UNIVERSITEIT TWENTE.

Effectonderzoek

Pilot startgroepen voor peuters

Eindrapportage 2016

Ilona Veer
Hans Luyten
Cathy van Tuijl
Peter Sleegers

Effectonderzoek

Pilot startgroepen voor peuters

Eindrapportage 2016

Inhoudsopgave

Inleiding

Aanleiding onderzoek

In Nederland klinkt steeds vaker de roep om kinderen reeds voorafgaande aan de basisschool, een educatief aanbod aan te bieden in een voorschoolse voorziening. Dit geldt in het bijzonder voor kinderen uit (niet-Nederlandstalige) achterstandsgezinnen. Deze voorschoolse voorziening zou mogelijk de schoolloopbaan van deze kinderen kunnen verbeteren (Onderwijsraad, 2008, 2010). Het belang dat gehecht wordt aan de voorschoolse periode is terug te voeren op actueel hersenonderzoek waarbij de voorschoolse periode, in vergelijking met alle andere perioden, gekenmerkt wordt door snelle ontwikkelingen die fundamenteel worden geacht voor verdere ontwikkeling (Shonkoff, 2010).

Vanuit een educatief standpunt is onderzocht of voorschoolse voorzieningen bijdragen aan sociale- en academische vaardigheden van kinderen voorafgaand aan de formele scholing. Academische vaardigheden verwijzen in deze context naar zaken als woordenschat, ontluikende geletterdheid en rekenvaardigheid alsmede naar ‘learning related skills’ (aandacht, executieve functies) (McClelland et al., 2007): kortom zaken die als voorwaardelijk worden beschouwd voor het latere formele leren. Er is empirische ondersteuning te vinden voor de educatieve bijdrage, maar tevens zijn er inconsistenties. Voor enkele hoogwaardige en intensieve programma’s worden stevige effecten gevonden op de cognitieve-, taal- en sociale ontwikkeling van achterstandskinderen (bijvoorbeeld Perry Preschool), terwijl voor heel veel andere (veelal minder intensieve) programma’s of voor specifieke groepen kinderen geen of veel kleinere effecten gevonden worden.

Voorschoolse voorzieningen in Nederland zijn divers en omvatten onder meer kinderopvang, peuterspeelzalen, en voorscholen (Veen et al., 2012). Deze typen voorschoolse voorzieningen verschillen in populatie (doelgroepen versus niet-doelgroepen), het aantal kinderen per groep, de startleeftijd van de kinderen, de duur van de opvang en het aantal uren dat kinderen er door brengen, alsmede in educatieve focus en aansluiting of integratie met het basisonderwijs. Onderzoek in Nederland laat zien dat voorschoolse voorzieningen in Nederland tot nu toe beperkt effectief zijn in het stimuleren van de cognitieve-, taal- en sociaal emotionele ontwikkeling (Broekhuizen, 2015; Slot, 2014) waarbij onduidelijk is of dat terug te voeren is op kwaliteit of kwantiteit van deze voorzieningen (of een combinatie).

De pilot startgroepen is opgezet met als doel de ontwikkeling van (doelgroep)kinderen op het gebied van taal, rekenen en learning related skills (bijvoorbeeld aandacht) te bevorderen door middel van het verbeteren van zowel kwantitatieve aspecten als kwalitatieve aspecten van voorschoolse educatie. Het verhogen van het aantal uren (kwantiteit) in een voorschoolse voorziening betekent echter niet automatisch dat er ook meer aandacht is voor de academische- en sociale ontwikkeling van kinderen. Tijdsbestedingsonderzoek (de Haan & van Tuijl, 2011) laat zien dat voorscholen verschillen in de mate waarin zij effectief gebruik maken van de beschikbare tijd. De resultaten van de voorschoolse educatie zijn in grote mate afhankelijk van de proceskwaliteit: de kwaliteit van het aanbod. Is er bijvoorbeeld sprake van stimulerende educatieve interacties in klein groepsverband? Wordt er effectief gebruik gemaakt van de beschikbare tijd?
In dit onderzoek naar de effectiviteit van startgroepen wordt onderzocht of bij een goede implementatie (gelet op structurele kenmerken als opleiding van de beroepskrachten, voldoende kwantiteit e.d.) de proceskwaliteit van het aanbod verhoogd wordt. Daarnaast wordt onderzocht in hoeverre de kenmerken van de startgroepen (zie onder) tezamen effect hebben op de ontwikkeling van doelgroepkinderen.

Pilot startgroepen

In september 2011 is een landelijke vierjarige proef begonnen met *startgroepen voor peuters (pilot startgroepen)*. In de startgroepen wordt minimaal vijf dagdelen van 2,5 uur (of **12,5 uur per week**) voorschoolse educatie aangeboden waarbij **opbrengstgericht** gewerkt wordt. Daarbij wordt intensief samengewerkt tussen een voorschoolse voorziening (peuterspeelzaal of kinderopvang) en een basisschool, waarbij de **basisschool de regie** heeft. Kinderen met een (taal)achterstand kunnen zich binnen een stimulerende omgeving ontwikkelen, onder regie van de basisschool. Het educatief aanbod van kinderopvang, peuterspeelzaal en basisschool dient goed op elkaar afgestemd te worden, zodat een **doorlopende lijn** ontstaat. De begeleiding van een startgroep is in handen van een beroepskracht voorschoolse educatie (mbo-3 niveau) en een begeleider voorschoolse educatie met een onderwijsbevoegdheid (**hbo-niveau**). De kinderen zitten fysiek op het kinderdagverblijf of de peuterspeelzaal. Voor de pilot startgroepen konden kinderdagverblijven en peuterspeelzalen zich aanmelden. Uit het aantal aanmeldingen zijn dertig pilots aangewezen, verspreid over verschillende typen gemeenten: G4, gemeenten uit krimpregio's en overige gemeenten.

Het ministerie van O, C en W wil de opbrengsten van de pilots laten onderzoeken. Het onderzoek bestaat uit twee delen: een beschrijvend onderzoek naar de implementatie van de startgroepen en een effectonderzoek. Het implementatie onderzoek wordt beschreven in een andere rapportage, uitgebracht door Oberon. In dit rapport wordt het effectonderzoek beschreven.

Centraal in het effectonderzoek staat de volgende vraag:

Wat is het effect van (de opzet van) de startgroep ten opzichte van de reguliere voor- en vroegschoolse educatie op de ontwikkeling van kinderen?

Deze eindrapportage bevat een beschrijving van de resultaten van het effectonderzoek waarin eveneens implementatiekenmerken van startgroepen (Oberon, 2015) worden gebruikt voor de analyses.
Achtereenvolgens worden de methode van onderzoek (opzet, respondenten, instrumenten en procedure) en de resultaten beschreven. In de discussie worden de belangrijkste resultaten en implicaties voor de praktijk beschreven waarna het rapport zal worden afgesloten met een algemene conclusie.

Methode

Opzet onderzoek

De pilot startgroepen kan opgevat worden als een interventie. Om het effect van deze interventie te meten, zal er een vergelijking gemaakt worden tussen een experimentele groep (een steekproef van kinderen uit de dertig startgroepen) en een controlegroep (een steekproef van kinderen die VVE-programma's volgen in reguliere voorschoolse voorzieningen). Door het vergelijken van de experimentele groep en de controlegroep kan het effect van de interventie (pilot startgroepen) op de ontwikkeling van kinderen worden nagegaan. Er is gekozen voor een quasi-experimentele onderzoeksopzet. Random toewijzing van respondenten aan de experimentele en controlegroep was op basis van ethische en praktische kwesties, ongewenst en onmogelijk. Verder was het, gezien het feit dat de pilot startgroepen al was gestart bij de

aanvang van het onderzoek, niet mogelijk kinderen aselect toe te wijzen aan de experimentele conditie, maar diende er op basis van leeftijd geselecteerd te worden.
Om toch een zuiver experimentele opzet van het onderzoek te benaderen, is ervoor gekozen om gebruik te maken van een propensity score matching. De propensity score is een manier om maximale gelijkheid tussen groepen (experimentele versus controle groep) te bewerkstelligen waardoor de condities van een “randomized controlled trial” kunnen worden benaderd. Het doel van de matching is het identificeren van vergelijkbare kinderen, afgezien van de condities (experimentele- versus controle groep) waarin ze zitten. Voor de propensity score matching worden achtergrondkenmerken geselecteerd waarvan op grond van eerder onderzoek verwacht mag worden dat ze de ontwikkeling van kinderen beïnvloeden. Een propensity score is de conditionele kans om toegewezen te worden aan een bepaalde interventiegroep (Rosenbaum & Rubin, 1983; Rubin & Thomas, 1996; Stürmer et al., 2006).
Een steekproef van kinderen van reguliere peuterspeelzalen en kinderdagverblijven die mee hebben gedaan met het Pre-COOL onderzoek is gebruikt voor de samenstelling van de controlegroep. De experimentele groep bestaat uit ongeveer 150 kinderen uit de startgroepen die in 2012 zijn gestart op de startgroep (cohort 1) en ongeveer 150 kinderen die in 2013 zijn gestart op de startgroep (cohort 2). Voor het onderzoek wordt een longitudinaal design gehanteerd met vier meetmomenten: rond de leeftijd van 2,5 jaar, 3 jaar, 4 jaar en 5 jaar wordt de ontwikkeling van kinderen gemeten op verschillende domeinen. Taken en toetsen op het gebied van taal, rekenen en selectieve aandacht staan centraal in deze rapportage.

