun eigen analyses uit te voeren en hierover te publiceren.

Abstract

The usefulness of cross-national surveys on student achievement

In recent years findings from cross-national surveys like PISA, TIMSS and PIRLS have become ever more prominent in the public debate on the quality of national education systems. The question is: what do these surveys tell us about the quality of education in the Netherlands, Flanders and the rest of the world? This is the main question to be addressed in the contributions that follow. Each contribution addresses different aspects of the basic question, but a recurring theme is the difference in dealing with research findings in the political and general public arena versus the discussion among scholars and researchers.

Internationaal vergelijkende studies van leerprestaties

P. W. van Rijn, J. H. Kordes en E. Gille

1 Doelstellingen van peilingonderzoek

Het internationaal vergelijken van leerprestaties bij taal (lezen), rekenen (wiskunde) en natuurwetenschappen is een complexe wetenschappelijke bezigheid. De doelen van zulk onderzoek zijn niet dezelfde als de doelen van een nationaal peilingonderzoek. Hoewel beide vormen van onderzoek informatie geven over vigerende onderwijssystemen, levert een nationaal opgezet onderzoek meer specifieke informatie over een specifiek onderwijssysteem op. Daarom zijn er vanuit een nationaal perspectief vragen te stellen over de relevantie (voor het onderwijsveld) van vergelijkende studies zoals PISA (Programme for International Student Assessment), TIMSS (Trends in International Mathematics and Science Study) en PIRLS (Progress in International Reading Literacy Study). Afgaande op de berichtgeving in de media, constateren we echter dat internationale ranglijstjes van leerprestaties meer impact hebben dan bevindingen in, bijvoorbeeld, typisch nationaal georiënteerde verslagen van de Periodieke Peiling van het Onderwijsniveau (PPON, Van der Schoot, 2008) of rapporten en overzichten van de Inspectie van het Onderwijs over de opbrengsten van het voortgezet- en het basisonderwijs[1].

PPON heeft directe relevantie voor het Nederlandse onderwijs. Het is in 1985 opgezet om inzichten te verkrijgen in het leeraanbod en de leeropbrengsten van het basisonderwijs. Het moet onder meer een empirische basis verschaffen voor de maatschappelijke onderwijsdiscussie. PISA, PIRLS en TIMSS zijn natuurlijk niet specifiek gericht op het Nederlandse onderwijsstelsel. Een opvallend resultaat – positief of negatief – in deze studies wil dus niet automatisch zeggen dat er iets aan de hand is met het onderwijs in het betreffende land. Een lagere plaats van Nederland op de internationale ranglijst voor wiskunde kan onder meer betekenen dat de onderwerpen die op de Nederlandse scholen behandeld worden, verder afstaan van het soort onderwerpen dat in genoemde studies getoetst wordt.

In deze discussiebijdrage wordt een aantal methodologische kwesties van internationaal vergelijkende studies besproken om de complexiteit, de mogelijkheden en de beperkingen van dergelijk onderzoek voor het voetlicht te brengen. Een drietal zaken komt achtereenvolgens aan de orde: de steekproef, meetinvariantie en contextvariabelen.

2 Steekproef

Om een representatieve steekproef voor een internationale populatie van leerlingen te verkrijgen is het noodzakelijk om een steekproefkader op te stellen waarin de doelpopulatie nauwgezet is afgebakend. Bij PISA bijvoorbeeld bestaat de doelpopulatie uit 15-jarige schoolgaande leerlingen. Maar wat wordt precies verstaan onder 15-jarige schoolgaande leerlingen? Vijftienjarige schoolgaande leerlingen worden volgens een steekproef getrokken voor deelname aan het onderzoek. Meningen kunnen verschillen over de mate waarin de steekproef representatief is voor alle vijftienjarige leerlingen in Nederland en in andere landen (OECD, 2009, p. 64). In PISA 2000 werden bijvoorbeeld in Nederland de leerlingen in het voorgezet speciaal onderwijs (vso) niet tot de populatie gerekend. Het vso hoorde toen bij het basisonderwijs en de leerlingen in het vso werden daardoor geacht niet tot de PISA-populatie te behoren. Kort daarna was er een stelselwijziging en behoorden deze leerlingen wel bij de populatie. Zij zaten nu in het svo (speciaal voortgezet onderwijs), dat bij het voortgezet onderwijs hoorde en ze maakten daardoor deel uit van de doelpopulatie. Ook wordt in deze studies een onderscheid gemaakt tussen de nationale en internationale doelpopulatie. Een verschil kan ontstaan, doordat er bijvoorbeeld meerdere talen in een land worden gesproken of dat het niet mogelijk is om in bepaalde geografische gebieden de toetsen af te nemen (IEA, 2008, p. 79), waardoor niet alle leerlingen die tot de populatie behoren, meedoen. De hierboven genoemde voorbeel-

den vormen een illustratie van de moeilijkheden bij het definiëren van een internationaal goed vergelijkbare doelpopulatie. Dit bemoeilijkt de interpretatie van de resultaten, omdat verschillen in leerprestaties op meerdere manieren zijn uit te leggen.

Een probleem met een populatiedefinitie op basis van leeftijd is dat niet iedere vijftienjarige leerling in de wereld evenveel onderwijs heeft genoten. In de Verenigde Staten zijn de zomervakanties bijvoorbeeld langer dan in Nederland. Een simpel rekenvoorbeeld: 11 weken vrij (grofweg in Nederland) of 13 weken vrij (grofweg in de VS) op jaarbasis levert na 10 jaar een verschil van 20 weken op, oftewel ongeveer een halfjaar onderwijs. Er is veel onderzoek gedaan naar de samenhang tussen de duur van zomervakanties en leerprestaties. In Amerikaans onderzoek is bijvoorbeeld gevonden dat toetsscores vlak na een zomervakantie aanzienlijk lager zijn dan vlak ervoor (Cooper, Nye, Charlton, Lindsay, & Greathouse, 1996). Een bijkomend probleem in grootschalig internationaal onderzoek is dat de zomervakanties op het noordelijke en zuidelijke halfrond niet in dezelfde maanden vallen.

Het is vaak niet goed mogelijk om over interessante deelpopulaties voldoende informatie te krijgen, omdat hiervan gewoonweg te weinig leerlingen in de steekproef terecht komen. Een oplossing hiervoor is om bewust wat meer leerlingen te trekken uit dergelijke deelpopulaties (zogeheten *oversampling*). Bijvoorbeeld, het uitsplitsen van de resultaten naar allochtone deelpopulaties is vaak interessant, maar niet goed mogelijk omdat dit slechts kleine groepen betreft.

In veel nationaal en internationaal onderwijsonderzoek wordt een steekproefdesign gebruikt dat bestaat uit twee fasen (zie bijv. Cochran, 1977). In de eerste fase wordt een steekproef van scholen getrokken waarbij doorgaans rekening wordt gehouden met de grootte van de school. In de tweede fase worden vervolgens de leerlingen aselect getrokken. Hierin verschilt PISA echter van TIMSS en PIRLS: gegeven dat een bepaalde school in de steekproef is opgenomen, wordt in PISA aselect een vast aantal leerlingen getrokken, terwijl bij TIMSS en PIRLS gehele schoolklassen willekeurig worden gekozen. In beide gevallen betekent het echter dat leerlingen in de steekproef die van verschillende scholen komen, anders moeten worden gewogen. Als dit niet wordt gedaan, dan zijn kleinere scholen oververtegenwoordigd in de steekproef. Dit kan worden geïllustreerd met een eenvoudig voorbeeld. Stel de populatie bestaat uit twee scholen: een school met 100 leerlingen en een school met 1.000 leerlingen. Als er nu 10 leerlingen van elk van beide scholen aselect worden getrokken, dan zitten er relatief veel meer leerlingen van de kleine school in de steekproef. Een leerling in de steekproef van de kleine school vertegenwoordigt namelijk 10 leerlingen en een leerling in de steekproef van de grote school vertegenwoordigd 100 leerlingen. Om ervoor te zorgen dat elke leerling een evenredig deel van de populatie vertegenwoordigt, kunnen de leerlingen in de steekproef van verschillende scholen verschillend worden gewogen. Nu zijn er veel manieren om deze weging te bewerkstelligen en zijn de methoden hiervoor steeds in ontwikkeling. Zo is de huidige wegingstechniek bij PISA niet meer exact dezelfde als in het begin.

Een meer praktisch probleem bij de steekproeftrekking voor PISA 2006, PIRLS 2006 en TIMSS 2007 was dat in sommige landen de onderzoeken gelijktijdig plaatsvonden. Om de belasting van scholen te verminderen, is er voor een aantal landen een controle voor eventuele overlap bij het trekken van de steekproef uitgevoerd (OECD, 2009, p. 76).

3 Meetinvariantie

Een enigszins onderbelicht thema in de rapportages van PISA, TIMSS en PIRLS is het concept van meetinvariantie (Meredith, 1993). In alle drie de onderzoeken wordt gebruik gemaakt van Item Respons Theorie (IRT) om de resultaten te analysen en te rapporteren. Kort samengevat behelst het gebruik van IRT in deze onderzoeken het opstellen van een passend statistisch model om de antwoorden van de leerlingen te beschrijven om vervolgens een meetschaal te maken waarop de resultaten van de landen kunnen worden gerapporteerd (Birnbaum, 1968; Rasch, 1960). Meetinvariantie heeft in het

algemeen betrekking op de vraag of voor verschillende subpopulaties hetzelfde statistisch model gebruikt kan worden. Met andere woorden, kunnen de antwoorden van de leerlingen uit de verschillende landen met hetzelfde statistisch model worden beschreven? Dit kan op veel verschillende manieren worden onderzocht.

Vaak wordt begonnen met onderzoek naar Differentieel Item Functioneren (DIF; Mellenbergh, 1989). Er kan dus worden bekeken of er items zijn aan te wijzen die zich anders gedragen in verschillende landen, waarbij wordt gecorrigeerd voor eventuele verschillen in vaardigheid tussen de landen. Dit principe kan ook worden gegeneraliseerd naar het functioneren van een hele toets (Raju, Van der Linden & Fleer, 1995). Bij zowel PISA als PIRLS en TIMSS worden vooraf alle items gescreend op mogelijke bias en vinden er pretests plaats om niet goed werkende items te detecteren (DIF). Deze opgaven worden dan niet meer gebruikt in het uiteindelijke onderzoek. Een serie opgaven die niet in aanmerking kwam voor het hoofdonderzoek van PISA had als onderwerp isolatie. Het maakt nogal uit of je in een land als Finland woont, of een land als Qatar. De leerlingen in Finland zullen bij raamisolatie denken aan dubbele beglazing. De leerlingen in Qatar denken aan gordijnen en andere zonneschermen. Met een item over fietsen zullen Nederlandse leerlingen beter overweg kunnen dan leerlingen in de VS. In de technische rapporten van deze projecten is dan ook veel te vinden over bijvoorbeeld hoe is omgegaan met het vertalen van de opgaven en het identificeren van potentiële bronnen van bias bij de richtlijnen voor het schrijven van opgaven, maar relatief minder over de statistische mogelijkheden om DIF te onderzoeken. Ook wordt maar beperkt gerapporteerd over de mate waarin de onderzoeksresultaten geldig zijn voor individuele landen. Het boek van Hambleton, Merenda en Spielberger (2005) bevat een verzameling papers over het aanpassen van toetsen ten behoeve van internationale vergelijkingen. Hierin is meer aandacht voor het controleren of het statistische model dat wordt gebruikt, geschikt is in alle landen.

4 Contextvariabelen

Het verzamelen van contextvariabelen binnen een internationaal vergelijkend onderzoek als PISA dient twee belangrijke doelen. Het eerste doel is dat contextvariabelen de vergelijkbaarheid van de gemeten kennis en vaardigheden van leerlingen uit verschillende landen vergroten door rekening te houden met de verschillen in de context van landen. Het tweede doel van het verzamelen van contextvariabelen is het vergroten van de bruikbaarheid van de data voor individuele landen. Landen zijn erbij gebaat als er contextvariabelen beschikbaar zijn waarmee beleid kan worden getoetst. Deze twee doelen kunnen een spanningsveld opleveren tussen vergelijkbaarheid tussen landen en bruikbaarheid binnen de eigen context van een land.

