

Evaluatie van Kaleidoscoop en Piramide

Eindrapportage

*A. Veen
J. Roeleveld
P. Leseman*

Evaluatie van Kaleidoscoop en Piramide

Eindrapportage

*A. Veen
J. Roeleveld
P. Leseman*

CIP-GEGEVENS KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Veen, A., Roeleveld, J., Leseman, P.

Evaluatie van Kaleidoscoop en Piramide, Eindrapport. Veen, A., Roeleveld, J., Leseman, P. Amsterdam: SCO-Kohnstamm Instituut van de Faculteit der Maatschappij- en Gedrags-wetenschappen, Universiteit van Amsterdam (SCO-rapport 576).

ISBN 90-6813-628-3

Alle rechten voorbehouden. Niets uit deze uitgave mag worden verveelvuldigd, opgeslagen in een geautomatiseerd gegevensbestand, of openbaar gemaakt, in enige vorm of op enige wijze, hetzij elektronisch, mechanisch, door fotokopieën, opnamen, of op enige manier, zonder voorafgaande schriftelijke toestemming van de uitgever.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher.

Uitgave en verspreiding:

SCO-Kohnstamm Instituut

Wibautstraat 4, Postbus 94208, 1090 GE Amsterdam

Tel.: 020-525 1357

<http://www.educ.uva.nl/sco>

© Copyright SCO-Kohnstamm Instituut, 2000

Inhoudsopgave

Ten geleide

1	Resultaten Kaleidoscoop en Piramide: technisch verslag	
1.1	Beschrijving van het landelijke experiment	7
1.2	Verschillen tussen de programma's.....	8
1.3	Doel van het evaluatie-onderzoek	9
2	Onderzoeksopzet en uitvoering	
2.1	Onderzoeksopzet	11
2.2	Onderzoeksprocedures	14
2.3	Meetinstrumenten	17
2.4	Uitval en doorstroming.....	26
2.5	Samenstelling van de onderzoeksgroepen	28
3	Resultaten	
3.1	Verschillen in taal, cognitie en sociaal-emotioneel gedrag in 1997	33
3.2	Verschillen in taal, cognitie en sociaal-emotioneel gedrag in 1998	35
3.3	Verschillen in taal, cognitie en sociaal-emotioneel gedrag in 1999	41
3.4	Verschillen in 'groei' tussen 1997-1998	44
3.5	Verschillen in 'groei' tussen 1997-1999	46
3.6	Verschillen naar etniciteit.....	48
4	Conclusies	53
4.1	Slotopmerkingen	57
	Literatuur	61

Ten geleide

In opdracht van de ministeries van Onderwijs, Cultuur en Wetenschappen en Volksgezondheid, Welzijn en Sport heeft het SCO-Kohnstamm Instituut onderzoek verricht naar de effectiviteit van twee educatieve programma's voor kinderen in maatschappelijke achterstandssituaties, Kaleidoscoop en Piramide. In dit eindverslag zijn de resultaten van dit onderzoek beschreven. Via deze weg willen wij degenen bedanken die een bijdrage aan het onderzoek hebben geleverd. In de eerste plaats natuurlijk de peuterspeelzalen, scholen, leerlingen en ouders die bereid waren om de voor dit onderzoek relevante informatie te verstrekken. In de tweede plaats de coördinatoren en andere bij de ontwikkeling van de programma's Kaleidoscoop en Piramide betrokkenen die bereid waren om de nodige organisatorische voorbereidingen te treffen en ons bij de programma- en de controle-locaties te introduceren.

Het onderzoek is begeleid door een commissie onder voorzitterschap van prof. dr. D.A.V. van der Leij, tevens voorzitter van de commissie Voor- en Vroegschoolse Educatie (VVE) die namens de ministeries de uitvoering van de programma's coördineert. We danken de begeleidingscommissie en leden van de VVE voor hun bijdrage aan het onderzoek en hun commentaar op een eerdere versie van dit eindrapport.

Tenslotte zijn in verschillende fasen van dit onderzoek en voor diverse werkzaamheden verschillende medewerkers van het SCO-Kohnstamm Instituut betrokken (Martha Otter, Brigit Triesscheijn, Ilona Koning, Dorothe Elshof, Edgar Tossijn), die wij via deze weg hartelijk willen bedanken voor het werk dat zij voor dit onderzoek hebben verzet.

De auteurs

Februari 2000

1 Resultaten Kaleidoscoop en Piramide: technisch verslag

Op het advies van de Commissie Voorschoolse Educatie (1994) is in 1995 begonnen met de ontwikkeling van twee experimentele voor- en voerschoolse educatieve programma's voor kinderen van 3-6 jaar in maatschappelijke achterstandssituaties, uit te voeren in peuterspeelzalen en kleutergroepen van basisscholen. Gekozen is voor het High/Scope-programma, dat onder de naam Kaleidoscoop door de Averroësstichting aan de Nederlandse situatie is aangepast, en het programma Piramide van het CITO. De programma's vormen een aanvulling op het aanbod gezinsgerichte stimuleringsprogramma's dat al enige jaren beschikbaar is, en bieden in principe een professioneel ontwikkeld alternatief voor de vele kleinschalige, sterk uiteenlopende en kortstondige projecten die de afgelopen jaren in het kader van het onderwijsvoorrangsbeleid in de voor- en voerschoolse periode zijn uitgevoerd.

In 1996 is begonnen met de implementatie en met ingang van het schooljaar 1996/1997 zijn de programma's op een aantal plaatsen voldoende in uitvoering genomen om evaluatie-onderzoek te kunnen starten naar de effectiviteit. Met dit onderzoek is in het voorjaar van 1997 door het SCO Kohnstamm Instituut een start gemaakt. Begin 1999 is een interim-rapport verschenen waarin verslag werd gedaan over de tussentijdse resultaten na twee van de drie geplande metingen (na ongeveer anderhalf jaar uitvoering van de programma's) (Leseman, Veen, Triesscheijn, Otter, 1999). In 1999 is de derde meting verricht. In deze eindrapportage wordt verslag gedaan van de resultaten over de drie achtereenvolgende metingen.

1.1 Beschrijving van het landelijke experiment

In verschillende opzichten is er sprake van een uniek experiment. Beide programma's bieden een omvattend, doorlopend educatief traject van circa drie jaar met brede cognitieve en sociaal-emotionele doelen. Het inhoudelijke curriculum en de pedagogisch-didactische werkwijzen berusten in beide gevallen op actuele wetenschappelijke theorievorming. Met de programma's wordt in principe op driejarige leeftijd begonnen en de interventie is, vanwege de eis om

Evaluatie Kaleidoscoop en Piramide

vier dagdelen deel te nemen en de duur van drie jaar, te kenschetsen als intensief. Daarmee voldoen beide programma's aan twee van de drie belangrijkste kenmerken van elders effectief gebleken voor- en voerschoolse educatieve programma's (Leseman, Otter, Blok, & Deckers, 1998). Wat betreft het derde kenmerk van effectieve benaderingen, namelijk ontwikkelingsadequate ontwikkelingsstimulering en toepassing van kindvolgende, op actief leren gerichte pedagogisch-didactische strategieën, kunnen beide programma's eveneens de toets der kritiek doorstaan, althans op papier. Hoe de uitvoering in praktijk verloopt en of dit tot de resultaten leidt die elders zijn gevonden, dient uiteraard in het lopende onderzoek te worden vastgesteld.

Naast het inhoudelijke educatieve programma, stellen beide programma's duidelijke en wetenschappelijk onderbouwde eisen aan de context waarin het inhoudelijke programma moet worden uitgevoerd. Deze eisen betreffen de zogenaamde structurele pedagogische kwaliteitscriteria (Howes, Philips, & Whitebook, 1992), zoals de leidster- en leerkracht-kindratio, de inrichting van de groepsruimte, de groepsgrootte, de betrokkenheid van de ouders, de deelnamegraad van de kinderen en voorzieningen voor meertalige opvang. De uitvoering vergt voorts nauwe samenwerking tussen peuterspeelzalen en basisscholen, behorend tot de sectoren welzijn en onderwijs. Dergelijke nauwe, integrale samenwerking tussen deze sectoren is tamelijk uniek in Nederland.

1.2 Verschillen tussen de programma's

Tussen de programma's bestaan behalve de genoemde overeenkomsten, ook verschillen en dat is een bijkomende reden waarom er sprake is van een uniek experiment, want de uitkomsten kunnen, hoe ze ook precies uitpakken, een belangrijke bijdrage leveren aan de verdere theorievorming en praktijkontwikkeling op het gebied van onderwijs voor jonge kinderen. De verschillen betreffen onder andere de implementatiestrategie, de manier waarop de uitvoering van het programma wordt bewaakt en de rol van de leidster en leerkracht.

In het programma Kaleidoscoop is de 'leerstof' –datgene wat de kinderen in de loop der tijd idealiter aan kennis en vaardigheden verwerven– minder geëxpliciteerd dan in het programma Piramide, dat expliciet via door de leidster of leerkracht geïnitieerde activiteiten de 'leertaken' van peuters en kleuters

inbrengt, zoals het leren kennen van kleuren, vormen, hoeveelheden, seriëren e.d.. De wijze waarop de activiteiten van de kinderen worden beïnvloed, is in Kaleidoscoop indirecter en zwakker gestuurd, en daarom wellicht meer 'natuurlijk' en 'spontaan', maar misschien ook riskanter gelet op de doelgroepen, dan in Piramide dat een meer gestructureerde opbouw kent met tweewekelijkse thema's en binnen de thema's hiërarchisch opgebouwde oriëntatie, verrijkings- en verdiepingsactiviteiten. In Kaleidoscoop is het vooral de dagelijkse cyclus van vooruitkijken, uitvoeren, terugkijken die aan de activiteiten van de kinderen een doelmatige, meta-cognitieve structuur geeft.

Een ander belangrijk verschil betreft de manier waarop de ontwikkeling van kinderen wordt gevolgd en hoe eventueel wordt ingegrepen. In Piramide wordt hiertoe gebruik gemaakt van genormeerde toetsen die frequent, twee maal per jaar, worden afgenomen. Blijkt een kind beneden een bepaalde norm te scoren, dan wordt individuele tutoring ingezet. Het voordeel van deze aanpak is de objectiviteit (gebruik van toetsen en normgegevens) en de koppeling van het volgsysteem aan een helder handelingsplan (tutoring). Het gevaar is dat de frequente toetsing stigmatiserend en ontmoedigend werkt, en dat de educatieve activiteiten geheel gericht worden naar het doel de toetsscore te verhogen. In Kaleidoscoop is het volgsysteem informeler, hoofdzakelijk gebaseerd op informele dagelijkse observaties van het kindgedrag in de natuurlijke speelwerksituatie. Hoewel er normen zijn gespecificeerd (de zogenaamde sleutelervaringen), is minder eenduidig af te leiden wanneer een kind daarbij achterblijft en wat daaraan gedaan zou moeten worden; het leren werken met het Child Observation Record (COR) is niet gemakkelijk gebleken. De aanpak is bij eerste indruk kindvriendelijker en minder ingrijpend in het natuurlijke proces van speelwerkactiviteiten, maar het risico bestaat dat in groepen met een grote vertegenwoordiging van achterstandskinderen normen en verwachtingen onbewust worden verlaagd.

1.3 Doel van het evaluatie-onderzoek

De programma's Kaleidoscoop en Piramide worden thans op proef uitgevoerd in een maatschappelijke context die gekenmerkt wordt door de alom hoge verwachtingen ten aanzien van vroege educatieve programma's. Toch is tevoren enige relativering over de potentie hiervan geboden. Wereldwijd is er namelijk

Evaluatie Kaleidoscoop en Piramide

geen sprake van onbetwistbaar succes van dergelijke programma's. Hoewel op korte termijn gemiddeld genomen matig-sterke tot sterke effecten worden gevonden, doven positieve resultaten doorgaans binnen twee jaar uit (Blok & Leseman, 1996). De ervaringen die in Nederland zijn opgedaan met zowel centrum- als gezinsgerichte programma's bieden evenmin reden tot groot optimisme. Aannemelijk is dat door het feit dat in Nederland het overgrote deel van de kinderen al vanaf vier jaar vijf dagen per week naar school gaat, de kans om effecten te vinden van voor- en vroegschoolse programma's ten opzichte van deze bestaande praktijk, kleiner is dan elders. Voor Kaleidoscoop en Piramide geldt in dit verband dat ze niet alleen moeten aantonen kwalitatief goede educatieve programma's te zijn, maar bovendien meerwaarde te hebben ten opzichte van de reguliere situatie.

Het evaluatie-onderzoek is opgezet met het doel kwaliteit en meerwaarde van beide programma's wetenschappelijk verantwoord te documenteren. Hiertoe is een onderzoeksopzet gekozen die, binnen de opzet en randvoorwaarden van het experiment zoals het thans wordt uitgevoerd, antwoord kan geven op de vraag of de programma's pedagogisch-didactisch van hoge kwaliteit en effectief zijn, dat wil zeggen, meerwaarde hebben in de bestaande situatie. In het navolgende verslag worden de resultaten van het evaluatie-onderzoek gerapporteerd. Het verslag is als volgt opgebouwd. In het nu volgende hoofdstuk wordt allereerst ingegaan op de onderzoeksopzet, dan wordt geschetst hoe het onderzoek in praktijk is uitgevoerd. Daarna worden de gebruikte meetinstrumenten besproken. In het derde hoofdstuk worden de onderzoeksresultaten gepresenteerd. Het verslag wordt afgesloten met enkele conclusies en slotopmerkingen.

2 Onderzoeksopzet en uitvoering

Het evaluatieonderzoek van het SCO Kohnstamm Instituut bestaat uit twee deelstudies: een breedtestudie en een dieptestudie. Daarnaast heeft het GION van de Rijksuniversiteit Groningen een implementatie-onderzoek uitgevoerd, waarover apart is gerapporteerd (Reezigt, 1999).

In de breedtestudie staat de klassieke vraag centraal naar de effecten van Kaleidoscoop en Piramide op taal- en cognitieve toetsen en op het sociaal-emotioneel gedrag van de kinderen in de groep. Verondersteld wordt dat de gekozen toetsen en beoordelingslijsten voorspellende waarde hebben voor de latere schoolprestaties. In de dieptestudie worden in ongeveer de helft van de speelzalen en kleutergroepen die aan de breedtestudie deelnemen, met behulp van een videocamera observaties verricht van de groepsprocessen, met name van de sociale interacties tijdens de belangrijkste onderdelen van het dagprogramma van Kaleidoscoop en Piramide. Deze observaties worden vergeleken met observaties in reguliere peuter- en kleutergroepen. Van de dieptestudie zijn op dit moment nog geen complete resultaten beschikbaar, zodat daarover nog niet gerapporteerd kan worden.

2.1 *Onderzoeksopzet*

De vraag van de breedtestudie naar de effectiviteit van Kaleidoscoop en Piramide heeft geleid tot een onderzoeksopzet met twee experimentele groepen en twee controlegroepen. In figuur 1 is hiervan een schematisch overzicht gegeven. Tabel 11 laat de cijfers zien van de deelname aan de drie achtereenvolgende metingen.