Respondenten

Voor het selecteren van het eerste cohort, zijn de startgroepen in het najaar van 2012 benaderd. In het voorjaar van 2013 is er gestart met het samenstellen van een tweede cohort. Aan beroepskrachten van startgroepen is gevraagd om de namen en geboortedata van alle kinderen in de groep door te geven. Vervolgens is er een selectie gemaakt op basis van leeftijd: alle 318 kinderen die op dat moment nog jonger waren dan drie jaar zijn geselecteerd voor het onderzoek. Door middel van een informatiebrief zijn ouders op de hoogte gesteld van het onderzoek en de mogelijkheid om af te zien van deelname aan het onderzoek.
Na het selecteren van een steekproef van 318 kinderen voor beide cohorten, hebben 4 ouders bezwaar gemaakt tegen deelname van hun kind aan het onderzoek. Vervolgens zijn 6 van de 314 kinderen nog voor het afnemen van de eerste meting uitgestroomd en bleek één van de

kinderen het VVE-programma minder dan 12,5 uur per week te volgen. Dit betekent dat er bij 307 kinderen daadwerkelijk een voormeting is gedaan.
Na de voormeting bleken nog eens 2 van de 307 kinderen het VVE-programma minder dan 12,5 uur per week te volgen. Vervolgens zijn er 21 kinderen uitgevallen door verhuizing, of overstap naar een andere voorschoolse voorziening. Hierdoor heeft bij 284 kinderen de tweede meting plaats gevonden. Na de tweede meting en vóór de overgang naar de basisschool zijn er door bovengenoemde redenen nog eens 12 kinderen uitgevallen. Bovendien was er één ouder die na twee metingen de toestemming tot het volgen van het kind op de basisschool alsnog introk. De derde meting is vervolgens afgenomen bij 271 kinderen uit groep 1. Op de 30 basisscholen waartoe de startgroepen behoren zijn 207 kinderen op vierjarige leeftijd getest. Op 52 andere basisscholen zijn de overige 64 vierjarige kinderen getest. In verband met de opzet van het onderzoek is er alleen bij cohort 1 ook een vierde meting uitgevoerd bij 128 kinderen uit groep 2. Daarvan zijn er 94 kinderen op de 30 basisscholen waartoe de startgroepen behoren getest. De overige 34 kinderen zijn op 31 andere basisscholen getest.

De kenmerken van de respondenten[1] worden weergegeven in tabel 5 in de bijlage. In deze tabel is te zien dat bijna de helft (47%) van de respondenten een jongen is. Met betrekking tot de sociaal economische achtergrond heeft ruim de helft (55%) van de respondenten minimaal één ouder die in een niet-Westers land is geboren, terwijl bij een derde van de respondenten thuis geen Nederlands wordt gesproken. Het netto inkomen van de hoofdverdiener (uit de gezinnen van de respondenten) is bij 56% minder dan €1500 netto per maand. Verder heeft bijna een kwart (24%) van de respondenten zeer laag opgeleide ouders (maximaal LBO). Uit deze gegevens blijkt dat kinderen met een lage sociaal economische status in grote mate vertegenwoordigd zijn in de startgroepen. Driekwart van de respondenten wordt dan ook door de beroepskrachten gekenmerkt als doelgroepkind, verwijzend naar de VVE-indicatie van het kind.

[1] de 307 respondenten uit de startgroepen waarbij een testafname op meetmoment 1 is gedaan minus één respondent waarbij de toestemming van ouders later in het onderzoek met terugwerkende kracht is ingetrokken.

Instrumenten

Individuele testafnames

In tabel 6 (zie bijlage) wordt weergegeven welke instrumenten[2] er zijn gebruikt om de ontwikkeling van de kinderen in de startgroepen op verschillende domeinen te analyseren door middel van individuele testafnames, alsmede de respons op de taken. De non-respons houdt in dat de taak niet of gedeeltelijk is afgenomen door gedrag van het kind, een taalprobleem van het kind, een technisch probleem of storende omgevingsfactoren.

Receptieve woordenschat (Verhagen, Mulder, & Leseman, 2015) wordt gemeten met een computertaak waarbij de kinderen verschillende items te zien krijgen, waarbij steeds één van vier plaatjes aangewezen dient te worden. Er wordt bijvoorbeeld gevraagd: “Waar is hond?” bij een scherm met vier plaatjes, waarvan één plaatje een afbeelding van een hond is.

Vroege rekenvaardigheden worden gemeten door een selectie uit de CITO-toets rekenen voor peuters (Op den Kamp, 2010). Met deze toets worden getalbegrip, meten en meetkunde gemeten. In deze taak krijgen kinderen verschillende items te zien waarbij een vraag gesteld wordt (bv: “Waar zie je drie vlinders?”) en ze één van drie plaatjes dienen aan te wijzen (in het genoemde voorbeeld: een plaatje met vier vlinders, een plaatje met drie vlinders en een plaatje met twee vlinders).

Selectieve aandacht wordt gemeten met een computertaak (Mulder, Hoofs, Verhagen, van der Veen & Leseman, 2014) waarbij kinderen een veld te zien krijgen met enkele target dieren: olifanten, en veel afleiders (beren en paarden). De dieren hebben allemaal dezelfde kleur en lijken erg op elkaar. Kinderen worden aangespoord zo snel mogelijk de olifanten aan te wijzen. De gevonden olifanten worden gemarkeerd zodat kinderen zien welke olifanten ze al gevonden hebben.

Voor een uitgebreide omschrijving van de taken verwijzen wij naar de technische rapporten van het Pre-COOL onderzoek (Veen et al., 2012; Veen et al., 2014; Pre-COOL consortium, 2016). Binnen het Pre-COOL onderzoek zijn de betrouwbaarheid en validiteit van de testbatterij onderzocht. De meetinstrumenten die zijn gebruikt voor de individuele testafnames tijdens meetmomenten 1 en 2 zijn betrouwbaar gebleken. Daarnaast is er convergente validiteit en predictieve validiteit aangetoond (zie Veen & Leseman, 2015).

[2] Er zijn veel meer taken afgenomen welke zijn beschreven in de tussenrapportages, echter deze taken zijn niet gebruikt voor de analyses voor het eindrapport en daardoor niet weergegeven in dit rapport.

Voor dit onderzoek hebben we ten behoeve van de betrouwbaarheid en validiteit daarnaast door middel van multilevel analyses[3] geanalyseerd in hoeverre de scores op de taken voorspellend zijn over de tijd[4]. Oftewel, in hoeverre voorspelt bijvoorbeeld selectieve aandacht op meetmoment 1, selectieve aandacht op meetmoment 4 (2,5 jaar later). Zoals te zien is aan de significantie waarden in de tabellen 13 tot en met 18 in de bijlage, is een meting van taal op meetmoment 1 (receptieve woordenschat mm1) zeer voorspellend voor taal op een later moment (receptieve woordenschat mm3 en mm4, CITO taal voor kleuters E1), zelfs wanneer er gecorrigeerd wordt voor achtergrondvariabelen. Zoals te zien is aan de bèta gewichten, is de meting van taal structureel de sterkste voorspeller van taal op een later meetmoment (sterker dan onder andere thuistaal en opleidingsniveau van ouders). Hetzelfde geldt voor rekenen (zie tabel 19, bijlage). De meting van rekenen op meetmoment 2 is significant voorspellend voor rekenen in groep 1 (CITO rekenen voor kleuters E1) en een betere voorspeller dan thuistaal en geboorteland of opleiding van ouders. Voor selectieve aandacht zijn de resultaten van de multilevel analyses weergegeven in de tabellen 20 tot en met 23 (zie bijlage). In tabel 22 is te zien dat selectieve aandacht op meetmoment 1 niet de beste voorspeller is van selectieve aandacht op meetmoment 4[5]. Selectieve aandacht is wel de sterkste voorspeller als er minder tijd tussen de metingen zit (zie tabellen 20, 21 en 23).
De resultaten in de tabellen bieden een extra bevestiging van de betrouwbaarheid en validiteit van de taken – zelfs bij zeer jonge kinderen met een lage sociaal economische status – maar laten ook zien dat het niveau op jonge leeftijd in zeer belangrijke mate (in grotere mate dan sociaal economische achtergrond) voorspellend is voor latere vaardigheid.

CITO-toetsen

Op een deel van de basisscholen zijn CITO-toetsen afgenomen in groep 1: **CITO Taal voor kleuters** M1 & E1 (Lansink, 2009) en **CITO Rekenen voor kleuters** M1 & E1 (Koerhuis, 2010). Met de CITO-toets voor kleuters taal E1 worden woordenschat en kritisch luisteren gemeten. In deze toets krijgen kinderen verschillende items te zien waarbij een vraag gesteld wordt (bv: “Waar zie je oprapen?”) en ze één van de plaatjes dienen aan te wijzen of aan te strepen.
Met de CITO-toets voor kleuters rekenen E1 worden getalbegrip, meten en meetkunde gemeten. In deze toets krijgen kinderen verschillende items te zien waarbij een vraag gesteld

[3] Tweezijdige toetsing.
[4] De analyses zijn uitgevoerd op basis van de data van de startgroepen (niet de data van de controlegroep).
[5] Bij éénzijdige toetsing is de relatie wel significant (P<,05).

wordt (bv: “Op welk plaatje staan de blokken van laag naar hoog?”) en ze één van de plaatjes dienen aan te wijzen of aan te strepen.
In tabel 7 (zie bijlage) wordt de respons van deze CITO-toetsen weergegeven. De respons is nagegaan voor de 269 respondenten die aan het einde van het onderzoek nog deelnamen aan het onderzoek (d.w.z. niet zijn uitgevallen vóór de leeftijd van vier jaar werd bereikt danwel daarna niet zijn verhuisd naar het buitenland). Kinderen uit cohort 2 zitten veelal pas net in groep 1 waardoor in 2015 nog geen CITO-toetsen zijn afgenomen. De CITO-toetsen zullen bij veel van deze kinderen pas in 2016 worden afgenomen (zie tabel 7, bijlage).

