

De toegevoegde waarde van vroeg testen

Paul Leseman

Inleiding

Hoewel de huidige minister van onderwijs geen grote daden-drang verweten kan worden als het gaat om het oplossen van schrijnende problemen in het onderwijs, zoals de versterkte segregatie om er maar eens een te noemen, dreigt op twee samenhangende punten ingrijpend beleid. Ondanks brede weerstand lijkt de minister vast te willen houden aan twee ideeën uit de vorige kabinetsperiode, namelijk de invoering in het basisonderwijs van een 'kleutertoets' en de invoering van een systeem om de 'toegevoegde waarde van scholen' te bepalen. De samenhang tussen deze ideeën is als volgt. Om de toegevoegde waarde van een school te kunnen bepalen is het belangrijk een beginmeting te hebben. De instroom verschilt immers enorm tussen scholen, en alleen afgaan op het eindresultaat miskent dat dit voor een deel terug te voeren is op de kenmerken van de leerlingen en hun thuismilieu. Het ligt voor de hand een kleutertoets als beginmeting te nemen, aangezien de basisschool in het Nederlandse bestel immers op vierjarige leeftijd begint.

Daarnaast, en misschien vooral, zou de kleutertoets een rol moeten gaan spelen in de toekenning van onderwijsachterstands- en zorgmiddelen aan basisscholen. Zo wordt het huidige systeem van toekenning van achterstandsmiddelen, gebaseerd op het sociaal-economische leerlinggewicht, als onbevredigend ervaren. Het is te weinig specifiek (er worden middelen gegeven voor leerlingen die dat helemaal niet nodig hebben) en te weinig sensitief (er worden geen extra middelen gegeven voor leerlingen die wel extra steun nodig hebben, maar die sociaal-economisch of etnisch gezien niet tot de achterstandsgroepen worden gerekend). Efficiëntere toedeling van middelen is niet alleen een sympathiek doel in economisch krappe tijden, maar ook nodig omdat het gewichtensysteem niet aantoonbaar effectief is, zo weten we al jaren.

In de discussies die recentelijk over deze plannen zijn gevoerd, in de Tweede Kamer, in adviescommissies en onder vakgenoten, is het idee van de kleutertest min of meer gesneuveld, maar is het idee van de toegevoegde waarde overeind gebleven met mijns inziens onwenselijke aanpassingen. Ik pleit ervoor juist het omgekeerde te doen. Een goed systeem van vroegte, op een breed spectrum van

Forum

Paul Leseman is als hoogleraar verbonden aan de Capaciteitsgroep Algemene Pedagogiek en Orthopedagogiek van de Universiteit Utrecht

Correspondentieadres: Paul Leseman, Capaciteitsgroep Algemene Pedagogiek en Orthopedagogiek, Postbus 80140, 3508 TC Utrecht. E-mail: p.p.m.leseman@fss.uu.nl

relevante vaardigheden en persoonlijkheidskenmerken gerichte systematische beoordeling, dat aanknopingspunten biedt voor preventief en vroeg-remediërend handelen, kan van grote waarde zijn voor het onderwijs. Aan een systeem van toegevoegde waarde om scholen 'objectief' op hun prestaties af te rekenen kleven echter te veel bezwaren. Laten we dit idee snel vergeten. Ik licht mijn standpunt toe.

De kleutertest en het stabiliteitsargument

Er zijn verschillende adviezen uitgebracht over de voornemens van de regering. Een commissie onder leiding van de voormalige Leidse hoogleraar ontwikkelingspsychologie Kohnstamm adviseerde enkele jaren geleden al negatief over de invoering van een kleutertoets. Aan dit advies lag een analyse van De Jong, Knol en Van der Leij (1996) ten grondslag van data uit een cohort-onderzoek met leerlingen van verschillende sociaal-economische en etnisch-culturele herkomst. De verschillende tests (onder andere Nederlandse woordenschat, non-verbale intelligentie, fonologische vaardigheden) die in dit onderzoek op vierjarige leeftijd en op verschillende momenten daarna waren afgenomen, bleken afzonderlijk en in combinatie slechts matig sterke voorspellers van de schoolprestaties in groep 3 te zijn. Veel leerlingen die later bleken uit te vallen en extra zorg nodig hadden, konden met de toetsen niet al vroeg als risicoleerling worden geïdentificeerd. Omgekeerd bleken leerlingen die bij de beginmetingen in ongunstige zin opvielen vaak 'spontaan herstel' te vertonen. Met andere woorden: de predictieve validiteit, sensitiviteit en specificiteit van de onderzochte toetsen was op z'n best matig. Met betrekkelijk eenvoudige tests kom je er dus kennelijk niet. Een rationele grondslag voor herverdeling van schaarse achterstands- en zorgmiddelen bieden ze, zo bezien, zeker niet.