Internationale vergelijkbaarheid wordt door het consortium dat het PISA-onderzoek uitvoert steeds nagestreefd, maar er zijn grenzen aan deze vergelijkbaarheid. Deze grenzen worden hieronder verduidelijkt met enkele voorbeelden. Het eerste voorbeeld is de vergelijkbaarheid van onderwijssystemen tussen landen. Het onderwijssysteem in een land is één van de belangrijkste contextvariabelen voor internationaal vergelijkend onderzoek naar kennis en vaardigheden. Een onderwijssysteem wordt binnen internationaal vergelijkend onderzoek uitgedrukt in ISCED-niveaus (International Standard Classification of Education; OECD, 1999). Grofweg bestaan er niveaus voor primair (ISCED 1), onderbouw secundair (ISCED 2), bovenbouw secundair (ISCED 3), tertiair (ISCED 4), hoger (ISCED 5) en postdoctoraal (ISCED 6) onderwijs. Voor al deze niveaus is beschreven welk niveau eraan vooraf gaat en op welk niveau (of de arbeidsmarkt) het voorbereidt. Het indelen van onderwijssystemen in deze ISCED-niveaus vergroot de vergelijkbaarheid van deze systemen aanzienlijk, maar de vergelijkbaarheid is niet optimaal. Neem Nederland als voorbeeld: Het Centraal Bureau voor de Statistiek (CBS) heeft het Nederlandse onderwijssysteem ingedeeld in ISCED-niveaus en bij internationaal vergelijkende onderzoeken zoals PISA en TIMSS wordt vanzelfsprekend aan deze indeling vastgehouden. Bij deze indeling is rekening

gehouden met het feit dat iedereen binnen Nederland een startkwalificatie dient te behalen alvorens de arbeidsmarkt te betreden. Deze startkwalificatie is gelijkgetrokken met het behalen van ISCED-niveau 3, met als gevolg dat de bovenbouw van het vmbo als ISCED-niveau 2 is gedefinieerd. Het behalen van een vmbo-diploma levert immers geen startkwalificatie op in Nederland. In vele landen waarmee Nederland wordt vergeleken in het PISA-onderzoek is bovenbouw secundair onderwijs gedefinieerd als ISCED-niveau 3. Uit een vergelijking van onderwijssystemen tussen landen zal blijken dat binnen Nederland relatief veel 15-jarigen onderwijsprogramma's op ISCED-niveau 2 volgen.

Een ander voorbeeld van een contextvariabele die niet optimaal berekend kan worden voor internationale vergelijkbaarheid is de leraar-leerlingratio. Het probleem zit in de verhouding tussen voltijd- en deeltijdleraren. In de berekening van deze ratio voor PISA is ervoor gekozen deeltijdleraren voor een half mee te tellen. In Nederland hebben deeltijdleraren over het algemeen een aanstelling van meer dan 50 procent. Dit betekent dat het aandeel van deeltijdleraren in een land de leraar-leerlingratio beïnvloedt, zodanig dat een groot aandeel deeltijders de leraar-leerlingratio onevenredig veel verhoogt (meer leerlingen per leraar). Nog een ander voorbeeld om de verschillen in cultuur tussen landen te illustreren en de invloed daarvan op internationale vergelijkbaarheid is het item *number of rooms with a bath or a shower*, dat een onderdeel vormt van de schaal voor de mate van welvaart. In eerdere cycli heeft dit item problemen opgeleverd, omdat de formulering die destijds gehanteerd werd in de Engelse bronversie die als basis voor vertaling in de verschillende testtalen dient (*bathrooms*) in sommige landen verwarring opleverde; in die landen wordt een toiletruimte zonder bad of douche ook *bathroom* genoemd. Dit is de reden voor de huidige uitgebreide formulering. Helaas wordt ook deze formulering soms verkeerd geïnterpreteerd, namelijk als kamers-en-suite. De problemen die de interpretatie van dit item oplevert voor landen geeft wellicht aan dat de badkamer in verschillende culturen een verschillend belang heeft. Men kan zich afvragen of het in dat geval wel een goede indicator is voor welvaart, althans in internationaal vergelijkend onderzoek.

Binnen PISA, waarvoor geldt dat landen voor deelname betalen, pakt het spanningsveld tussen vergelijkbaarheid tussen landen en bruikbaarheid binnen de eigen context van een land vaak uit in het voordeel van bruikbaarheid voor de eigen context van een land. Vooral in het aanpassen van vragen uit de contextvragenlijsten wordt veel vrijheid aan landen gegeven. Onderstaande voorbeelden laten zien welke problemen hierdoor kunnen ontstaan voor de internationale vergelijkbaarheid.

Variabelen waarvoor landen zelf mogen kiezen hoeveel responsopties ze gebruiken zijn *geboorteland van de leerling* en *diens ouders* en *thuistaal van de leerling*. De keuze van het aantal landen en talen is een politieke kwestie; het ene land besteedt meer aandacht aan de integratie van allochtonen dan het andere. Dit heeft echter wel gevolgen voor de internationale vergelijkbaarheid. Deze variabelen worden voor een internationale vergelijking gehercodeerd naar slechts twee responsopties: *land van testafname* en *ander land* voor de eerste variabele en *toetstaal* en *andere taal* voor de tweede variabele. Een land als Nederland, waar de regering voor kiest veel landen op te nemen om de invloed van land van herkomst beter te kunnen bepalen, wordt wat herkomst betreft op een basaal niveau met andere landen vergeleken: is de student een autochtoon, een eerste- of een tweedegraadsallochtoon. Het onderscheid tussen westerse en niet-westerse allochtonen kan in internationale vergelijkingen hierdoor niet worden gemaakt (zie ook de bijdrage van Dronkers in deze uitgave van Pedagogische Studiën).

Er zijn vele voorbeelden te geven van aanpassingen die landen maken in vragenlijsten. We geven hier twee opvallende voorbeelden voor PISA. Eén ervan is de opsplitsing van vragen over *test language lessons* voor verschillende soorten lessen; literatuur en taal. Voor de vragen over *test language lessons* geven de responsopties een frequentie van voorkomen aan: 1) *never or hardly ever*, 2) *in some lessons*, 3) *in most lessons* en 4) *in all lessons*. Als de vragen in twee versies in de

vragenlijst zijn opgenomen en voor internationaal vergelijk moeten worden samengevoegd tot één score per item, dan moeten er keuzes gemaakt worden die consequenties hebben voor dit vergelijk. In de laatste PISA-cyclus (2009) is ervoor gekozen een afgerond gemiddelde te nemen: een 1 voor de ene les en een 3 voor de tweede les levert voor internationaal vergelijk een 2 op, een 2 voor de ene les en een 3 voor de tweede les levert een 3 op. Omdat het aantal lessen literatuur en taal hoogstwaarschijnlijk in een land niet gelijk is, is deze oplossing niet optimaal en gaat de vergelijking mank, maar hetzelfde geldt voor andere oplossingen. Het tweede voorbeeld betreft de vragen over natuurwetenschappen. In veel landen bestaat er één natuurwetenschappelijk vak *science*, maar in een aantal landen – waaronder Nederland – zijn de natuurwetenschappen opgedeeld in drie of meer vakken. Eén van de vragen die betrekking heeft op natuurwetenschappelijke vakken is *How many minutes, on average, are there in a class period for the following subjects?* Het is goed mogelijk dat in sommige landen het aantal minuten in een lesuur per natuurwetenschappelijk vak verschilt. Hierdoor zullen de leerlingen uit die landen een gewogen gemiddelde moeten berekenen; een opdracht die als PISA-opgave niet zou misstaan.

Na deze voorbeelden moeten we ook weer niet te somber worden over de vergelijkbaarheid van contextvariabelen. Voor heel veel variabelen geldt dat ze prima te vergelijken zijn tussen landen, bijvoorbeeld het aantal leerlingen op een school (schoolgrootte) en het feit of een leerling ooit een leerjaar heeft gedoubleerd (al komt dit in sommige landen niet voor). Het is echter gepast om bij het vergelijken van resultaten tussen landen de beperkingen van contextvergelijkingen in gedachten te houden.

5 Conclusies

Internationaal vergelijkende studies van leerprestaties leveren zeer bruikbare informatie over de kennis en vaardigheden van leerlingen in het onderwijs. De informatie uit dergelijke studies levert een belangrijke aanvulling op de informatie uit nationaal onderzoek zoals PPON en onderzoek van de Inspectie van het Onderwijs. De kracht ervan zit in het vanuit verschillende perspectieven benaderen en uitvoeren van het onderzoek naar leerprestaties. De methodologische haken en ogen zijn voor internationale vergelijkende studies wat substantiëler dan voor nationaal opgezet onderzoek, maar geeft de deelnemende landen een inzicht in hun leerprestaties in een internationaal perspectief. Men kan zich natuurlijk wel afvragen of een onderwijssysteem aangepast dient te worden om hoger in de internationale ranglijst te eindigen, maar als tegenvallende resultaten uit verschillende soorten onderzoek convergeren, valt er waarschijnlijk iets te verbeteren. Een ander aspect is dat bij PISA, PIRLS en TIMSS de landen zich steeds beter met zichzelf over tijd kunnen vergelijken naarmate de onderzoeken langer lopen. Dit maakt het mogelijk om eventuele onderwijsvernieuwingen te evalueren met behulp van dergelijke studies. Een interessante optie is bijvoorbeeld om de geleidelijke invoering van de referentieniveaus voor taal en rekenen in het funderend onderwijs naast de longitudinale resultaten van Nederland in de genoemde internationale studies te zetten.

Noot

1 zie www.onderwijsinspectie.nl

Literatuur

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395-480). Reading, MA: Addison-Wesley.

Cochran, W. G. (1977). *Sampling techniques*. New York: John Wiley.

Cooper, H., Nye, B., Charlton, Lindsay, J., & Greathouse, S. (1996). The effects of summer vacation on achievement test scores: A narrative and meta-analytic review. *Review of Educational Research, 66*, 227-268.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (2005). *Adapting educational and psy-*

chological tests for cross-cultural assessment. Mahwah, NJ: Lawrence Erlbaum.
IEA. (2008). *TIMSS 2007. Technical report*. Chestnut Hill, MA: IEA.
Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational* Research*, 13*, 127-143.
Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 525-543.
OECD. (1999). *Classifying educational programmes. Manual for ISCED-97 Implementation in OECD Countries.* Paris: OECD.
OECD. (2009). *PISA 2006 Technical report*. OECD: Paris.
Raju, N. S., Linden, W. J. van der, & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19*, 353-368.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: The Danish Institute of Educational Research. (Expanded edition, 1980. Chicago: The University of Chicago Press.)
Schoot, F. van der. (2008). *Onderwijs op peil? Een samenvattend overzicht van 20 jaar PPON*. Arnhem, Nederland: Cito.

Manuscript aanvaard: 23 januari 2011

Auteurs

Peter van Rijn is werkzaam bij ETS Princeton (VS) en **Joke Kordes** en **Erna Gille** zijn werkzaam bij Cito in Arnhem.

Correspondentieadres: Erna Gille, Cito, Postbus 1034, 6801 MG Arnhem. Email: erna.gille@cito.nl.

De politieke waarde van internationale prestatie-indicatoren

S. Karsten

1 Inleiding

1.1 Historische achtergrond

In het verleden hebben staatslieden regelmatig alarm geslagen over de stand van het onderwijs van hun land. Vaak waren het oorlogen die een dergelijke schok teweeg brachten. Beroemd is de uitspraak van de Britse premier Lloyd George na de Eerste Wereldoorlog: "The most formidable institution we had to fight in Germany was not the arsenal of Krupps or the yards in which they turned out submarines, but the schools of Germany*"*. Tijdens de Koude Oorlog leidde de lancering van de Russische Spoetnik in 1954 tot een grootschalige vernieuwing van het Amerikaanse wiskundeonderwijs. In de jaren tachtig kwam een door president Reagan ingestelde commissie, onder invloed van de economische wedloop met Japan en Korea, tot de veelzeggende metafoor: "If an unfriendly foreign power had attempted to impose on America the mediocre educational performance that exists today, we might well have viewed it as an act of war". Ten slotte is ook de Lissabon-agenda van de Europese Unie uit 2000 te lezen als een angstkreet, namelijk, dat zonder verbetering van het onderwijs Europa achterblijft als grootmacht in de wereld.