Vragenlijsten

Door middel van een **oudervragenlijst** (Veen et al., 2012) is de taalvaardigheid, het geheugen, de aandacht, zelfcontrole, inhibitie, motoriek, sociale vaardigheden en gedrag gemeten. Daarnaast zijn onafhankelijke variabelen gemeten: sociaal economische variabelen (opleiding, inkomen, etniciteit), opvoedingsstijl, opvoedingssatisfactie, sociale steun, risicofactoren, cognitieve stimulering, relatie met de startgroep en verleden van voorschoolse educatie. De respons op de vragenlijsten is 96% (zie tabel 8, bijlage). De sociaal economische variabelen zijn gebruikt voor de matching, waarbij de informatie van het CBS altijd als uitgangspunt werd genomen[6].
Andere variabelen zijn gemeten met de **vragenlijst voor pedagogisch medewerkers** (Veen et al., 2012). Hierin zijn achtergrondgegevens van de beroepskrachten en structurele kenmerken van de startgroepen bevraagd. Daarnaast is het spelaanbod, educatief aanbod, spelmateriaal en inrichting van de groepsruimte, emotionele ondersteuning, groepsorganisatie, omgaan met verschillen, ouderbetrokkenheid, teamoverleg en samenwerking met de basisschool bevraagd. De variabele “groep splitsen” is gebruikt voor de analyses ten behoeve van de subvragen (zie analytische strategie).
De vragenlijst voor pedagogisch medewerkers is in het schooljaar 2012-2013 ingevuld door alle beroepskrachten die destijds werkzaam waren op de startgroepen en na de tweede observatie ook door de meeste beroepskrachten die bij die observatie aanwezig waren (drie beroepskrachten hebben de vragenlijst niet ingevuld). Bij de tweede observatie waren een aantal nieuwe beroepskrachten aanwezig en enkele invalkrachten.

[6] Wanneer informatie over bijvoorbeeld opleidingsachtergrond van moeder gegeven was in zowel de CBS data als de data o.b.v. de oudervragenlijst en er was geen overeenstemming, is uitgegaan van de data van het CBS.

Observaties

Met behulp van observaties is informatie verzameld over de proceskwaliteit op de groepen. Met het observatie-instrument Classroom Assessment Scoring System Toddler: **CLASS-toddler** (Le Paro, Hamre, & Pianta, 2011, aangepast en bewerkt door Veen et al., 2012) wordt de proceskwaliteit op pedagogisch en educatief gebied gemeten. Op pedagogisch gebied wordt emotionele regulatie geobserveerd aan de hand van de items: positieve- en negatieve sfeer, sensitieve responsiviteit en de aandacht voor het perspectief van het kind. Groepsorganisatie wordt geobserveerd aan de hand van het item gedragsregulatie. De mate van educatieve ondersteuning wordt gemeten met de items: faciliteren van leren en ontwikkeling, kwaliteit van feedback en stimuleren van taalontwikkeling.

De Early Childhood Environment Rating Scale – Extension: **ECERS-E** (Sylva, Sammons, Siraj-Blatchford, & Taggart, 2008, vertaald door Veen et al., 2012) is een observatie-instrument (waarvan de schalen geletterdheid en rekenen zijn gebruikt) dat tezamen met de CLASS-toddler is afgenomen en waarmee zowel structurele kwaliteit als proceskwaliteit op het educatieve vlak gemeten wordt. Er worden zes items op het domein geletterdheid gescoord: o.a. de aanwezigheid van geschreven letters en woorden en de mate waarin er gepraat wordt met en geluisterd wordt naar kinderen. Daarnaast worden er drie items op het domein Rekenen gescoord: o.a. tellen en het toepassen van tellen, en lezen en schrijven van eenvoudige cijfers. Zie tabel 10 in de bijlage voor een beschrijving van alle items.

De observaties zijn tweemaal uitgevoerd op alle startgroepen, namelijk aan het begin van de pilot (najaar 2012) en halverwege de pilot (voorjaar 2014). De implementatie van de startgroepen bevond zich in 2014 in een meer gevorderd stadium dan tijdens de observatie van 2012 het geval was. Daarom worden alleen de scores van de observaties halverwege de pilot meegenomen in de analyses en weergegeven in dit rapport.

Procedure

Individuele testafnames

Ten einde de taken op gestandaardiseerde wijze te kunnen afnemen bij individuele kinderen is er een training gevolgd (Hanna Mulder en Josje Verhagen, Universiteit Utrecht). In het kader

van deze training diende er een video gemaakt te worden waarin gedemonstreerd werd hoe de taken op gestandaardiseerde wijze zijn afgenomen.
De eerste auteur had in 2010 en 2011 reeds trainingen gevolgd voor het afnemen van tests van meetmomenten 1 en 2, twee voldoende video's afgeleverd en bij veel jonge kinderen taken afgenomen ten behoeve van het Pre-COOL onderzoek. Zij is dus zeer ervaren in het afnemen van de individuele testen die gebruikt worden voor het onderzoek naar startgroepen.
De testafnames van meetmomenten 3 en 4 heeft zij tezamen met drie onderzoeksassistenten uitgevoerd (zie tabel 9, bijlage). De onderzoeksassistenten zijn allen eerder werkzaam geweest voor het Pre-COOL onderzoek: onderzoeksassistent 1 als observator en onderzoeksassistenten 2 en 3 als testleiders. In 2014 heeft de eerste auteur samen met onderzoeksassistent 1 een training gevolgd (Babs de Haas, Universiteit Utrecht) om ook de testbatterij van meetmomenten 3 en 4 op gestandaardiseerde wijze te kunnen afnemen. Vervolgens is er door beiden een video gemaakt waarin de gestandaardiseerde werkwijze werd gedemonstreerd. Op grond van deze video-opnames, zijn zij geschikt bevonden voor het afnemen van de tests.
Onderzoeksassistenten 2 en 3 hadden dezelfde training al eerder gevolgd in het kader van hun bijdrage aan het Pre-COOL onderzoek. Wel hebben zij voor het effectonderzoek startgroepen een nieuwe video gemaakt van een testafname ten behoeve van de betrouwbaarheid (Babs de Haas, Universiteit Utrecht). Deze video liet zien dat zij de gestandaardiseerde werkwijze goed beheersten.
Voordat de eerste auteur als testleider aan de slag kon gaan met de eerste meting, was het van belang dat vooral de kinderen, maar ook hun ouders en de beroepskrachten een vertrouwd gevoel hadden bij haar en de condities waaronder de testen zouden worden afgenomen. Voor de kinderen was het vooral van belang dat zij de testafname als leuk en prettig zouden ervaren. Om die reden heeft de testleider de tijd genomen om kennis te maken met beroepskrachten, ouders en kinderen op alle startgroepen. Ten tijde van de testafnames kregen kinderen veel complimentjes (mede in het kader van de gestandaardiseerde afname). Na afloop van elke testafname kregen kinderen bovendien een cadeautje. Bij de tweede en daarop volgende metingen wilden kinderen over het algemeen dan ook graag mee voor de testafnames. Het belang en plezier van het kind stond altijd centraal bij de testafnames.

Vragenlijsten

Bij de kennismaking met de ouders is hen gevraagd of zij ten behoeve van het onderzoek een oudervragenlijst wilden invullen. Gezien de variatie in kennis van de Nederlandse taal is de

mogelijkheid aangeboden om met behulp van een tolk de vragenlijst in te vullen. De ouders van vier respondenten hebben van deze mogelijkheid gebruik gemaakt. Daarnaast hebben de beroepskrachten van de startgroepen ouders heel vaak geholpen bij het invullen van de oudervragenlijsten. De eerste auteur heeft de ouders van drie respondenten geholpen bij het invullen van de vragenlijst.
De vragenlijst voor pedagogisch medewerkers is gedigitaliseerd en eind 2012 verstuurd naar alle beroepskrachten die destijds werkzaam waren op de startgroepen. Na de tweede observatie zijn ook de nieuwe beroepskrachten benaderd met de vraag de vragenlijst digitaal in te vullen.

Observaties

Eind 2012 zijn de beroepskrachten van alle startgroepen benaderd om een observatie gericht op het meten van de proceskwaliteit op pedagogisch en educatief gebied (zie instrumentensectie), in te plannen. Onderzoeksassistent 1 – zij is werkzaam geweest als observator binnen het Pre-COOL onderzoek – is hiervoor gecontracteerd. Zij heeft destijds een intensieve training gevolgd (Pauline Slot, Universiteit Utrecht) voor het gebruik van de CLASS-toddler en ECERS-E, een live-observatie van de CLASS-toddler gescoord en een betrouwbaarheidstest uitgevoerd[7]. Voor het startgroepen onderzoek heeft zij in het najaar van 2012 opnieuw een betrouwbaarheidstest uitgevoerd en goed afgelegd. Zij heeft vervolgens op 30 startgroepen de proceskwaliteit met behulp van de CLASS-toddler gemeten: vier cycli[8] van 15 minuten gedurende een ochtend. Daarnaast heeft zij in de groep aanvullende informatie verzameld ten behoeve van de ECERS-E. In het voorjaar van 2014 zijn de observaties uitgevoerd door onderzoeksassistent 1 en een tweede observator. De tweede observator is eveneens binnen het Pre-COOL onderzoek werkzaam geweest als observator. Beide observatoren hebben begin 2014 een betrouwbaarheidstest uitgevoerd en behaald. Zij hebben vervolgens allebei op 15 startgroepen vier activiteiten (cycli) van 15 minuten geobserveerd met behulp van de CLASS-toddler. Daarnaast is door hen beiden de groepscontext beoordeeld aan de hand van de ECERS-E.