De eerste controlegroep, controlegroep A in figuur 1 bestaat uit kinderen van dezelfde leeftijd en sociaal-economische en etnisch-culturele herkomst die reguliere peuterspeelzalen en basisscholen bezoeken in dezelfde of vergelijkbare wijken. Omdat er in het kader van het experiment geen mogelijkheid was om kinderen door loting aselekt toe te wijzen aan één van de drie condities, zijn in alle drie de condities uit de aanwezige groepen kinderen steekproeven samengesteld die naar belangrijke kind- en achtergrondkenmerken als gelijkwaardig beschouwd kunnen worden.

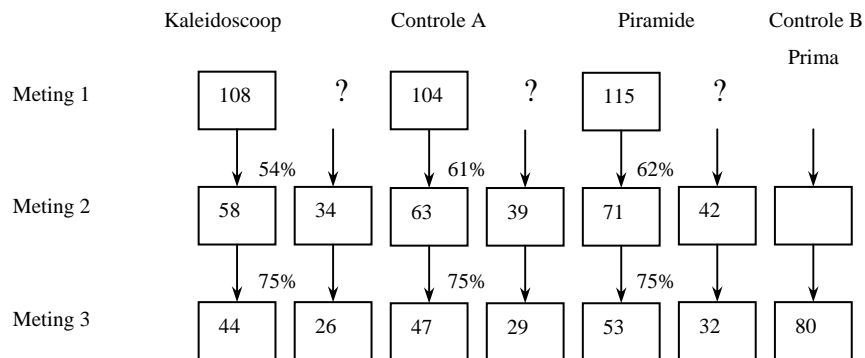
Evaluatie Kaleidoscoop en Piramide

In de structuur van het experiment, dat uitvoering behelst van driejarige educatieve programma's in combinaties van samenwerkende peuterspeelzalen en basisscholen, is rekening gehouden met het risico van uitval. Verwacht werd dat de uitstroom met name bij de overgang van de peuterspeelzaal naar de basisschool groot zou zijn, vanwege de vrijheid van schoolkeuze. Daarom is per conditie begonnen met relatief grote onderzoeksgroepen, zodat bij een uitval van 50% (in het meest ongunstige geval) voor statistische analyses nog voldoende kinderen zouden overblijven die het gehele driejarige traject hadden meegedaan. Daarnaast was aan het begin van de onderbouwperiode een aanvulling van de steekproeven gepland met nieuwe kinderen die geen van de deelnemende peuterspeelzalen hadden bezocht (maar misschien wel andere). Deze kinderen nemen dus niet aan het gehele driejarige traject van de programma's deel, maar alleen aan het tweejarige deeltraject dat in de basisschool wordt uitgevoerd.

De eerste controlegroep betreft kinderen die een regulier peuter- en kleuterprogramma volgen. Het evalueren van de effectiviteit van Kaleidoscoop en Piramide in vergelijking met deze controlegroep komt er op neer dat bepaald wordt wat het meer-effect is van beide programma's ten opzichte van de reguliere aanpak. Dit is op zichzelf een verdedigbare strategie, omdat het reguliere peuterspeelzaal aanbod en kleuteronderwijs tot de normale, thans bestaande situatie moet worden gerekend en omdat Kaleidoscoop en Piramide bedoeld zijn iets extra's op te leveren.

Deelname aan peuterspeelzalen en vergelijkbare activiteiten in de voorschoolse periode is echter minder 'normaal' en wijd verbreid. Hoewel in gebieden met veel onderwijsachterstand de peuterspeelzaal al langer als middel voor schoolvoorbereiding en ontwikkelingsstimulering wordt ingezet, is zelfs daar de deelname van kinderen uit achterstandsgroepen verhoudingsgewijs nog klein.

Opzet en (verwachte) aantallen kinderen
breedte-studie



Veel kinderen die een risico lopen op onderwijsachterstand wanneer ze eenmaal in het basisonderwijs zitten, blijven in de voorschoolse periode thuis en nemen aan geen enkel omvattend stimuleringsprogramma deel. Dit is de reden waarom een tweede controlegroep, controlegroep B in figuur 1, in de onderzoeksopzet is opgenomen, bestaande uit kinderen die geen voorschools programma volgden of een peuterspeelzaal bezochten; uiteraard volgden deze kinderen wel regulier kleuteronderwijs. Deze tweede controlegroep wordt achteraf gevormd uit het databestand van het zogenaamde Primacohort. Het Primacohort-onderzoek is een landelijk, tweejaarlijks in verschillende groepen van de basisschool uitgevoerd onderzoek naar schoolprestaties. Er zijn enkele tienduizenden leerlingen bij betrokken. De metingen van het Primacohort vinden onder andere plaats in groep 2, waardoor een vergelijking mogelijk is met de onderzoekskinderen. Behalve schoolprestaties en kindkenmerken als leeftijd en sekse, worden in het Primacohort ook gezinskenmerken (sociaal-economische status (SES), etniciteit, gezinsklimaat) en gebruik van voorschoolse voorzieningen uitgebreid geïnventariseerd, zodat op grond van deze gegevens een equivalente, gematchte *post hoc* controlegroep kan worden samengesteld. Gegevens over gebruik van voorschoolse voorzieningen kunnen dienen om alleen kinderen te selecteren die geen peuterspeelzaal of kinderdagverblijf bezochten voorafgaande aan de school. Deze vergelijking hebben we helaas niet helemaal op de geplande manier kunnen uitvoeren (zie par. 2.3).

Evaluatie Kaleidoscoop en Piramide

In het onderzoek zijn drie achtereenvolgende jaren metingen verricht: in 1997, 1998 en 1999. Alle metingen vonden plaats vlak vóór de zomervakantie, in de maanden mei tot en met juli. De eerste meting, in 1997, is niet te beschouwen als voormeting, omdat de programma's in het najaar van 1996 zijn begonnen en de kinderen die voor de eerste meting waren geselecteerd gemiddeld al een half jaar aan de interventie waren blootgesteld. Tijdens de derde meting, aan het eind van het educatieve traject van Kaleidoscoop en Piramide in de kleutergroepen, is het meetinstrumentarium enigszins uitgebreid ten opzichte van de beide voorgaande metingen en overlapt een deel van het meetinstrumentarium met de instrumenten die in het Primacohort-onderzoek wordt gebruikt, met het oog op de vergelijking van de experimentele groepen met de tweede controlegroep.

2.2 Onderzoeksprocedures

Het onderzoek is in april 1997, direct na toekenning van de onderzoeksopdracht, gestart. In korte tijd zijn contacten gelegd met de peuterspeelzalen van de eerste en tweede tranche waar de programma's Kaleidoscoop en Piramide volgens de beschikbare gegevens al sinds het najaar van 1996 in uitvoering waren. Bepaald was dat in principe alle locaties uit de eerste en tweede tranche in het onderzoek zouden worden opgenomen. Om logistieke redenen (afstand en kosten om testleiders naar deze locatie laten gaan) en vanwege het feit dat het aantal Piramide- het aantal Kaleidoscoop-settings toch al overtrof, is in overleg met de opdrachtgevers / de begeleidingscommissie voor het onderzoek besloten om de Piramide-locatie Vroomshoop niet in het onderzoek mee te nemen.

Met de plaatselijke coördinatoren van de overige programmaspeelzalen is overlegd welke speelzalen in hun werkgebied het beste benaderd zouden kunnen worden voor de controleconditie (controlegroep A). Daarbij werden de volgende criteria gespecificeerd. Het moest gaan om speelzalen die:

- 1) zich bevonden in dezelfde wijk, of anders in een sociaal-economisch vergelijkbare wijk, dorp of stad in de regio, en
- 2) werkten met een naar sociaal-economische en etnische achtergrond zoveel mogelijk vergelijkbare groep kinderen, maar die
- 3) geen Kaleidoscoop of Piramide, of een vergelijkbaar educatief totaalprogramma uitvoerden.

Als volgende stap werden bij de coördinatoren van alle speelzalen, ook van de controlegroep, gegevens opgevraagd over de kinderen: leeftijd, etniciteit vader en moeder, sekse, deelnamegraad (aantal dagdelen), deelnameduur (aantal maanden dat het kind de speelzaal bezocht), aanwezigheid door de week, doorstroming naar een basisschool voor zover al bekend. Nadat in overleg met de coördinatoren bepaald was welke kinderen allemaal voor de geplande testronde in aanmerking kwamen, is de procedure voor het verkrijgen van toestemming van de ouders gestart. Op enkele uitzonderingen na werd deze toestemming verleend. In de periode mei, juni, juli 1997 met een kleine uitloop naar de maanden augustus en september van dat jaar is de eerste meetronde uitgevoerd. Van het cohort van 327 kinderen, konden er 243 volledig worden getest. De belangrijkste reden waarom sommige geselecteerde kinderen niet konden worden getest was afwezigheid door ziekte op het moment van testafname. Daarnaast speelde ook de naderende vakantieperiode een rol.

Om de kinderen bij hun overstap van de peuterspeelzaal naar de basisschool te kunnen volgen, zijn gedurende het schooljaar 97-98 maandelijks antwoordkaartjes aan de peuterspeelzalen gestuurd met de naam van het kind dat die maand vier jaar zou worden. De leidsters werd verzocht hierop te noteren naar welke basisschool het betreffende kind zou uitstromen en het ingevulde kaartje te retourneren. Vervolgens konden de basisscholen benaderd worden met de vraag medewerking aan het onderzoek te verlenen.

Voor wat betreft de controlegroep zijn alle basisscholen benaderd waar twee of meer kinderen uit een bepaalde controlespeelzaal naartoe zijn uitgestroomd. Scholen waar slechts één kind uit de oorspronkelijke onderzoeksgroep naartoe ging, zijn om logistieke en financiële redenen niet benaderd. Als responsverhogende maatregel werden boekenbonnen in het vooruitzicht gesteld. Alle voor de controlegroep benaderde scholen bleken bereid om aan het onderzoek deel te nemen. Tussen de eerste en de tweede meting vielen twee scholen, met een relatief groot aantal 'controlekinderen' uiteindelijk toch af omdat zij inmiddels met het Piramide-programma waren gaan werken en dus niet meer als controlescholen beschouwd konden worden. Daarnaast viel een groep Kaleidoscoop-kinderen af, omdat de betreffende basisschool niet voldeed aan de eisen die aan de programma-settings gesteld worden. Dit gebeurde in overleg met de Averroësstichting. Tussen de tweede en de derde meting zijn geen scholen afgevallen maar wel bleken drie controlescholen niet bereid om mee te doen aan de afname van de Primacohort-instrumenten.

Evaluatie Kaleidoscoop en Piramide

Ter aanvulling van het verlies van kinderen uit de steekproef bij de overgang van de peuterspeelzaal naar de basisschool, zijn zowel de experimentele condities als de controlegroep aangevuld met nieuwe kinderen, tot een totaal van ongeveer 100 per conditie. De scholen werd hiertoe gevraagd om alle kinderen uit de kleutergroepen waarin zich onderzoekskinderen bevonden, van dezelfde leeftijd als de onderzoekskinderen (geboren tussen 1 juni 1993 en juni 1994) aan ons door te geven. Op basis daarvan zijn nieuwe kinderen geselecteerd. Bij deze selectie is in de eerste plaats rekening gehouden met de etnische herkomst van de kinderen. Voorts werd gestreefd naar een goede spreiding over de locaties en de condities. De bereidheid tot deelname was opnieuw goed.

De tests werden evenals in 1997 afgenomen door getrainde testleiders van Nederlandse, Marokkaanse en Turkse herkomst. De testafname vond plaats in de periode half mei-half juli 1998. Een aanvullende testronde vond plaats na de zomervakantie. Van de 307 geselecteerde kinderen konden er uiteindelijk 283 volledig worden getest. Opnieuw bleek afwezigheid door ziekte de belangrijkste reden waarom sommige kinderen niet konden worden getest.

Voor de derde meting in 1999 zijn, voor wat betreft de individuele testafname, geen nieuwe kinderen aan het bestand toegevoegd. Ten behoeve van de organisatie van de testafnames voor dit derde meetmoment zijn de scholen benaderd met de vraag in welke groep en bij welke leerkracht de onderzoekskinderen zaten. Op grond van deze inventarisatie konden 261 kinderen in de derde meting betrokken worden. Ook in deze derde meetronde zijn de individuele toetsen afgenomen door getrainde testleiders in de eigen taal van de kinderen.

In aanvulling op de individuele testafname heeft een groepsgewijze afname plaatsgevonden van toetsen die in het Primacohort-onderzoek worden gebruikt in groep 2 van het basisonderwijs. Deze toetsen zijn afgenomen in kleutergroepen waarin de onderzoekskinderen zitten. Dit betekent dat, behalve de onderzoekskinderen, ook andere kinderen aan deze toetsafname hebben meegedaan. Ter voorbereiding hierop hebben de scholen formulieren ingevuld (analoog aan het in het Primacohort-onderzoek gebruikte 'leerlingformulier') met namen en achtergrondkenmerken (sekse, opleidingsniveau ouders, leerlinggewicht, etnische herkomst) van kinderen van dezelfde leeftijd als de onderzoekskinderen, dat wil zeggen geboren tussen 1 juni 1993 en 1 juni 1994. Geprobeerd is om de afnamecondities zoveel mogelijk overeen te laten stemmen

met die van het Primacohort-onderzoek. Dat wil zeggen dat de toetsen groepsgewijs (in groepjes van 5-10 kinderen) zijn afgenomen, in het Nederlands, door getrainde, ervaren toetsleiders, die even tevoren¹ betrokken waren bij de afname van de testafname in het kader van het Primacohort-onderzoek.

Behalve de testafname bij de kinderen is met de ouders van alle onderzoekskinderen een kort interview gehouden. De meeste interviews werden gedaan in 1997 en, voor zover het ging om nieuwe kinderen, in 1998. In 1999 zijn nog enkele gesprekken gevoerd met de ouders waarmee dat in de voorgaande jaren niet gelukt was. De oudervragenlijst werd afgenomen in de eigen taal en was gericht op het verzamelen van informatie over achtergrondkenmerken van de kinderen zoals het opleidingsniveau van de ouders, hun geboorteland, de thuistaal en de gezinssamenstelling. De vragenlijst is met opzet kort gehouden omdat hij werd afgenomen bij het brengen en halen van de kinderen en niet te veel tijd in beslag mocht nemen. De ouders konden dit moment eventueel benutten om vragen over het onderzoek te stellen.

2.3 Meetinstrumenten

In alle drie de meetronden is voor wat betreft de individuele testafname hetzelfde instrumentarium gebruikt, maar in de derde meetronde, in 1999 is het instrumentarium enigszins aangevuld en uitgebreid. De tests zijn afgenomen in de eigen taal van de kinderen, waarbij gebruik is gemaakt van Nederlandse, Turkse en Marokkaans-Arabischtalige en Marokkaans-Berber-talige testleiders. Als indicator voor de taalontwikkeling van de kinderen werd gebruikt gemaakt van de test Passieve Woordenschat uit de Diagnostische Toets Tweektaligheid (DTT, Narain & Verhoeven, 1994). Deze test is zowel in het Nederlands als in het Turks en Marokkaans-Arabisch beschikbaar en bevat 60 items. De test is opgedeeld in even en oneven items. Bij de Turkse en Marokkaanse kinderen werden de 30 even items in de eigen taal afgenomen en de 30 oneven in het Nederlands, om zowel het niveau in het Nederlands als in de eigen taal te kunnen bepalen. Bij de Nederlandse kinderen werden eveneens twee testhelften afgenomen, maar de 'eigen taal' was nu Nederlands. Op alle drie de meetmomenten werden steeds

¹ De toetsafname t.b.v. de derde ronde van het Prima-cohort-onderzoek vond plaats in de maanden januari-maart 1999.