Wanneer we voor politieke doeleinden de opbrengsten van verschillende nationale onderwijssystemen willen vergelijken, zou het natuurlijk mooi zijn om over een aantal objectieve en universele indicatoren te beschikken. Wanneer economen bijvoorbeeld over nationale opbrengsten spreken, dan gebruiken zij indicatoren als het Bruto Nationaal Product. Op het terrein van het onderwijs is het niet waarschijnlijk dat we een maat vinden die daarmee te vergelijken is. Toch bestaat er een lange traditie in het verzamelen van statistische gegevens over verschillende aspecten van nationale onderwijssystemen. Na de Tweede Wereldoorlog was het primaire doel daarvan het verzamelen van informa-

tie voor de planning van het *aanbod*. Tot in de jaren tachtig was er nog betrekkelijk weinig aandacht voor de *opbrengsten* van onderwijsstelsels. Aanvankelijk werden alleen de participatiegraad en het aantal diploma's (*attainment*) als indicatoren voor opbrengst gebruikt.

1.2 IEA als eerste

Hoewel nog steeds gebruikt, zoals bij de Lissabon-agenda, bevredigen die *attainment*-indicatoren niet helemaal. De groei van het aantal diploma's kan immers ook een teken zijn van inflatie (Dore, 1976). Daarom werd de uitdaging om ook prestaties (*achievement*) te vergelijken steeds groter. De eerste poging daartoe kwam uit de wetenschappelijke hoek. Al in 1958 werd de International Association for the Evaluation of Educational Achievement (IEA) opgericht door wetenschappers die elkaar in kringen van de Unesco tegenkwamen. In de woorden van een van de bekendste oprichters, Torsten Husén, was het primaire doel niet politiek maar wetenschappelijk: "We simply wanted to take advantage of the international variability with regard to both outcomes of the educational systems and the factors that caused differences in those countries" (Husén, 1979). Onderwijssystemen werden door hen gezien als natuurlijke 'laboratoria' om uit te vinden "wat werkt en waarom" (een opvallend moderne uitspraak). De IEA startte in de jaren zestig met de eerste internationaal vergelijkende onderzoeken op het terrein van het wiskundeonderwijs. In de jaren negentig groeide dit initiatief uit tot studies met herhaalde metingen, met meer vakken en ook meer niveaus in het onderwijs (bijvoorbeeld leesonderwijs, exacte vakken en burgerschap). De twee bekendste voorbeelden zijn: Progress in International Reading Literacy Study (PIRLS) en Trends in International Mathematics and Science Study (TIMMS).

1.3 OESO

Met een uitgesproken politiek doel voor ogen zette de Organisatie voor Economische Samenwerking en Ontwikkeling (OESO) in de jaren negentig de eerste stappen voor een uitgebreid stelsel van prestatiemeting. Deze organisatie, opgericht door de Verenigde Staten op het hoogtepunt van de Koude Oorlog, was aanvankelijk bedoeld om de rijke, niet-communistische landen van advies te dienen bij de stimulering van hun economische groei. Omdat onderwijs werd gezien als belangrijke groeifactor (*human capital*) werd ook aan bewindslieden en ambtenaren op het terrein van het onderwijs een podium verschaft voor informatie-uitwisseling en beleidsgerichte rapportages van deskundigen over het onderwijs. Op aandringen van de VS werd een nieuwe, veel krachtigere koers ingeslagen en kwam het Program for International Student Assessment (PISA) tot stand. Met de uitbreiding binnen en buiten de OESO-lidstaten (momenteel ongeveer zestig landen) is vooral de gestandaardiseerde en driejaarlijks herhaalde meting van competenties van 15-jarigen op het gebied van lezen, wiskunde en natuurwetenschappen een krachtig instrument in handen van het nationale en internationale beleid geworden.

1.4 De politieke betekenis

Over de politieke invloed van internationale prestatiemetingen – in het bijzonder van PISA – wordt onder voorstanders en critici weinig getwist, maar over de waarde van dergelijke internationale prestatiegegevens voor het beleid des te meer. De vraag is wat je ermee kan en wat niet. Daar wil ik in mijn discussiebijdrage wat meer bij stilstaan. Daarbij zal ik eerst ingaan op de voor- en nadelen van internationale prestatiemeting voor onderwijsbeleid. Vervolgens zal ik kort stilstaan bij wat we weten hoe deze gegevens feitelijk in de politiek gebruikt worden. Ik sluit af met een pleidooi voor een kritisch gebruik van dergelijke gegevens.

2 Internationale prestatie-indicatoren als beleidsinstrument

Leren van elkaar, ook op het terrein van het onderwijs, is een oud gebruik in de beleids- en onderwijspraktijk. Veel onderwijsvernieuwers en ook beleidsadviseurs trokken in het verleden (maar ook nu steeds) naar het buitenland om daar het onderwijslandschap te verkennen, nieuwe ideeën op te doen en vooral om te leren. Toch verschilt die werk-

wijze fundamenteel van het leren door middel van prestatie-indicatoren of andere kengetallen. Als een groot voordeel van objectieve en grootschalig verworven kwantitatieve gegevens wordt gezien dat men niet meer afhankelijk is van de subjectieve indrukken en soms ook mooie praatjes van de gesprekspartners bij een bezoek. In het verleden kon het gebeuren dat bij een OESO-review door internationale deskundigen de ambtenaren van het te bezoeken land van te voren een zorgvuldige balans op maakten, de gesprekspartners selecteerden en soms ook gewoon de mythes over het eigen bestel herhaalden (zoals in Nederland het geval was in 1990 over de vrijheid van het onderwijs; zie daarvoor Karsten, 2008).

2.1 Valkuilen en dilemma's

Voor een zinvol gebruik van internationale prestatiemeting moet men zich bewust zijn van een aantal dilemma's en valkuilen. In de eerste plaats dient men zich af te vragen waarvoor men de indicatoren wil gebruiken. De ervaring van prestatiemeting in de publieke sector leert dat naarmate men er meer functies (leren, beoordelen, afrekenen) aan toekent, des te meer het systeem zichzelf uitholt (De Bruijn, 2001)[1]. Deze paradox tekent zich bij internationale prestatiemeting voornamelijk af op het niveau van landen, maar kan in de toekomst ook gevolgen hebben voor de lagere beleidsniveaus. Zo trok Frankrijk zich terug uit de eerste internationale studie naar geletterdheid onder volwassenen (IALS), toen duidelijk werd dat het land slecht zou scoren. Maar ook de weigering van sommige landen om bepaalde – meestal politiek gevoelige – gegevens te verzamelen (bijvoorbeeld over de herkomstlanden van migranten waardoor nu alle migrantengroepen op één hoop gegooid worden) is een voorbeeld van het streven van beleidsmakers om zo goed mogelijk voor de dag te komen. Uiteindelijk kunnen ook scholen hun medewerking gaan weigeren als zij zich bedreigd voelen door wat er met de gegevens gebeurt op politiek niveau. Momenteel is in mijn ogen de politieke invloed bij de dataverzameling bij het werk van de OESO (PISA, ALL en PIAAC) eigenlijk te groot (zie uitvoeriger bijdrage van Dronkers in dit nummer). Dat is een belangrijk nadeel, want hoe groter het politieke belang, des te strategischer het gedrag wordt van degenen die de gegevens moeten leveren.

In de tweede plaats kan een te sterke oriëntatie op opbrengsten zonder aandacht voor de processen daarachter de gebruikswaarde verminderen. Belangrijke vraag daarbij is in hoeverre het niveau van leeropbrengsten is te danken aan het onderwijs of aan andere factoren. Opvallende scores en ook veranderingen in de tijd kunnen zeker een signalerende werking hebben. De opbrengstindicatoren echter onthullen niet direct de oorzaken achter een bepaalde score of plaats op de ranglijst. Zij kunnen daarom vaak moeilijk op zich beschouwd worden als goed of slecht. Wat zit er achter een indicator? Dat vraagt om een nadere analyse. Wijst een bepaalde score op succes of falen van overheidsbeleid? Een hoge score heeft een plafondeffect (men kan alleen nog maar dalen) en kan leiden tot verstarring of zelfingenomenheid. Niet alle indicatoren zijn eenduidig. Is er bijvoorbeeld een uitruil mogelijk tussen bepaalde scores (bijvoorbeeld een hoge gemiddelde score of een kleine spreiding)? Ligt een bepaalde uitkomst alleen aan het onderwijssysteem of aan factoren daarbuiten?

Het belang van extrinsieke factoren wordt direct duidelijk als we subpopulaties binnen eenzelfde bestel vergelijken. Een voorbeeld daarvan zijn de twee taalgemeenschappen in Finland: Fins en Zweeds. Zo was in PISA 2000 de gemiddelde score op *lezen* voor de eerstgenoemde gemeenschap 548, maar voor de andere gemeenschap slechts 513 (iets lager dan het gemiddelde in Zweden). Als een dergelijke achtergrondvariabele van de leerling al zo veel uitmaakt in een vrij homogeen land als Finland, kunnen we ons afvragen wat dan de betekenis is van internationale prestatiegegevens voor de werking van een nationaal onderwijsbestel als geheel; in dit geval een bestel dat momenteel als een groot voorbeeld wordt genoemd. Een ander voorbeeld zijn de uitkomsten van wiskundetesten uit PISA 2003: wanneer we de eerste en tweede generatiemigranten uit de nationale gemiddelden verwijderen, dan wordt de toppositie van Finland ingenomen door Neder-

land en België (Wuttke, 2007). Welke politieke keuzes kunnen we hier uit afleiden? Geduld, migratiestop, spreiding of onderwijsverbetering? Beleidsalternatieven veronderstellen causale inzichten in de bestudeerde problemen en in de mogelijke consequenties van bepaalde beleidsinterventies. Voor een dergelijke analyse zijn de meeste internationale prestatiegegevens (nog) niet geschikt. PISA bijvoorbeeld laat slechts cross-sectionele analyses toe en heeft betrekking op een heel specifiek moment in de onderwijsloopbaan van leerlingen.

Dit laatste punt brengt mij bij het derde dilemma. Internationale prestatiemeting geeft geen kant-en-klare beleidrecepten. Zelfs al zou er in het buitenland een helder antwoord te vinden zijn op beleidsvragen die hier aan de orde zijn, dan is het nog onzeker of dit antwoord hier ook zal passen. Bij beleid gaat het immers niet alleen om de vraag of iets werkt (Hemerijck, 2003). Uiteraard moet het beleid doeltreffend en zo doelmatig mogelijk zijn (instrumentele doelmatigheid). Naast instrumentele doelmatigheid kent het beleid echter nog drie andere kwaliteitseisen: institutionele slagvaardigheid, constitutionele rechtmatigheid en maatschappelijke aanvaardbaarheid. *Slagvaardigheid* heeft betrekking op politieke haalbaarheid en bestuurlijke uitvoerbaarheid. Bepaalde beleidsmaatregelen zijn bijvoorbeeld moeilijk te passen binnen de institutionele structuur van een politiek of onderwijssysteem. Wij kennen bijvoorbeeld in de politiek coalitieregeringen, in het onderwijs redelijk autonome besturen en een historisch gegroeid stelsel met keuzevrijheid en afzonderlijke schooltypes. De ervaring met de basisvorming heeft geleerd dat daar moeilijk iets in radicale zin te veranderen valt en in de uitvoering vaak een doelverschuiving optreedt.

Verder behoren beleidsbeslissingen constitutioneel rechtmatig te zijn. Landen verschillen in de mate van regulering en subsidiëring van het religieus gefundeerde onderwijs; verschillen die tot behoorlijke prestatieverschillen tussen en binnen stelsels kunnen leiden (Dronkers, 2004). In veel landen is het grondwettelijk moeilijk om daarin verandering te brengen, ook al zou dat tot verbetering van hun positie leiden. Tot slot moeten beleidsinterventies in overeenstemming zijn met algemeen aanvaarde normen en waarden in de maatschappij. Ook daarvoor geldt dat maatregelen niet simpel over te nemen zijn. Het zou best kunnen zijn dat zij botsen met heersende normen en waarden en als zodanig negatieve neveneffecten oproepen die het beoogde effect uiteindelijk teniet doen. Kortom, wat vanuit internationale gegevens evident lijkt, hoeft dat voor nationaal beleid niet te zijn. Het moet binnen een nationale context nog blijken of het gaat werken, passen, mogen of behoren. Dit zijn geen hindernissen die voor altijd vastliggen, maar maken het wel heel moeilijk om ontwikkelingen via doelgerichte acties te sturen.