Het is achteraf te betreuren dat het buiten de opdracht van De Jong e.a. viel om na te gaan of er met andere instrumenten en beoordelingsprocedures misschien betere resultaten geboekt zouden kunnen worden. Nu is het beeld ontstaan dat toetsing, of breder: systematische op theorie gestoelde beoordeling, op jonge leeftijd helemaal niet nuttig is voor de onderwijspraktijk. Het belangrijkste argument is dat de cognitieve vaardigheden van jonge kinderen, en mogelijke problemen in de ontwikkeling daarvan, nu eenmaal nog te weinig stabiel, maar juist "zo veranderlijk als het leven zelf" zijn, zoals de titel van het rapport van de commissie Kohnstamm luidt. Vanuit een klassiek psychometrisch gezichtspunt klopt dit stabiliteitsargument misschien, maar hoe geldig is dit gezichtspunt? Is het veranderlijke en veranderbare van jonge kinderen nu juist niet het meest interessante vanuit een pedagogisch-didactisch gezichtspunt?

Wanneer een kind wordt getest, bijvoorbeeld begin groep 1, met een traditionele, eenmalig afgenomen test – een intelligentietest, een woordenschattest, een rekenbegrippentest – worden veel verschillende dingen tegelijk gemeten. Dit is onlangs helder uit de doeken gedaan door Sternberg en Grigorenko (2001). Je meet onder andere het *expliciete voorafgaande leren* van het kind dat het resultaat is van zijn of haar 'leervermogen' én het aanbod aan leerervaringen tot dan toe. Het laatste hangt bij jonge kinderen sterk samen met de sociaal-economische, culturele en pedagogische kenmerken van het gezin. Wat onder 'leervermogen' verstaan moet worden is trouwens ook niet eenvoudig. Het is zeker niet alleen maar de voornamelijk door aanleg bepaalde 'zuivere intelligentie' van een kind. Ook zintuiglijke, motorische en cognitieve uitvoeringsvaardigheden zoals aandacht en werkgeheugen, en emotionele kenmerken beïnvloeden wat een kind leert van een aanbod. Voorafgaand leren voorspelt de latere schoolprestaties rede-

lijk goed. Dat komt niet alleen doordat het leervermogen als een tamelijk stabiele eigenschap in de testcores tot uitdrukking komt. Het is ook een gevolg van het feit dat de voorafgaande leerervaringen in de gezinscontext meestal een goede aanwijzing geven van de toekomstige ondersteuning die het gezin het kind zal bieden. Ten overvloede zij vermeld dat stabiliteit hier betekent: stabiliteit van de *grote verschillen in ondersteuning* die gezinnen bieden.

Het is interessant het punt van de stabiliteit en voorspellende waarde verder te exploreren. Onderzoekers hebben geprobeerd tests van het leervermogen te ontwikkelen die minimaal gevoelig zijn voor voorafgaand expliciet leren en daarmee voor verschillen in leerervaringen naar sociaal milieu, om zo zuiver mogelijk het leervermogen – de intelligentie in strikte zin – te kunnen meten. Met één type test, de elementaire cognitieve decisietaak (ECT), komt men dicht in de buurt. Zulke tests correleren hoog met metingen van de snelheid van de neurologische informatieverwerking, zijn al vanaf heel jonge leeftijd zeer stabiel en hebben een hoge erfelijkheidsfactor h^2 , omdat er door de testprocedure slechts minimale invloed is van leerervaringen, sociaal milieu, uitvoeringsvaardigheden en emotionele reacties. Het probleem is echter dat testresultaten met ECT's slechts zwak correleren met gewone intelligentietests (correlaties in de orde van grootte van .30) en nog minder met schoolprestatieingen. Voor de milieufactor is dit precies omgekeerd. Door de meting van de informele leerervaringen in het gezin te verbeteren, bijvoorbeeld door gegevens over het opleidingsniveau van de ouders te combineren met observatiegegevens van sociale interacties in het gezin, worden de verbanden met gewone intelligentietests en vooral met latere schoolprestaties juist sterker. In eigen onderzoek vond ik correlaties van gecombineerde metingen van rond de .50 met technisch lezen tot .60 met begrijpend lezen einde groep 3; de correlatie met de Nederlandse woordenschat in groep 3 was zelfs .80. Misschien is dat wat stabiel blijkt bij een vroege kleutertest in relatie tot latere schoolprestaties vooral te danken aan de stabiliteit van de gezinscontext.