Evaluatie Kaleidoscoop en Piramide

dezelfde items van de DTT afgenomen. Op het derde meetmoment is de DTT aangevuld met de sub-toets Passieve Woordenschat uit de Taaltoets Allochtone Kinderen (TAK; Verhoeven, Vermeer en Van de Guchte, 1986), die iets moeilijker is en bedoeld is voor een oudere leeftijdsgroep. Deze aanvulling was nodig vanwege het mogelijk bereiken van een plafond in de DTT. De TAK bestaat uit 98 items, waarvan de eerste 40 geselecteerd zijn. De test is op dezelfde manier als de DTT opgedeeld in twee testhelften, en ook hier zijn de even items afgenomen in de eigen taal en de oneven items in het Nederlands. Al met al (DTT en TAK samen) hebben de kinderen bij de derde meting dus 50 woorden in de eigen taal voorgelegd gekregen en 50 woorden in het Nederlands.

Voor het meten van de cognitieve ontwikkeling werd gebruik gemaakt van de Vragenlijst Denkprocessen (VLDP; 34 items; Groenendaal & Colthof, 1983), het onderdeel Plaatjes Sorteren van de Kleuter-SON (6 items; Sniijders & Sniijders-Oomen, 1975) en de test Blokpatronen van de Leidse Diagnostische Test (LDT; 15 items; Schroot's & van Alphen de Veer, 1976). Voorbeeld-items uit de verschillende tests zijn opgenomen als bijlage bij dit rapport.

De VLDP is afgenomen in drie taalversies (Nederlands, Turks en Marokkaans-Arabisch), afhankelijk van de taaldominantie van het kind. De VLDP kent drie 'moeilijkheidsniveaus'. In 1997 en 1998 zijn beide keren de niveaus I en II afgenomen (22 items), in 1999 zijn de resterende 12 items van niveau III hieraan toegevoegd.

De SON en de LDT kennen minimale verbale instructie en werden eveneens in de dominante taal afgenomen. Marokkaanse kinderen met een Berber-achtergrond werden zoveel mogelijk in het Berber geïnstrueerd, maar in het Nederlands of Marokkaans-Arabisch getest omdat er geen Berbersversies van de toetsen beschikbaar waren. Ook hier geldt dat op alle drie de meetmomenten dezelfde test werd afgenomen.

Gegevens over de sociaal-emotionele ontwikkeling van de kinderen werden verkregen met behulp van de R-Schobl (Bleichrodt, Resing & Zaal, 1993), die, bij de eerste meting door de peuterleidsters en bij de tweede en derde meting door de leerkrachten voor elk onderzoekskind werd ingevuld. Hieruit zijn vier schalen geconstrueerd, te weten extraversie, werkhouding, aangenaam sociaal gedrag en emotionele stabiliteit. De schaal extraversie verwijst naar externaliserend gedrag (negatief aandachttrekken) en een hoge score wijst op probleemgedrag.

In de derde meetronde zijn, naast de individuele testafnamen, groepsgewijs toetsen afgenomen die gebruikt worden in het Primacohort-onderzoek. Het gaat om de toetsen die deel uitmaken van het leerlingvolgsysteem van het Cito: Ordenen, Platenboek Oudste Kleuters en de Begrippentoets. Ordenen meet vaardigheden die betrekking hebben op het voorbereidend rekenen bij kleuters. De Begrippentoets meet kennis van begrippen die zowel voor het lees- als het rekenonderwijs van belang zijn. De toetsen zijn bedoeld voor afname bij oudste kleuters (eind groep 2). Als voorbereiding op deze testafname is per onderzoekskind bij de scholen geïnventariseerd in welk kleuterjaar het betreffende kind zich op dat moment bevond, groep 1 of groep 2. Hieruit bleek dat een behoorlijk deel van de kinderen (ook kinderen die op grond van hun leeftijd al in groep 2 hadden kunnen zitten) in groep 1 zat. Desondanks is besloten om de toetsafname van deze voor afname bij oudste kleuters bedoelde Prima-toetsen door te laten gaan. Onder andere om tegemoet te komen aan de bezwaren van sommige scholen die de toetsen op voorhand te moeilijk achtten voor de wat jongere onderzoekskinderen, is besloten om de moeilijkste toetsonderdelen weg te laten. Bij de Begrippentoets is deel I (40 items) afgenomen. Deel II (de laatste 20 items) waarin woorden worden getoond in plaats van plaatjes, is niet afgenomen. Bij Ordenen (totaal 42 items) zijn 36 items afgenomen. De laatste 6 items, waarin getallen centraal staan, werden niet afgenomen. Uiteindelijk hebben we de toetsgegevens echter toch niet kunnen benutten voor het doel waarvoor ze primair verzameld werden, namelijk de vergelijking met gegevens uit het Primacohort-onderzoek (controlegroep B, figuur 1). Bij nadere analyse bleken namelijk teveel onderzoekskinderen jonger (en mede hierdoor lager te scoren) dan de kinderen die regulier meedoen aan het Primacohort-onderzoek. De groepen bleken op grond van hun leeftijd dermate onvergelijkbaar dat presentatie van deze gegevens in deze rapportage weinig zinvol is. We gaan ervan uit dat de bedoelde vergelijking een jaar later, wanneer de onderzoekskinderen zich eind groep 2 bevinden, wel te maken is. In overleg met de opdrachtgevers/de begeleidingscommissie voor het onderzoek is daarom besloten om de toetsen in het voorjaar van 2000, in het kader van de follow-up-studie², opnieuw bij de

² In opdracht van de Ministeries van OCenW en VWS zal in 2000 een follow-up-onderzoek worden uitgevoerd in de groepen 2 en 3 van de programma- en controlescholen, waarin zoveel mogelijk de huidige onderzoekskinderen maar ook nieuwe kinderen betrokken zullen worden.

Evaluatie Kaleidoscoop en Piramide

onderzoekskinderen af te nemen en de vergelijking met de Primacohort-gegevens nog eens uit te voeren.

De verzamelde toetsgegevens van de Begrippentoets en de toets Ordenen zijn wel gebruikt in de analyses die betrekking hebben op de vergelijking tussen de experimentele groepen en de controlegroep (A). De resultaten hiervan zijn te vinden in de volgende paragrafen.

Tabel 1: Schaalbetrouwbaarheden van tests en beoordelingsschalen (Cronbachs α) bij de eerste (1997, tweede (1998) en derde afname (1999).

Test	Items	Alle kinderen
Woordenschat Eigen Taal	30	.66-.85
Woordenschat Nederlands	30	.82-.90
Vragenlijst Denkprocessen	22 /33	.73-.80
SON-Sorteren 2-7	6	.48-.84
LDT-Blokpatronen	15	.72-.76
Extraversie	7	.89-.90
Werkhouding	12	.93-.93
Aangenaam Gedrag	10	.88-.89
Emotionele Stabiliteit	7	.73-.77
Test nieuw 1999		
TAK eigen taal	20	.74
Tak Nederlands	20	.74
Begrippen	40	.88
Ordenen	36	.81

In tabel 1 is een overzicht gegeven van de schaalbetrouwbaarheden (Cronbachs α 's) van de toetsen en gedragsbeoordelingsschalen, gebaseerd op de afnamen in 1997, 1998 en 1999. In de eerste kolom staat het aantal afgenomen items per instrument. Bij de VLDP zijn de eerste en tweede meting 22 items afgenomen, bij de derde meting zijn er 11 items bijgekomen.

De betrouwbaarheden zijn voor de meeste toetsen bevredigend, verschillen weinig per meetmoment en wijken niet wezenlijk af tussen de verschillende etnische groepen. Opvallend is de lage betrouwbaarheid van de sorteertaak uit de

SON. Omwille van de vergelijkbaarheid met de metingen uit voorgaande jaren hebben we besloten om meetinstrument te handhaven.

In de tabellen 2, 3 en 4 zijn enkele basisgegevens over de toetsen en beoordelingslijsten weer-gegeven.

Tabel 2: Gemiddelden, standaardafwijkingen en maximum-score van de tests bij eerste (1997), tweede (1998) en derde afname (1999).

Test	WSeig 97 n=243	Wsned 97 n=235	VLDP 97 n=243	Sorteren 97 n=236	Blok 97 n=241
Gemiddelde	14.09	12.13	11.46	1.75	2.38
Standaardafwijking	5.72	6.52	6.94	1.91	3.13
Maximumscore	30	30	44	6	30
Test	WSeig 98 n=283	Wsned 98 n=278	VLDP 98 n=283	Sorteren 98 n=283	Blok 98 n=283
Gemiddelde	19.87	16.65	21.90	3.41	8.55
Standaardafwijking	4.43	7.14	6.97	1.85	4.19
Maximumscore	30	30	44	6	30
Test	Wseig 99 N=254	Wsned 99 N=253	VLDP 99 N=235	Sorteren 99 N=238	Blok 99 N=210
Gemiddelde	22.49	21.71	29.14	4.82	18.83
Standaardafwijking	4.63	6.17	6.54	1.31	4.68
Maximumscore	30	30	68	6	30

Daarnaast zijn er vier nieuwe cognitieve toetsen in 1999. Twee toetsen als aanvulling/verlenging van de woordenschattoetsen in eigen taal en het Nederlands (40 items uit de TAK) en de toetsen Begrippen en Ordenen uit het Cito-leerlingvolgsysteem.

Evaluatie Kaleidoscoop en Piramide

Tabel 3: Gemiddelden, standaardafwijkingen en maximum-score van de in 1999 nieuw toegevoegde tests

Test (nieuw in 99)	Takeig99 (n = 246)	Takned99 (n = 245)	Begrippen (n = 212)	Ordenen (n = 178)
Gemiddelde	14.85	15.22	26.19	21.09
Standaardafwijking	3.77	3.59	7.66	6.45
Maximum score	20	20	40	36

Tabel 4: Gemiddelden, standaardafwijkingen en maximum-score van de SCHOBL-schalen bij eerste (1997) tweede (1998) en derde (1999) afname.

Test	Extraversie 97* n=199	Werkhouding 97 n=199	Aangenaam 97 n=199	Emot Stabiel 97 n=199
Gemiddelde	25.08	44.50	34.79	22.94
Standaardafwijking	8.92	14.71	9.88	6.10
Maximumscore	42	72	60	42
Test	Extraversie 98* n=271	Werkhouding 98 n=271	Aangenaam 98 n=271	Emot Stabiel 98 n=271
Gemiddelde	24.01	42.68	35.88	22.17
Standaardafwijking	7.96	13.06	8.52	5.84
Maximumscore	42	72	60	42
Test	Extraversie 99* N=220	Werkhouding 99 N=218	Aangenaam 99 N=216	Emot Stabiel 99 N=224
Gemiddelde	24.64	45.08	34.87	22.54
Standaardafwijking	7.75	12.94	9.01	5.43
Maximumscore	42	72	60	42

*De schaal extraversie verwijst naar externaliserend gedrag en een hoge score wijst op probleemgedrag.

Met het oog op de vergelijkbaarheid met de interimrapportage (Leseman e.a., 1999) presenteren we eerst de correlaties tussen de eerste en de tweede testafname en voegen hieraan vervolgens de berekeningen van de derde testafname toe. De

scores op de eerste en de tweede testafname zijn weergegeven in tabel 5 en 6. De correlaties van de eerste meting staan onder de diagonaal, die van de tweede meting erboven. Op de diagonaal staan de correlaties tussen de meetmomenten. De correlaties zijn in het algemeen volgens verwachting en wijzen er op dat de toetsen intern valide zijn en voldoende betrouwbaar zijn afgenomen. Voor wat betreft de woordenschat eigen taal, de woordenschat Nederlands en de vragenlijst denkprocessen zijn ook de cross-lag correlaties tussen beide meetmomenten goed (resp. .53, .61 en .58). De sorteertaak uit de SON en de blokpatronentest wijken echter af. Ook de correlaties van deze toetsen met de andere toetsen zijn overwegend laag.

Tabel 5: Correlaties tussen de tests; onder de diagonaal de tests van de eerste meetronde (1997), boven de diagonaal de tests van de tweede meetronde (1998), en op de diagonaal de correlaties tussen de eerste en tweede afname per test.

Test	1998	WSeig	Wsned	VLDP	Sort	Blok
1997						
Woordenschat Eigen Taal	.53	.53	.60	.37	.39	
Woordenschat Nederlands	.56	.61	.42	.30	.35	
Vragenlijst Denkprocessen	.60	.54	.58	.43	.50	
Sorteren SON 2-7	.28	.16	.39	.25	.21	
Blokpatronen LDT	.32	.12	.34	.43	.36	

Tabel 6: Correlaties tussen de SCHOBL-schalen; onder de diagonaal de schalen van de eerste meetronde (1997), boven de diagonaal de schalen van de tweede meetronde (1998), en op de diagonaal de correlaties tussen de afname in 1997 en 1998 per schaal.

Schaal	1998	Extraversie	Werkhouding	Aangenaam	Emot Stabiel
1997					
Extraversie	.57	-.43	-.67	.05	
Werkhouding	-.26	.51	.45	-.08	
Aangenaam gedrag	-.75	.33	.35	.05	
Emotioneel Stabiel	-.01	-.10	.08	.17	

Een mogelijke verklaring is dat de afname van SON en blokpatronen in 1997 minder betrouwbaar is geweest door de relatief korte tijd die er toen was om testleiders in

Evaluatie Kaleidoscoop en Piramide

het gebruik van deze moeilijker af te nemen toetsen te instrueren. Ook kan de moeilijkheidsgraad van deze tests een rol spelen, waardoor ze bij de eerste meting onvoldoende differentiëren. De lage cross-lag correlaties voor aangenaam gedrag en emotionele stabiliteit zouden er op kunnen wijzen dat dit type gedrag moeilijk door een leidster of leerkracht is te beoordelen; het is gedrag dat nauwkeuriger in een gezinssituatie kan worden vastgesteld (G.A. Kohnstamm, persoonlijke mededeling). Tabel 10 laat overigens zien dat de cross-lag correlaties voor aangenaam gedrag en emotionele stabiliteit tussen 98 en 99 duidelijk hoger zijn dan tussen 97 en 98. Mogelijk wordt dit type gedrag met het ouder worden manifester en stabiel en hiermee beter te beoordelen; mogelijk is ook dat het instrument zich beter leent voor gebruik in de basisschool dan in de peuterspeelzaal.