2.2 Wat wordt gemeten?

Bij de voorafgaande dilemma's ben ik ervan uitgegaan dat wat als prestaties wordt gemeten niet omstreden is. Dat is bij PISA, maar ook bij PIRLS en TIMMS niet echt het geval. Ik doel hier niet zozeer op de methodologische problemen van de metingen (het model, de geldigheid en betrouwbaarheid van de maten), maar op wat men feitelijk probeert te meten. Wat is daarvan de politieke waarde? De ambitie van PISA is het toetsen van "real-life skills and competencies in authentic contexts" (Schleicher, 2007). Men zou kunnen beargumenteren dat dit alleen al per definitie onmogelijk is. Politiek gezien is echter het belangrijkste gegeven dat PISA – in tegenstelling tot TIMMS – *geen* schoolse kennis meet. Het model en ook de items die gebruikt worden, gaan ervan uit dat er geen verband is met de verschillende curricula. De toetsen van TIMMS proberen wel kennis te toetsen; namelijk kennis die min of meer gemeenschappelijk is in de curricula van de deelnemende landen. Dit heeft als voordeel dat we iets te weten komen over doelbereiking, maar het betekent ook dat het 'getoetste curriculum' van TIMMS traditionele trekken vertoont (veel items zouden ook zestig jaar geleden bruikbaar zijn geweest)[2].

Als voordeel van PISA wordt over het algemeen gezien dat het los staat van het aangeboden curriculum en ruimte biedt voor vernieuwing. PISA benadrukt daarbij ook dat de getoetste competenties ook buiten de school verworven kunnen worden door informeel

leren. Dit laatste past bij de filosofie van de OESO, maar is niet het beeld dat beleidsmakers en het brede publiek hebben bij wat er getoetst is en zou moeten worden. In de VS is er bijvoorbeeld al de nodige kritiek gekomen op het ideologische karakter van sommige toetsitems (Loveless, 2009). Ook andere onderzoekers hebben het karakter van de toetsitems onder vuur genomen. Zo kwam Koretz (2008) bijvoorbeeld tot de bevinding dat in de wiskundetoets van PISA slechts 11% van de items aan algebra was gewijd. Wat is dan de betekenis van een dergelijke toets voor de kwaliteit van het wiskundeonderwijs op scholen, laat staan voor de effectiviteit van het onderwijs in het geheel? In welke fase van de schoolloopbaan zijn de gemeten vaardigheden al dan niet verworven?

2.3 Relatief of absoluut

Belangrijk in dit verband is ook hoe de uitkomsten worden gepresenteerd: relatief of absoluut. Dit bepaalt mede de perceptie van de stand van het onderwijs door politici en het publiek. De scores van landen kunnen heel dicht bij elkaar liggen en niet-significant van elkaar verschillen. De rangorde op basis van dergelijke scores – en zelfs ook de absolute score – echter kan dan over de tijd sterk wisselen zonder dat er een fundamentele verandering heeft plaatsgevonden. Zo 'duikelde' Japan in de rangorde op de PISA-wiskundetoets tussen 2000 en 2003 (mede omdat Nederland en Hongkong in 2000 niet meetelden en wel in 2003) zonder dat er een significante verandering in de gemiddelde score had plaatsgevonden, en er geen ander land in die twee jaren significant beter scoorde dan Japan. Een ander leuk voorbeeld van publieke vertekening geeft wetenschapsjournalist Van Manen (2009) over de mediaontvangst van de PISA-uitkomsten van 2003. Hij vergelijkt de toetsresultaten van Finland, Nederland en Vlaanderen met gemiddelde lichaamslengte en gewicht. Zouden we ons, zo stelt hij, in allerlei verklaringen verdiepen wanneer blijkt dat Finse jongens gemiddeld vier millimeter langer zijn dan Nederlandse jongens (namelijk vergelijkbaar met het verschil in score op PISA)? Of zouden we de Vlaamse frietcultuur als schuldige aanwijzen voor het feit dat Vlaamse jongeren 250 gram op een gemiddelde van 70 kilo zwaarder zijn dan de rest van de wereld? Waarschijnlijk niet. Wat kunnen we dan opmaken uit de gegevens voor de kwaliteit van ons (wiskunde)onderwijs?

3 Politiek gebruik van internationale prestatiegegevens

Al eerder heb ik opgemerkt dat PISA verschilt van andere internationale vergelijkende studies vanwege het duidelijk politieke doel: "it aims to provide a new basis for policy dialogue and for collaboration in defining and implementing educational goals" (Figazollo, 2009). Zelfs, zo stelt de OESO (Figazollo, 2009), wanneer "school and system characteristics cannot provide precise policy prescriptions, they can address educational policies correlated to high performance". In allerlei aanpalende studies en studiebijeenkomsten heeft de OESO de afgelopen jaren een consistente, vooral economisch gefundeerde, boodschap uitgezonden, te weten vergroting van de *productiviteit van het onderwijs* door sterkere *marktwerking*, grotere *autonomie* voor scholen gekoppeld aan een systeem van verantwoording en rekenschap met behulp van *externe standaarden*. Toch heeft ook de OESO naast doelmatigheidsargumenten altijd veel aandacht besteed aan het vraagstuk van gelijkheid. Ook daarin is de boodschap al jaren consistent: landen waar leerlingen pas op latere leeftijd kiezen doen het verhoudingsgewijs beter. Het feit dat Nederland daarin een wat vreemde eend in de bijt is, heeft de OESO er niet van weerhouden om de voorkeur voor een geïntegreerd stelsel uit te spreken en geregeld Nederlandse bewindslieden daar op aan te spreken.

3.1 Gemotiveerd scepticisme

In de reacties op de kritiek van de OESO op het selectieve karakter van het voortgezet onderwijs in Nederland, zien we een meer algemeen patroon van politiek gebruik van wetenschappelijke evidentie: een zogenoemd gemotiveerd scepticisme (Kunda, 1990). Het concept van gemotiveerd scepticisme helpt ons bij de verklaring voor de manier waarop politici informatie verwerken. Daarin verschillen zij niet van andere mensen, zoals ge-

wone burgers en ook wetenschappers. Onderzoek dat onze vermoedens of hypotheses bevestigt, wordt vrijwel direct geaccepteerd. Wanneer wij echter met contra-evidentie worden geconfronteerd, zijn wij 'gemotiveerde sceptici'. Dan worden wij kritisch: er worden alternatieven gezocht, methodologische gebreken uitvergroot, en variabelen anders geïnterpreteerd. Alleen bij aanhoudend 'bewijs' zijn wij misschien geneigd om onze overtuiging (geloof) te wijzigen. Experimenten naar dit verschijnsel laten zien dat in situaties van sterke polarisering van standpunten de neiging om gemotiveerde sceptici te worden heel groot is. Dit verschijnsel is goed terug te vinden in de wijze waarop de PISA-uitkomsten in verschillende landen door politiek en publiek ontvangen zijn.

3.2 Schrikreacties

In veel landen is er naar aanleiding van de publicaties van de driejaarlijkse PISA-resultaten de nodige turbulentie ontstaan en hebben regeringen van zeer uiteenlopende politieke kleur de resultaten aangegrepen om hervormingen te bepleiten (zie voor een overzicht Figazollo, 2009). In Duitsland veroorzaakten de resultaten van PISA 2000 een ware schok. Voor die tijd werd in dat land aangenomen dat men tot de beste onderwijssystemen ter wereld behoorde. In 2000 echter bleek Duitsland op de twintigste plaats te staan wat rekenen, lezen en natuurwetenschappen betreft. Later werd wel enige nuancering aangebracht tussen de verschillende bondsstaten, maar de schok was er niet minder om. In 2002 organiseerden de ministers van de verschillende deelstaten een conferentie waar tot een grondige hervorming van het Duitse onderwijs werd besloten. Dit land is een voorbeeld van het patroon dat zich op meer plaatsen heeft voorgedaan: eerst kritiek op de uitkomsten en gebruikte toetsen, maar vervolgens acceptatie van de gegevens en maatregelen die in de lijn liggen van de door de OESO voorgestelde hervormingen. Opvallend is ook dat niet alleen beleidsmakers in de ban van de gegevens zijn geraakt, maar ook de publieke opinie er sterk door beïnvloed wordt (Pongratz, 2006), zodat er nu zelfs PISA-testjes op de markt van gezelschapsspelen zijn. Dit heeft als gevolg dat de cijfers een grote strategische waarde in de politiek hebben gekregen.

3.3 Strategisch gebruik

Hoe strategisch de gegevens worden, laten de ontwikkelingen in Japan zien. Dit land behoorde nog in 2000 en ook in de voorafgaande peilingen van IEA tot de top op het terrein van de wiskundeprestaties. Juist tegen het einde van de vorige eeuw vond in Japan een grootschalige hervorming plaats, bekend als de *yutori*hervorming dat wil zeggen *minder prestatiedruk* en *minder stampwerk* (voor die tijd het handelsmerk van het Japanse onderwijs). Deze hervorming lijkt in veel opzichten op de ideeën rond Het Nieuwe Leren in Nederland: meer leerlinggericht, probleemgestuurd en vakoverstijgend onderwijs. Daarmee paste deze hervorming, die in 2002 werd ingevoerd, binnen de internationale trend van onderwijsinnovaties die juist vanuit de OESO in de jaren tachtig en negentig verkondigd was. De hervorming stuitte echter op veel weerstand en stond al bij invoering onder grote druk. De vermeende neergang van Japan als koploper op het terrein van het wiskundeonderwijs tussen 2000 en 2003 werd dan ook prompt ingezet als bewijs voor het falen van deze hervorming. De kleine, maar niet significante terugval in de wiskundeprestaties werd opgeblazen tot een regelrechte crisis tot in 2005 het Ministerie van Onderwijs ook tot de constatering kwam dat zij 'op het verkeerde spoor zat'. Vervolgens werden de resultaten van PISA 2003 gebruikt om de hervorming terug te draaien en ditmaal het Japanse onderwijs om te buigen in een meer marktgericht, neoliberale richting (Knipprath, 2010; Takayama, 2008).

3.4 Rol media

In de beleidsarena is de laatste jaren de rol van de media enorm toegenomen. De Raad voor Maatschappelijke Ontwikkeling spreekt in een rapport van 2003 zelfs van een 'medialogica' dat wil zeggen een sterke verwevenheid van de media en de politiek. Aan de ene kant vervullen de media een belangrijke rol in het signaleren van maatschappelijke problemen en dragen op die manier bij aan de agendavorming in de politiek. Aan de andere kant beschikken zij over een zekere macht

om ook de problemen in een bepaald perspectief te plaatsen (*framing*). Dat is geen 'neutrale' activiteit: zij vormen de publieke opinie en in zekere mate ook de perceptie van beleidsmakers van een bepaald probleem. Wanneer de media aandacht besteden aan de uitkomsten van de internationale prestatie-indicatoren dan is de toonzetting ook zeer belangrijk. Volgens een onderzoek van Figazzolo (2009) onder vertegenwoordigers van de onderwijsbonden komt naar voren dat de media zich vooral concentreren op de rangorde van de verschillende landen en dat veruit de meeste bonden door de pers om commentaar gevraagd is "waarom leraren zo slecht presteren". In een analyse van *Meltwater News* (aangehaald in Figazollo, 2009) van 12.000 artikelen die wereldwijd zijn gepubliceerd in de periode december 2007 en oktober 2008 komt naar voren dat eenderde van de artikelen alleen betrekking heeft de rangordes zonder enige verklaring, in bijna eenderde de resultaten gebruikt worden om een hervorming van het onderwijs te bepleiten en slechts een paar procent van de artikelen de leraren de schuld geeft voor de slechte resultaten. Wat er bepleit wordt aan hervormingen gaat vooral in de richting wat de OESO zelf ook steeds naar voren brengt: meer marktmechanismen in het onderwijs en meer rekenschap (*accountability*).