Impliciet leren en dynamische tests

Met een traditionele test wordt volgens Sternberg en Grigorenko ook het zogenaamde impliciete leren tijdens de testafname gemeten. Een kind leert als het ware tussen de regels door wat van hem of haar wordt verwacht en hoe bepaalde items aangepakt moeten worden. In ontwikkelingspsychologische termen is er in de testsituatie sprake van 'micro-ontwikkeling van cognitieve vaardigheid'. Doe je een tweede testafname onmiddellijk na de eerste, dan zullen de scores in de meeste gevallen hoger zijn: het kind heeft van de eerste test geleerd, oplossingsstrategieën bedacht, het heeft, kortom, een hoger vaardigheidsniveau ontwikkeld. Dat vaardigheidsniveau is misschien nog instabiel en zal na korte tijd vervallen, maar geeft wel een potentie aan. Dit heet *impliciet* leren, omdat het in een traditionele testsituatie juist *niet* de bedoeling is dat kinderen van de test leren. Impliciet leren is daarom een speciaal soort leren in een heel speciaal soort leersituatie. De leertaak is vreemd en staat wat inhoud en vorm betreft meestal ver af van de toekomstige leersituaties in de klas. De 'didactische begeleiding' is bovendien bijna absurd, want de testleider moet een neutrale houding aannemen, en alleen standaard aanwijzingen en vooral geen inhoudelijke of strategische informatie geven. De testleider stelt een arbitraire tijdslimiet en brengt – bewust of onbewust – prestatiedruk in. Het impliciete leren bij een testafname is daarmee het resultaat van een complex dynamisch samenspel van verschillende situationele en psychologische factoren: het leervermogen van het kind, de aard van de test (vorm en inhoud), het executieve functioneren van het kind (aandacht,

werkgeheugen, inhibitie) en de emotionele reactie van het kind op de testsituatie.

Hoe kinderen leren in een testsituatie, hoe het proces van 'micro-ontwikkeling' verloopt, is in principe erg interessant. Het kan namelijk een beeld geven van hoe kinderen zich in de onderwijssituaties nadien zullen ontwikkelen. Of ze bijvoorbeeld hun aanvankelijke achterstand snel of juist moeizaam zullen inlopen, of hun leerresultaten binnen één leerstofdomein blijven of juist vanzelf generaliseren naar andere leerstofdomeinen. De veronderstelling is dat de micro-ontwikkeling gelijkenis heeft met de macro-ontwikkeling van vaardigheden over langere perioden. Bewijs hiervoor wordt gerapporteerd door Siegler en Svetina (2002). Hoe kinderen leren in een testsituatie is in principe vooral interessant, omdat het informatie kan geven over de vraag hoe een bepaald kind pedagogisch-didactisch het beste ondersteund kan worden bij het toekomstige leren. Sommige kinderen hebben bijvoorbeeld meer tijd nodig voor leertaken, maar kunnen goed zelfstandig werken en komen uiteindelijk op een adequaat niveau. Bij andere kinderen moet de aandacht sterker gestuurd worden door een directieve leerkracht die hen voortdurend controleert en aanspoort. Bij weer andere kinderen is het van groot belang dat de leerkracht veilige sociale situaties creëert en prestatiedruk zoveel mogelijk vermijdt.