Tabel 7: Correlaties tussen toetsscores uit 1999

Test	Wseig 99	Wsned 99	Takeig 99	Takned 99	VLDP 99	Sor- teren 99	Blok 99	Orde- nen	Begrip- pen
Wseig99	1.00								
Wsned99	.71	1.00							
Takeig99	.73	.72	1.00						
Takned99	.69	.83	.70	1.00					
VLDP99	.58	.60	.52	.52	1.00				
Sorteren99	.08	.00	.02	.05	.01	1.00			
Blok99	.17	.12	.03	.08	.37	.25	1.00		
Ordenen	.39	.42	.40	.40	.48	.15	.24	1.00	
Begrippen	.54	.68	.57	.56	.65	.19	.21	.67	1.00

Ook nu weer komen vooral de SON en in iets mindere mate de blokpatronen naar voren als toetsen die laag correleren met de andere toetsen. De hoge correlaties tussen de (verbaal-cognitieve) toetsen, de woordenschattoetsen en de Vragenlijst Denkprocessen enerzijds en de lage correlaties tussen de nonverbaal-performale toetsen, de SON en de Blokpatronen en de andere toetsen anderzijds, zouden erop kunnen duiden dat deze toetsen in feite een onderliggende verbaal-cognitieve factor meten.

Tabel 8 geeft de correlaties tussen toetsscores op dezelfde test over de 3 achtereenvolgende meetjaren. Over het algemeen zijn de correlaties tussen de jaren bevredigend. Bij de SON zijn ze echter aan de lage kant.

Tabel 8: Correlaties tussen toetsscores op dezelfde test over de eerste, tweede en derde meting

Test	97-98	98-99	97-99
Wseig	.53	.53	.37
Wsned	.62	.73	.59
VLDP	.58	.64	.54
Sorteren	.20	.30	-.02
Blok	.28	.73	.34

Voor de volledigheid geven we in de tabellen 9 en 10 de correlaties weer tussen respectievelijk de gedragsoordelen uit 1999 en de correlaties tussen de gelijke gedragsoordelen over drie meetmomenten.

Tabel 9: Correlaties tussen gedragsoordelen uit 1999

Gedragsoordeel	Extraversie	Werkhouding	Aangenaam gedrag	Emotionele stabiliteit
Extraversie	1.00			
Werkhouding	-.47	1.00		
Aangenaam gedrag	-.71	.49	1.00	
Emotionele stabiliteit	-.04	.14	.16	1.00

Tabel 10: Correlaties gelijke gedragsoordelen over 3 meetmomenten

Gedragsoordeel	97-98	98-99	97-99
Extraversie	.57	.70	.30
Werkhouding	.51	.60	.45
Aangenaam gedrag	.35	.55	.23
Emotionele stabiliteit	.17	.54	.10

2.4 Uitval en doorstroming

Door de complexe situatie van het experiment, waarbij peuterspeelzalen en basisscholen tezamen een driejarig educatief traject aanbieden, zonder dat doorstroming binnen het samenwerkingsverband is gegarandeerd, is van het begin af aan de mogelijke uitstroom vanuit de peuterspeelzaal naar een andere dan de aan het experiment deelnemende basisschool een grote zorg geweest. Bij de selectie van proeflocaties was de doorstroom een belangrijk aandachtspunt en ook op de locaties zelf is op verschillende manieren geprobeerd ouders te binden aan het gehele traject en te stimuleren hun kinderen naar de samenwerkende basisschool te laten gaan. Uitval betekent verdunning van de interventie en diffusie van de ingezette middelen. Ook voor het evaluatie-onderzoek is uitval een ernstige complicatie, omdat het leidt tot reductie van het aantal kinderen dat gedurende de hele looptijd van een programma gevolgd kan worden en ook tot vertekeningen wanneer de uitval selectief is.

Uit de onderzoeksresultaten blijkt dat de zorg om de uitval niet misplaatst was. In tabel 6 staan de cijfers vermeld. Hoewel de van tevoren vastgelegde plafonds van resp. 50%-uitval bij de overgang van peuterspeelzaal naar basisschool en 25%-uitval van het eerste naar het tweede kleuterjaar gelukkig niet bereikt worden, is de gemiddelde uitval van plm. 41% in 1998 en 15% in 1999 hoog te noemen. In de Kaleidoscoop-conditie was de uitval bij de overgang peuterspeelzaal-basisschool zelfs 46%. Om de uitval te compenseren zijn voor de 1998-meting nieuwe kinderen geworven. Het aandeel in 1998 nieuw toegevoegde kinderen is in de drie condities ongeveer gelijk, namelijk zo'n 37%.

Tabel 11 Uitval, instroom en doorstroom van peuters en kleuters, naar programma

	Kaleidoscoop	Piramide	Controle	Totaal
In 1997 in het onderzoek	108	115	104	327
Uitgevallen	50	44	41	135
Doorgestroomd	58	71	63	192
In 1998 in het onderzoek	92	113	102	307
Doorgestroomd	58	71	63	192
nieuw ingestroomd	34	42	39	115
In 1999 in het onderzoek	80	96	85	261
vanaf 1997 doorgestroomd	50	59	52	161
in 1998 nieuw ingestroomd	30	37	33	100
uitgevallen in 1999 (t.o.v. 1998)	12 (13%)	17 (15%)	17 (17%)	46 (15%)

Voor de analyse van de effecten van de volledige drie-jarige programma's, met inbegrip dus van de peuterperiode (vanaf 1997), resteren in 1999 161 leerlingen. Voor 100 leerlingen geldt dat zij pas tijdens de basisschool, vanaf 1998, in de programma's zijn ingestroomd.

De analyses gaan vaak over wat lagere aantallen leerlingen, doordat bij alle meetrondes een aantal kinderen niet of onvolledig is getest (meestal vanwege ziekte). Verder doet een aantal kinderen wel mee vanaf 1997 maar is in dat jaar niet getest. Een andere reden voor lagere aantallen is dat in 1998 een aantal kinderen pas na de zomervakantie getest kon worden. Uit de analyses is gebleken dat de scores van deze kinderen niet meer met die van andere kinderen te vergelijken zijn.

Van de kinderen, die in 1998 betrokken waren in het onderzoek is in 1999 gemiddeld 15% uitgevallen.

Er bestaan tussen de kinderen die in het onderzoek blijven en die uitvallen geen significante verschillen in achtergrondkenmerken. Er is hooguit een zeer lichte tendens dat de uitvallers iets minder vaak van Turkse of Marokkaanse herkomst zijn en dat de thuistaal iets vaker Nederlands is.

Bij de eerdere scores op toetsen treedt één significant verschil op: uitvallers scoorden in 1998 iets lager op de SON ($p < .10$). Op 4 van de 5 cognitieve toetsen scoorden uitvallers iets lager; alleen op woordenschat eigen taal een fractie hoger.

Evaluatie Kaleidoscoop en Piramide

Op de gedragoordelen is er een wisselend beeld en geen van de verschillen tussen uitvallers en blijvers is significant.

2.5 Samenstelling van de onderzoeksgroepen

De samenstelling van de onderzoeksgroepen, met name de vraag of zij als equivalent –vergelijkbaar– kunnen worden beschouwd, is in een quasi-experimentele onderzoeksopzet (zonder de mogelijkheid kinderen door loting toe te wijzen aan condities) van cruciale betekenis. Het onderzoek heeft daarop echter slechts ten dele greep. De keuze van proeflocaties was al gemaakt en daarbij werden criteria gehanteerd die met het oog op equivalentie tot complicaties kunnen leiden. Dit betreft onder andere de selectie van locaties met een grote vertegenwoordiging van achterstandskinderen. Een andere factor waarop het onderzoek geen greep heeft, is de uitstroom bij de overgang van peuterspeelzaal naar de basisschool. Die uitstroom bleek vrij groot en selectief, met name doordat in de Kaleidoscoopconditie –om ons onbekende redenen– een hele groep Marokkaanse kinderen uitviel.

De strategie die is toegepast bij het rekruteren van controle speelzalen en basisscholen is hiervóór toegelicht. Ook de criteria die zijn gehanteerd bij het selecteren van kinderen in de speelzalen en scholen voor het onderzoek, en bij het werven van nieuwe kinderen om de uitstroom te compenseren, zijn toegelicht. Het resultaat hiervan wordt in de volgende tabellen samengevat. In de nu volgende tabellen is de samenstelling naar de meest relevante kind- en achtergrondkenmerken van de drie condities weergegeven. Per meting wordt een overzicht gegeven. De genoemde achtergrondvariabelen zijn als volgt geconstrueerd (de gegevens zijn verzameld via de oudervragenlijsten). De sociaal-economische status (SES) van het gezin is een index gebaseerd op het opleidingsniveau en het aantal jaren onderwijs van beide ouders. De etnische herkomst is gebaseerd op het geboorteland van het kind, de ouders en grootouders aan moederszijde. In het onderzoek worden drie etnische groepen onderscheiden: Nederlanders, Turken en Marokkanen. Surinaamse en Antilliaanse kinderen kwamen in zeer geringe aantallen voor in de onderzoeksgroepen en zijn toegevoegd aan de groep Nederlanders. De thuistaal is een index gebaseerd op een viertal vragen naar het taalgebruik in het gezin. De maximum-score 1 betekent dat door ouders en andere kinderen in gezin altijd Nederlands met het

onderzoekskind wordt gesproken; de minimumscore 0 betekent dat thuis altijd een andere taal, de eigen taal, wordt gesproken.

Tabel 12 geeft de samenstelling naar de meest relevante kind- en achtergrondkenmerken van de drie condities tijdens de eerste meting in 1997 weer. Tussen de condities bestaan kleine verschillen die, behoudens het verschil in gemiddelde leeftijd van maximaal 1.2 maand, niet significant zijn, ondanks de relatief grote aantallen kinderen.

Tabel 12: Samenstelling in 1997 van de totale steekproef, inclusief latere uitstromers, naar achtergrondkenmerken.

Kenmerk	Kaleidoscoop N=105	Piramide n=113	Controle n=91	η^2
Leeftijd in maanden bij eerste meetronde	42.62 (3.73)	42.65 (3.55)	41.44 (3.20)	.02*
Percentage meisjes	52%	58%	49%	.01
Percentage Turkse kinderen	23%	32%	24%	.01
Percentage Marokkaanse kinderen	17%	12%	9%	.01
Index Thuistaal Nederlands	.52 (.46)	.56 (.42)	.59 (.42)	.00
Sociaal-economische Status	.09 (.70)	-.05 (.47)	.01 (.62)	.01
Opleidingsniveau Vader	3.89 (1.79)	3.54 (1.17)	3.67 (1.42)	.01

$p < .10$; * $p < .05$; ** $p < .01$

Zoals genoemd is bij de overgang naar de basisschool ongeveer 41% van de oorspronkelijke groep uitgevallen en zijn, om de onderzoeksgroep voor de resterende interventieperiode aan te vullen, in groep 1 nieuwe kinderen geworven. In tabel 13 is de samenstelling weergegeven naar kind- en achtergrondkenmerken van de onderzoeksgroep ten tijde van de tweede meting in de zomer van 1998. De groepen blijken op twee aspecten significant te verschillen, namelijk het aandeel Turkse kinderen (het grootst in de Piramide-

Evaluatie Kaleidoscoop en Piramide

conditie) en de sociaal-economische status van het gezin (het hoogst in de Kaleidoscoop-conditie). Ook zijn er in de Kaleidoscoop-conditie naar verhouding nog steeds meer Marokkaanse kinderen, is in de controlegroep Nederlands vaker de thuistaal en zijn er kleine verschillen in gemiddelde leeftijd. De conclusie is dat de groepen redelijk equivalent zijn, maar dat in verdere statistische analyses met name voor de etnische samenstelling (en daarmee samenhangend thuistaal en SES) moeten worden gecontroleerd.

Tabel 13: Samenstelling in 1998 van de totale steekproef, inclusief nieuwe instroom, naar achtergrondkenmerken.

Kenmerk	Kaleidoscoop N=91	Piramide n=112	Controle n=92	η^2
Leeftijd in maanden bij tweede meetronde	55.32 (3.41)	55.62 (3.37)	54.98 (3.38)	.01
Percentage meisjes	51%	53%	48%	.01
Percentage Turkse kinderen	27%	36%	18%	.03*
Percentage Marokkaanse kinderen	25%	17%	18%	.01
Index Thuistaal Nederlands	.49 (.43)	.50 (.42)	.58 (.43)	.01
Sociaal-economische Status	.20 (.64)	.03 (.51)	-.04 (.79)	.02*
Opleidingsniveau Vader	3.87 (1.97)	3.62 (1.36)	3.47 (1.74)	.01

$p < .10$; * $p < .05$; ** $p < .01$

In de Kaleidoscoop conditie zijn meer Marokkaanse kinderen, in de Piramide-conditie meer Turkse, maar het gemiddelde SES-niveau is in Kaleidoscoop iets hoger dan in beide andere condities. In de gezinnen van de controlekinderen is Nederlands iets vaker de belangrijkste spreektaal dan in de andere condities. Aan Piramide nemen naar verhouding meer meisjes deel. De verschillen zijn echter betrekkelijk klein en niet significant. De conclusie is daarom dat de opzet om te komen tot equivalente groepen is geslaagd. Toch is het raadzaam in de verdere

statistische analyses de variabelen waarop de condities verschillen op te nemen als co-variabelen, om de equivalentie statistisch te versterken.

Tabel 14: Samenstelling in 1998 van de groep kinderen die vanuit de speelzalen doorstroomden naar deelnemende basisscholen, naar achtergrondkenmerken.

Kenmerk	Kaleidoscoop N=55	Piramide n=69	Controle n=50	η^2
Leeftijd in maanden bij tweede meetronde	55.40 (3.64)	55.21 (3.35)	54.00 (2.88)	.03†
Percentage meisjes	52%	55%	48%	.01
Percentage Turkse kinderen	18%	33%	26%	.02
Percentage Marokkaanse kinderen	15	13%	10%	.00
Index Thuis taal Nederlands	.64 (.44)	.51 (.43)	.55 (.44)	.02
Sociaal-economische Status	.20 (.56)	-.01 (.42)	-.03 (.59)	.04*
Opleidingsniveau Vader	4.26 (1.74)	3.72 (1.24)	3.53 (1.24)	.04*

$p < .10$; * $p < .05$; ** $p < .01$

Met het oog op longitudinale groei-analyses van belang te weten hoe de groepen doorgestroomde kinderen zijn samengesteld. Tabel 14 geeft de resultaten. Er blijkt een klein, op $p < .10$ significant verschil in gemiddelde leeftijd, en wat grotere significante verschillen wat betreft SES. In de Kaleidoscoop-conditie is verder het aandeel Turkse kinderen lager dan in beide andere condities en wordt thuis vaker Nederlands gesproken. Uit een vergelijking van de oorspronkelijke onderzoeksgroep met de groep doorstromers blijkt dat de uitval in de Kaleidoscoopconditie selectief was. Het aandeel allochtone kinderen is verminderd en in samenhang hiermee zijn de gemiddelde scores op de index Nederlands als thuistaal en sociaal-economische status verhoogd. Dit komt doordat een complete groep met Marokkaanse kinderen uitviel. Toch is de equivalentie van de drie condities ook in de groep doorstromers redelijk te

Evaluatie Kaleidoscoop en Piramide

noemen. Door gebruik van covariantie-analyse kan afdoende voor de verschillen worden gecontroleerd.

Tabel 15 geeft de samenstelling naar kind- en achtergrondkenmerken van de onderzoeksgroep ten tijde van de derde meting in de zomer van 1999 weer.