Propensity score matching

Zoals eerder beschreven, is gekozen voor een onderzoeksopzet waarbij respondenten uit de experimentele groep (de startgroepen), op basis van achtergrondkenmerken (door middel van

[7] De betrouwbaarheidstest wordt behaald indien ten minste 80% van de items maximaal 1 punt afwijkt van de score van de trainer.
[8] Een cyclus verwijst naar een activiteit, bv: eet-drinkmoment, voorleesactiviteit, activiteit in de kleine groep, etc.

propensity scores) gematcht worden met respondenten uit een controlegroep (Pre-COOL) die een VVE-programma hebben gevolgd. Door het corrigeren voor achtergrondkenmerken worden de verschillen tussen beide groepen zo klein mogelijk. Daartoe kunnen de gegevens (zoals sociaal economische status van het gezin waarin het kind opgroeit) die verzameld zijn in dit onderzoek en in het kader van Pre-COOL, gebruikt worden om kinderen uit de experimentele groep (startgroepen) te koppelen aan kinderen uit de controlegroep (reguliere VVE kinderen). De onderzoeksdata van Pre-COOL zijn in D.A.N.S. (Data Archiving and Network Services) opgeslagen. In 2012 en 2013 is deze databank geraadpleegd om de achtergrondgegevens van de respondenten uit het Pre-COOL onderzoek te achterhalen ten einde de matching te kunnen uitvoeren. Deelname aan een VVE-programma is nagegaan voor alle respondenten. Helaas bleek dit van slechts ongeveer de helft van de kinderen bekend te zijn. Zodoende is besloten te streven naar de selectie van een zo groot mogelijk percentage kinderen uit een VVE-instelling voor de controlegroep. Uiteindelijk bestaat de gewogen controlegroep voor 66% uit kinderen die een VVE-programma hebben gevolgd, voor 32% uit kinderen waarvoor het onbekend is en voor 3% uit kinderen die geen VVE-programma hebben gevolgd. Dat betekent dat van de kinderen waarvan bekend is of ze een VVE-programma hebben gevolgd, 96% een VVE-programma heeft gevolgd[9].

Verder bleek de respons op de oudervragenlijst in het instellingencohort van Pre-COOL laag te zijn, voornamelijk bij de kinderen met een lagere sociaal economische status: de groep die in grote mate vertegenwoordigd is in de startgroepen. Daarom is besloten aanvullende data op te vragen bij het CBS (Centraal Bureau voor de Statistiek, Den Haag). Allereerst is bij het Kohnstamm Instituut een verzoek ingediend om de postcodes en geboortedata van de Pre-COOL respondenten te verkrijgen. Op basis van deze informatie kon voor veel respondenten door het CBS worden nagegaan om welke personen het ging. Vervolgens heeft het CBS een versleuteling van de privacygevoelige informatie uitgevoerd en het bestand met (rinpersoon) identificatienummers beschikbaar gesteld. Hierdoor was het mogelijk om binnen de beveiligde omgeving van het CBS te werken met de versleutelde data. Deze omgeving is zodanig beveiligd, dat informatie over personen niet achterhaald kan worden uit de data en niet gekoppeld kan worden aan eigen bestanden. Bovendien kan er alleen met de data gewerkt worden op een door het CBS beschikbaar gestelde computer, met gebruik van een pasje en regelmatige identiteitscontrole (ongeveer elk half uur) door middel van vingerafdruk. De ruimte waar de computer op de Universiteit Twente staat is afgesloten voor derden en het is verboden

[9] Van 69% is bekend of er een VVE-programma is gevolgd. 66% (VVE) van 69% (VVE en niet-VVE) houdt in dat 96% van de groep waarbij geen missings zijn op de VVE-variabele een VVE-programma heeft gevolgd.

om afbeeldingen te maken van het scherm. Het is ook niet mogelijk om bestanden te mailen, op te slaan of af te drukken. De privacy van de respondenten is zodoende goed gewaarborgd.
Nadat het CBS de koppeling had uitgevoerd, kon er begin 2015 gestart worden met het achterhalen van de achtergrondkenmerken. Voor elke variabele moest nagegaan worden welke informatie er beschikbaar was en hoe deze informatie was gecategoriseerd. Het heeft zodoende enkele maanden geduurd voordat de data van alle achtergrondkenmerken op een correcte manier gekoppeld was aan de respondenten uit Pre-COOL en de startgroepen. Uiteindelijk is er een bestand gemaakt waarin respondenten uit de startgroepen gematcht konden worden met respondenten uit het Pre-COOL onderzoek.
Er bleken vervolgens nog een aantal knelpunten te zijn, waardoor de matching nog niet uitgevoerd kon worden. Allereerst bleek het leeftijdsverschil tussen respondenten van de startgroepen en van Pre-COOL zeer groot te zijn, respectievelijk gemiddeld 2,7 jaar en 2,3 jaar op het eerste meetmoment (zie tabel 2). Bovendien liep de tijd tussen meetmoment 1 en meetmoment 2 nogal uiteen: bij de startgroepen zat er gemiddeld slechts een half jaar tussen de metingen terwijl dat bij Pre-COOL ruim een jaar was. Op het tweede meetmoment waren de verschillen kleiner: 3,3 jaar (startgroepen) en 3,5 jaar (Pre-COOL). Er is zodoende besloten om meetmoment 1 in eerste instantie buiten beschouwing te laten en leeftijd op meetmoment 2 te gebruiken als één van de covariaten voor de propensity score matching. Zodoende kunnen er vergelijkingen tussen de kinderen uit de startgroepen en de controlegroep gemaakt worden op dat meetmoment.
Een tweede probleem dat zich voordeed was dat er in het Pre-COOL bestand (ook na de koppeling door het CBS) relatief weinig respondenten met een niet-Westerse etnische achtergrond en een laag opleidingsniveau van ouders in het bestand zaten. Om dit probleem te ondervangen is besloten een weging toe te passen. Dit houdt in dat respondenten met een lagere SES uit Pre-COOL in een aantal gevallen meerdere keren gebruikt worden voor de matching. Nadeel hiervan is dat individuele trajecten zwaar meetellen en zeer bepalend kunnen zijn voor de uitkomsten.
Het statistisch programma SPSS bleek niet toereikend voor deze manier van matchen. Het statistische programma ‘R’ (versie, 3.0.2) bleek wel geschikt om een gewogen matching uit te voeren. Met dit programma zijn de propensity scores en gewichten voor alle respondenten berekend. Zodoende bleef er uit het grote Pre-COOL bestand met duizenden respondenten nog

slechts een kleine groep respondenten over voor de vergelijking (206 respondenten met een gewicht tussen 0,37 en 9,33[10]).

De feitelijke matching is vervolgens uitgevoerd op basis van de volgende variabelen: geslacht, thuistaal (alleen Nederlands, Nederlands en andere taal, geen Nederlands), geboorteland moeder, geboorteland vader (beiden gecategoriseerd naar oude- en nieuwe immigranten[11]), éénoudergezin (0,1), belangrijkste inkomensbron van het huishouden (loon, bijstandsverzekering, arbeidsongeschiktheidsverzekering, overig geen loon), opleiding moeder (indien onbekend, opleiding vader) en leeftijd op meetmoment 2. Na de matching werd het mogelijk om een eerste vergelijking te maken tussen de experimentele groep en de controlegroep op basis van scores op meetmoment 2.

Om een vergelijking te kunnen maken op basis van meer meetmomenten was het noodzakelijk om een beroep te doen op meetmomenten 3 en 4 en de CITO-toets scores uit groep 1. Deze scores zijn recentelijk beschikbaar gekomen.

Bij Pre-COOL zijn de gegevens van meetmoment 3 en 4 en de CITO-scores uit groep 1 weliswaar enige tijd geleden verzameld, maar nog niet beschikbaar in D.A.N.S. Door middel van een verzoek aan Paul Leseman (Universiteit Utrecht), wiens vakgroep, samen met o.a. het Kohnstamm Instituut, verantwoordelijk is voor het uitvoeren van het Pre-COOL cohortonderzoek, is er in november 2015 voortijdig toegang verkregen tot de Pre-COOL data van meetmomenten 3 en 4. Meetmomenten 3 en 4 zijn bij het startgroepen onderzoek eind oktober 2015 afgerond. In november 2015 zijn al deze verzamelde gegevens getransporteerd naar SPSS.

Betreffende de CITO-toetsscores van het Pre-COOL onderzoek, is in oktober 2015 een verzoek ingediend bij het Kohnstamm Instituut. Eind oktober 2015 zijn de CITO-toetsscores uit groep 1 beschikbaar gesteld: Taal voor kleuters (Lansink, 2009), Rekenen voor kleuters (Koerhuis, 2010). In oktober en november 2015 zijn alle scholen benaderd waar respondenten uit de startgroepen onderwijs volgen voor het verkrijgen van diezelfde CITO-toetssscores. Nadat vrijwel alle aanwezige scores zijn verkregen (herhaaldelijke rappels, zie tabel 7 in de bijlage voor de respons) is op 14 december 2015 het bestand geprepareerd dat aanvullend naar het CBS verstuurd kon worden[12]. Vanaf 22 december 2015 is het nieuwe bestand door het CBS beschikbaar gesteld voor de onderzoekers. Op basis daarvan werd het mogelijk om ook

[10] Het gewicht staat voor het aantal keer dat de respondenten uit de controlegroep gematcht worden met respondenten uit de experimentele groep. Respondenten met een gewicht van 5,0 wegen vijf keer zwaarder dan de respondenten met een gewicht van 1,0.

