Samenvatting (Summary in Dutch)

Introductie

Sinds de kleuterschool in 1985 bij het basisonderwijs is gevoegd als groep 1 en 2 is er frictie ontstaan tussen de van oudsher ontwikkelings-georiënteerde benadering van het kleuteronderwijs en de programma-georiënteerde benadering in het primair onderwijs. De eerste benadering legt de nadruk op kind gestuurd en speels leren, met observatie als de belangrijkste methode om ontwikkeling te volgen. Daartegenover legt de programma-georiënteerde benadering de nadruk op het behalen van norm-gerelateerde ontwikkelingsdoelen, met vooraf geplande instructie en regelmatige toetsing. Centraal in deze discussie staat het gebruik van de kleutertoetsen van het Centraal Instituut Toets Ontwikkeling (Cito). Zeker in de laatste jaren zijn de spanningen rondom het gebruik van gestandaardiseerde en genormeerde toetsen bij kleuters toegenomen. Zo werd in de Tweede Kamer in 2013 een motie geaccepteerd die het gebruik van een landelijk genormeerde kleutertoets niet langer verplicht stelt. De ontwikkeling van kleuters zou te grillig verlopen om ze betrouwbaar te toetsen. In 2017 besloot de regering dat kleutertoetsen, zoals die van Cito, per 2021 afgeschaft worden. Het toetsen in opgaveboekjes en met name de vorm van normering doet, volgens minister Van Engelshoven, onvoldoende recht aan de sprongsgewijze ontwikkeling van kleuters.

Evenals in Nederland is het gebruik van gestandaardiseerde toetsen in de kleuterjaren een belangrijk onderwerp van discussie in veel landen. Een belangrijke motivatie voor het gebruik van toetsen op deze leeftijd komt voort uit onderzoek wat aantoont dat ontluikende schoolse (i.e. taal- en reken)vaardigheden de sterkste voorspellers zijn voor latere schoolse vaardigheden. Als toetsen gebruikt kunnen worden om problemen in de ontwikkeling van ontluikende schoolse vaardigheden te onderkennen, zouden potentiële problemen op latere leeftijd mogelijk al vroeg ondervangen kunnen worden. Meerdere onderzoeken geven daarbij aan dat vroegtijdig ingrijpen over het algemeen tot betere resultaten leidt dan op een latere leeftijd remediëren. Hoewel dit belangrijke argumenten zijn voor het toetsen van taal- en rekenvaardigheden bij kleuters is betrouwbaar toetsen op deze leeftijd geen gemakkelijke opgave. Zo verloopt de ontwikkeling bij kleuters snel en sprongsgewijs, waardoor testresultaten op één moment mogelijk niet zoveel zeggen over latere prestatie. Daarnaast zijn jonge kinderen over het algemeen minder goed toetsbaar met klassieke meerkeuzetoetsen waarbij individuele prestatie centraal staat: ze zijn vaak niet gewend om in een toetsopstelling met papier en potlood te werken, zijn snel afgeleid, of begrijpen nog niet goed wat van hen verwacht wordt. Tenslotte zijn er op deze leeftijd grote ontwikkelingsverschillen tussen

kinderen, enerzijds door het grillige en snelle verloop van de vroege ontwikkeling en anderzijds door verschillen in de thuiscontext. Hierdoor is het lastig om te bepalen wanneer een vermeende achterstand als voorspeller van een problematische ontwikkeling moet worden gezien. Het risico bestaat daarom dat toetsen op deze leeftijd kinderen ten onrechte als 'risicoleerlingen' identificeren of juist geen problemen in de taal- en rekenontwikkeling laten zien terwijl deze op latere leeftijd alsnog aan het licht komen. Enerzijds kan dit leiden tot stigmatisering en gevoelens van incompetentie bij kinderen, anderzijds krijgen kinderen hierdoor niet de hulp die ze nodig hebben. De vraag die vaak gesteld wordt is dan ook of de resultaten van dit soort toetsen stabiel genoeg zijn om problemen bij individuele kinderen te onderkennen.