Tabel 15: Samenstelling in 1999 van de totale steekproef, inclusief nieuwe instroom, naar achtergrondkenmerken.

	Kaleidoscoop	Piramide	Controle	eta-kwadraat
Leeftijd in maanden bij de derde meetronde	67.05 (3.61)	67.74 (3.36)	66.75 (3.32)	.02
Percentage meisjes	54%	47%	44%	.01
Percentage Turkse kinderen	29%	41%	18%	.05 **
Percentage Marokkaanse kinderen	23%	17%	19%	.00
Index thuistaal Nederlands	.48 (.43)	.44 (.42)	.61 (.42)	.03 *
Sociaal-economische Status	.12 (.66)	.03 (.48)	-.03 (.78)	.01
Opleidingsniveau Vader	3.71 (1.94)	3.54 (1.45)	3.54 (1.72)	.00

Doordat de kinderen die in 1999 uitgevallen zijn nauwelijks verschillen van de kinderen die in het onderzoek blijven, is het beeld in deze tabel vergelijkbaar met de samenstelling van de onderzoeksgroep in 1998 (Tabel 13). In de Piramide-conditie zijn meer Turkse kinderen en in de gezinnen van de controlekinderen is Nederlands wat vaker de belangrijkste spreektaal dan in de andere condities.

3 Resultaten

In deze paragraaf komen de onderzoeksresultaten aan de orde. Allereerst wordt ingegaan op verschillen in gemiddelde toets- en gedragsbeoordelingsscores tussen de beide programma's en de controlegroep in 1997, 1998 en in 1999. Dan worden door middel van zogenaamde residuele verschillen tussen de programma's en de controleconditie bepaald in gemiddelde 'groei' tussen de eerste (1997) en de laatste meting (1999). Bij de analyses onderscheiden we de groep van 'alle kinderen' en de 'doorstromers'. Bij alle kinderen gaat het om de kinderen die sinds 1997 aan het onderzoek meedoen én de kinderen die in 1998 nieuw in het onderzoek zijn ingestroomd. Bij de doorstroomgroep gaat het om kinderen die in het peuterjaar deelnamen aan de programma's of aan een peuterspeelzaal uit de controlegroep en niet tussentijds uit het onderzoek zijn uitgestroomd.

Vervolgens presenteren we uitkomsten van analyses die betrekking hebben op het vaststellen van eventuele verschillen tussen de verschillende etnische groepen in de mate waarin zij profiteren van de programma's.

Zoals in 2.3 opgemerkt zullen we op dit moment de gegevens die verzameld zijn met de toetsen Ordenen en Begrippen van het Cito-leerlingvolgsysteem alleen gebruiken voor de vergelijking van de programma-kinderen met de controleconditie. De verzamelde gegevens zullen niet worden gebruikt voor een vergelijking tussen de programmakinderen en de groep-2-leerlingen uit het Primacohort-onderzoek op de toetsen Begrippen en Ordenen. We hebben vastgesteld dat de onderzoekskinderen op grond van hun leeftijd dermate onvergelijkbaar zijn met de Primacohort-kinderen dat rapportage over deze gegevens weinig zinvol is.

3.1 *Verschillen in taal, cognitie en sociaal-emotioneel gedrag in 1997*

Tabel 16 geeft de gemiddelde scores op de toetsen en de sociaal-emotioneel-gedragsbeoordelingsschalen in 1997 voor de kinderen die binnen de programma's doorstroomden naar de basisschool. Met behulp van covariantie-analyses is nagegaan of de verschillen significant zijn. De weergegeven gemiddelden zijn de

Evaluatie Kaleidoscoop en Piramide

zogenaamde aangepaste ('adjusted') gemiddelden, na controle voor de effecten van de co-variaten. De d-scores voor de paarsgewijze vergelijking van de programmacondities met de controleconditie zijn berekend door de verschillen in aangepaste gemiddelden te delen door de overall standaardafwijking.

Er zijn bij de eerste meetronde in 1997, na gemiddeld een half jaar implementatie, geen duidelijke verschillen tussen de programmacondities en de controleconditie. Op de toetsen woordenschat eigen taal, denkprocessen, sorteren en blokpatronen scoren Piramidekinderen iets hoger dan controlekinderen, maar de effecten zijn in termen van Cohen (1988) zwak ("weak")³.

Uit de gedragsbeoordelingsscores blijkt geen consistent patroon. Kaleidoscoopkinderen zijn volgens hun leidsters minder extravert dan controlekinderen, dat wil zeggen, ze vertonen minder problematisch externaliserend gedrag in de klas. Op de schaal emotionele stabiliteit wijken zowel de Kaleidoscoop- als Piramidekinderen in negatieve zin af.

De vraag of de (kleine) verschillen tussen de condities wijzen op effecten van de programma's, dan wel op verschillen in beginniveau tussen de kinderen, is gegeven de onderzoekopzet en het feit dat de eerste meetronde pas laat kon worden uitgevoerd, niet te beantwoorden. De kinderen namen globaal al ongeveer een half jaar deel aan de programma's, terwijl sommigen al langer de betreffende speelzalen bezochten. De eerste meting in 1997 is derhalve niet te beschouwen als een voormeting. Omdat de drie onderzoeksgroepen naar achtergrondkenmerken redelijk equivalent zijn, is de voorlopige conclusie dat de resultaten van de meting in 1997 nog geen duidelijke effecten van de programma's laten zien ten opzichte van reguliere peuterspeelzalen. Een mogelijke verklaring is dat de deelnameduur daarvoor nog te kort is. Een andere verklaring is dat de implementatie van de programma's in dit eerste jaar nog niet maximaal kon zijn, omdat ingrijpende vernieuwingen nu eenmaal tijd vergen.

³ Als criteria voor effectgrootte hanteert Cohen (1988): weak (d=0.20); medium (d=0.50); strong (d=0.80). In deze rapportage zijn deze begrippen vertaald met 'zwak', 'matig' en 'sterk'. Mogelijke alternatieve vertalingen voor 'zwak' en 'matig' zijn 'klein' en 'middelgroot' (Cohen, J. (1988) *Statistical power analysis for the behavioural sciences*. New York: Academic Press).

Tabel 16: Resultaten van de 1997-meting, gecontroleerd voor covariaten; alle kinderen die in 1997 zijn getest, inclusief uitstroom; gemiddelden, SD's, en d-scores voor paarsgewijze vergelijkingen tussen programma's en controlegroep.

Test	Kaleidoscoop (N=60)	Piramide (N=95)	Controle (N=75)	d (K-C) (N=135)	d (P-C) (N=170)
WS eigen taal	12.75 (5.80)	14.80 (5.94)	13.92 (5.15)	-.21	.16
WS Nederlands	11.23 (7.34)	12.85 (6.62)	12.11 (5.35)	-.14	.11
Denkprocessen	10.24 (6.48)	12.86 (7.74)	10.74 (6.03)	-.07	.30†
Sorteren	1.31 (1.88)	2.02 (1.98)	1.75 (1.84)	-.23	.14
Blokpatronen	2.39 (2.72)	2.69 (3.56)	2.00 (2.99)	.12	.22
Extraversie ^a	24.89 (9.05)	24.25 (9.57)	25.75 (8.56)	-.10	-.17
Werkhouding	43.43 (15.19)	44.35 (16.11)	45.49 (13.55)	-.14	-.08
Aangenaam gedrag	33.38 (8.67)	35.18 (10.27)	35.64 (10.51)	-.23	-.05
Emotionele stabiliteit	21.11 (5.26)	22.50 (6.54)	24.51 (6.25)	-.55	-.33

^a Een hoge score op deze schaal betekent negatief / probleemgedrag

Covariaten: Turks, Marokkaans, thuistaal, SES, sekse.

$p < .10$; * $p < .05$; ** $p < .01$

3.2 Verschillen in taal, cognitie en sociaal-emotioneel gedrag in 1998

Al bij de tweede meting (d.w.z.: naar aanleiding van de analyses die werden uitgevoerd ten behoeve van de interimrapportage) bestond het vermoeden dat de controlegroep niet in alle opzichten vergelijkbaar was met beide experimentele groepen. Naar individuele kenmerken zijn er weliswaar weinig verschillen (zie Tabel 13), maar het vermoeden bestond dat de scholen die door de controleleerlingen bezocht werden wel verschilden van de experimentele scholen.

Evaluatie Kaleidoscoop en Piramide

Door de schoolscores van de controle- en de experimentele scholen te vergelijken kunnen we nagaan of dit het geval is. De schoolscore of SE-score heeft betrekking op de sociaal-economische status en de etnische achtergrond van de leerlingpopulatie. Naarmate de school meer allochtone leerlingen heeft en het opleidingsniveau van de ouders lager is, is de SE-score hoger. In de schoolscores, die gebaseerd zijn op de zgn. leerlinggewichten, weerspiegelt zich dus de samenstelling van de leerlingbevolking.

In de volgende tabel geven we gemiddelde schoolscores voor de controle- en de experimentele scholen⁴.

Tabel 17: Gemiddelde schoolscores voor de controle- en de experimentele scholen

	Kaleidoscoop	Piramide	Controle
gem. schoolscore	141.2	140.8	122.7
st. dev.	22.8	21.4	24.6
aantal scholen	6	12	13

Uit de tabel blijkt dat de controlescholen qua samenstelling van de leerlingbevolking systematisch afwijken van beide groepen experimentele scholen. Hoewel de kinderen uit de controle en experimentele groepen dus wel op elkaar lijken, krijgen zij onderwijs op scholen die minder vergelijkbaar zijn. En deze verschillen tussen de scholen kunnen hun invloed hebben op de ontwikkeling die de kinderen doormaken en daarmee het bepalen van het effect van de programma's bemoeilijken.

Als aanvulling op de controle voor verschillen in individuele kenmerken (covariaten), zoals die in de interimrapportage heeft plaatsgevonden, zullen we daarom nu ook controleren voor de schoolscore van de school waarop de leerling zit⁵. De toevoeging van deze nieuwe covariaat brengt een wijziging teweeg in de

⁴ Cijfers zijn afkomstig van de afdeling Cfi van het Ministerie van Onderwijs, Cultuur en Wetenschappen.

⁵ Idealiter zou hier een multiniveau analyse moeten worden uitgevoerd. Vanwege de lage aantallen (er zijn bijvoorbeeld verschillende scholen met maar 1 leerling) hebben we hier vanaf gezien.

analyses van de meting 1998 die werden uitgevoerd ten behoeve van de interimrapportage (en daarmee in de tabellen 11, 12 en 13 uit de interimrapportage).

Tabel 18 geeft de gemiddelde scores op de toetsen en de gedragsbeoordelingschalen in 1998 voor alle kinderen, inclusief de nieuwe instromers die niet in de peuterfase aan een van de programma's deelnamen of tot de controlegroep behoorden en ongeveer 37% van de totale onderzoeksgroep uitmaken. Ten tijde van de 1998-meting hebben de doorgestroomde kinderen gemiddeld ongeveer anderhalf jaar aan een van de programma's deelgenomen, de nieuw ingestroomde kinderen gemiddeld ongeveer een half jaar. Er blijken tussen de condities geen noemenswaardige verschillen te bestaan. De toets sorteren vormt een uitzondering met een klein effect ten gunste van Piramide. Ook wat betreft het sociaal-emotioneel gedrag doen zich geen duidelijke verschillen voor.

Evaluatie Kaleidoscoop en Piramide

Tabel 18: Resultaten van de 1998-meting, gecontroleerd voor covariaten inclusief de schoolscore: alle kinderen, inclusief nieuwe instroom uit 1998; gemiddelden, SD's en d-scores voor paarsgewijze vergelijkingen tussen programma's en controlegroep.

	Kaleidoscoop	Piramide	Controle	d (K-C)	d (P-C)
<i>Toetsen</i>	(N=84)	(N=104)	(N=83)		
WSeig98	19.76	20.31	19.38		
(sd)	(4.4)	(3.7)	(5.3)	.09	.21
WSned98	16.49	16.80	16.14		
(sd)	(7.2)	(7.1)	(7.3)	.05	.09
VLDP98	21.86	22.05	21.73		
(sd)	(7.0)	(6.8)	(7.3)	.02	.05
SON98	3.46	3.64	3.05		
(sd)	(1.7)	(1.8)	(2.0)	.22	.32 *
ldt blok 98	8.35	8.84	8.35		
(sd)	(3.9)	(4.5)	(3.9)	.00	.12
<i>Gedragsoordelen</i>					
Extraversie ^a	24.90	24.00	23.09		
(sd)	(8.8)	(7.1)	(8.2)	.23	.11
Werkhouding	43.26	42.68	42.20		
(sd)	(13.2)	(13.7)	(12.5)	.08	.04
Aangenaam gedrag	36.79	35.49	36.60		
(sd)	(9.3)	(7.6)	(8.5)	.02	.13
Emotionele stabiliteit	22.66	22.20	21.73		
(sd)	(5.8)	(5.8)	(6.0)	.16	.08

^a Een hoge score op deze schaal betekent negatief / probleemgedrag
 Covariaten: Turks, Marokkaans, thuistaal, SES, sekse; schoolscore
 # p<.10; * p<.05; ** p<.01

In tabel 19 staan de resultaten van de meting in 1998, maar nu beperkt tot de kinderen die vanuit de peuterspeelzaal binnen de programma's doorstromden naar de basisschool. De resultaten zijn, wanneer de analyse tot de doorstromers wordt beperkt, aanmerkelijk gunstiger. Zowel de Kaleidoscoop als Piramide-kinderen halen op de cognitieve en taaltoetsen gemiddeld betere resultaten. De effecten van Piramide zijn daarbij over het geheel genomen wat sterker. Met uitzondering van de toets blokpatronen, zijn alle verschillen tussen Piramide en de controlegroep significant op $p < .05$ en $p < .01$. De gemiddelde effectgrootte is

.40, dat wil zeggen, matig sterk. Op de toetsen woordenschat eigen taal en sorteren zijn ook de verschillen tussen Kaleidoscoop en de controlegroep significant. De gemiddelde effectgrootte voor Kaleidoscoop bedraagt .25, dat wil zeggen, een zwak effect.

Wat betreft het sociaal-emotionele gedrag zijn de verschillen tussen de condities ook in deze groep niet noemenswaardig. Kaleidoscoopkinderen worden als iets extravertter, dus als iets meer probleemgedrag vertonend, beoordeeld (nota bene: dat is anders dan in de peutertijd!). Kaleidoscoopkinderen hebben - daarentegen echter ook- gemiddeld een iets betere werkhouding. De Piramidekinderen scoren iets hoger op het kenmerk aangenaam gedrag, maar worden juist weer iets minder gunstig beoordeeld dan de kinderen in de controlegroep op het kenmerk emotionele stabiliteit. De effecten zijn echter klein.

Evaluatie Kaleidoscoop en Piramide

Tabel 19: Resultaten van de 1998-meting, gecontroleerd voor covariaten inclusief de school-score: alleen de kinderen die vanuit de peuterspeelzaal doorstroomden naar de basisschool; gemiddelden, SD's en d-scores voor paarsgewijze vergelijkingen tussen programma's en controlegroep.