4 Besluit

Samengevat kunnen we constateren dat de internationale prestatie-indicatoren in de afgelopen decennia een steeds belangrijker rol in het politieke beleidsproces zijn gaan spelen. Het is fascinerend om te zien hoe een, in vergelijking met wetten en regels zacht, beleidsinstrument als prestatiegegevens zo dominant kan worden. We dienen wel te beseffen dat internationale vergelijkingen aan de ene kant geen onomstreden evidentie aandragen voor specifieke beleidsbeslissingen en aan de andere kant een belangrijk instrument zijn geworden op het politieke strijdtoneel. Politiek wordt wel omschreven als de gezaghebbende toedeling van waarden. Onderwijsbeleid is dan ook de uitdrukking van de waardeoriëntaties van die politieke actoren die over de meeste macht beschikken. Dit geldt ook voor het gebruik van de uitkomsten van internationale prestatiemeting. Veel hangt af van wat erin gestopt wordt (wie bepaalt wat en hoe er wordt gemeten?) en van wie er, gezien zijn machtspositie, in slaagt op basis van de uitkomsten zijn interpretatie doorslaggevend te laten zijn. De Franse progressieve sociologen Baudelot en Establet, bekend om hun radicale boek *l'École capitaliste en France* uit 1971, schetsen een alternatief scenario. In een recente publicatie (Baudelot & Establet, 2009) gebruiken zij de PISA-resultaten voor een felle aanklacht tegen het elitaire karakter van het Franse onderwijs. Ook dat is mogelijk op basis van internationale prestatie-indicatoren. Wanneer we het beleid zien als een constructief proces van overleg en argumentatie, dan kunnen de internationale prestatiegegevens een waardevolle bijdrage leveren voor een zinvol debat. Dat vraagt wel om een voorzichtig en kritisch gebruik van de gegevens. Dat kan zeker als onderzoekers meer internationale samenwerking zoeken om de verschillende internationale onderzoekingen verder te ontwikkelen en benutten. Zij moeten dan ook alert en kritisch blijven over de resultaten en zeker ook de neveneffecten beter in kaart brengen. Wie wordt eventueel schade toegebracht en waarom? Maar ook, wie kunnen er profiteren van de gegevens om de werking van het eigen stelsel te verbeteren?

4.1 Internationaal gemiddelde

Dit betekent in de eerste plaats dat de notie van een internationaal gemiddelde met grote voorzichtigheid moet worden bezien (Koretz, 2008). Het gemiddelde kan van onderzoek tot onderzoek verschillen (ook tussen PISA in het ene jaar en PISA in het andere jaar). Die verschillen ontstaan omdat er verschillende landen meedoen en verschillende steekproeven worden getrokken. Dus een eerste stap zou zijn om vergelijkingen met het gemiddelde met omzichtigheid te bezien. Het is beter zich te concentreren op specifieke vergelijkingen, door bijvoorbeeld landen te vergelijken die sterk verschillen of juist weinig verschillen. Die vergelijkingen zijn vaak veel informatiever.

4.2 Verschillen in resultaat

De resultaten hangen ook af welke toetsen zijn afgenomen. Zo is de rangorde in PISA verschillend van die in TIMMS. Daarom is het beter kleine verschillen in rangorde tussen verschillende studies (PISA en TIMMS) en in de tijd te negeren. Het feit dat een verschil significant is, is ook geen voldoende garantie. Het is veel veiliger om te constateren dat een land systematisch bij een ander land in score achterblijft dan dat twee landen significant verschillen. Soms is het verschil weliswaar significant, maar is de grootte van het verschil veel kleiner dan een systematisch verschil tussen twee landen.

4.3 Nadere studie nodig

Ten slotte moet er voor gewaakt worden om een bepaalde studie voor een specifiek jaar als het definitieve antwoord op de vraag naar prestatieverschillen te beschouwen. Nadere studie blijft vereist. Aanvullende bronnen en studies kunnen vaak een beter licht werpen op een eenmaal gevonden verschil. Als deze gegevens nog ontbreken, dan moeten we voorzichtig blijven met het trekken van conclusies. Een flink verschil is een duidelijke aanwijzing, maar een schijnexactheid tot twee of drie decimalen achter de komma is ronduit gevaarlijk. Bovendien ligt met de toename van toetsen waar in politieke zin veel van afhangt altijd het spook van toetsinflatie op de loer. Naar dat verschijnsel wordt zowel nationaal als internationaal nog te weinig onderzoek gedaan. Een dergelijk onderzoek zou ons wel eens heel wat voorzichtiger maken met het doorzetten van de huidige trend van rekenschap en verantwoording op basis van (internationale) prestatie-indicatoren.

Noten

1 Dit staat ook bekend als de wet van Campbell:"The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor" (Koretz, 2008, p. 237).

2 Overigens correleren beide toetsen hoog, hetgeen aangeeft dat vrijwel hetzelfde gemeten wordt (Hanushek & Woessman, 2010). Eerder hebben ook Rindermann en Ceci (2009) laten zien dat noch de inhoud van de toets, de aard van de getoetste kennis, het jaar van afname of de populatie veel uitmaakt voor de score van een land op internationale prestatie-indicatoren. Hieruit kan men ook concluderen dat de gemeten vaardigheden zeker niet uitsluitend op school zijn verworven en de score niet louter opgevat kan worden als een resultaat van meer of minder effectief onderwijs.

Literatuur

Baudelot, C., & Establet, R. (2009). *L'élitisme républicain. L'école francaise à l'epreuve des comparaisons internationals*. Paris: Seuil.

Bruijn, H. de.(2001). *Prestatiemeting in de publieke sector*. Utrecht, Nederland: Lemma.

Dore, R. P. (1976). *The diploma disease*. London: Allen & Unwin.

Dronkers, J. (2004). Do public and religious schools really differ? Assessing the European evidence. In P. J. Wolf & S. Macedo (Eds), *Educating citizens. International perspectives on civic values and school choice* (pp. 287-314). Washington: Brookings Institution Press.

Figazollo, L. (2009). *Impact of PISA 2006 on the education policy debate*. Brussels: Education International.

Hanushek, E. A., & Woessman, L. (2010). *The economics of international differences in educational achievement*. NBER Working Paper 15949. Cambridge: NBER.

Hemerijck, A. (2003). Vier kernvragen van beleid. *Beleid en Maatschappij, 30*(1), 3-19.

Husén, T. (1979). An international research venture in retrospect: The IEA Surveys. *Comparative Education Review, 23*, 371-385.

Karsten, S. (2008). De mythe van de vrijheid van onderwijs. In S. Goorhuis-Brouwer et al. (red.), *Mythes in het onderwijs* (pp. 143-158). Amsterdam: SWP.

Knipprath, H. (2010). What PISA tells us about quality and inequality of Japanese education in mathematics and science. *International Journal of Science and Mathematics Education, 9*, 389-408.

Koretz, D. (2008). *Measuring up. What educational testing really tells us*. Cambridge/London:

Harvard University Press.
Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 1008*, 480-498.
Loveless, T. (2009). The use and misuse of international assessments. *The Brown Center Report on American Education, 2*(3), 8-18.
Manen, H. van. (2009). *Goochelen met getallen*. Amsterdam: Boom.
Pongratz, L. (2006) Voluntary self-control: education reform as a governmental strategy. *Education Philosophy and Theory, 38*, 471-482.
Rindermann, H., & Ceci, S. J. (2009). Educational policy and country outcomes in international cognitive competence. *Perspectives on Psychological Science, 4*, 551-568.
Schleicher, A. (2007). Can competencies assessed by PISA be considered the fundamental school knowledge 15-year-olds should possess? *Journal of Educational Change, 8*, 349-357.
Taykayama, K. (2008). The politics of international league tables: PISA in Japan's achievement crisis debate. *Comparative Education, 44*, 387-407.
Wuttke, J. (2007). Uncertainties and Bias in PISA. In S.T. Hopmann, G. Brinek & M. Retzl (Eds.), *PISA zufolge PISA – PISA According to PISA. Hält PISA, was es verspricht? Does PISA Keep What It Promises?* (pp. 241-264). Wenen: Lit-Verlag.

Manuscript aanvaard: 23 januari 2011

Auteur

Sjoerd Karsten is als bijzonder hoogleraar werkzaam aan de Universiteit van Amsterdam.

Correspondentieadres: Sjoerd Karsten, Afdeling Child Development and Education, Faculteit der Gedrags- en Maatschappijwetenschappen, Universiteit van Amsterdam, Nieuwe Prinsengracht 130, 1018 VZ Amsterdam. E-mail: s.karsten@uva.nl.

De maatschappelijke en wetenschappelijke waarde van internationale data over onderwijsprestaties

J. Dronkers

1 Inleiding

In deze bijdrage ga ik speciaal in op de maatschappelijke en wetenschappelijke betekenis van internationale onderwijsprestaties. Daarbij heb ik mij vooral gebaseerd op mijn persoonlijke ervaringen met het verrichten van secundaire analyses met de hieronder te bespreken internationale data over onderwijsprestaties en op het begeleiden van een aantal internationaal vergelijkende studies.

2 Maatschappelijke waarde

Crossnationale gegevens, die wetenschappers maar ook beleidsmakers in staat stellen onderwijsuitkomsten uit verschillende landen met elkaar te vergelijken, komen niet in de eerste plaats beschikbaar door wetenschapsinterne oorzaken, maar door een aantal maatschappelijke, politieke en technologische ontwikkelingen.

2.1 Maatschappelijke ontwikkelingen

Die maatschappelijke ontwikkelingen, die leidden tot meer crossnationale data en vergelijkingen, kunnen aangeduid worden met het gegroeide (toegekende) belang van *human capital*. Na de afloop van de Tweede Wereldoorlog werd in de meeste democratische en kapitalistische landen het belang van een goed opgeleide beroepsbevolking en de noodzaak van de mobilisatie van alle beschikbare talenten erkend. Ook werd vanaf die tijd de rol van de overheid voor de schepping van de condities voor een goedopgeleide bevolking en voor de ontginning van alle beschikbare talenten voluit erkend. Een goed opgeleide beroepsbevolking en het gebruik van alle talenten waren nodig voor de wederopbouw na de oorlog, en voor de onderlinge concurrentieverhoudingen tussen kapitalistische samenlevingen. Vooral als hun onderlinge economische barrières verlaagd werden,

zoals in de Europese Gemeenschap gebeurde. Economieën zouden niet meer concurreren door protectie van interne markten met hoge tariefmuren, maar door de kwaliteit van hun producten en dus door de kwaliteit en inzet van hun productiemiddelen, waaronder de opgeleide beroepsbevolking. Deze maatschappelijke noodzaak van een goede opleiding voor allen, leidde dan ook al snel tot international vergelijkingen, zowel door intergouvermentele als academische organisaties. De OESO (Organisatie van Economische Samenwerking en Ontwikkeling, de denktank van de westerse geïndustrialiseerde landen, beter bekend onder het Engelse afkorting OECD, gevestigd in Parijs) hield zich daarom al vroeg bezig met het beoordelen van de kwaliteit van de opleiding van die beroepsbevolking, naast hun meer strikt economisch studie- en advieswerk. Zo maakt de OESO al sinds de jaren zeventig evaluaties van de onderwijssystemen van de aangesloten landen. De OESO besloot ook, na de erkenning dat het aantal en het niveau van de nationale diploma's onbruikbaar is voor de vergelijking van de kwaliteit van opleidingen tussen landen (Martens, Rusconi, & Leuze, 2007), eind van de jaren negentig tot de crossnationale meting van het *human capital* van de toekomstige beroepsbevolking. Dat leidde tot de invloedrijke PISA-studies (Programme for International Student Assessment) van de OESO. Deze studies meten "how far students near the end of compulsory education have acquired some of the knowledge and skills that are essential for full participation in society". Deze omschrijving laat duidelijk zien dat het bij PISA gaat om de hoeveelheid kennis en vaardigheden, en dat de plaats waar die kennis en vaardigheden verworven zijn (familie, *peer*groep, verenigingen, school, etc.) secundair is. Omdat PISA in haar enquêtes aan andere socialisatiekaders dan de school weinig aandacht besteedt (het meest nog aan het ouderlijk gezin, maar de *peer*groep van de leerlingen blijft geheel buiten beschouwing), dreigt snel het gevaar dat variaties in bruikbare kennis en vaardigheden aan het onderwijs worden toegeschreven, terwijl de oorzaak heel goed kan liggen in een ander socialisatiekader.