Het probleem van een traditionele test als 'kleutertest' is onmiddellijk duidelijk. Zo'n test onderscheidt niet tussen voorafgaand expliciet leren en het impliciete leren in de testsituatie, onderscheidt niet tussen kindgebonden (neurobiologische) kenmerken, zoals het elementaire leervermogen, en gezinskenmerken, onderscheidt niet tussen cognitieve, executieve en emotionele kindkenmerken. De traditionele testsituatie, als een situatie van impliciet leren, is bovendien met de beste wil van de wereld geen goede afspiegeling te noemen van de reële leertaken en pedagogisch-didactische begeleiding waarmee een kind daarna te maken zal krijgen, en is daarom maar zeer beperkt informatief. Een traditionele test of testbatterij als kleutertest moeten we afwijzen, daar schiet niemand wat mee op. Maar alternatieven zijn denkbaar en zouden gebruikt kunnen worden om de onderwijskwaliteit te verbeteren.

Er is in dit verband ervaring opgedaan met zogenaamde leerpotentieeltests, waarbij het leren tijdens de test door gestandaardiseerde hulp te bieden of door tussen twee testmomenten vaardigheden te trainen, als het ware is geformaliseerd. Leerpotentieelscores zouden een betere indruk kunnen geven van het leervermogen dan traditionele statische tests. Helaas is deze belofte nog niet ingelost, aldus Sternberg en Grigorenko. Zo blijkt in het algemeen de voorspellende en diagnostische waarde niet veel beter te zijn dan die van een traditionele test. Een van de redenen daarvoor is dat veel typen leerpotentieeltests inhoudelijk niet anders zijn dan een gangbare intelligentietest. Ze staan dus ook 'op afstand' van de toekomstige leertaken en de sociale situaties waarin die leertaken gedaan moeten worden. Het belangrijkste verschil is dat er hulp wordt gegeven. Ook over die hulp is echter discussie mogelijk. In één variant worden bij testitems die niet goed worden beantwoord, aanwijzingen gegeven volgens tevoren vastgelegde formuleringen. In een andere variant wordt een uniform, op de testinhoud afgestemd trainingsprogramma (bijvoorbeeld inductief redeneren) aangeboden tussen een voor- en natest. In beide gevallen is de belangrijkste kritiek dat de geboden hulp niet optimaal op de veelvoudige verschillen tussen kinderen is afgestemd.

Veelbelovend, gelet op voorgaande beschouwing, zijn de zogenaamde dynamische beoordelingsprocedures ('dynamic assessments'). Hierbij wordt op een systematische manier onderzocht hoe kinderen in verschillende authentieke - 'schoollevensechte' - pedagogisch-didactische situaties relevante leertaken uitvoeren, hoe ze reageren op gevarieerde instructie, en hoe het leerproces, gezien

als micro-vaardigheidsontwikkeling, in zo'n authentieke situatie verloopt. De beoordelingsprocedure is *dynamisch* omdat de aandacht gericht wordt op de veranderlijkheid van kinderen als gevolg van variaties in taken, situaties en didactische ondersteuning. Wat geldt als authentiek en relevant is ten dele afhankelijk van de leeftijd, het ontwikkelingsniveau en daarmee samenhangend de schoolfase van het kind. Een serie proefleeslessen is op vierjarige leeftijd om verschillende redenen nog niet aan de orde, maar zou in de tweede helft van groep 2 een prima situatie kunnen zijn om te beoordelen hoe kinderen individueel zullen reageren op de leesinstructie in groep 3 en waar ze, preventief, eventueel extra ondersteuning bij nodig hebben. Wat geldt als authentiek en relevant dient daarom voor elke ontwikkelingsfase afzonderlijk bepaald te worden. De latere criteriumvaardigheden zoals lezen, schrijven en rekenen in groep 3 vormen het referentiepunt. Inzicht in de ontwikkelingstrajecten die tot deze criteriumvaardigheden leiden is hierbij van groot belang. Daarover is inmiddels veel bekend. Het voert echter buiten het bestek van dit artikel om er in detail op in te gaan.

Inderdaad, alle leerkrachten in groep 1 beoordelen nieuwe leerlingen, ze observeren ze in verschillende situaties, vormen zich een oordeel over deze leerlingen en hun gezinsachtergrond, en passen eventueel (maar lang niet altijd) hun pedagogisch-didactisch handelen aan. De vraag is echter hoe systematisch dit gebeurt, of de leerkrachtbeoordelingen goed geworteld zijn in actuele wetenschappelijke kennis en of er een directe relatie is met adaptief pedagogisch-didactisch handelen. In toenemende mate worden door leerkrachten voor vroege beoordeling tests uit een leerlingvolgsysteem gebruikt. Dat zijn meestal eenvoudige traditionele tests die, zoals ik betoogd heb, pedagogisch-didactisch te weinig informatief en in de traditionele psychometrische zin onvoldoende sensitief en specifiek zijn. Mijn vermoeden is dat hier met dynamische beoordelingsprocedures veel winst te behalen is.