	Kaleidoscoop	Piramide	Controle	d (K-C)	d (P-C)
<i>toetsen</i>	(N=52)	(N=67)	(N=47)		
WSeig98	20.48	20.83	18.81		
(sd)	(3.8)	(3.6)	(5.0)	.40 *	.48 **
WSned98	16.65	17.97	15.41		
(sd)	(7.1)	(6.9)	(7.2)	.18	.36 *
VLDP98	20.77	22.87	20.14		
(sd)	(6.3)	(6.2)	(7.2)	.10	.42 *
SON98	3.21	3.43	2.30		
(sd)	(1.5)	(1.8)	(1.9)	.51 *	.63 **
ldt blok 98	8.24	8.52	7.93		
(sd)	(3.6)	(4.7)	(3.9)	.07	.14
<i>gedragsoordelen</i>					
extraversie ^a	25.29	23.53	23.57		
(sd)	(9.4)	(6.6)	(8.3)	.21	-.00
werkhouding	43.79	42.02	41.87		
(sd)	(14.1)	(13.5)	(11.6)	.14	.01
aangenaam	36.45	35.91	36.97		
gedrag					
(sd)	(9.9)	(7.6)	(8.2)	.06	.12
emotionele	23.01	21.71	22.43		
stabiliteit					
(sd)	(6.0)	(5.4)	(6.1)	.10	-.13

^a Een hoge score op deze schaal betekent negatief / probleemgedrag

Covariaten: Turks, Marokkaans, thuistaal, SES, sekse; schoolscore

p<.10; * p<.05; ** p<.01

3.3 Verschillen in taal, cognitie en sociaal-emotioneel gedrag in 1999

In tabel 20 geven we de gemiddelde scores op de toetsen en de gedragsbeoordelingsschalen in 1999 voor alle kinderen. In de tabel zijn tevens de scores opgenomen op de nieuw opgenomen toetsen Begrippen en Ordenen en de scores op de items uit de Taaltoets Allochtone Kinderen (TAK). Bij Ordenen en Begrippen is het probleem dat relatief veel kinderen uit de controlegroep niet aan deze beide toetsen hebben deelgenomen, terwijl deze ‘uitgevallen’ kinderen op de overige toetsen wat hoger dan gemiddeld scoren. De resultaten van de controlekinderen op Ordenen en Begrippen zijn dus mogelijk te laag. Om deze reden is bij het vergelijken van de gemiddelden van de drie groepen op deze twee toetsen niet alleen voor de gebruikelijke reeks covariaten gecontroleerd, maar ook nog voor de scores op een reeks andere toetsen (de vier woordenschattoetsen en de Blokpatronen uit 99). Op deze manier is geprobeerd zo goed mogelijk te corrigeren voor de selectieve uitval van kinderen bij de afname van deze toetsen.

Evaluatie Kaleidoscoop en Piramide

Tabel 20: Resultaten van de 1999-meting, gecontroleerd voor covariaten inclusief de schoolscore: alle kinderen, inclusief nieuwe instroom uit 1998; gemiddelden, SD's en d-scores voor paarsgewijze vergelijkingen tussen programma's en controlegroep.

	Kaleidoscoop	Piramide	Controle	d (K-C)	D (P-C)
<i>Toetsen</i>	(N=80)	(N=96)	(N=85)		
WSeig99	23.69	22.61	21.23		
(sd)	(3.8)	(5.0)	(4.9)	.53 **	.30 *
WSned99	22.46	22.09	20.77		
(sd)	(5.7)	(6.7)	(5.9)	.27 *	.21 #
Takeig99	15.40	14.75	14.45		
(sd)	(3.8)	(3.9)	(3.7)	.25 #	.08
Takned99	16.14	15.06	14.61		
(sd)	(2.9)	(4.1)	(3.5)	.42 **	.12
Begrippen ^a	28.16	28.12	28.43		
(sd)	(7.7)	(6.3)	(6.2)	-.04	-.05
Ordenen ^a	22.53	24.33	20.95		
(sd)	(6.2)	(5.1)	(5.8)	.26	.56 **
VLDP99	29.57	29.58	28.39		
(sd)	(6.2)	(6.8)	(6.6)	.18	.18
SON99	4.78	4.94	4.76		
(sd)	(1.3)	(1.4)	(1.3)	.01	.13
ldt blok	18.23	19.81	18.77		
(sd)	(4.2)	(5.5)	(4.1)	-.11	.23
<i>Gedragsoordelen</i>					
Extraversie ^b	25.60	24.22	24.20		
(sd)	(8.3)	(8.0)	(6.8)	.18	.00
Werkhouding	45.62	43.57	47.15		
(sd)	(12.2)	(13.5)	(12.7)	-.12	-.28
Aangenaam gedrag	34.73	35.67	33.60		
(sd)	(10.0)	(8.6)	(8.3)	.13	.23
Emotionele stabiliteit	23.31	22.49	21.15		
(sd)	(4.8)	(5.7)	(5.3)	.40 *	.25

^a Hierbij is ook nog gecontroleerd voor de score op vijf andere toetsen. ^b Een hoge score op deze schaal betekent negatief / probleemgedrag. Covariaten: Turks, Marokkaans, thuistaal, SES, sekse; schoolscore. # p<.10; * p<.05; ** p<.01

Het algemene beeld is dat de beide experimentele condities bij de cognitieve toetsen over de hele linie wat hoger scoren, maar de effecten zijn bescheiden (en ook lang niet altijd significant).

Er is geen duidelijk verschil tussen de beide experimenten: op de ene toets scoort het ene programma beter, op de andere toets het andere programma. Effecten van Kaleidoscoop zijn vooral te vinden op de taaltoetsen, de gemiddelde effectgrootte is .37, terwijl bij Piramide de hoge scores op Ordenen opvallen (effectgrootte .56).

Tabel 21: Resultaten van de 1999-meting, gecontroleerd voor covariaten inclusief de school-score: alleen de kinderen die vanuit de peuterspeelzaal doorstroomden naar de basisschool; gemiddelden, SD's en d-scores voor paarsgewijze vergelijkingen tussen programma's en controlegroep.

	Kaleidoscoop	Piramide	Controle	d (K-C)	d (P-C)
<i>Toetsen</i>	(N=50)	(N=59)	(N=52)		
Wseig99	23.80	23.33	22.13		
(sd)	(3.7)	(5.1)	(3.4)	.40 *	.29 #
Wsned99	22.75	22.84	21.57		
(sd)	(5.6)	(6.7)	(5.9)	.19	.21
Takeig99	15.69	15.30	14.71		
(sd)	(3.3)	(4.0)	(3.5)	.27	.16
Takned99	16.06	15.56	15.20		
(sd)	(2.8)	(4.0)	(2.9)	.26	.11
Begrippen ^a	29.00	30.00	28.36		
(sd)	(7.3)	(5.7)	(6.4)	.10	.24
Ordenen ^a	24.19	25.59	20.20		
(sd)	(6.0)	(5.0)	(6.2)	.65 **	.88 **
VLDP99	28.53	29.85	27.54		
(sd)	(6.2)	(6.9)	(6.4)	.15	.36 *
SON99	4.48	4.69	4.56		
(sd)	(1.4)	(1.6)	(1.3)	-.05	.09
ldt blokp	17.61	19.50	18.63		
(sd)	(4.1)	(5.4)	(4.1)	-.22	.19

Evaluatie Kaleidoscoop en Piramide

Vervolg tabel 21

	Kaleidoscoop	Piramide	Controle	d (K-C)	d (P-C)
<i>Gedragsoordelen</i>					
Extraversie ^b	26.04	23.89	23.65		
(sd)	(8.1)	(7.4)	(6.5)	.32	.03
Werkhouding	45.17	43.96	47.41		
(sd)	(13.2)	(13.8)	(11.5)	-.17	-.26
Aangenaam gedrag	33.94	36.58	34.52		
(sd)	(10.0)	(8.2)	(9.0)	-.06	.23
Emotionele stabiliteit	23.33	22.90	22.25		
(sd)	(4.8)	(5.6)	(5.4)	.20	.12

^a Hierbij is ook nog gecontroleerd voor de score op vijf andere toetsen.

^b Een hoge score op deze schaal betekent negatief/probleemgedrag. Covariaten: Turks, Marokkaans, thuistaal, SES, sekse; schoolscore . # $p < .10$; * $p < .05$; ** $p < .01$

De overige effecten blijven doorgaans van dezelfde orde, al zijn ze minder vaak significant doordat de aantallen bij deze doorstroomgroep kleiner zijn. Wanneer we kijken naar de resultaten op de meting 1999 van de doorstroomgroep (tabel 21), dan blijkt dit laatste effect voor Piramide groter te worden, tot een, in termen van Cohen, sterk effect (.88; $p < .01$). Dan zien we ook bij Kaleidoscoop een effect op Ordenen (.65, $p < .01$). De effecten op Begrippen daarentegen zijn zeer bescheiden en niet significant. Voorts vinden we binnen deze doorstroomgroep een zwak tot matig (.36) effect van de Piramidekinderen op de Vragenlijst Denkprocessen.

3.4 Verschillen in 'groei' tussen 1997-1998

Met behulp van multiple regressie-analyse zijn de zogenaamde *residuele verschillscores* bepaald, na aanpassing voor de covariaten. Residuele verschillscores worden bepaald door per toets en beoordelingsschaal een regressievergelijking op te stellen met de toetsscores van de eerste meting en de al genoemde covariaten (SES, etniciteit enz.) als prediktoren (zie ook Hessels, 1993). De door deze prediktoren verklaarde variantie in de scores van de tweede meting wordt vervolgens afgetrokken van de totale variantie; er resteert een

constante (de intercept) en een residuele variantie. De constante staat voor de gemiddelde toename in scores, de resterende residuele variantie voor individuele verschillen in toename. Anders dan gewone verschillscores, gelden residuele verschillscores als betrouwbaar. Residuele verschillscores verwijzen naar verschillen tussen kinderen in 'groei', in dit geval tussen de eerste en derde meting in respectievelijk 1997 en 1999.

Tabel 22 geeft de resultaten voor de drie condities. De verschillscores zijn hierin weergegeven als afwijkingen van de gemiddelde verschillscore, met de bijbehorende standaardafwijking. De gegevens, gebaseerd op relatief kleine aantallen ($N_{\text{totaal}} = 117$), tonen dat de kinderen die de programma's Kaleidoscoop en Piramide volgden op het merendeel van de cognitieve en taaltoetsen tussen 1997 en 1998 sterker vooruitgingen dan de controlekinderen (de meeste d-scores zijn positief). De effectgrootten zijn voor Piramide in termen van Cohen 'bescheiden' oftewel 'matig sterk' (de gemiddelde d-score is .46) en op een na significant op $p < .10$ of $p < .05$. Voor Kaleidoscoop zijn ze kleiner en is alleen de SON significant ($p < .10$). De Vragenlijst Denkprocessen wijkt enigszins af. De prestaties van de Kaleidoscoopkinderen nemen op deze toets gemiddeld minder snel toe dan in de controlegroep, terwijl in Piramidegroep de toename bij deze toets kleiner is dan bij de andere toetsen.

Wat betreft het sociaal-emotioneel gedrag zijn er geen duidelijke (positieve) verschillen in ontwikkeling vast te stellen tussen de drie condities. De werkhouding en emotionele stabiliteit van de Kaleidoscoopkinderen is, volgens de leerkrachten, tussen 1997 en 1998 sterker verbeterd dan van de controle- en Piramidekinderen, maar de effecten zijn zwak.

Evaluatie Kaleidoscoop en Piramide

Tabel 22: Residuele vershilscores 1998-1997, uitgedrukt als afwijking van de gemiddelde residuele vershilscore, gecontroleerd voor covariaten inclusief de schoolscore; gemiddelden, SD's en d-scores voor paarsgewijze vergelijkingen tussen programma's en controlegroep.

	Kaleidoscoop	Piramide	Controle	d (K-C)	d (P-C)
<i>Toetsen</i>	(N=32)	(N=55)	(N=30)		
WSeig98	.36	.42	-1.25		
(sd)	(3.4)	(3.3)	(4.1)	.43	.45 #
WSned98	-.49	1.07	-1.48		
(sd)	(4.5)	(4.6)	(5.1)	.19	.49 #
VLDP98	-1.33	1.02	-.53		
(sd)	(6.5)	(3.7)	(6.1)	-.15	.29
SON98	.11	.28	-.69		
(sd)	(1.3)	(1.8)	(1.9)	.45 #	.55 *
ldt blok	-.36	.78	-1.21		
(sd)	(3.9)	(3.9)	(3.2)	.21	.51 *
<i>Gedragsoordelen</i>					
Extraversie ^a	.50	-.78	.59		
(sd)	(6.4)	(5.5)	(7.3)	-.01	-.24
Werkhouding	1.18	-.74	.03		
(sd)	(12.1)	(9.3)	(9.4)	.10	-.07
Aangenaam gedrag	1.01	-.76	.21		
(sd)	(6.7)	(7.4)	(8.1)	.10	-.12
Emotionele stabiliteit	1.43	-.61	-.21		
(sd)	(5.7)	(5.8)	(5.9)	.27	-.07

^a Een hoge score op deze schaal betekent negatief / probleemgedrag

Covariaten: Turks, Marokkaans, thuistaal, SES, sekse; schoolscore

p<.10; * p<.05; ** p<.01

3.5 Verschillen in 'groei' tussen 1997-1999

Tabel 23 presenteert de resultaten van de 'groei-analyse' voor de drie condities in de periode 1997-1999. Verschillen in groei tussen de programmakinderen en de controlegroep doen zich voor op de Woordenschattoetsen, zowel Nederlands als eigen taal, waarbij het bij Kaleidoscoop gaat om significante verschillen ($p < .10$)

en op de Vragenlijst Denkprocessen. Op de SON is de groei van de programmakinderen ten opzichte van controlegroep iets minder, maar niet significant. Piramidekinderen gaan ten opzichte van de controlegroep iets vooruit op de Blokpatronen, evenmin significant.

Voor wat betreft het sociaal-emotionele gedrag zijn zowel de Piramide- als de Kaleidoscoopkinderen significant sterker vooruit gegaan op de schaal emotionele stabiliteit, waarbij het verschil voor Kaleidoscoop significant is ($p < .10$).

Tabel 23 Residuele verschilcores 1999-1997, uitgedrukt als afwijking van de gemiddelde residuele verschilcore, gecontroleerd voor covariaten inclusief de schoolscore; gemiddelden, SD's en d-scores voor paarsgewijze vergelijkingen tussen programma's en controlegroep.