Naast een betrouwbare identificatie van risicoleerlingen is het belangrijk dat toetsen informatie bieden dat gebruikt kan worden in de remediëring van een onderkend probleem. In dit opzicht wordt er vaak onderscheid gemaakt tussen twee mogelijke rollen van toetsen. Ten eerste, het bieden van een waardeoordeel dat gebruikt kan worden in de controleerbaarheid van het leerproces. Ten tweede, het bieden van informatie dat gebruikt kan worden voor de verbetering van het leerproces. Deze twee rollen van toetsen zijn veelvoudig terug te vinden in de literatuur en komen onder meer tot uitdrukking in de termen 'summatief' en 'formatief' toetsen. Gestandaardiseerde normatieve toetsen worden vaak gezien als instrumenten van controle, aangezien het waardeoordeel in de vorm van een normscore bij dit soort toetsen vaak centraal staat. Aan de andere kant sluit dit waardeoordeel het pedagogisch gebruik van deze instrumenten niet uit. Zo bieden de toetsen van Cito bijvoorbeeld subscores die inzicht kunnen geven in specifieke typen opgaven waar een kind op uitvalt. De vraag is of deze toetsen voldoende aanknopingspunten bieden voor de leerkracht om passende aanpassingen te maken in het onderwijsaanbod voor individuele kinderen.

Hoewel Cito overtuigend onderzoek heeft gedaan naar de betrouwbaarheid en validiteit van de kleutertoetsen is dit onderzoek voornamelijk cross-sectioneel. Daarbij geeft Cito aan dat onderzoek naar de predictieve validiteit niet nodig is, omdat de toetsen niet bedoeld zijn voor voorspellend gebruik. Hoewel deze instrumenten voornamelijk gemaakt zijn om de ontwikkeling van kinderen te volgen, worden de toetsen in de handleiding omschreven als instrumenten voor het signaleren van problemen in de taal- en rekenontwikkeling ten behoeve van interventie. Het signaleren van problemen in de ontwikkeling heeft een inherent voorspellend karakter. De basis voor signalering aan de hand van een testscore is namelijk niet zozeer gegrond in wat de score zegt over de huidige of voorafgaande ontwikkeling van het kind, maar in wat dit betekent voor de verdere ontwikkeling. Aangezien de meeste instrumenten gemaakt zijn om een beslissing te ondersteunen met verwachte uitkomsten in de nabije of verafgelegen toekomst, vormen deze uitkomsten een belangrijk onderdeel in de evaluatie van een instrument.

Hoewel toetsing in de kleuterjaren kan leiden tot vroegtijdige onderkenning en remediering van problemen zijn er vraagtekens bij de stabiliteit en bruikbaarheid van toetsresultaten op deze leeftijd. In dit proefschrift proberen we te achterhalen hoe scores op de kleutertoetsen zich verhouden tot latere uitkomsten en hoe leerkrachten de bruikbaarheid van deze toetsen ervaren. De hoofdvragen van dit proefschrift luidden dan ook 'Hoe ervaren leerkrachten de bruikbaarheid van de Cito kleutertoetsen toetsen bij hun dagelijkse activiteiten?', 'Wat is de stabiliteit van vroege toetsscores van het Cito Leerling- en Onderwijs Volgsysteem?' en 'Hoe beïnvloedt de stabiliteit van deze scores door de toets gesteunde beslissingen over individuele kinderen?'

Bevindingen

Hoofstuk 2 beschrijft hoe leerkrachten de rol van de kleutertoetsen ervaren in hun dagelijks lesgeven. Hiervoor is een vragenlijst verspreid onder 97 leerkrachten die hun visie op de verschillende rollen van toetsen meet. De vragenlijst maakt onder meer onderscheid in de rol van de toets als waardeoordeel over het onderwijsproces en als instrument ter verbetering van het onderwijsproces. De resultaten geven aan dat veel leerkrachten de toets niet uitsluitend als instrument van controle zien. Hoewel leerkrachten de rol van de kleutertoets als waardeoordeel en als instrument voor verbetering erkennen, laten diepte-interviews met zes leerkrachten zien dat beide doeleinden van de toets wezenlijk anders ervaren worden. Dit is veelal afhankelijk van de context waarin ze lesgeven. Omdat lage scores vaak als onvoldoende worden ervaren, kan het toetsresultaat door de leerkracht als een afstraffing ervaren worden als de populatie waaraan een leerkracht lesgeeft onder gemiddeld scoort. Des te meer omdat deze kinderen in de toetshandleiding als 'risicoleerlingen' worden bestempeld en in het toetssysteem letterlijk 'in het rood' blijven scoren, zelfs wanneer ze een gemiddelde ontwikkeling doormaken. Leerkrachten die minder druk ervaren, doordat ze zich meer gesteund voelen door het managementteam (MT) of niet vaak kinderen in de 'rode zone' hebben, zien de toetsen meer als prettige ondersteuning van hun eigen observaties.