Toegevoegde waarde en het stabiliteitsargument

Ik kom nu terug op het bepalen van de toegevoegde waarde van scholen. Ook hierover is een advies uitgebracht. De Onderwijsraad adviseert na wikken en wegen en met veel relativeringen positief, maar stelt enkele aanpassingen voor. Een van de belangrijkste daarvan is de beginmeting niet al in groep 1 (met een 'kleutertoets'), maar pas op zijn vroegst begin groep 3 uit te voeren. Eerder meten leidt tot grotere onbetrouwbaarheid in het bepalen van toegevoegde waarde, omdat de voorspellende waarde van vroege toetsen voor latere schoolprestaties beperkt is. De bevindingen van De Jong e.a. spelen een belangrijke rol in de afweging van de Onderwijsraad. De keuze van begin groep 3 als meetmoment is echter tamelijk arbitrair, omdat de leerlingen begin groep 3 nog maar in het beginstadium verkeren van de ontwikkeling van de later geteste schoolvaardigheden. Begrijpend lezen bijvoorbeeld is nog niet aan de orde aan het begin van groep 3, omdat het technisch lezen nog niet is verworven; zelfs eind groep 3 is voor begrijpend lezen nog een lastig ijkpunt. Een groot bezwaar is dat de onderwijsperiode voorafgaande aan de beginmeting (groep 1 en 2; en bij steeds meer scholen ook een voorschoolperiode) simpelweg niet mee beoordeeld wordt, terwijl er alle reden is juist deze periode te beschouwen als een periode waarin preventieve en remediërende interventies, adaptieve instructiestrategieën in algemene zin, nuttig en mogelijk zijn, zoals ik hiervoor heb betoogd.

De paradox wordt nog duidelijker als bedacht wordt dat de betrouwbaarheid van de toegevoegde waarde per school groter is, naarmate de correlatie (op leerling-niveau) tussen begin- en eindmeting sterker is, terwijl tegelijkertijd de ver-

schillen tussen scholen daardoor steeds kleiner worden. Zo blijkt uit analyses van het Cito dat schoolprestatie metingen met toetsen uit het Cito-Leerlingvolgsysteem medio groep 4 beter als beginpunt genomen kunnen worden om de scores in de Eindtoetsen te voorspellen en om betrouwbaar de toegevoegde waarde van scholen te kunnen bepalen dan scores in groep 3. Pas vanaf groep 4 is er, volgens de psychometrische redenering, bij begin- en eindmeting sprake van dezelfde onderliggende vaardigheden op de gebieden van lezen en rekenen (Wijnstra, Ouwens & Béguin, 2003). De consequentie is echter dat de variantie tussen scholen, die wijst op verschillen in toegevoegde waarde, halveert. Hoe betrouwbaarder de beginmeting, des te geringer de verschillen in toegevoegde waarde, zo lijkt het.

Een bijkomend probleem is het volgende. In het zogenaamde schooleffectiviteitsonderzoek is geprobeerd de variantie tussen scholen te verklaren uit een groot aantal onderwijskundige kenmerken, zoals bijvoorbeeld de tijd die aan instructie wordt besteed, de methoden die worden gebruikt, de teamgeest en het onderwijskundig management van de schoolleider. Tot op heden kon echter nooit meer dan de *helft*, meestal minder, van die tussen-schoolvariantie verklaard kon worden. Bovendien waren de effecten van de onderzochte variabelen vaak niet eenduidig. Anders gezegd: de *validiteit* van de toegevoegde waarde, als indicator van onderwijskundige kwaliteit, is twijfelachtig. Voor scholen die zich onder de streep bevinden en dolgraag hun toegevoegde waarde willen vergroten, biedt de onderwijskunde helaas weinig houvast.