[11] Oude immigrant is geboren in: Turkije, Marokko, Suriname, Antillen of Aruba. Nieuwe immigrant is geboren in: Irak, Afghanistan, Somalië, of overig niet-Westers land.

[12] In verband met de procedure bij het CBS ter bescherming van de privacy van respondenten, is het niet mogelijk om zelf aanvullende data aan het bestand te koppelen. Dit wordt door het CBS gedaan aan de hand van de versleutelde privacy-gevoelige variabelen.

vergelijkingen te maken op basis van de meetmomenten 3 en 4 en de CITO-toetsscores van rekenen en taal uit groep 1.
Ondanks deze pogingen om tot een goede matching te komen, blijft een structureel probleem het verschil in leeftijd. Op meetmoment 2 is er weliswaar gematcht op leeftijd, maar door het verschil in de periodes van dataverzameling tussen Pre-COOL en het onderzoek naar startgroepen, blijft het leeftijdsverschil aanzienlijk. Ten einde ook het eerste meetmoment te kunnen gebruiken en ontwikkelingspaden goed te kunnen vergelijken, is het dan ook van belang dat de scores op de taken (tests) een vergelijkbare onderliggende vaardigheid weergeven. Dit kan bepaald worden door middel van item respons theorie (IRT) schaling. Een dergelijk schaling brengt de moeilijkheidsgraad van de items in kaart. Zodoende is het mogelijk om scores op toetsen die niet exact dezelfde items omvatten onderling vergelijkbaar te maken en kan de groei van de kinderen in kaart worden gebracht. Jan Boom, Hanna Mulder en Josje Verhagen (Universiteit Utrecht) hebben hiertoe een IRT schaling uitgevoerd op de data met betrekking tot receptieve woordenschat en selectieve aandacht van het Pre-COOL bestand. Zij hebben tevens een vergelijkbare schaling uitgevoerd op de data van het onderzoek naar startgroepen. Met behulp van deze vaardigheidsscores is het mogelijk geworden om voor de twee groepen (de experimentele groep en de controlegroep) ontwikkelingspaden te schatten en deze te vergelijken. Zodoende kan met meer precisie geanalyseerd worden of er verschillen zijn in groei tussen respondenten uit de experimentele groep en respondenten uit de controlegroep en daarmee kan de hoofdvraag uit dit onderzoek optimaal worden beantwoord.
In dit rapport zullen dan ook de resultaten worden beschreven die betrekking hebben op verschillen (tussen kinderen uit startgroepen en controlekinderen) op taal, rekenen en selectieve aandacht tijdens en na de interventie (meetmomenten 2, 3, 4 en CITO-toetsscores groep 1). Bovendien worden ontwikkelingsverschillen tussen kinderen uit de experimentele groep en de controlegroep op taal (receptieve woordenschat) en selectieve aandacht van ongeveer 2,5 jaar tot 5 jaar in kaart gebracht.

Analytische strategie

Proceskwaliteit

Zoals eerder beschreven bestaat de interventie ‘pilot startgroepen’ uit vijf onderscheidende kenmerken in vergelijking met reguliere VVE: 1) een aanbod van minimaal 12,5 uur per week (meer uren), 2) opbrengstgericht werken aan de hand van doelen en voortdurende evaluatie van doelen, 3) regie van de basisschool, 4) een doorlopende lijn wat betreft het VVE programma en 5) de aanstelling van een hbo-geschoolde begeleider voorschoolse educatie met een onderwijsbevoegdheid naast de (minimaal) mbo-geschoolde beroepskracht. Verondersteld wordt dat bovengenoemde kenmerken samenhangen met een verschil in proceskwaliteit tussen de experimentele groep en de controlegroep. Bovendien draagt het in kaart brengen van de proceskwaliteit bij aan het inzichtelijk krijgen van de praktijk van de interventie.
Voordat de resultaten naar het effect van de startgroepen op de ontwikkeling van kinderen worden beschreven (hoofdvraag), zal daarom eerst een beschrijving worden gegeven van de proceskwaliteit binnen de startgroepen en de verschillen tussen startgroepen. Vervolgens wordt middels een t-test een vergelijking gemaakt tussen de proceskwaliteit in de experimentele groep en die in de controlegroep.

Hoofdvraag

Voor het beantwoorden van de centrale vraag van dit onderzoek, namelijk wat *het effect is van de startgroep, ten opzichte van de reguliere voor- en vroegschoolse educatie, op de ontwikkeling van kinderen*, worden scores van kinderen uit de experimentele groep (steekproef van kinderen uit de startgroepen) vergeleken met scores van kinderen uit de controlegroep (gewogen steekproef samengesteld uit Pre-COOL met behulp van propensity score matching). Er worden verschillende analyses uitgevoerd. Er worden vergelijkingen gemaakt op korte termijn, op lange termijn en over de gehele looptijd van het onderzoek.
Voor de vergelijkingen op korte termijn, wordt er middels een t-test een vergelijking gemaakt op driejarige leeftijd (het tweede meetmoment) tussen kinderen uit de experimentele groep en de controlegroep. Immers, door de propensity score matching op onder andere leeftijd (van kinderen tijdens de taken van het tweede meetmoment) zijn de kinderen uit de twee groepen dan gemiddeld nagenoeg even oud. De vergelijkingen vinden plaats op de volgende

afhankelijke variabelen: receptieve woordenschat, CITO rekenen voor peuters en selectieve aandacht.

Voor de vergelijkingen op de langere termijn worden de CITO-toetsscores uit groep 1 (*CITO taal voor kleuters E1 en CITO rekenen voor kleuters E1)* vergeleken tussen de experimentele groep en controlegroep middels regressie analyses.[13] Gezien het kleine verschil in leeftijd tussen kinderen uit de experimentele groep en controlegroep tijdens het maken van de toetsen, wordt de leeftijd die kinderen hadden tijdens het maken van de CITO-toetsen in de analyses meegenomen als covariaat.

Tot slot worden er vergelijkingen gemaakt over de gehele looptijd van het onderzoek. Hiertoe wordt de gemiddelde latente vaardigheid van kinderen uit de experimentele groep vergeleken met die van kinderen uit de controlegroep. Dit houdt in dat de extern verkregen vaardigheidsscores[14] (zie evt. pagina 18) voor receptieve woordenschat en selectieve aandacht worden geplot voor beide groepen (op meetmoment 1, 2, 3 en 4). Zodoende wordt inzichtelijk hoe de groei op deze vaardigheden over tijd verloopt voor de experimentele groep in vergelijking met de controlegroep.

De hypothese is dat de kinderen uit de startgroepen significant (P<.05, eenzijdige toetsing) hoger scoren dan kinderen uit de controlegroep op alle bovengenoemde taken en toetsen, zowel op de korte termijn (meetmoment 2) als op de lange termijn (meetmomenten 3 en 4, groep 1).

Subvragen

Zoals eerder beschreven, is er naast het effectonderzoek tevens een beschrijvend onderzoek naar de implementatie van startgroepen door Oberon uitgevoerd (Oberon, 2015). In dit implementatieonderzoek is informatie verzameld over enkele typische kenmerken van startgroepen. Gegevens over deze implementatie kenmerken van de startgroepen worden bij het effectonderzoek gebruikt om een beter inzicht te krijgen in de werking van deze kenmerken voor de ontwikkeling van kinderen in startgroepen. Aangezien reguliere VVE-instellingen deze kenmerken niet hebben en daarom deze kenmerken ook niet zijn gemeten bij de gematchte controlegroep van het Pre-COOL cohort, zijn de analyses over de implementatie alleen gericht

[13] In verband met een lage groepsvariantie (mogelijk mede door gewichten van respondenten uit de controlegroep) worden er geen multilevel analyses uitgevoerd, maar reguliere regressie analyses.

[14] De oorspronkelijke scores zijn ruwe scores, gebaseerd op het percentage goed (receptieve woordenschat, rekenen voor peuters mm2) of aantal items correct (selectieve aandacht). De oorspronkelijke ruwe scores kunnen onderling niet vergeleken worden. De vaardigheidsscores demonstreren een latente vaardigheid, waardoor scores over de tijd vergeleken kunnen worden.

op de data van de startgroepen. In dit verband zullen de volgende subvragen worden beantwoord:

1. Wat is de invloed van de onderwijssetting op de ontwikkeling van de kinderen in startgroepen? Meer specifiek: a) de inhoudelijke regie door de schoolleider, b) opbrengstgericht werken c) doorgaande ontwikkel- en leerlijn

2. Wat is de invloed van de inzet van een begeleider voorschoolse educatie met een lesbevoegdheid voor het basisonderwijs (pabo-niveau) op de ontwikkeling van kinderen in startgroepen?

3. Wat is de invloed van groepssamenstelling op de ontwikkeling van kinderen in startgroepen?

4. *Wat is de invloed van de betrokkenheid van ouders op de ontwikkeling van kinderen in startgroepen?*

Om deze subvragen te beantwoorden zullen multilevel analyses worden uitgevoerd, waarbij verschillen binnen de startgroepen (zie tabel 11, bijlage) gerelateerd worden aan diverse scores op taken die bij de kinderen zijn afgenomen. Hierbij wordt steeds gecorrigeerd voor de volgende covariaten: geslacht, geboorteland ouders niet-westers, hoogste opleidingsniveau ouders, thuistaal (geen Nederlands, Nederlands en andere taal/talen, alleen Nederlands) en taak- of toets score op de voormeting.
Voor de beantwoording van de eerste subvraag worden de volgende onafhankelijke variabelen gebruikt die door Oberon zijn gemeten:

a) Effectief leiderschap: aantal wisselingen van schoolleiders, inhoudelijk-, faciliterend-, delegerend- en democratisch leiderschap.
b) Opbrengstgericht werken: gegevens verzamelen, doelen formuleren, groepsplannen opstellen, evalueren en mate van implementatie van opbrengstgericht werken.
c) Doorgaande lijn: leerkracht intern aangetrokken, zelfde VVE-programma in startgroep als in groep 1 en 2. Aanvullende meting door UT: doorgaande lijn (zijn de kinderen uit de steekproef doorgestroomd naar de startgroep basisschool).