Dat perspectief is verschillend bij de concurrent: de IEA (International Association for the Evaluation of Educational Achievement), die in 1958 startte met een bijeenkomst van een groep geleerden, onderwijspsychologen, sociologen en psychometrici in het UNESCO Institute for Education in Hamburg, en dat sinds die tijd data verzameld over onderwijsresultaten in specifieke vakken (taal: PRILS (Progress in International Reading Literacy Study), wiskunde en natuurwetenschappen: TIMSS (Trends in International Mathematics and Science Study), en burgerschap: ICCS (International Civic and Citizenship Education Study)). Deze veel meer onderwijskundig samengestelde organisatie wil internationale standaarden geven die beleidsmakers helpen de relatieve zwakke en sterke kanten van hun onderwijsstelsels vast te stellen. Maar zelfs de stichters van de IEA meenden al dat de meeste landen vergelijkbare omschrijvingen hebben van de optimale onderwijsresultaten en dat ze vooral verschilden in de middelen om dat te bereiken. De leerprestaties die de IEA op de verschillende domeinen meet, zijn nauwer verbonden met de inhoud van de feitelijke curricula in de aangesloten landen en de IEA meet dan ook veel nauwkeuriger dan PISA of in de bedoelde en aangeboden curricula wel de gemeten kennis en vaardigheden worden geleerd. Wel moet hierbij opgemerkt worden dat correcties voor verschillen of overlap in feitelijke curricula tussen landen nauwelijks invloed hebben op de posities die landen innemen in de internationale rangorde. Kennelijk maakt het niet zo veel uit wat er precies getest wordt aan schoolse kennis.

2.2 Politieke ontwikkelingen

De politieke ontwikkelingen, die leidden tot meer crossnationale data en vergelijkingen, kunnen aangeduid worden met het trefwoord legitimiteit. In moderne samenlevingen ontlenen overheden een deel van hun legitimiteit aan de kwaliteit van de door hen verzorgde voorzieningen. Onderwijs hoort zeker daarbij (Meyer, Boli, Thomas, & Ramirez, 1997). Een goed functionerend onderwijs legitimeert zowel de bestaande politieke structuren als de maatschappelijke ongelijkheid. Een in de ogen van de burgers 'eerlijke' selectie en hoge kwaliteit in het onderwijs legi-

timeert de ongelijke verdeling van mensen over de ongelijke posities in de maatschappij (beroep, inkomen, gezondheid, partners). In de loop van de 20ste eeuw kreeg 'eerlijke' selectie de betekenis dat alleen cognitieve prestaties en motivatie van de leerling doorslaggevend mogen zijn in het onderwijs, terwijl aangeboren eigenschappen zoals ouderlijk milieu, geslacht en etnische herkomst geen rol meer behoren te spelen (de zogenaamde gelijke startkansen in het onderwijs). Overheden maar ook politiek betrokkenen hebben daarom behoefte aan informatie over de mate van 'eerlijke' selectie en de kwaliteit in hun onderwijs. Dat helpt overheden een beter beleid uit te zetten (het overnemen van de *best practice* van de buurlanden), een perspectief te hebben voor de mate van 'oneerlijkheid' in hun onderwijs ("in andere landen is het nog veel erger"), maar dat helpt ook protesterende burgers en partijen ("wij lopen ver achter bij'). Bovendien speelden de mate van 'eerlijke' selectie en kwaliteit in het onderwijs ook een rol tijdens de koude oorlog, toen zowel communistische als kapitalistische landen claimden het meest meritocratische en beste onderwijsstelsel te hebben. Dit belang van de 'legitimiteit' van het nationale onderwijsstelsel verklaart ook de gevoeligheid van overheden voor uitkomsten van internationale vergelijkingen, zowel de rangordeningen (Nederland scoort goed dankzij ons beleid) als de effecten van sommige achtergrondvariabelen (Nederland kent een sterk effect van ouderlijk milieu op onderwijsprestaties, zie OECD, 2004). Hoe meer nationale overheden moeten opereren in een geglobaliseerde omgeving, zoals de Europese Unie, waardoor hun handelingsbereik kleiner is geworden (invoering van de euro; Schengenakkoord voor vrij verkeer van personen), des te belangrijker hun legitimatie door een 'eerlijk' en goed onderwijsstelsel wordt. In dit verband is het nuttig erop te wijzen dat het Verdrag van Maastricht, waardoor de euro mogelijk werd, tegelijkertijd het onderwijs tot exclusieve verantwoordelijkheid van de lidstaten verklaarde. De enige taak, die de Europese Commissie ten aanzien van het onderwijs heeft, is die van informatieverzameling en -uitwisseling. Deze legitimiteit maakte in Duitsland de publicatie van de PISA-data zo explosief. Onderwijs is in Duitsland een zaak van de deelstaten (*Länder*), niet van de federale overheid (*Bund*). Als gevolg van de politieke verdeeldheid tussen deze deelstaten over het onderwijs (in noordelijke deelstaten bevorderde een SPD-overheid een soort middenschoolontwikkeling, terwijl in zuidelijke deelstaten een CDU-overheid dat tegenhield) bestaat er nog geen nationale dataverzameling over de in het onderwijs geleerde hoeveelheid geleerde kennis en vaardigheden.[1] Dat betekende dat onderwijsresultaten tussen deelstaten nauwelijks onderling vergeleken konden worden, en veel deelstaten deden en doen alles om dat zo te laten. De nationaal representatief verzamelde PISA-data maakten plotseling een publieke vergelijking van de verschillen in onderwijsresultaten van deelstaten wel mogelijk. De relatief lage rangorde van Duitsland te midden van de andere OESO-landen bleek vooral te wijten te zijn aan het gemiddeld lage kennis- en vaardighedenniveau in de noordelijke deelstaten, hoewel in die deelstaten de effecten van achtergrondvariabelen op de onderwijsprestaties kleiner bleek dan in de zuidelijke deelstaten. Dit verlies van de legitimiteit van het onderwijs in de noordelijke deelstaten door deze PISA-data leiden tot een golf van onderwijshervormingen (waaronder de invoering van een centraal examen in Noordrijn-Westfalen en Hessen), allen bedoeld om het vertrouwen in de 'eerlijke' selectie en de kwaliteit in het onderwijs terug te winnen.

2.3 Technologische ontwikkelingen.

De technologische ontwikkelingen, die leidden tot meer crossnationale data en vergelijkingen, hebben vooral betrekking op de invoering van de computer. Zowel de hardware als de software, maar ook de groei in gedigitaliseerde administratieve data, maken crossnationale vergelijkingen mogelijk, waarvan vroegere geleerden alleen maar konden dromen. Ook maakten deze technologische ontwikkelingen het mogelijk dat data een veel ruimere verspreiding krijgen dan ooit voor mogelijk werd gehouden. Zowel IEA als PISA stellen bijna alle data vrij beschikbaar voor onderzoekers, omdat geïnteresseerde onderzoekers ze van hun internet-

pagina's kunnen downloaden en hun eigen analyses uitvoeren. Het betekende echter ook dat het aantal 'ongecontroleerde' secundaire analyses van deze data tegenwoordig groter is dan vroeger, en dat in deze analyses geheel andere onderzoeksvragen aan de orde komen dan de oorspronkelijke verzamelaars en opdrachtgevers hadden gedacht.

3 Wetenschappelijke waarde

Hierboven blijkt al dat de crossnationale data niet ontstaan zijn om een bepaalde diepzinnige theorie te toetsen met behulp van onderwijsdata. Het doel van bijna alle belangrijke datasets van IEA en PISA/OESO is bepaalde beleidstheorieën met betrekking tot onderwijsstelsels (en haar onderdelen) van een empirische onderbouwing te voorzien. De wetenschappelijkheid van die onderbouwing is beperkt omdat onderwijsstelsels (en haar onderdelen) niet via reële experimenten met voldoende gecontroleerde condities gevarieerd kunnen worden. Als men uitsluitend experimenteel dubbelblind getoetste hypothesen als wetenschappelijk bewijs toelaat, zijn deze crossnationale data onbruikbaar. Maar datzelfde geldt dan voor grote en belangrijke wetenschapsgebieden zoals de biologie met de evolutietheorie of de astronomie met al haar waarnemingen voorbij het zonnestelsel. In deze en andere wetenschapsgebieden (en dat geldt ook voor de onderwijswetenschappen) zal men niet veel verder kunnen komen dan zorgvuldige veldobservaties, gecombineerd met uitgekiende methoden om ruis en vertekeningen die voortvloeien uit veldobservaties (zoals selectiviteit; omgekeerde causaliteit; samenhangende factoren en processen) te neutraliseren. De hierboven genoemde wetenschapsgebieden, die met dezelfde beperkingen worden geconfronteerd, laten ook zien dat dit probleem niets met de tegenstelling tussen harde en zachte wetenschap heeft te maken. Als men zorgvuldige veldobservaties, gecombineerd met uitgekiende methoden om ruis en vertekeningen te neutraliseren, accepteert als middel in de wetenschap dan kunnen crossnationale data bruikbaar zijn voor wetenschappelijke doeleinden.

De crossnationale data van IEA en PISA hebben een andere belangrijke beperking: het zijn dwarsdoorsneden op een bepaalde leeftijd van de leerlingen (PISA) of een bepaalde niveau (IEA) in plaats van longitudinale metingen van steeds dezelfde leerlingen. Dit is een belangrijke beperking voor onderwijsonderzoekers, omdat onderwijs een onderdeel is van de ontwikkeling van menselijke embryo's tot volwassen mensen. Het processmatige, longitudinale karakter van deze socialisering (en waarvan onderwijs slechts het geïnstitutionaliseerde deel vormt) maakt dat dwarsdoorsneden op een bepaald moment van die ontwikkeling altijd beperkt en tot op zekere hoogte zelfs vertekend en dus misleidend zullen zijn. Deze beperking van alle crossnationale data tot dwarsdoorsneden komt in verschillende wetenschapsgebieden voor, maar is echter voor het onderwijs ernstiger dan voor bijvoorbeeld stemgedrag, omdat onderwijs en socialisatie bij uitstek een proceskarakter heeft, in tegenstelling tot stemgedrag.

Deze beperking wordt versterkt doordat zowel de huidige PISA- als de IEA-vragenlijsten weinig retrospectieve vragen aan de leerlingen bevatten (vragen over feit uit het verleden van de respondent). Nu is de betrouwbaarheid op retrospectieve vragen altijd discutabel, maar dat is geen reden ze dan maar niet te stellen. Een technische oplossing zou zijn meer historische gegevens uit de school- en onderwijsadministratie over de leerling te koppelen aan de antwoorden van de leerlingen. Op langere termijn is een dergelijke koppeling gemakkelijker door te voeren als gevolg van technologische ontwikkelingen (in Nederland zou het onderwijsnummer hiervoor bruikbaar zijn). Wel blijft er dan nog het ingewikkelde probleem om deze historische gegevens uit de school- en onderwijsadministratie internationaal vergelijkbaar te maken. Bovendien zullen niet alle deelnemende landen de noodzakelijke gegevens uit school- en onderwijsadministratie kunnen (ze bestaan niet) of willen (privacy) leveren.

Er zijn pogingen om deze beperking op te heffen door gebruik te maken van bepaalde technieken, meestal afkomstig uit de econometrie. Zo bestaat er de techniek van *(fuzzy)*

regression-discontinuity, die gebruik maakt van bepaalde breekpunten in de tijd (zoals de datumgrens bij de vorming van jaargroepen: 1 oktober in Nederland). In Nederland is deze techniek succesvol toegepast door Luyten (2006). Een andere techniek is de Differences-in-Differences-methode, waarbij twee dwarsdoorsneden van vergelijkbare leerlingpopulaties aan elkaar worden gekoppeld, maar waarbij de eerste, jongere dwarsdoorsnede nog niet is blootgesteld aan een bepaalde *treatment,* terwijl de tweede, oudere dwarsdoorsnede daar al wel is aan blootgesteld. Hanushek en Wössmann (2006) hebben deze techniek toegepast om de effecten van de mate van externe differentiatie in het voortgezet onderwijsvast te stellen. Een derde techniek is *propensity score matching*, waarbij geprobeerd wordt de ongemeten effecten van zelfselectie te elimineren, bijvoorbeeld schoolkeuze (Dronkers & Avram, 2010). Hoewel deze econometrische technieken tot op zekere hoogte de beperkingen van dwarsdoorsneden kunnen opheffen, vereisen ze vaak sterke, maar niet erg realistische, aannames waardoor de resultaten niet erg robuust zijn (Jakubowski, 2010) of zijn ze alleen bruikbaar voor heel specifieke, maar beperkte, onderzoeksvragen.