Staan we nog eens stil bij de psychometrische redenering. Is het meten van beginvaardigheden in groep 3 of 4 nu echt zoveel anders dan in de kleuterleeftijd? Ook een toets technisch lezen of begrijpend lezen of een tempotoets rekenen doet een beroep op een complexe structuur van taal- en cognitieve vaardigheden, executieve vaardigheden en emotionele zelfregulatie. De toets meet niet alleen het voorafgaande leren, als resultante van deze vaardigheden, de instructie op school en de ondersteuning vanuit het gezin, maar ook hoe een leerling ter plekke de taak oplost. Ik veronderstel nu het volgende. De vaardigheidsontwikkeling is in midden- en bovenbouw van het basisonderwijs *in principe* net zo 'veranderlijk als het leven zelf' als in de onderbouw. Een complexe vaardigheid als begrijpend lezen is in groep 3 of 4 *structureel* anders dan in groep 8, alleen al omdat bijvoorbeeld grammaticale structuren als zinsbouw, voegwoordgebruik en werkwoordsvervoeging in teksten voor groep 8 (en in de eindtoets) veel complexer zijn en omdat de woorden die erin voorkomen, vaak technischer en specialistischer zijn en naar abstractere entiteiten verwijzen. Het is daarom eerder een wonder dan een vanzelfsprekendheid dat toetscores in groep 3 of 4 relatief sterk correleren met de eindmeting in groep 8 en dat verschillen tussen leerlingen, onder andere naar hun sociaal-cultureel milieu, zo stabiel blijken. De vraag is: waar komt die relatief grote stabiliteit vandaan?

Een deel van het antwoord is natuurlijk dat de psychologische eigenschappen van leerlingen in groep 3 of 4 stabiel zijn geworden, met name executieve vaardigheden, zelfbeeld en emotionele zelfregulatie. Ook disfunctionele patronen hierin zijn gestabiliseerd, voorzover ze niet in een eerdere fase zijn voorkomen of gecompenseerd. Een ander deel van het antwoord is dat het onderwijssysteem zelf de stabiliteit creëert. Vanaf groep 3 of 4 vormen de leersituaties - en leertaken daarbinnen - en de instructie die erbij gegeven wordt, een stabiel systeem dat de verdere ontwikkeling van leerlingen op de verschillende vaardigheidsgebieden normatief kanaliseert. Die stabiliteit ontstaat door het gebruik van meerjarige methoden met een geleidelijke opbouw van de complexe vaardigheden die in groep 8 worden getest, door de vaste roosters, lesurentabellen en het jaarklas-

sensysteem waarmee de factor tijd wordt gereguleerd, door de frequente toepassing van methodegebonden toetsen en toetsen uit meerjarige leerlingvolgsystemen, enzovoort. Het systeem is afgestemd op een 'modale meerderheid' van leerlingen die binnen het systeem een 'normale ontwikkeling' tonen. Waarschijnlijk is ons onderwijsstelsel, zo globaal bezien, voldoende tot goed. Er zijn immers, zo nemen we aan, geen aanwijzingen dat basisscholieren in Nederland gemiddeld veel slechter rekenen of lezen dan in andere westerse landen of onderpresteren ten opzichte van algemeen aanvaarde standaards.

Afwijkende ontwikkeling

Voor leerlingen buiten de modale groep, de hoogbegaafden, de 'zwakkere' leerlingen en de leerlingen met specifieke cognitieve of emotionele problemen of met sociaal-culturele achterstanden, is het systeem minder geschikt, omdat de afstemming op de alternatieve ontwikkelingstrajecten van deze leerlingen niet optimaal is. Het aandeel atypisch lerende en ontwikkelende leerlingen kan sterk variëren per klas en school en per cohort, met niet gemakkelijk te voorspellen consequenties voor het eindresultaat. Net als in de onderbouw kunnen aanvankelijke achterstanden soms spontaan verdwijnen als er op een of andere manier een compensatie optreedt, of als er simpelweg extra tijd en geduld is. Soms zal er een relatieve achteruitgang zijn die met de huidige hulpmogelijkheden juist niet te voorkomen is, zelfs als deze maximaal worden ingezet. Soms zou een aangepaste methode met andere inhoudelijke accenten en extra tijd wonderen kunnen doen om aanwezig potentieel te realiseren. Laat ik dit illustreren.