Van bijna alle bovengenoemde variabelen wordt een positief effect verwacht op de ontwikkeling. Echter, van het aantal wisselingen van schoolleiders wordt een negatief effect verwacht.
Voor de tweede subvraag naar de invloed van de inzet van een bevoegde leerkracht naast een mbo-geschoolde beroepskracht wordt allereerst gekeken naar de analyses van de proceskwaliteit. Verwacht wordt dat de inzet van een pabo (hbo-)geschoolde beroepskracht samenhangt met een hogere proceskwaliteit in de startgroepen in vergelijking met de

gemiddelde proceskwaliteit in de controlegroep. Daarnaast worden de volgende door Oberon gemeten variabelen gebruikt: ervaren competentie van de leerkracht, ervaren competentie van de pedagogisch medewerker, samenwerking tussen beiden, aantal wisselingen van de leerkracht, aantal wisselingen van de pedagogisch medewerker. Verwacht wordt dat ervaren competentie en een goede samenwerking samenhangen met de ontwikkeling van kinderen, terwijl er een negatieve samenhang wordt verwacht als er sprake is van veel wisselingen van beroepskrachten.

Voor de beantwoording van de derde subvraag wordt de onafhankelijke variabele 'groepssamenstelling' gebruikt. Voor de analyses wordt deze variabele opgesplitst in twee dichotome variabelen[15], zodat alle groepscombinaties kunnen worden getoetst. Op grond van eerder verricht Nederlands onderzoek mag verwacht worden dat kinderen uit gemengde groepen meer ontwikkeling laten zien (de Haan, Elbers, Hoofs, & Leseman, 2012) dan kinderen uit groepen met alleen doelgroepkinderen.

Twee variabelen die sterk samenhangen met de variabele groepssamenstelling zijn: percentage thuistaal niet-Nederlands en percentage kinderen met leerlinggewicht (beiden sociaal economische groepskenmerken). Verwacht wordt dat de respondenten uit de groepen die hoog scoren op deze variabelen, minder ontwikkeling laten zien.

Voor de laatste subvraag naar de effecten van ouderbetrokkenheid worden de volgende variabelen gebruikt die door Oberon zijn gemeten: ouderparticipatie en educatief partnerschap: thuisopdrachten voor ouders, individuele oudergesprekken (10-minuten gesprekken) en (thematische) groepsbijeenkomsten voor ouders. Van alle variabelen wordt een positief effect op de ontwikkeling verwacht.

Omdat een goede implementatie van de kenmerken van de startgroepen tijd vraagt, kan verwacht worden dat het effect van deze kenmerken groter is op de ontwikkeling van kinderen uit cohort twee dan uit het eerste cohort. Ook van het algehele succes van de implementatie (zoals vastgesteld in het door Oberon uitgevoerde implementatieonderzoek: Oberon, 2015) wordt een positief effect op de ontwikkeling verwacht.

Naast deze implementatie kenmerken die door Oberon zijn gemeten, zijn ten behoeve van het effectonderzoek ook structurele- en procesgerelateerde kwaliteitskenmerken van de startgroepen gemeten. Van een aantal *structurele kwaliteit*skenmerken wordt een effect op de ontwikkeling van kinderen verwacht. Zo wordt een positief effect verwacht van het aantal uren per week, maar daarentegen een negatief effect van het aantal middaguren. Jongere kinderen kunnen 's middags nog behoefte aan slaap hebben en erg moe zijn, ofwel 's middags niet altijd

[15] 1. Groepen met (bijna) uitsluitend niet-doelgroep kinderen. 2. Gemengde groepen.

door hun ouders gebracht zijn. Ook van het aantal kinderen per startgroep wordt een negatief effect verwacht omdat in een grote groep minder mogelijkheden zijn om interacties aan te gaan en het bovendien moeilijker is om het niveau af te stemmen op individuele kinderen (de Haan, Leseman & Elbers, 2011).
Verder worden effecten verwacht van bepaalde aspecten van de proceskwaliteit. Zo wordt verondersteld dat de aanstelling van een stagiaire op de groep, samenhangt met positieve effecten. Immers, de aanwezigheid van een stagiaire maakt het makkelijker om de groep in kleine groepjes te splitsen. Dit zal de ontwikkeling van kinderen in de startgroepen naar verwachting ten goede komen. Bovendien wordt verwacht dat de mate waarin pedagogische ondersteuning (adviezen en feedback op pedagogisch handelen) op de werkvloer wordt gegeven (door bv. collega's, leidinggevenden, of een pedagoog), een positieve invloed heeft op de ontwikkeling van kinderen. Een andere belangrijke indicator van proceskwaliteit is de geobserveerde kwaliteit, zoals gemeten met de CLASS-toddler en ECERS-E. Onderzoek heeft laten zien dat betekenisvolle interacties een belangrijke rol spelen in de ontwikkeling van kinderen (Howes et al., 2008; Sylva et al., 2011). Zo bleek uit het Pre-COOL onderzoek (waar dezelfde instrumenten zijn gebruikt) dat een hogere mate van emotionele regulatie samenhangt met ontwikkeling in receptieve woordenschat en dat een hogere mate van educatieve ondersteuning samenhangt met ontwikkeling in selectieve aandacht (Slot, 2014). Op grond van deze resultaten kan verwacht worden dat de kwaliteit op pedagogisch en educatief gebied de ontwikkeling van kinderen positief zal beïnvloeden. Om dit na te gaan zullen in dit onderzoek de gemiddelde scores op de drie domeinen van de CLASS-toddler (emotionele regulatie, gedragsorganisatie en educatieve ondersteuning) en de gemiddelde scores op taal en geletterdheid (ECERS-E) worden gebruikt als onafhankelijke variabelen. Daarnaast zijn vanwege de variatie op proceskwaliteit tussen activiteiten de maximum scores op emotionele regulatie en educatieve ondersteuning (gemiddeld over de dimensies van elk domein) als onafhankelijke variabelen in de analyses gebruikt (zie tabel 11 in de bijlage).

Resultaten

Allereerst zal de proceskwaliteit op de startgroepen worden beschreven, zoals gemeten met de CLASS-toddler en ECERS-E. Deze resultaten geven inzicht in de verschillen tussen startgroepen ten aanzien van de proceskwaliteit. Daarna wordt er een vergelijking gemaakt

tussen de (gemiddeld ervaren) proceskwaliteit van kinderen uit de experimentele groep en die van kinderen uit de controlegroep. De vergelijking vormt een eerste antwoord op de tweede subvraag (het effect van de inzet van een hbo-geschoolde beroepskracht met onderwijsbevoegdheid).
In het tweede gedeelte zullen gegevens gepresenteerd worden die bijdragen aan de beantwoording van de hoofdvraag. Er wordt een beschrijving gegeven van de variabelen op grond waarvan de propensity score matching is uitgevoerd. Vervolgens worden de resultaten beschreven van de t-tests en regressie analyses (zie analytische strategie). Om een goed beeld te geven van de verschillen tussen de experimentele groep en controlegroep, worden de gemiddelde scores in een tabel weergegeven. Daarnaast wordt de groei in kaart gebracht op receptieve woordenschat en selectieve aandacht. De groeicurves van de experimentele groep en controlegroep zullen in figuren worden afgebeeld.
In het laatste gedeelte worden de resultaten beschreven die antwoorden zullen geven op de subvragen. Hierbij worden de verschillen tussen startgroepen gerelateerd aan verschillen tussen kinderen op taal, rekenen en selectieve aandacht.

Proceskwaliteit

Om meer inzicht te verkrijgen in de proceskwaliteit op de startgroepen, wordt in tabel 10 (zie bijlage) de range, het gemiddelde en de spreiding van de scores op alle dimensies van de CLASS-toddler en de items van de ECERS-E weergegeven. Voor de CLASS-toddler wordt tevens weergegeven hoe vaak elke beroepskracht sterker was op de dimensie dan haar collega.[16]
Hoewel activiteiten sterk kunnen variëren in vorm en inhoud en een eetmoment niet te vergelijken is met een voorleesactiviteit[17], geven de minimum en maximum scores toch een goed beeld van de kwaliteit van het aanbod. Zoals weergegeven in tabel 10, is de geobserveerde sfeer op de groep bijna altijd midden of hoog positief[18]. Hetzelfde geldt voor sensitieve responsiviteit[19]. De scores variëren meer op de mate waarin beroepskrachten ruimte bieden voor het perspectief van het kind (zie sd). Een hoge score houdt in dat zij de kinderen de activiteiten veelal laten bepalen en leiden, flexibel zijn in plannen tijdens deze activiteiten en

[16] Deze scoring is ten behoeve van dit onderzoek aan het instrument toegevoegd. Betrouwbaarheid en validiteit van de meting kunnen echter niet worden onderzocht. De scores dienen daarom met enige voorzichtigheid geïnterpreteerd te worden.
[17] Een eet- en drinkmoment kan door beroepskrachten benut worden voor leren en ontwikkeling, maar de educatieve kwaliteit kan ook laag zijn. Bij voorlezen is de type activiteit op zichzelf al educatief stimulerend.
[18] op 3 laag-midden scores van 3 na (van in totaal 120 activiteiten of cycli).
[19] Op 5 laag-midden scores na (van in totaal 120 activiteiten of cycli).