Hoewel het in principe mogelijk is dat IEA en PISA in de toekomst longitudinale gegevens in plaats van dwarsdoorsneden gaan verzamelen, is de kans daarop niet erg groot. Privacybelemmeringen en de enorme kosten van een dergelijke longitudinale meting maken het niet waarschijnlijk dat de financierende en subsidiërende overheden daarin geld willen steken. Bovendien hebben longitudinale data vanuit het perspectief van beleidsmakers een belangrijke beperking: ze hebben meestal betrekking op het onderwijsbeleid van een aantal jaren geleden en dus zijn de resultaten met deze 'verouderde' data niet relevant voor de huidige beleidskeuzen. Deze opvatting mag wellicht gebaseerd zijn op een – vanuit wetenschappelijk oogpunt – overschatting van het belang van het onderwijsbeleid voor de uitkomsten van onderwijsprocessen, maar dat verandert weinig aan de politieke realiteit. Daarom blijken de huidige dwarsdoorsneden van PISA en IEA de best mogelijke data om de effecten van (delen van) onderwijsstelsels op geïnstitutionaliseerde socialisatie te meten.

Samenvattend ligt de kracht van deze crossnationale databestanden vooral bij de analyse van de effecten van onderwijsstelsels op het functioneren van het onderwijs bij het verwerven van kennis en vaardigheden. Ook zijn deze gegevens wel bruikbaar bij het meten van schooleffecten (de internationale data kennen een grotere variatie in schoolkenmerken dan nationale data), maar de meting van die schoolkenmerken is zeker niet optimaal. Bij de PISA-metingen zijn alleen de meningen van de directeur over het functioneren van de betrokken school en de leerkrachten beschikbaar, en zijn er geen middelen om de betrouwbaarheid van die mening van de directeur vast te stellen. Bij de IEA-metingen ligt het probleem precies anders om. De meningen van een aantal docenten over het functioneren van de betrokken school zijn bekend, maar het is onduidelijk of die gemiddeld het juiste beeld geven. Voor een wetenschappelijke analyse van onderwijsleerprocessen zijn de IEA- en PISA-data door hun dwarsdoorsnede karakter ongeschikt.

Zoals gebruikelijk zijn er beperkingen in de aard en het aantal verzamelde gegevens, maar dat is niet specifiek voor deze crossnationale dwarsdoorsneden. Wel is het mij opgevallen dat de IEA-datasets minder rijk zijn in het verzamelen van achtergrondkenmerken van de leerlingen dan de PISA-data. Wellicht ligt dit aan de onderwijskundige achtergrond van de belangrijkste deelnemers in IEA, maar ook de grotere politieke correctheid bij IEA kan een verklaring zijn. In de PISA- en OESO-organisatie spelen economen een grotere rol en die beroepsgroep heeft in vergelijking met onderwijskundigen minder last van politieke correctheid en een grotere interesse in de maatschappelijke omstandigheden van het onderwijs.

4 Problemen met de data

Door de hoge kosten van crossnationale data en door de noodzakelijke medewerking van veel overheidsinstanties is er sprake van een duidelijke politieke beïnvloeding van de data-

verzameling en -beschikbaarstelling, maar ook op de daarmee uitgevoerde analyses. Daarmee wil ik echter niet zeggen dat er bij nationale dataverzamelingen geen sprake zou zijn van politieke beïnvloeding. Hieronder zal ik een aantal voorbeelden behandelen, waarbij de nadruk op de PISA-data zal liggen omdat die data de grootste maatschappelijke impact gehad hebben.

Er is sprake van politieke invloed op dataverzameling en beschikbaarstelling. Het sterkste voorbeeld is Frankrijk: sinds de meting in 2003 zijn via PISA de antwoorden van de schoolleiders over de kenmerken van Franse scholen niet meer beschikbaar voor analyse door onafhankelijke wetenschappers. Australië en Canada geven het onderscheid tussen openbare en bijzondere scholen (juister geformuleerd tussen *public*, *private government-dependent* en *private government-independent* scholen) in hun landen niet in de publieke data set. Canada, Engeland[2], Frankrijk, Nederland[3], USA, Zweden, en andere landen met veel immigranten vragen niet het specifieke geboorteland van leerlingen en ouders. Vanaf de meting in 2003 hadden de aan PISA deelnemende landen afgesproken dat elke land naar de belangrijkste geboortelanden van leerlingen en ouders zou vragen. Maar een groot aantal landen heeft dat niet gedaan: zij maken nog steeds alleen maar het onderscheid tussen wel of niet in het testland geboren, ondanks het feit dat het geboorteland van leerlingen of hun ouders meer variantie in onderwijsprestaties bindt dan hun land van bestemming (Heus & Dronkers, 2010; Levels, Dronkers, & Kraaykamp, 2006). In de meting van 2006 is de gezinsvorm, waarin de leerling op dat moment opgroeit (gezin met beide ouders, moedergezin, vadergezin, etc.) verdwenen als een van de achtergrondkenmerken, ondanks het feit dat gezinsvorm veel meer variantie verklaart dan de meeste schoolkenmerken (bijvoorbeeld het aantal computers in de school). Maar in de meting van 2009 is het weer terug.

Wie echter denkt dat de data van IEA minder beperkingen kennen, omdat dit een organisatie van wetenschappers is, komt bedrogen uit. Aan de meting van burgerschap (1999) doet Nederland niet mee. Toen vonden het Nederlandse Ministerie van Onderwijs en ook de vele belangengroepen die altijd de mond vol hebben over de maatschappelijke taak van het onderwijs het niet nodig om daar meer over te weten[4]. De opstellers hebben in de IEA-studies alle gezinsvormen op een veel te simpele driedeling geperst (twee ouders, inclusief stiefouder), moedergezin of vadergezin. Naar het specifieke geboorteland van leerlingen en hun ouders wordt in het geheel niet gevraagd. Ook de godsdienst van de leerlingen of de ideologische, religieuze of pedagogische achtergrond van de scholen ontbreekt in het IEA-onderzoek naar burgerschap. Zelfcensuur is vaak effectiever dan het afdwingen van politiek correcte opvattingen.

Het zal niet verbazen dat een dergelijke politieke beïnvloeding tot politiek correcte rapporten leidt. Het beste voorbeeld is het OESO-rapport "Where immigrants succeed. Pisa 2003" (OECD, 2006). In dit officiële rapport worden de onderwijsresultaten van leerlingen met een immigrantenachtergrond in de deelnemende PISA-landen geanalyseerd met behulp van de meting in 2003 van PISA. Dat was de eerste PISA-meting waarin een 15-tal deelnemende landen naar het specifieke geboorteland van leerlingen en ouders hadden gevraagd. Maar daarvan is in het rapport zo goed als niets terug te vinden. Leerlingen met een immigrantenachtergrond in Finland worden zonder meer vergeleken met leerlingen met een immigrantenachtergrond in Duitsland, ondanks het feit dat eerstgenoemde hoofdzakelijk uit Zweden en Rusland (waaronder het taalverwante Estland) afkomstig zijn, terwijl de laatstgenoemde uit Turkije, Joegoslavië, etc. De conclusie van dit rapport, dat leerlingen met een immigrantenachtergrond het beter doen in het Finse onderwijs dan in het Duitse, is dus misleidend, omdat hier prestaties van leerlingen uit zeer verschillende herkomstlanden vergeleken worden, zonder dat de kenmerken van die landen er bij betrokken zijn. Ook kregen de schrijvers van dit rapport van hun opdrachtgevers een feitelijk verbod om een analyse uit te voeren van de relatie tussen het in de deelnemende landen gevoerde beleid ten aanzien van leerlingen met een immigrantenachtergrond en hun scores op de kennis en vaardigheden toetsen. Merkwaardigerwijs zijn

er uit de onderzoekswereld maar weinig protesten te horen geweest over deze gang van zaken met dit OESO rapport en is het nog steeds invloedrijk.

Maar de crossnationale data van IEA en PISA hebben twee belangrijke kenmerken waardoor de politieke beïnvloeding door nationale overheden geen ernstige gevolgen hoeft te hebben. Het ontbreken van een land binnen een internationale vergelijking van bijvoorbeeld Europese onderwijsstelsels is niet leuk, maar ook niet ernstig. Als een bepaald effect in alle Europese landen gevonden wordt (bijvoorbeeld een positief effect van *private government-dependent* scholen op onderwijsprestaties; zie Dronkers & Robert, 2004) en dat effect is min of meer even groot in al die Europese landen, dan is het onwaarschijnlijk dat dit effect in Frankrijk niet zou bestaan. Met andere woorden, het crossnationaal karakter van deze data biedt voldoende mogelijkheden om nationale politieke beïnvloeding te omzeilen, want het is in de meeste gevallen onwaarschijnlijk dat het onderwijs in het ene land een heel ander effect heeft dan in buurlanden of in politiek-economisch verwante landen (zoals de OESO-landen)[5]. Een tweede kenmerk waardoor de gevolgen van politieke beïnvloeding beperkt kan blijven, is de onmiddellijke terbeschikkingstelling van de data. Dat vergroot de mogelijkheden om politiek correcte rapporteren te controleren en eventueel te corrigeren, maakt het mogelijk dat politiek ongewenste vragen toch empirisch worden aangepakt door onafhankelijke onderzoekers, en biedt de mogelijkheid om geheel nieuwe vragen aan te pakken. Wat dat betreft voldoen de IEA- en de PISA-data goed aan een belangrijke eis binnen het wetenschappelijk bedrijf: de publieke beschikbaarheid van de data. De meeste schoolloopbaandata en andere onderwijsdata in Nederland voldoen in veel mindere mate aan dit criterium. Meestal is eerst toestemming van het CBS of een andere dataverzamelaar nodig, voordat een onafhankelijke wetenschapper aan de slag mag. Bovendien verloopt er vaak een lange tijd tussen de eerste publicatie en de beschikbaarstelling van data, teneinde de (onderzoeks-)belangen van dataverzamelaar maar ook opdrachtgever te beschermen.[6]

5 Onzinnige bezwaren

In de twee andere bijdragen over crossnationale data over leerprestaties komen een groot aantal problemen met deze data aan de orde. Uit het voorafgaande zal reeds duidelijk zijn dat die problemen reëel zijn. Maar soortgelijke problemen komen in veel nationale data ook voor, zonder dat er een haan naar kraait. Blijkbaar zijn de belangen bij de uitkomsten van crossnationale vergelijkingen groter, en wordt er aan die problemen plotseling zwaarder getild. Gelukkig geldt zowel voor IEA als PISA het spreekwoord “De honden blaffen, maar de karavaan trekt verder”, want een aantal bezwaren zijn onzinnig. Daarvan wil ik er drie hier bespreken.