Leerlingen met ernstige leesproblemen, 10% tot 25% van een cohort afgaande op een gangbare classificatie als het D- en E-niveau van het Cito-leerlingvolgsysteem, vormen een interessante groep in dit verband, omdat bij dieper gravend diagnostisch onderzoek blijkt dat deze groep intern erg heterogeen is, met uiteenlopende pedagogisch-didactische behoeften en evenzeer uiteenlopende prognoses voor de latere schoolprestaties en eindtoetsresultaten (Snowling, Bishop & Stothard, 2000). Uit de scores in groep 3 en zelfs groep 4 op gangbare toetsen voor technisch lezen en begrijpend lezen blijkt die heterogeniteit vooralsnog niet; deze toetsen differentiëren te weinig en de scores van de bedoelde groep zijn allemaal laag tot zeer laag.

Leerlingen met uitsluitend problemen in de zogenaamde oppervlakkige fonologische informatieverwerking (bewust herkennen en kunnen manipuleren van fonemen) zullen echter door hun normale woordkennis, taalbegrip, intelligentie en andere vaardigheden naar verhouding goed scoren op taken voor begrijpend lezen in groep 8 en hun problemen met het technische aspect van lezen grotendeels overwinnen. Dit komt door wat wel *top-down* compensatie genoemd wordt. Het begrip van de tekst stuurt als het ware van bovenaf het ontcijferen van de gedrukte woorden via opgeroepen inhoudelijke woordkennis en vergroot zo de snelheid en accuratesse van het technische lezen. De toenemende kennis van de betekenis en het schriftbeeld van woorden maakt deze groep aanvankelijk zwakke lezers bij het lezen minder afhankelijk van het letter-voor-letter analyseren, waar ze niet goed in zijn. Onderwijskundig is het van belang dit alternatieve ontwikkelingstraject te ondersteunen, er tijd voor te geven en er voor te waken dat er secundaire problemen ontstaan, bijvoorbeeld wat betreft motivatie en zelfbeeld en zorgen bij de ouders. De praktijk op scholen is trouwens lang niet altijd in overeenstemming met deze principes, waardoor slechte lezers aan het begin, vaker slechte lezers aan het eind blijven.

Leerlingen met specifieke taalproblemen, op hun beurt, hebben naast fonolo-

gische informatieverwerkingsproblemen ook met een afwijkende ontwikkeling van de grammaticale kennis te kampen. Deze groep zal naar verhouding juist slechter gaan scoren op latere toetsen van begrijpend lezen, ondanks een normale non-verbale intelligentie en ondanks het feit dat het grammaticale probleem aanvankelijk niet opvalt. Dit komt omdat de leesteksten vanaf groep 4 taalkundig steeds moeilijker worden en voor tekstbegrip en het verwerven van kennis uit de tekst een steeds groter beroep doen op grammaticale vaardigheden. Dit effect zal bovendien uitstralen naar andere leergebieden, zoals rekenen-wis-kunde waarop aanvankelijk normaal gepresteerd werd, omdat in dit vaardigheidsgebied vergelijkbare structurele veranderingen optreden tussen groep 3 en groep 8.

Van huis uit anderstalige allochtone leerlingen scoren gemiddeld ook laag op leestoetsen in groep 3 of 4, hoofdzakelijk door achterstand in kennis van het Nederlands in alle facetten (klankstructuren, grammatica, woordkennis, tekstorganisatie). Door de intensieve, kwalitatief goede instructie in technisch lezen, die kenmerkend is voor het onderwijs in de onder- en middenbouw van Nederlandse basisscholen, zijn de prestaties op toetsen voor technisch lezen na verloop van tijd verhoudingsgewijs goed. Tijd en intensieve instructie compenseren de aanvankelijke achterstand op dit gebied. Door de onderdompeling in het Nederlands in de klas verdwijnt ook de achterstand in woordenschat, althans in de zogenaamde oppervlakkige woordenschat voor alledaagse interpersoonlijke communicatie. Maar met begrijpend lezen en parallel daaraan met de verwerking van een diepe en specialistische woordenschat gaat het verhoudingsgewijs minder goed. Ook dit heeft waarschijnlijk te maken met het toenemende beroep dat in leestaken gedaan wordt op de kennis van specialistische woorden en complexe grammaticale structuren, die kenmerkend zijn voor formele, geschreven taal. Daarin wordt op Nederlandse basisscholen zeer waarschijnlijk onvoldoende instructie gegeven. Het zit niet expliciet in de gebruikte methoden. De extra tijd die voor onderricht hierin nodig is voor deze groep, kan niet binnen de beschikbare uren vrijgemaakt worden zonder andere leergebieden in het gedrang te brengen. Er is op dit vlak in het algemeen weinig ondersteuning vanuit het gezin. Bovendien wordt op scholen nog nauwelijks beseft dat dit een kern van het probleem zou kunnen zijn.