	Kaleidoscoop	Piramide	Controle	d (K-C)	d (P-C)
<i>Toetsen</i>	(N=33)	(N=54)	(N=47)		
WSeig99	.76	.12	-.85		
(sd)	(3.2)	(5.0)	(3.1)	.40 #	.24
WSned99	1.06	.31	-1.11		
(sd)	(3.5)	(4.6)	(5.0)	.48 #	.32
VLDP99	-.04	.39	-.39		
(sd)	(6.4)	(5.9)	(4.4)	.08	.14
SON99	-.32	.07	.19		
(sd)	(1.6)	(1.5)	(1.4)	-.34	-.08
ldt blokp	-1.01	.89	-.34		
(sd)	(4.3)	(5.4)	(3.1)	-.15	.27
<i>Gedragsoordelen</i>					
Extraversie ^a	.37	-.78	.72		
(sd)	(7.4)	(7.2)	(7.0)	-.05	-.21
Werkhouding	-2.93	.13	3.32		
(sd)	(10.7)	(12.8)	(10.7)	-.54	-.28
Aangenaam gedrag	-1.83	1.37	.12		
(sd)	(8.7)	(7.9)	(9.7)	-.22	.14
Emotionele stabiliteit	1.12	.41	-1.88		
(sd)	(4.9)	(5.4)	(5.4)	.58 #	.44

^a Een hoge score op deze schaal betekent negatief / probleemgedrag

Covariaten: Turks, Marokkaans, thuistaal, SES, sekse; schoolscore

$p < .10$; * $p < .05$; ** $p < .01$

3.6 Verschillen naar etniciteit

Tenslotte is nagegaan of kinderen van Turkse en Marokkaanse herkomst mogelijk extra profiteren van de programma's Kaleidoscoop en Piramide. Daartoe is allereerst onderzocht hoe deze leerlingen scoren op de uitgebreide reeks toetsen van het laatste meetmoment (1999) en welke verschillen daarbij optreden met kinderen van Nederlandse herkomst.

In deze variantie-analyse is in het bijzonder getoetst:

- of er verschillen in gemiddelde scores bestaan tussen de drie condities (de twee programma's en de controlegroep);
- of er verschillen in gemiddelden bestaan tussen de 2 etnische groepen; en
- of er sprake is van een interactie; d.w.z. of het verschil tussen de 2 etnische groepen anders is bij de verschillende programma's.

Om deze interactie is het ons dus te doen: als Turkse en Marokkaanse leerlingen in het bijzonder profiteren van de speciale programma's zou er sprake moeten zijn van een (significant) interactie-effect. In tabel 24 worden de uitkomsten van de analyses samengevat. (nota bene: het gaat hier om 'ruwe' gemiddelden, waarbij niet gecontroleerd is voor covariaten; dat is bij deze vergelijking van etnische groepen uiteraard niet goed mogelijk).

Tabel 24: 'Achterstand' Turkse/Marokkaanse in vergelijking met Nederlandse leerlingen, ruwe scores (d.w.z.: niet gecontroleerd voor covariaten)

	Kaleidoscoop		Piramide	Controle	Significantie van:	
(gr.mean)	Achterstand Turks / Marokkaanse leerl.		prog	etn	intera	
WSeig99 (22.5)	4.0	5.0	5.8	#	**	ns
WSned99 (21.7)	7.7	8.2	8.5	ns	**	ns
Takeig99 (14.9)	3.5	4.7	4.2	#	**	ns
Takned99 (15.2)	3.3	4.7	5.0	*	**	ns
VLDP99 (29.1)	3.5	6.1	5.0	ns	**	ns
SON99 (4.8)	-1.0	-0.6	-0.3	ns	**	ns
ldt blokp (18.8)	-0.7	-0.1	-1.3	ns	ns	ns
Begrippen (26.2)	7.7	8.7	4.3	ns	**	ns
Ordenen (21.1)	6.1	2.2	2.0	**	**	ns
Extraversie ^a (24.6)	1.0	-1.7	0.9	ns	ns	ns
Werkhouding (45.1)	4.3	1.9	3.0	ns	#	ns
Aangenaam gedrag (34.9)	-0.1	5.0	2.1	ns	*	ns
Emotionele stabiliteit (22.5)	0.1	1.8	1.6	#	ns	ns

^a Een hoge score op deze schaal betekent negatief / probleemgedrag
p<.10; * p<.05; ** p<.01

Evaluatie Kaleidoscoop en Piramide

Bij een groot deel van de toetsen zien we een (significant) verschil tussen de etnische groepen: de Turkse en Marokkaanse leerlingen blijven achter bij de Nederlandse leerlingen. Maar er is geen sprake van een interactie-effect, dat wil zeggen dat het verschil tussen de beide etnische groepen in alle drie condities op vergelijkbare wijze terugkomt. Er zijn dus geen aanwijzingen dat leerlingen met een Turkse of Marokkaanse achtergrond meer of minder profiteren van de speciale programma's. De verschillen tussen de programma's hebben bij deze leerlingen dezelfde effecten als bij leerlingen zonder Turkse of Marokkaanse achtergrond.

Bij de voorgaande analyses is alleen gekeken naar de resultaten van 1999, en niet naar eerdere metingen. We kunnen echter, voor kinderen die al in 1997 bij het onderzoek betrokken waren, weer een residuele groeiscore bepalen. Ook bij deze scores kunnen we weer variantie-analyses uitvoeren en nagaan of de gemiddelden verschillen tussen de condities en/of tussen de twee etnische groepen (Turks/Marokkaans versus Nederlands) en of er sprake is van een interactie-effect. Wel worden nu de aantallen erg klein, zodat we voorzichtig moeten zijn met conclusies. De volgende tabel geeft de uitkomsten van de analyses (wanneer gemiddelden over minder dan 10 kinderen zijn berekend, wordt dit apart vermeld).

Tabel 25: 'Achterstand' van Turkse/Marokkaanse in vergelijking met Nederlandse leerlingen in residuele groeiscoringen

	Kaleidoscoop		Piramide	Controle	Significantie van:	
(gr.mean)	Achterstand Turks/ Marokkaanse leerl.		prog	etn	intera	
Wseig (0.0)	3.3	4.6	4.0 (nTuma = 6)	ns	**	ns
Wsned (0.0)	1.5	3.1	9.8 (nTuma = 4)	ns	**	**
VLDP (0.0)	4.6	5.7 (nTuma = 7)	1.8 (nTuma = 4)	ns	**	ns
SON (0.0)	-0.8	-0.7	-0.5 (nTuma = 5)	ns	*	ns
ldt blok (0.0)	-3.8	0.4	-1.7 (nTuma = 5)	ns	ns	ns

p<.10; * p<.05; ** p<.01

Naar aanleiding van tabel 25 concluderen we dat er verschillen zijn in groeiscoringen tussen Nederlandse en Turks/Marokkaanse leerlingen. Deze laatste blijven achter op de verbaal-cognitieve toetsen (Woordenschat eigen taal, Nederlands en Vragenlijst Denkprocessen; $p < .01$), maar scoren op de sorteertoets van de SON wat beter ($p < .05$). De verschillen in groeiscoringen bij de LDT blokpatronen zijn niet significant. Bij de meeste toetsen is er geen sprake van een interactie, dat wil zeggen dat het verschil tussen Nederlandse en Turks/Marokkaanse leerlingen van conditie tot conditie ongeveer gelijk is. De groeiscoringen voor Woordenschat Nederlands van de 4 leerlingen met een Turkse of Marokkaanse achtergrond, die zowel in 97 als in 99 in de controlegroep getoetst zijn blijven achter bij die van Turks/Marokkaanse leerlingen in de beide programma's: hier is sprake van een significant interactie-effect. Deze groep is echter dermate klein dat we hier geen algemene conclusies aan durven te verbinden.

Evaluatie Kaleidoscoop en Piramide

4 Conclusies

In deze rapportage wordt verslag gedaan van het evaluatie-onderzoek naar de programma's Kaleidoscoop en Piramide. Het onderzoek is uitgevoerd met als doel vast te stellen of deze programma's een meerwaarde hebben ten opzichte van regulier kleuteronderwijs. Ten behoeve van dit onderzoek zijn in drie achtereenvolgende jaren (1997, 98 en 99) metingen verricht bij drie groepen kinderen: kinderen die deelnemen aan Kaleidoscoop of Piramide of een controle-peuterspeelzaal/basisschool bezoeken. Na de tweede meting (1998) zijn de tussentijdse resultaten beschreven in een interimrapportage (Leseman e.a., 1999). De onderhavige rapportage heeft betrekking op de resultaten over drie metingen.

Naar aanleiding van de in het voorgaande gepresenteerde resultaten concluderen we dat de programma's al met al gunstig naar voren komen in vergelijking met de controleconditie. Over de gehele linie scoren de programma's hoger dan de controlegroep, al dan niet significant. De effecten zijn echter bescheiden en wisselend. Hierbij is het van belang onderscheid te maken tussen de in het interimverslag gepresenteerde resultaten (het tweede meetmoment, na ongeveer anderhalf jaar deelname aan de programma's; in feite de periode van de peuterspeelzaal) en de resultaten op de eindmeting (1999). Bovendien blijkt dan het beeld voor de doorstroomgroep (de kinderen die vanaf de peuterspeelzaal in 1997 deelnemen) weer wat anders dan voor de gehele groep (alle kinderen inclusief de nieuwe instromers uit 1998). We zullen dit hieronder toelichten.

De onderzoeksresultaten hebben ertoe geleid dat een aantal in de interimrapportage gepresenteerde analyses op grond van nieuwe gegevens in uitgebreide vorm zijn herhaald. Dit naar aanleiding van het vermoeden dat de kinderen uit de controleconditie, ondanks pogingen scholen in dezelfde of vergelijkbare buurten te selecteren, in groepen zitten die wat betreft educatieve achterstand gunstiger zijn samengesteld dan de groepen waarin de programma's worden uitgevoerd. Een indicatie hiervoor vormde het gegeven dat de in 1998 nieuw ingestroomde controlekinderen op de meeste toetsen beter scoorden dan de nieuw ingestroomde programmakinderen. Door de schoolscores van de experimentele en de controlescholen met elkaar te vergelijken is nagegaan of de

Evaluatie Kaleidoscoop en Piramide

scholen qua samenstelling van de leerlingpopulatie van elkaar verschillen. Uit de analyses is gebleken dat deze samenstelling –inderdaad- verschilt en dat de kinderen in de controlegroep waarschijnlijk in een gunstiger leeromgeving vertoeven dan de kinderen uit de programmascholen., een gegeven dat moet worden meegewogen bij de interpretatie van de onderzoeksgegevens.

De resultaten in de interimrapportage gaven een tamelijk positief beeld te zien van de effecten van de programma's, met name de resultaten van de zogenaamde 'doorstroomgroep'. Op het tweede meetmoment hadden deze kinderen maximaal anderhalf jaar deelgenomen aan één van beide programma's. In 1998 scoorden de programma-kinderen uit de doorstroomgroep op alle taal- en cognitieve toetsen hoger dan de controlekinderen. De verschillen waren in een aantal gevallen significant: voor de Piramidekinderen was het verschil met de controlegroep significant ($p < .05$ en $p < .01$) voor de beide Woordenschattoetsen (eigen taal en Nederlands), de Vragenlijst Denkprocessen en de SON-sorteertaak (ook op Blokpatronen scoorde Piramide iets hoger maar niet significant). De Kaleidoscoopgroep scoorde significant hoger ($p < .05$) op de toets Woordenschat eigen taal en op de SON-sorteertaak. Het waren dus vooral de Piramidekinderen die op de taal- en cognitieve toetsen hoger scoorden dan de controlegroep.

Deze resultaten beperkten zich tot de doorstroomgroep; analyses op de gegevens van de hele groep lieten geen noemenswaardige verschillen zien tussen de programmacondities en de controleconditie op de taal- en cognitieve toetsen naar aanleiding van de resultaten op de eerste meting (97) en de tweede meting (98). Een uitzondering vormde in 1998 de toets SON-sorteren, met een matig sterk effect ten gunste van Piramide. Dit beeld is, ook na invoering van het covariaat schoolscore, hetzelfde gebleven.

Naar aanleiding van de meting in 1999 concluderen we dat het beeld dat de programma's over de gehele linie wat hoger scoren dan de controlegroep in grote lijnen overeind blijft. Uit de analyses over alle kinderen komen zowel voor Kaleidoscoop als voor Piramide significante verschillen met de controlegroep naar voren. Kaleidoscoop scoort op alle vier de woordenschattoetsen (Woordenschat DTT-eigen taal en Nederlands en Woordenschat TAK-eigen taal en Nederlands) hoger dan de controlegroep. De gemiddelde effectgrootte voor Kaleidoscoop op deze vier toetsen is .37, dat wil zeggen zwak tot matig.

Piramide scoort hoger op twee van de vier toetsen (Woordenschat DDT eigen taal en Nederlands) en voorts op Ordenen (effectgrootte .56).

Bij de doorstroomgroep zijn de verschillen met de controleconditie kleiner dan voor de hele groep, en alleen significant voor de woordenschattoets DDT-eigen taal (de effectgrootten voor Kaleidoscoop en Piramide zijn resp. .40 en .29) en op Ordenen. De effectgrootte op Ordenen voor Kaleidoscoop is .65. Voor Piramide is het effect op Ordenen bij de doorstroomgroep in vergelijking met 'alle kinderen' sterker (.88; in termen van Cohen een sterk effect). Daarnaast scoren de Piramidekinderen uit de doorstroomgroep hoger op de Vragenlijst Denkprocessen (.36) in vergelijking met de controlegroep.

Vonden we in de interimrapportage vooral binnen de doorstroomgroep verschillen met de controleconditie ten gunste van de programma's, in de meting van 1999 vinden we meer verschillen in de groep 'alle kinderen'. Dit wijst er enerzijds op dat de peuterperiode van beide programma's relevant zou kunnen zijn als het gaat om het boeken van effecten; de meeste effecten vinden we immers bij de kinderen vlak na de instroom in de basisschool. Uit het feit dat de verschillen binnen de doorstroomgroep tussen de programma-condities en de controleconditie later afnemen leiden we af dat de doorstroomgroep deze voorsprong niet weet vast te houden. Deze 'afbuiging' is overigens op zich iets waar we op grond van de gegevens niet zonder meer een verklaring voor kunnen geven, zie verder onder slotopmerkingen.

Als we de drie meetmomenten samennemen, ons beperken tot de toetsen en de gedrags-beoordelingen even buiten beschouwing laten, concluderen we dat de resultaten van de programma's beter zijn in het verbaal-conceptuele domein (woordenschattoetsen, Vragenlijst Denkprocessen, Ordenen) dan in het nonverbale-performale domein (Blokpatronen en SON-sorteren). Eén toets (Begrippen) wijkt in dit opzicht af, een resultaat dat we niet kunnen verklaren. Als het gaat om Kaleidoscoop zijn de verschillen met de controle-conditie in het verbaal-conceptuele domein meestal statistisch significant; bij Piramide zijn de resultaten op het laatste meetmoment iets minder gunstig. De effectgrootten zijn, in de terminologie van Cohen overwegend zwak tot bescheiden, met een enkele uitschieter naar boven (er is een sterk effect op Ordenen) en naar beneden (geen effecten op Begrippen, SON-sorteren en Blokpatronen).

Evaluatie Kaleidoscoop en Piramide

Een mogelijke verklaring voor het differentiële effect van de programma's is dat de vaardigheden die met nonverbaal-performale toetsen worden gemeten, sterker met aangeboren en moeilijker te veranderen cognitieve structuren verband houden (Leseman & van den Boom, 1999). Een feit is ook dat in het onderzoek van de Jong, Klapwijk en Van der Leij (1995) onder kleuters dezelfde Blokpatronentest die in dit onderzoek werd gebruikt geen noemenswaardige en significante verschillen naar etniciteit liet zien.