1. De PISA-scores komen tot stand door bedrog. Scholen sturen hun slechtste leerlingen weg tijdens de PISA-testen en laten alleen hun beste leerlingen meedoen. Veel empirische bewijzen worden meestal niet aangedragen, maar laten wij voor het gemak aannemen dat scholen inderdaad alleen de betere leerlingen laten deelnemen. Als alle scholen dat in de zelfde mate doen is dat voor de internationale vergelijking van leerprestaties niet bezwaarlijk. In alle landen worden door dit bedrog de leerprestaties in dezelfde mate overschat. Het zou alleen bezwaarlijk zijn als in het ene land de mate van bedrog door scholen groter zou zijn dan in een ander land. Als dat waar zou zijn, is de logische conclusie dat Finse scholen het best zijn in dit bedrog. Ook zouden scholen in Noord-Italië beter zijn in dit bedrog dan scholen in Zuid-Italië. Het lijkt allemaal niet erg waarschijnlijk. Het is plausibeler om uit te gaan van de aanname dat scholen in de OESO-landen een vergelijkbaar niveau van bedrog hebben. Wel kunnen de verschillen in de omvang van het speciaal onderwijs in de deelnemende landen de vergelijkbaarheid beïnvloeden. Leerlingen in het speciaal onderwijs behoren niet tot de onderzoekspopulaties van PISA en IEA, maar de percentages leerlingen in het speciaal onderwijs variëren wel van land tot land. De hoge score van Nederland in de PISA-meting van 2000 lijkt te wijten aan het feit dat

leerlingen op lom- en mlk-scholen niet meededen (Knecht-Van Eekelen, Gille, & Van Rijn, 2007). Ook kan de feitelijke deelname van 15-jarigen aan het onderwijs niet in overeenstemming zijn met de leerplichtwetgeving in een land. Dit verschil tussen de juridische verplichting tot schoolbezoek en het werkelijke volgen van onderwijs is groot in deelnemende landen zoals Brazilië en Turkije, en in mindere mate ook voor andere landen (OECD, 2007).

2. De uitkomstverschillen tussen landen komen door de onderwijsprestaties van leerlingen met een immigratieachtergrond. Het niveau van het Duitse onderwijs is daardoor in PISA en IEA onderschat. Dat lijkt gezien het niet opnemen van het herkomstland (zie hierboven) een plausibel argument, en inderdaad wordt het verschil in de gemiddelde Duitse en Finse score kleiner, als men rekening houdt met de percentages en de verschillen in herkomst van migrantenleerlingen in die landen. Maar het percentage leerlingen met een immigrantenachtergrond is te klein[7] en de verschillen in scores tussen de inboorlingen zijn te groot, om een substantiële verandering in de onderlinge rangorde te veroorzaken.
3. Uitkomsten voor taal zijn heel anders dan voor wiskunde of natuurwetenschappen. Het lijkt wellicht een voor de handliggend bezwaar, maar dit wordt niet gesteund door de feitelijke uitkomsten met PISA-data. Er blijkt tussen de drie domeinen (taal, wiskunde, natuurwetenschappen) een grote gezamenlijke variantie te bestaan. Blijkbaar meten de testen voor deze drie domeinen in eerste instantie de manifeste intelligentie van leerlingen en pas in tweede instantie specifieke kennis en vaardigheden in een bepaald leerdomein. Deze gezamenlijke variantie is op het nationaal niveau nog sterker dan op het individuele niveau. Rindermann en Ceci (2009) hebben laten zien dat noch de inhoud van de toets (naar vakgebied), de aard van de getoetste kennis (curriculum of minder curriculum georiënteerd), het jaar van de meting of de onderzoekspopulatie (basisschool- of VO-leerlingen) veel uitmaken voor de internationale rangorde[8]. De IEA-data laten een dergelijke conclusie niet toe, omdat IEA de testen voor de verschillende leerdomeinen niet bij dezelfde leerlingen afneemt. Sinds de meting in 2003 doet PISA dat wel en toen pas werd de grote gezamenlijke variantie zichtbaar. Omdat PISA burgerschapkennis en -vaardigheden niet meet, weten wij niet of deze overlap in variantie ook voor burgerschap geldt. Maar de effecten van de gebruikelijke achtergrondkenmerken zijn bij burgerschap niet erg verschillend van die van taal, wiskunde of natuurwetenschappen.

6 Gebruik van crossnationale data in Nederland

Alvast in Nederland is er onvoldoende wetenschappelijk toezicht op de PISA-dataverzameling. Alleen het Nederlandse Ministerie van Onderwijs is verantwoordelijk voor de dataverzameling (die in opdracht wordt uitgevoerd door het Cito) zonder overleg met onafhankelijke wetenschappers. Dat wreekt zich zowel bij de beslissingen over de invulling van de vragen (het beste voorbeeld hiervan is het niet-opnemen van het geboorteland in de metingen van 2003 en 2006[9]). Bij IEA lijkt de academische inbreng groter, maar ook hier heeft het ministerie van onderwijs een grote rol (lid van het *standing committee* en nationaal vertegenwoordiger in de algemene IEA-vergadering).

Daar komt bij dat verschillende politici (zoals Ritzen of Netelenbos) verwijzen naar de hoge scores van Nederlandse scholieren in de internationale rangordes van IEA en PISA om het succes van hun politiek te bewijzen. Maar zij hebben nooit de opdracht gegeven voor een onafhankelijk onderzoek naar de verklaring van de relatief hoge scores van Nederlandse scholieren in vergelijking met die van de andere Europese landen. Blijkbaar vonden deze politici het niet nodig hun gelijk empirisch te onderbouwen, of wilden zij niet weten waaraan Nederlandse leerlingen die hoge score te danken hebben. Daarom stelden zij geen enkele beurs voor postdocs of aio's beschikbaar voor een dergelijke secun-

daire analyse van de PISA- of IEA-data.

De grote invloed die nationale regeringen hebben op IEA en PISA is waarschijnlijk de onvermijdelijke prijs die de wetenschappelijke wereld moet betalen voor deze kostbare dataverzameling. Maar het rechtvaardigt niet de afwezigheid van wetenschappelijk betrokkenheid bij de dataverzameling en de analyses. Echter wij kunnen niet alleen de overheid de schuld hiervan geven. Ook de beroepsvereniging is op dit punt niet actief genoeg. Ik herinner mij geen serieuze aandacht aan internationale vergelijkingen op bijeenkomsten van de VOR. De VOR heeft nooit geprotesteerd tegen de machtspositie van het Ministerie van Onderwijs in PISA en IEA. Ook bestaat er geen afzonderlijk NWO- of PROO-programma meer voor de analyse van de crossnationale data van IEA of PISA.

Het is daarom niet verbazingwekkend dat de wetenschappelijke oogst van analyses van de IEA- en PISA data, die verder gaan dan een beschrijving van de primaire uitkomsten tegenvalt. Onderwijskundigen en sociologen, die – gegeven de mogelijkheden van crossnationale data voor hun belangrijkste onderzoeksvragen – het voortouw hadden moeten nemen, hebben het vaak laten afweten. Die rol lijkt overgenomen te zijn door economen, die de PISA- en de IEA-data ontdekt lijken te hebben. Zelfs in Nederland is dit goed te illustreren: het door economen beheerste Centraal Planbureau heeft meer met de PISA-data gedaan dan het door sociale wetenschappers gedomineerde Sociaal en Cultureel Planbureau. Er is dus voor Nederlandse onderzoekers nog een wereld te winnen.

Noten

1 Men is pas afgelopen jaar begonnen met een dergelijke nationale dataverzameling, die door zowel de *Bund* als de *Länder* betaald wordt, maar waarbij rechtstreekse vergelijking van onderwijsresultaten van deelstaten niet toegestaan is.

2 Schotland vraagt wel naar geboorteland, maar dat komt omdat het Schotse onderwijs buiten de eeuwigdurende Unie met Engeland is gebleven, net als de Schotse rechtspraak.

3 Vanaf de PISA-wave 2009 vraagt Nederland wel naar het specifieke geboorteland van leerlingen en ouders, dankzij een intensieve lobby.

4 Nederland doet wel mee aan de ICCS-wave 2009.

5 Het is even waarschijnlijk dat in een bepaalde school het normale positieve effect van ouderlijk milieu op onderwijsprestaties niet zou bestaan maar dat op die bepaalde school juist een tegengesteld negatief effect zou voorkomen.

6 Overigens zijn de extra data en variabelen die Duitsland in elke PISA meting toevoegde, nog slechter bereikbaar voor onafhankelijke onderzoekers. Een gedetailleerd onderzoeksvoorstel wordt anoniem gerefereerd, en afwijzing is een realistische uitkomst, bij voorbeeld omdat analyse van effecten van herkomstlanden in Duitsland nog uit den boze is. Als een andere onderzoeker al met een verwante analyse bezig is, wordt er ook geen toestemming geven voor een andere secundaire analyse. Analyses die neerkomen op een vergelijking van onderwijsstelsels van deelstaten zijn ook niet toegestaan. Het kan dus altijd nog erger dan in Nederland. De webpagina van dit Duitse data-archief geeft meer (maar te optimistische) informatie over het verwerven van de Duitse PISA-data: www.iqb.hu-berlin.de\FDZ.

7 PISA en IEA meten door het gebruik van geboorteland van leerlingen en ouders bovendien alleen maar eerste en tweede generatie, en de derde generatie blijft onzichtbaar. Op de lange termijn zal men toch moeten overgaan tot de meting van subjectieve etnische identiteit, zoals ook in de VS gebruikelijk is.

8 Als dit juist is, vereist dit wel een nieuwe doordenking van de validiteit van de taal, wiskunde en natuurwetenschappen toetsen. Wat meten wij met de PISA- en IEA-toetsen? Manifeste intelligentie? (Rindermann & Ceci, 2009).

9 Het geboorteland is in de meting van 2009 wel gemeten in Nederland, maar dat is het gevolg van langdurig lobbyen.

Literatuur

Dronkers, J., & Avram, S. (2010). A cross-national analysis of the relations of school choice and effectiveness differences between private-

dependent and public schools. *Educational Research and Evaluation, 16*,151-175.

Dronkers, J., & Robert, P. (2004). De effectiviteit van openbaar en bijzonder onderwijs: een crossnationale analyse. *Mens en Maatschappij, 79*, 170-192.

Hanushek, E., & Wössmann, L. (2006). Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *Economic Journal, 11*, C63-C76.

Heus, M. de, & Dronkers, J. (2010). De schoolprestaties van immigrantenkinderen in 16 OECD-landen. De invloed van onderwijsstelsels en overige samenlevingskenmerken van zowel herkomst als bestemmingslanden. *Tijdschrift voor Sociologie, 30*, 260-294

Jakubowski, M. (2010). Institutional tracking and achievement growth: Exploring Differences-in-Differences approach to PIRLS, TIMSS, and PISA data. In J. Dronkers (Ed.). *Quality and inequality of education. Cross-national perspectives* (pp. 41-82). New York: Springer.

Knecht-van Eekelen, A., Gille, E., & Rijn, P. van. (2007). *Resultaten PISA-2006. Praktische kennis en vaardigheden van 15-jarigen. Nederlandse uitkomsten van het OESO Programme for International Student Assessment (PISA) op het gebied van natuurwetenschappen, leesvaardigheid en wiskunde in het Jaar 2006*. Arnhem, Nederland: CITO.

Levels, M., Dronkers, J., & Kraaykamp, G. (2006). Het belang van herkomst en bestemming voor de schoolprestaties van immigranten. Een crossnationale vergelijking. In F. van Tubergen & I. Maas (red.). *Allochtonen in Nederland in internationaal perspectief* (pp. 137-160). Amsterdam: Amsterdam University Press.

Luyten, H. (2006). An empirical assessment of the absolute effect of schooling: regression-discontinuity applied to TIMSS-95. *Oxford Review of Education*, *32*, 397-429.

Martens, K., Rusconi, A., & Leuze, K. (Eds.) (2007). *New arenas of education governance – The impact of international organisations and markets on educational policymaking*, London: Palgrave

Meyer, J. W., Boli, J., Thomas, G. M., & Ramirez, F. O. (1997). World society and the nation state. *American Journal of Sociology, 103*, 144-181.

OECD. (2004). *Learning for tomorrow's world: First results from PISA 2003*. Paris: OECD.

OECD. (2006). *Where immigrant students succeed. Pisa 2003*. Paris: OECD.

OECD. (2007). *PISA 2006 Science Competencies for Tomorrow's World. Analysis*. Paris: OECD.

Rindermann, H., & Ceci, S. J. (2009). Educational policy and country outcomes in international cognitive competence. *Perspectives on Psychological Science, 4*, 551-568.

Manuscript aanvaard: 23 januari 2011

Auteur

Jaap Dronkers is als hoogleraar werkzaam aan het Researchcentrum voor Onderwijs en Arbeidsmarkt (ROA) van de Universiteit Maastricht.

Correspondentieadres: Researchcentrum voor Onderwijs en Arbeidsmarkt (ROA), Universiteit Maastricht, Postbus 616, 6200 MD Maastricht. E-mail: j.dronkers@maastrichtuniversity.nl.