De heterogeniteit van leerlingen naar taal-, cognitieve, executieve en emotionele vaardigheden, en naar hulpbronnen thuis, vormt een fundamenteel probleem voor het bepalen van de toegevoegde waarde van scholen en voor de onderwijskundige interpretatie ervan. Het staat namelijk niet vast dat de effecten van het geboden onderwijs voor iedere leerling hetzelfde zijn. Dat is mijns inziens in de midden- en bovenbouw niet wezenlijk anders dan in de onderbouw. Het statistische stabiliteitsargument is om verschillende redenen geen goed argument, zoals ik heb betoogd. Suboptimale ondersteuning van afwijkende leer- en ontwikkelingstrajecten zal in het algemeen de voorspellende waarde van de beginmeting versterken, maar niet uit alle kinderen halen wat erin zit. Spontane vooruitgang door compensatie of onvermijdelijke achteruitgang vanwege veranderende taakeisen, maakt het toerekenen van schooleffecten hachelijk. Het tegenargument dat de heterogeniteit van leerlingen wellicht beperkt is, omdat bijvoorbeeld het aandeel zeer slechte lezers misschien maar klein is (de officiële prevalentie van dyslexie is ongeveer 4% en van specifieke taalstoornissen ongeveer 5%; het aandeel anderstalige kinderen nadert echter op sommige scholen de 100%), miskent dat het hier om pragmatische, maar arbitraire classificaties gaat, waarachter een veel grotere heterogeniteit schuil kan gaan.

Tot slot

Het lijkt bij dit alles overigens geen twijfel dat sommige scholen beter zijn dan andere, dat sommige leerkrachten er meer van maken dan andere leerkrachten en dat sommige groepen leerlingen op de ene school beter af zijn dan op de andere. Om daarin inzicht te krijgen is het van belang preciezer te weten welke pedagogisch-didactische behoeften bestaan en of daaraan, naar beste weten met de pedagogisch-didactische middelen die beschikbaar zijn, wordt tegemoetgekomen door leerkracht en school, waar nodig in samenwerking met ouders en zorginstellingen. Goed onderwijs begint al vroeg: goed inzicht in de verschillen tussen kinderen en hun behoeften moet al vroeg worden gewonnen en benut. De vraag is waar de onderwijspraktijk meer behoefte aan heeft: aan een systeem van outputsturing, zodat er meer druk op de scholen uitgeoefend kan worden om binnen de school uit te vinden hoe het beter kan, waarbij gangbare onderwijskundige kennis weinig houvast biedt, of aan een systeem van ondersteuning van de dagelijkse onderwijspraktijk en van verdere professionalisering van het pedagogisch-didactisch handelen. Ik kies voor het laatste.

Literatuur

- De Jong, P.F., Knol, D.L. & Leij, A. van der, (1996). *Predictie van onderwijsachterstanden. Rapport voor de Commissie Indicatiestelling 4-jarigen*. Amsterdam: Vrije Universiteit, Faculteit Psychologische en Pedagogische Wetenschappen.
- Siegler, R.S. & Svetina, M. (2002). A microgenetic/cross-sectional study of matrix completion. *Child Development, 73*, 793-809.
- Snowling, M., Bishop, D.V.M. & Stothard, S.E. (2000). Is preschool language impairment a risk factor for dyslexia in adolescence? *Journal of Child Psychology and Psychiatry, 41*, 587-600.
- Sternberg, R.J. & Grigorenko, E. (2001). All testing is dynamic testing. *Issues in Education, 7*, 137-170.
- Wijnstra, J., Ouwens, M. & Béguin, A. (2003). *De toegevoegde waarde van de basisschool*. Arnhem: Cito-groep. (intern rapport)