proberen de autonomie van kinderen te vergroten. Tijdens sommige activiteiten wordt dit veelal gedaan, terwijl in andere activiteiten weinig ruimte is voor het perspectief van het kind. Ook de mate waarin gedrag adequaat wordt gereguleerd varieert sterk tussen de verschillende activiteiten, maar gemiddeld wordt hierop een midden-hoge score behaald. Op het educatieve domein zijn de scores een stuk lager. Hoewel er momenten zijn van een relatief hoge mate van educatieve ondersteuning, is het gemiddelde lager dan de middenscore. Wel zijn er op alle items activiteiten waarbij de educatieve kwaliteit midden-hoog tot hoog was met een score van 6. Op alle items zijn er echter ook activiteiten waarbij de absolute minimum score is behaald.
In tabel 10 is verder zichtbaar dat leerkrachten op alle items van de CLASS-toddler vaker kwalitatief sterker zijn dan de pedagogisch medewerkers. Leerkrachten waren 80% tot 86% van de tijd sterker op het educatieve domein volgens de observatoren.
Tijdens de groepsobservaties is ook de ECERS-E afgenomen. Met dit instrument is geobserveerd welke materialen er aanwezig waren op de groepen en in welke mate de omgeving educatief stimulerend was voor de kinderen (Sylva et al., 2008). Zo werd er met betrekking tot geletterdheid bijvoorbeeld geobserveerd in welke mate er geschreven letters en woorden aanwezig waren in de groepsruimte en wat de aard van de boeken was die beschikbaar waren voor de kinderen. Zoals zichtbaar is in tabel 10 (zie bijlage) is de mate van geletterdheid op de meeste items gemiddeld onvoldoende tot minimaal. Er waren op sommige groepen echter wel uitschieters, waardoor er een 4 (tussen minimaal en goed) tot een 7 (uitstekend) gescoord werd. De score 6 bij het item ‘boeken en leeshoek’ houdt bijvoorbeeld in dat er een grote verscheidenheid aan boeken met een variërende moeilijkheidsgraad aanwezig was, de kinderen de leeshoek zelfstandig gebruikten en de leeshoek comfortabel was ingericht. De mate waarin er gepraat en geluisterd werd gedurende de observatie, vertoonde een ander beeld. In bijna een derde van de groepen werd hierop goed tot uitstekend gescoord. Een voorbeeld hiervan is dat kinderen aangemoedigd werden om vragen op uitgebreidere wijze te beantwoorden en er regelmatig open vragen gesteld werden om de taal van kinderen uit te breiden. Gemiddeld genomen was de mate waarin kinderen aangemoedigd werden om te praten en luisteren tussen minimaal en goed, te vergelijken met een midden score. Op bijna een kwart van de startgroepen werd echter onvoldoende tot minimaal gescoord op dit item. Dit houdt in dat er slechts enkele conversaties met de kinderen waren gedurende de ochtend.
Met betrekking tot het aanbod op het gebied van rekenen scoren de meeste startgroepen onvoldoende tot minimaal. Een uitzondering vormt het item ‘sorteren, vergelijken en matchen’ waarop gemiddeld minimaal tot goed is gescoord. Dit houdt in dat beroepskrachten sorteren, vergelijken en matchen demonstreren aan de kinderen en hen toestaan te participeren.

Er blijkt enige variatie te zijn tussen de startgroepen, en binnen startgroepen gedurende de dag, op het gebied van proceskwaliteit. De emotionele regulatie en groepsorganisatie blijken op de startgroepen van redelijk hoge kwaliteit te zijn. Over het geheel gezien is de educatieve kwaliteit gemiddeld echter vrij laag.

Wanneer we de proceskwaliteit op de verschillende domeinen van de CLASS-toddler en ECERS-E vergelijken tussen kinderen uit de startgroepen en kinderen uit de controlegroep, blijken er op alle domeinen van de CLASS-toddler significante verschillen te bestaan (zie tabel 1). Wat betreft emotionele regulatie en groepsorganisatie wordt er op de startgroepen significant hoger gescoord, terwijl op educatieve ondersteuning juist iets hoger wordt gescoord binnen de controlegroep. Een opvallend resultaat, omdat verwacht mag worden dat, vanwege de aanwezigheid van een bevoegde hbo-kracht binnen de setting van de startgroepen, de educatieve ondersteuning juist van hoge(re) kwaliteit is. Op de ECERS-E worden geen significante verschillen gevonden. Wel valt op dat voor zowel kinderen uit de startgroepen als voor kinderen uit de controlegroep gemiddeld genomen een lage kwaliteit op het gebied van educatie (taal en rekenen) wordt geobserveerd.

Tabel 1 Proceskwaliteit startgroepen en controlegroep (mm2)

		CLASS-toddler			**Ecers-E**	
		Emotionele regulatie	Groeps-organisatie	Educatieve ondersteuning	Geletterd-heid	Rekenen
Startgroepen	**Mean**	**5.44**	**5.09**	**2.88**	**2.69**	**1.89**
	N	276	276	276	276	276
	Sd	0.45	0.62	0.54	0.47	0.77
Controlegroep	**Mean**	**5.20**	**4.37**	**3.20**	**2.76**	**1.77**
	N	134	134	134	136	136
	Sd	0.40	0.59	0.63	0.78	0.62
Sign.		**.000**	**.000**	**.000**	**N.S.**	**N.S.**

Resultaten hoofdvraag

Voor de propensity score matching is een aantal covariaten geselecteerd. In tabel 12 (zie bijlage) wordt een percentuele verdeling weergegeven van de waarden op deze covariaten, waarbij de experimentele groep (kinderen uit de startgroepen) vergeleken wordt met de controlegroep (de gematchte en gewogen subsample uit het Pre-COOL onderzoek). In de tabel is te zien dat er op de variabelen na de matching geen significante verschillen bestaan tussen

de groepen. Ook is te zien dat de leeftijd op meetmoment 2 is meegenomen in de matching waardoor de leeftijd niet verschilt tussen de groepen (3,26 en 3,27 jaar). Zodoende kunnen de scores op meetmoment 2 goed met elkaar vergeleken worden.

Tabel 2 Vergelijking van experimentele groep en controlegroep

		Startgroepen (N=276)	**Controlegroep** (N=206 gematchte en gewogen groep)	**P-waarde ***	**Effect-grootte ≥ 0.20**
		Mean (sd)	**Mean (sd)**		
Leeftijd taken (in jaren)	mm1	2.73 (0.18)	2.29 (0.22)	0.000	
	mm2	3.26 (0.17)	3.27 (0.18)	N.S.	
	mm3	4.31 (0.19)	4.73 (0.22)	0.000	
	mm4	5.28 (0.18)	5.80 (0.27)	0.000	
Leeftijd CITO groep 1 (in jaren)	Taal E1	5.00 (0.33)	4.90 (0.36)	0.007	
	Rekenen E1	5.00 (0.33)	4.89 (0.33)	0.003	
Receptieve woordenschat (perc. goed)	mm2	46.44 (20.13)	47.83 (17.48)	N.S.	
Selectieve aandacht (1-8)	mm2	5.65 (1.27)	5.41 (1.10)	0.017	0.20
CITO rekenen voor peuters (perc. goed)	mm2	50.84 (20.93)	43.21 (17.03)	0.000	0.40
CITO groep 1 (vaardigheidsscores)	Taal E1	55.12 (13.61)	51.75 (12.55)	0.016	0.26
	Rekenen E1	70.52 (13.68)	68.65 (9.29)	N.S.	

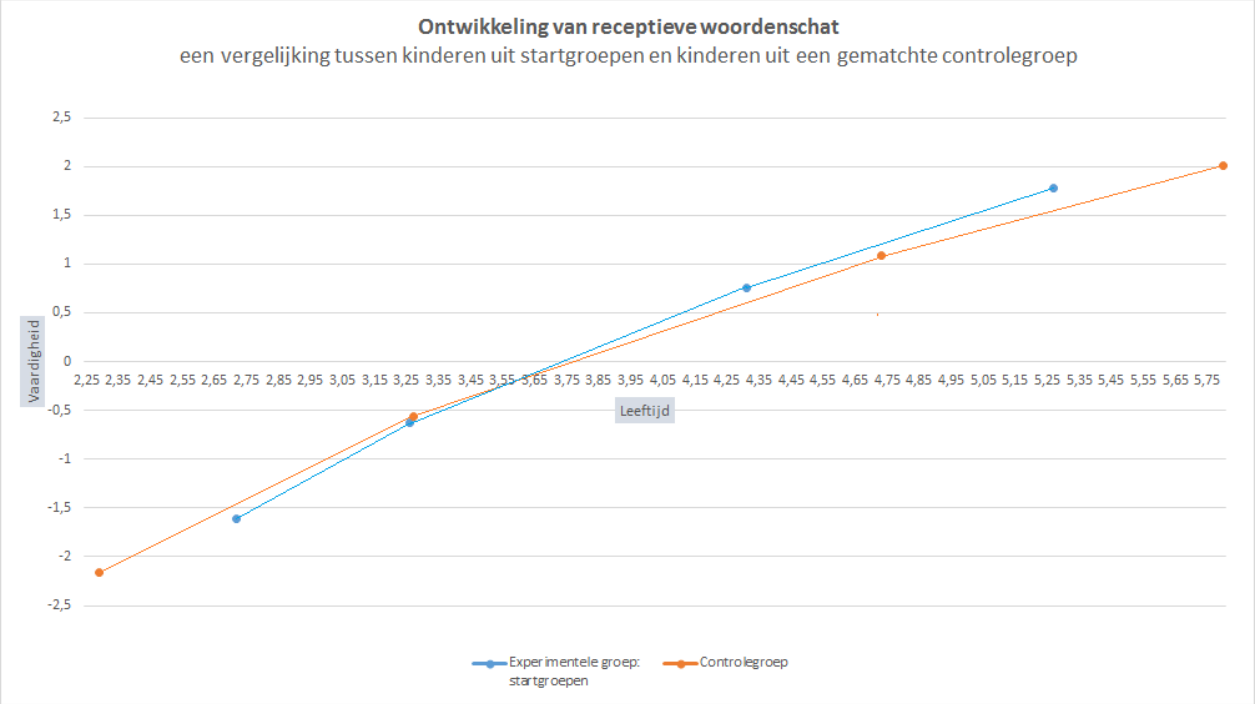
* De p-waarde (significantieniveau) bij eenzijdige toetsing