De scores op de gedragsbeoordelingsschalen van de R-Schobl laten in geen van de metingen betekenisvolle verschillen zien tussen de groepen. Het enige significante verschil zien we in 1999, op de toets emotionele stabiliteit. Beide programma's scoren hoger op dit kenmerk dan de controlegroep en voor Kaleidoscoop is dit verschil significant op $p < .05$. In 1997 constateerden we juist ten aanzien van dit kenmerk een (weliswaar niet significant) lagere score van de programmakinderen in vergelijking met de controlekinderen, vooral voor de Kaleidoscoopgroep. Naar de oorzaak van de verandering in dit beeld kunnen we slechts gissen: zijn de programmakinderen (en dan met name de Kaleidoscoopkinderen) in vergelijking met de controlekinderen inderdaad emotioneel stabiel geworden over de jaren of is het instabiele patroon terug te voeren op wisselende oordelen van de leidsters en de leerkrachten over de kinderen? De beoordelingslijsten zijn immers op twee en soms op alle drie de meetmomenten door verschillende personen ingevuld. In elk geval laten de groeiscoringen (tabel 23) voor Kaleidoscoop een significante toename zien op dit kenmerk ten opzichte van de controlegroep ($p < .10$).

Over de sociaal-emotionele kenmerken kunnen we verder opmerken dat er geen sprake is van een consistent patroon, noch binnen de schalen, noch over de schalen heen. Op de schaal extraversie bijvoorbeeld zien we dat Kaleidoscoop en Piramide in 1997 lager scoorden dan de controlegroep (d.w.z.: volgens de leidster minder problematisch externaliserend gedrag vertoonden), terwijl ze in 1998 en 1999 hoger scoren op deze schaal dan de controlegroep. Bij 'werkhouding' komen de programma's in 1997 en 1999 lager uit dan de controlegroep maar in 1998 scoren ze juist iets hoger. Kortom, de scores vertonen een grillig patroon, waaraan we moeilijk conclusies kunnen verbinden.

Tenslotte zijn analyses uitgevoerd om te kunnen bepalen of alle kinderen, ongeacht etnische achtergrond, in gelijke mate profiteren van de programma's. De programma's zijn bedoeld voor kinderen die de doelgroep vormen van het Onderwijs Achterstandenbeleid: Nederlandse en allochtone kinderen uit de lagere sociaal-economische milieus. Uit de toegepaste analyses blijkt echter dat er geen aanwijzingen zijn dat Turkse en Marokkaanse kinderen meer of minder van de programma's profiteren. Voor zover we op grond van de vrij kleine aantallen kunnen concluderen nemen zij dus evenveel mee van de programma's als de Nederlandse kinderen in het onderzoek.

4.1 Slotopmerkingen

Alles overziend zijn de resultaten van de programma's overwegend positief te noemen, zeker wanneer we het volgende in overweging nemen.

Op basis van een overzicht van buitenlandse, overwegend Amerikaanse 'centerbased' programma's uit de periode 1985-1996, waarin experimentele condities werden vergeleken met zowel reguliere condities (zoals in het onderhavige onderzoek) als met condities waarin geen enkele interventie plaatsvond, worden gemiddelde d-scores in het cognitieve domein gevonden van .41 (algemene intelligentie, rekenen) tot .49 (taal en geletterdheid). Effecten in het sociaal-emotionele domein bleken aanzienlijk kleiner (.20) (Leseman, Blok, Otter & Deckers, 1998). Effectgrootten van rond de .50 gelden als redelijk (Blok & Leseman, 1996).

De grootte van de effecten van met name Kaleidoscoop en in iets mindere Piramide zijn bij de eindmeting dus goed in overeenstemming met wat in vergelijkbaar buitenlands onderzoek, waarin ook programma's met reguliere preschool of kindergarten werden vergeleken, is gevonden, althans voor zover het verbale-conceptuele domein betreft.

Ook in vergelijking met een ander, in Nederland uitgevoerd programma voor dezelfde doelgroep, het programma Opstap Opnieuw, is het beeld positief. In beide onderzoeken zijn deels dezelfde toetsen gebruikt. De scores van Kaleidoscoop en Piramide in de groep doorstromers op de toets Ordenen zijn dermate groot, dat de effectgrootten in termen van Cohen sterk te noemen zijn. In de evaluatie van Opstap Opnieuw blijkt de winst van de Turkse kleuters ook het

Evaluatie Kaleidoscoop en Piramide

grootst op deze en een verwante toets ($d=.50$ respectievelijk $.42$). De scores op dezelfde Woordenschattoetsen Eigen Taal tonen bij Opstap Opnieuw echter kleinere effecten ($d's < .28$) en op dezelfde Woordenschattoetsen Nederlandse Taal geen effecten ($d's < .03$); van Tuijl, Leseman & Rispens, in druk) dan met name Kaleidoscoop ($d's > .27$) te zien geeft. Een ander belangrijk verschil lijkt te zijn dat Marokkaanse kleuters in de programma's Kaleidoscoop en Piramide evenveel profijt hebben als Turkse en Nederlandse kleuters (door de kleine aantallen is er geen definitief uitsluitel te geven; er zijn geen aanwijzingen voor het tegendeel), terwijl er bij de Marokkaanse kleuters in Opstap Opnieuw geen enkel positief effect werd gevonden, waarschijnlijk als gevolg van het feit dat de implementatie van het programma in Marokkaanse gezinnen op vele obstakels stuitte.

Door toevoeging van de schoolscore als covariaat wordt duidelijk hoezeer de school, maar in het bijzonder de groepssamenstelling van invloed is op de individuele ontwikkeling en leerresultaten (zie Terwel, 1997). In het huidige onderzoek zijn de Kaleidoscoop en Piramide-scholen vergeleken met reguliere basisscholen die in sociaal-economisch, etnisch en sociolinguïstisch opzicht waarschijnlijk aanmerkelijk gunstiger omgevingen boden. De SE-score van de Kaleidoscoop- en Piramidescholen was bijna 1 SD hoger⁶. Vanuit het belang om de kinderen uit de controlespeelzalen te volgen was dit onvermijdelijk en de toevoeging van het covariaat schoolscore heeft de onvergelykbaarheid die hier door ontstond enigszins onder controle gebracht.

De school- of SE-score is echter slechts een globale maat voor een hele school en weinig precies als het gaat om de rol van groepsecologische kenmerken op individuele leerprestaties (hierover is inmiddels de nodige literatuur verschenen, waarin bijvoorbeeld gekeken is naar het aantal verschillende taalgroepen, het aandeel etnische minderheden, het gemiddelde cognitieve (begin) niveau, verschillen in gedragstijlen etc. zie Terwel, 1997; Cohen & Lotan, 1995).

⁶ De schoolscore of SE-score heeft betrekking op de sociaal-economische status en de etnische achtergrond van de leerlingpopulatie. Naarmate de school meer allochtone leerlingen heeft en het opleidingsniveau van de ouders lager is, is de SE-score hoger. Een standaarddeviatie (SD) verschil kan aanzienlijk genoemd worden.

Eerder is gesteld dat de programma's Kaleidoscoop en Piramide niet alleen maar goed moeten zijn, maar meerwaarde moeten tonen ten opzichte van reguliere peuterspeelzalen en scholen. De vraag is nu ten opzichte welke reguliere scholen: die met een gemiddelde of zelfs ondergemiddelde SE-schoolscore, zoals de controlescholen in dit onderzoek of de scholen met een vergelijkbare gemiddelde schoolscore als de huidige Kaleidoscoop en Piramidescholen. Wij menen het laatste omdat in de huidige praktijk de doelgroep van het Onderwijs Achterstanden Beleid vooral in deze scholen is geconcentreerd. Uit het Prima-cohort kan opgemaakt worden dat een verschil in SE-score van 1 SD correspondeert met een verschil in prestatiescores van 0.5 SD op toetsen als ordenen. Hieruit kan afgeleid worden dat de effecten van de programma's ten opzichte van reguliere, maar wat betreft etnische en sociaal-economische samenstelling vergelijkbare basisscholen in dit onderzoek waarschijnlijk onderschat zijn. De mate waarin dit het geval is, is op dit moment niet vast te stellen, maar in het follow-up onderzoek zal expliciet aandacht besteed worden aan dit punt en zal, in aanvulling op de toetsafname, geprobeerd worden om meer informatie te verkrijgen over bepaalde schoolkenmerken, zoals kenmerken van de leerlingpopulatie ('student body characteristics') op de aan het onderzoek deelnemende scholen.

Ten opzichte van de resultaten van de tussenmeting zijn de verwachtingen over de effectiviteit van de programma's ten dele wel en ten dele niet bewaarheid. Wat betreft Kaleidoscoop heeft de voorspelde 'inhaalslag' inderdaad plaatsgevonden. Ter verklaring toen voor de iets mindere resultaten van Kaleidoscoop wezen we o.m. op de moeilijkere implementatie van dit programma en als gevolg daarvan de tragere start. Die is, blijkens de huidige gegevens, overwonnen. Het valt op dat vooral het effect op de 'nieuwe instromers' bijdraagt aan de effectiviteit: met die kinderen wordt in Kaleidoscoop in de kleuterperiode de grootste winst geboekt, terwijl in de controleconditie de nieuw ingestroomde kinderen juist slechter af zijn en minder profijt hebben van het curriculum aldaar. Wat betreft Piramide is het succes van de peuterperiode niet in dezelfde mate in de kleuterfase voortgezet. Dit resultaat bevreemdt en zou kunnen wijzen op implementatieproblemen, met name de functie van de tutor. Hoewel wij niet over precieze gegevens beschikken is uit verschillende berichten duidelijk geworden dat de tutor niet altijd beschikbaar was voor de taken die deze in het kader van het programma moest

Evaluatie Kaleidoscoop en Piramide

vervullen (zie ook Reezigt, 1999). Omdat tutors in de kleuterperiode in dienst waren van basisscholen werden zij ook ingezet om het bekende, veel voorkomende probleem van het ziekteverzuim op te vangen. Verder speelde een rol, zo werd vernomen, dat het moeilijk bleek om gekwalificeerde personen te werven voor de tutor-functie. Dit zou ertoe hebben kunnen bijdragen dat het Piramide-programma in de kleuterklassen niet optimaal kon worden uitgevoerd, met eventuele gevolgen voor de effectiviteit van het programma in de kleuterperiode. Deze inmiddels 'gangbare' situatie op basisscholen - het hoge ziekteverzuim en de moeilijkheid om gekwalificeerde krachten aan te trekken - speelden uiteraard ook Kaleidoscoop parten, maar wellicht is dit programma door zijn aard daar minder gevoelig voor.

Literatuur

- Bleichrodt, N., Resing, W., & Zaal, J. (1993). *Beoordeling van schoolgedrag (SCHOBL-R)*. Amsterdam: Vrije Universiteit.
- Blok, H., & Leseman, P.P.M. (1996). Effecten van voorschoolse stimuleringsprogramma's: een review van reviews. *Pedagogische Studiën*, 73, 184-197.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd edition)*. Hillsdale, NJ: Erlbaum.
- Cohen, E.G., & Lotan, R.A. (1995). Producing equal-status interaction in the heterogeneous classroom. *American Educational Research Journal*, 32, 1, 99-120.
- Commissie Voorschoolse Educatie (1994). *Allochtone kleuters meer aandacht*. Rijswijk/Zoetermeer: Ministerie van WVC/Ministerie van O&W.
- Groenendaal, H.J., & Colthof, J. (1983). *Vragenlijst Denkprocessen, Beknopte Handleiding*. Utrecht: Rijksuniversiteit, IPAW.
- Groenendaal, H.J., van der Horst, J.M.H., & Palmen, C.M.A. (1987). Taalvaardigheid en algemeen cognitief functioneren van kleuters in verschillende sociaal-economische milieus. *Pedagogische Studien*, 64, 381-389.
- Haskins, R. (1985). Public school aggression among children with varying day-care experience. *Child Development*, 56, 689-703.
- Hegland, S.M., & Rix, M.K. (1990). Aggression and assertiveness in Kindergarten children differing in day care experiences. *Early Childhood Research Quarterly*, 5, 105-116.
- Hessels, M.G.P. (1993). *Leerpotentieel test voor etnische minderheden*. Utrecht: Proefschrift Universiteit Utrecht.
- Howes, C., Philips, D.A., & Whitebook, M. (1992). Thresholds of quality: implications for the social development of children in center-based child care. *Child Development*, 63, 449-460.

- Hox, J.J. (1997). *Er is nieuws onder de zon: nieuwe oplossingen voor oude problemen*. Inaugurale rede Universiteit Utrecht. Amsterdam: TT-Publikaties.
- Laevers, F. (1998). Het leren van het jonge kind. In L. Verschaffel & J. Vermunt (Red.), *Onderwijskundig Lexicon III: Het leren van leerlingen* (pp. 133-153). Alphen a/d Rijn: Samsom H.D. Tjeenk Willink.
- Leseman, P.P.M., Otter, M.E., Blok, H., & Deckers, P. (1998). Effecten van voor- en vroegschoolse stimuleringsprogramma's: een meta-analyse van evaluatiestudies 1985-1996. *Nederlands Tijdschrift voor Opvoeding, Vorming en Onderwijs*, 14, 2, 134-154.
- Leseman, P.P.M., A. Veen, B. Triesscheijn, M. Otter (1999). *Evaluatie van Kaleidoscoop en Piramide. Verslag van tussentijdse resultaten*. SCO-rapport 539. Amsterdam: SCO-Kohnstamm Instituut.
- Leseman, P.P.M., & van den Boom, D.C. (1999). Verschillen tussen leerlingen: erfelijkheid en omgeving. In P.P.M. Leseman (Red.), *Achtergronden van leerlingen. Onderwijskundig Lexicon* (Vol. III) (pp. 13-34). Alphen a/d Rijn: Samsom.
- Narain, G. en L. Verhoeven (1994). *Ontwikkeling van tweetaligheid bij allochtone kleuters*. Tilburg: Tilburg University Press.
- Ramey, C.T., & Ramey, S.L. (1998). Early intervention and early experience. *American Psychologist*, 53, 2, 109-120.
- Reezigt, G.J. (1999). *De implementatie van Kaleidoscoop en Piramide*. Groningen, GION.
- Schroots, J.J.F. en Alphen de Veer, R.J. van (1976). *LDT, Leidse Diagnostische Test. Deel I Handleiding*. Amsterdam: Swets en Zeitlinger.
- Slavin, R.E., & Madden, N.A. (1999). Success for All: Effects of prevention and early intervention on elementary students' reading. In L. Elderling & P.P.M. Leseman (Eds.), *Effective early education: cross-cultural perspectives* (pp. 306-331). New York: Garland Francis & Taylor.

Snijders, J.Th. en Snijders-Oomen, N. (1975). *Snijders-Oomen niet verbale intelligentieschaal (SON 21/2-7)*. Groningen: Wolters-Noordhoff.

Terwel, J. (1997). *Grenzen aan de groep*. Amsterdam: Inaugurele rede.

Tuijl, C. van, Leseman, P.P.M., & Rispens, J. (in press). Efficacy of an intensive home-based educational program for 4-6 year old ethnic minority children in the Netherlands. *International Journal of Behavioral Development*.

Verhoeven, L., A. Vermeer, C. van de Guchte (1986). *Taaltoets Allochtone Kinderen: toetspakket*. Tilburg: Zwijssen.