De nadruk op de normscores als criterium voor voldoende leidt al snel tot aanpassingen in het onderwijsaanbod om 'onvoldoendes' te verhelpen of te voorkomen. Deze aanpassingen bestaan bijvoorbeeld uit het specifiek aanbieden van woorden in de toets, of oefenen met het format waarin de toetsopgaven aangeboden worden. De invloed van de toets op het handelen van de leerkracht gebeurt meestal onbewust en vanuit een, voor de leerkracht, logische redenering. Leerkrachten vinden het bijvoorbeeld niet eerlijk om kinderen te toetsen op iets wat niet is aangeboden, of zien de toetsopgaven als stof wat kinderen moeten kennen. Het resultaat is echter dat kinderen systematisch hoger scoren dan de oorspronkelijke normgroep, wat ook terug te zien is in de kwantitatieve gegevens van de toetsen. Hierdoor zijn de normen niet langer een juiste weergave van het niveau van een kind ten opzichte van landelijke prestatie.

Summary

In **hoofdstuk 3** beschrijven we de consistentie van percentielscores op de taal- en rekentoetsen van 431 kinderen. Hierbij werd vooral gekeken naar de laagst scorende 25% van de kinderen op toetsafnames tussen groep 1 en 3, aangezien kinderen met deze scores vaak als risicoleerlingen gezien worden. De resultaten laten zien dat slechts een klein percentage van deze kinderen – 11% en 17% voor taal en rekenen respectievelijk – consistent in deze laagste scorecategorie scoorden. Daarentegen behaalde een hoog percentage – 47% en 35% voor taal en rekenen respectievelijk – van de kinderen die in latere jaren bij de 25% laagst scorende kinderen horen, bovengemiddelde scores op de kleutertoetsen. Gemiddelde correlaties tussen de kleutertoetsen onderling ($r = .3$), en tussen de kleutertoetsen en spellings- en rekentoetsen vanaf groep 1 ($r = .2$) laten eveneens zien dat deze toetsscores minder sterk samenhangen dan de scores vanaf groep 1 ($r = .6$). Dit zou kunnen wijzen op grote variabiliteit in de ontwikkelingstrajecten van jonge kinderen.

In **hoofdstuk 4** bouwen we de definitie van stabiliteit uit hoofdstuk 3 verder uit. Hoewel de consistentie van percentielscores een vorm van stabiliteit beschrijft, laat een korte verkenning van de literatuur zien dat de term op veel verschillende manieren gebruikt wordt. We nemen in dit hoofdstuk de brede definitie van Wohlwill over, welke stabiliteit definieert als de mate waarin eerdere scores latere scores voorspellen. Door onderscheid te maken in de manier waarop voorspellingen over latere scores kunnen worden gedaan aan de hand van vroege testscores, definieert Wohlwill ten minste vier typen stabiliteit. Drie van deze typen zijn door Tisak en Meredith uitgewerkt in geneste structural equation modellen. In dit hoofdstuk bouwen we de modellen van Tisak en Meredith om naar multilevel modellen en voegen we een vierde definitie toe. We gebruiken twee van deze modellen van stabiliteit in de evaluatie van de toetsen van Cito. Het eerste model neemt aan dat kinderen een gelijke ontwikkelingen doormaken en zodoende hun rangscore behouden over de tijd. Deze aanname wordt ‘lineaire stabiliteit’ genoemd en gaat ervan uit dat een kind dat in de laagste 25% scoort, deze score behoudt als er niet ingegrepen wordt. Het tweede model neemt aan dat elk kind zijn/haar eigen groei doormaakt. Deze aanname wordt ‘functie stabiliteit’ genoemd en betekent dat het belangrijk is om rekening te houden met de eerdere groei van een kind om latere voorspellingen te doen. Als kinderen bijvoorbeeld stagneren in scores wordt er onder deze aanname van uitgegaan dat deze stagnatie doorzet als er niet wordt ingegrepen.

Beide aannames worden in dit hoofdstuk gepresenteerd in modellen van de taal- en rekenscores van 1402 kinderen tussen groep 2 en groep 5. De resultaten laten zien dat de sterkere aanname van functie stabiliteit de scores van de gehele groep iets beter beschrijft dan lineaire stabiliteit. De verschillen in de overeenstemming met de data van beide aannames zijn echter klein en de scores van een groot deel van de kinderen worden adequaat beschreven onder de aanname van lineaire stabiliteit. Een kleine groep kinderen – 10.7% en 12.1% voor taal en rekenen

respectievelijk – laat aanzienlijk afwijkende groei zien van de rest van de steekproef en zijn duidelijk beter te beschrijven met de aanname van functie stabiliteit. De grote intra-individuele variatie in individuele testscores maakt het echter moeilijk om deze kinderen te identificeren aan de hand van enkele toetsresultaten. Schijnbare afwijkingen van de gemiddelde groei zijn vaak tijdelijk van aard en zetten niet structureel door. Pas na vijf testafnames lijkt het dat men met enige zekerheid kan zeggen of een daling in scores het resultaat is van systematisch afwijkende groei. Deze resultaten suggereren dat de toetsen niet sensitief genoeg zijn om structurele afwijkingen in groei te onderscheiden van willekeurige fluctuaties in de scores.

In **hoofdstuk 5** kijken we naar de voorspellingen die leerkrachten zouden maken op basis van verschillende aannames van stabiliteit. We gebruiken de twee stabiliteitsaannames die in hoofdstuk 4 geëvalueerd zijn om voorspellingen te maken voor de volgende testscore van 911 kinderen. Hierbij maken we zowel voorspellingen met informatie van alle voorgaande testafnames, als met informatie van de laatste paar testafnames. De resultaten laten zien dat voorspellingen op basis van een gemiddelde groei (lineaire stabiliteit) accurater zijn dan voorspellingen die rekening houden met anders dan gemiddelde groei tussen de testafnames (functie stabiliteit). De gemiddelde percentielscore die een kind behaald heeft vormt volgens deze resultaten de beste schatting van de volgende score. Ook de laatst behaalde score geeft een redelijke indicatie van het niveau van het kind op de volgende afname. Het meenemen van de behaalde groei van individuele kinderen leidt over het algemeen tot slechtere voorspellingen, vooral wanneer naar de groei over de laatste twee metingen gekeken wordt. Kinderen als risicoleerlingen identificeren wanneer ze stagneren lijkt tot veel onterechte identificaties te leiden, aangezien het naar alle waarschijnlijkheid om een tijdelijke daling in de resultaten gaat. Sterker nog, 60% van alle kinderen in deze steekproef laat ten minste één stagnatie zien binnen de gemeten periode. Dit zijn vaak niet de kinderen die structureel verminderde groei laten zien over de hele meetperiode. Hoewel men er voor een meer accurate voorspelling beter vanuit kan gaan dat de scores willekeurig fluctueren rond het gemiddelde niveau van het kind, zijn ook hier grote afwijkingen te verwachten. Voor taal en rekenen week de helft van de beste voorspellingen met meer dan 16 respectievelijk 13 percentielpunten af van de verwachte score.

Conclusies

Om de stabiliteit van de scores te beschrijven is het belangrijk om onderscheid te maken tussen verschillende interpretaties van de scores en de bijbehorende typen stabiliteit. Hoewel de toetsen een inschatting geven van het niveau van een kind ten opzichte van andere kinderen in de populatie, laten scores een hoge mate van ongestructureerde intra-individuele variabiliteit zien. Dit maakt het lastig om het niveau van een kind betrouwbaar vast te stellen. Daarbij is het nog moeilijker

om iets te zeggen over de groei van individuele kinderen. Vaak zijn schijnbare stijgingen of dalingen slechts tijdelijk van aard en laten de scores van de meeste kinderen een eenduidige groei zien over een langere periode. Zelfs wanneer dit niet het geval is, is het lastig om kinderen met een afwijkende groei betrouwbaar te identificeren aan de hand van een of twee stagnaties. Aangezien identificatie aan de hand van toetsresultaten gestoeld is op een betrouwbare voorspelling van de toekomstige ontwikkeling van een kind, hebben de scores op deze toetsen slechts een beperkte bruikbaarheid in de signalering van vroegtijdige taal- en rekenproblemen.

De resultaten in hoofdstuk 2 laten zien dat leerkrachten de toetsen niet uitsluitend zien als instrumenten van controle. Sommige leerkrachten ervaren de toetsen zelfs als een prettige ondersteuning van hun eigen oordeel. Het is wel belangrijk om hierbij te vermelden dat dit sterk afhankelijk lijkt van de context waarin de leerkracht lesgeeft. Lage scores worden vaak als onvoldoende gezien en kunnen als afstraffing ervaren worden wanneer kinderen niet boven een laag niveau uitstijgen ondanks een gemiddelde groei. Dit idee dat onder gemiddeld gelijk is aan onvoldoende wordt bekrachtigd door de brede definitie van risicoleerlingen in de handleiding – de laagst scorende 25% – en het bijbehorende kleurenschema. Dit systeem motiveert leerkrachten om kinderen uit de rode zone te krijgen of te houden, wat gepaard gaat met aanpassingen in het onderwijsaanbod en onvermijdelijk met norminflatie. Samen met de reactie van Cito om verouderde normen bij te stellen creëert dit mogelijk een onderwijssysteem wat steeds meer gericht is op de inhoud en vorm van de toets. Hoewel zowel Cito als leerkrachten vanuit verdedigbare principes handelen, lijken ze een tegenstrijdig doel na te streven met betrekking tot de normen. Leerkrachten willen – soms onder druk van ouders of het MT – lage scores vermijden terwijl Cito representatieve normen wil.

Dit laat een belangrijk probleem zien met het gebruik van een normatieve score als een criterium voor risicoleerlingen. De manier waarop de toetsen ontwikkeld zijn plaatst 20% of 25% van de kinderen per definitie in een risicogroep. Dit heeft weinig te maken met de ‘problematische’ prestatie van het kind en nog minder met een problematische ontwikkeling in de taal- en/of rekenvaardigheden, maar meer met hoe deze prestatie zich verhoudt tot andere kinderen. Waar je vervolgens de scheidingslijn legt tussen risico- en niet-risico-leerling is arbitrair. Cito ontwikkelt inmiddels een nieuw observatie instrument ‘Kleuters in beeld’. De focus op een normscore ten opzichte van andere kinderen zou hierbij ondergeschikt worden aan het bieden van diagnostische informatie. Het gebruik van observaties van jonge kinderen heeft het voordeel dat dit over het algemeen dichter bij het curriculum staat in vorm en inhoud. Daarnaast zou dit systeem gebruik kunnen maken van meer frequente kleine observaties, waardoor een momentopname die niet representatief is voor het kind minder invloed heeft. Het is echter belangrijk dat dit instrument een duidelijk doel voor ogen heeft en afgestemd is op dit doel. Het ‘volgen van de ontwikkeling’ van

individuele kinderen is hierbij geen op zichzelf staand doel. Als identificatie van mogelijke leer- en ontwikkelingsproblemen het doel is zou het instrument specifiek gericht moeten zijn op deze doelgroep met een evidence based criterium waaraan leerlingen getoetst worden. Een norm-gerefereerde toets lijkt hier niet de meest geschikte keuze en leidt al snel tot onderwijs wat nauwer gericht is op specifieke testitems in plaats van de achterliggende vaardigheid. Na uitgave is het belangrijk dat het instrument continu geëvalueerd wordt. Niet alleen op de doelen en interpretaties die in de handleiding aangegeven zijn, maar ook op de bredere impact die het instrument heeft op het onderwijssysteem. Hoewel de resultaten in dit proefschrift een momentopname geven van een ingewikkeld en veranderend proces, zullen de bevindingen relevant zijn in elke context waarin een eenduidige normscore wordt gebruikt om ontwikkelingsproblemen bij kleuters te identificeren.