In tabel 2 worden verschillende resultaten van de analyses weergegeven, alsmede de leeftijden van de kinderen uit de steekproef. De resultaten laten zien dat er op korte termijn (meetmoment

2) geen effect voor taal (receptieve woordenschat) wordt gevonden. Wel worden er kleine korte termijn effecten[20] voor rekenen (0,40) en selectieve aandacht (0,20) gevonden.
Om na te kunnen gaan of er ook lange termijn effecten zijn, zijn er regressie analyses uitgevoerd, waarbij leeftijd op toetsdatum als covariaat is gebruikt. Er wordt een klein lange termijn effect voor taal (0,26) gevonden. Oftewel, kinderen uit de startgroepen scoren in groep 1 significant hoger op de "CITO toets voor kleuters taal E1" dan kinderen uit de controlegroep. Op de "CITO toets voor kleuters rekenen E1" wordt geen significant verschil gevonden tussen scores van kinderen uit de startgroepen en die van kinderen uit de controlegroep. Oftewel, de voorsprong op rekenen op meetmoment 2 (zie hierboven) lijkt te zijn uitgedoofd aan het einde van groep 1.
Om verschillende ontwikkelingspaden over tijd in kaart te brengen, zijn er vervolgens vaardigheidsscores berekend (Boom, Mulder, & Verhagen, 2016). Met deze vaardigheidsscores kan de (latente) groei over langere tijd in kaart worden gebracht. Bovendien kunnen de vaardigheidscores van verschillende meetmomenten in de tijd vergeleken worden tussen de experimentele en de controlegroep. Daartoe is allereerst nagegaan wat de invloed is van uitval op de vergelijkbaarheid van de experimentele groep met de controlegroep. Hiertoe zijn alleen de respondenten geselecteerd die op alle meetmomenten een score hebben behaald. Op receptieve woordenschat gaat het in totaal om N=191 respondenten en voor selectieve aandacht om N=176 respondenten. Wanneer de gemiddelde propensity scores vergeleken worden, blijken deze niet significant van elkaar te verschillen. Oftewel, kinderen uit de startgroepen die op alle meetmomenten een score hebben behaald zijn wat betreft achtergrondkenmerken waarop gematcht is vergelijkbaar met kinderen uit de controlegroep die op alle meetmomenten een score hebben behaald.
Vervolgens zijn de gemiddelde scores (alsmede de leeftijd waarop deze score is behaald) van de experimentele groep en de controlegroep in Excel geïmporteerd. Deze scores zijn geplot. In figuur 1 is de ontwikkelingsvergelijking op receptieve woordenschat weergegeven. Zoals te zien is in figuur 1, lijkt de receptieve woordenschat van de kinderen in de startgroepen sterker te groeien, in vergelijking met die van kinderen in de controlegroep[21]. In tabel 3 worden de vaardigheidsscores weergegeven die samen deze figuur vormen (zwartgedrukt). Tevens worden de effecten weergegeven. Deze effecten zijn gebaseerd op het uitgangspunt van

[20] Effectgroottes: Cohen's D: 0.20 klein effect, 0.50 medium effect, 0.80 groot effect (Cohen, 1992).
[21] Alleen de stippen uit figuur 1 zijn vastgesteld op basis van gemeten data. De door ons getrokken lijnen tussen elke twee punten hebben weliswaar een lineair verloop, het is niet duidelijk of dit correspondeert met de werkelijkheid. Het is bijvoorbeeld mogelijk dat er in werkelijkheid sprake is van een kwadratische groei tussen meetmoment 1 en 2 binnen de controlegroep. In dat geval zou het goed mogelijk zijn dat de blauwe stip op de lijn van de controlegroep ligt.

Figuur 1



lineaire groei tussen twee meetmomenten. Zodoende was het mogelijk om de gemiddelde scores te berekenen van de controlegroep op de leeftijden waarop de metingen zijn verricht

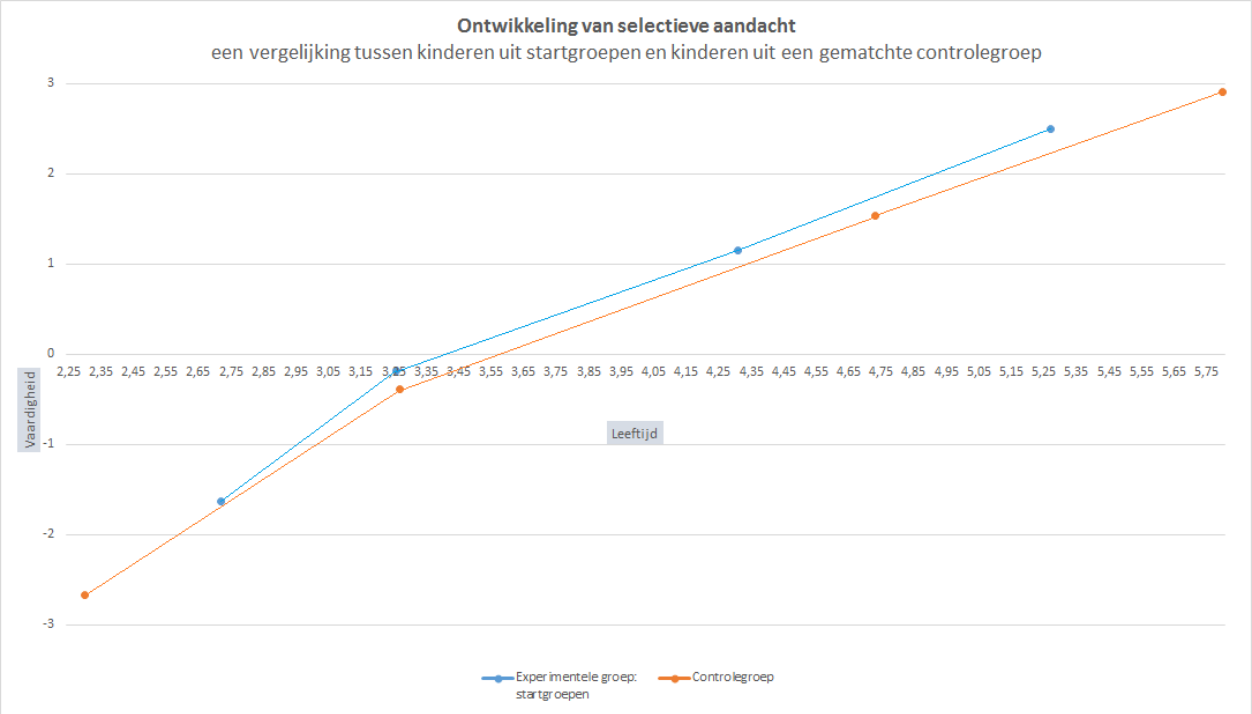
Tabel 3 Gemiddelde vaardigheid receptieve woordenschat

Leeftijd	Experimentele groep			Controlegroep			t	P waarde*	Effect-grootte ≥ 0.20
	N	Mean	Sd	N	Mean	Sd			
2.29				201	-2.16	1.03			
2.72	269	-1.61	0.86	204	-1.46	0.89	-1.86	N.S.	
3.26	275	-0.63	0.85	206	-0.58	0.75	-0.62	N.S.	
3.27				206	-0.57	0.75			
4.31	262	0.76	0.70	142	0.61	0.70	2.08	0.019	0.21
4.73				77	1.08	0.64			
5.27	125	1.78	0.59	82	1.55	0.62	2.68	0.004	0.38
5.80				86	2.00	0.60			

* De p-waarde (significantieniveau) bij eenzijdige toetsing

binnen de experimentele groep. Deze getallen zijn (grijs gedrukt) eveneens afgebeeld in tabel 3. In de tabel is verder zichtbaar dat er op de lange termijn (na doorstroom vanuit de startgroep) een klein effect van de startgroepen (interventie) zichtbaar is op meetmomenten 3 en 4. Zoals eerder werd aangegeven lijkt er een trend zichtbaar waarbij er sprake is van een inhaalslag van kinderen uit de startgroepen. Deze conclusie dient echter met voorzichtigheid getrokken te worden, omdat de verschillen op de eerste twee meetmomenten niet significant zijn (zie tabel 3). Wel is duidelijk dat de kinderen uit startgroepen op de langere termijn een grotere receptieve woordenschat hebben en dat dit verschil lijkt toe te nemen.

Figuur 2



In figuur 2 is de ontwikkeling op selectieve aandacht voor zowel de experimentele als controlegroep weergegeven. In de figuur lijkt er sprake te zijn van een gelijke start, waarbij er tussen meetmoment 1 en 2 een voorsprong ontstaat van kinderen uit de startgroepen in vergelijking met kinderen uit de controlegroep. Dit beeld komt overeen met de p-waarden uit tabel 4. Op meetmoment 1 is er geen significant verschil. Vervolgens ontstaat er bij meetmoment 2 een significante voorsprong van kinderen uit startgroepen en deze voorsprong houdt aan tot en met het laatste meetmoment. Het kleine gevonden effect op meetmoment 2 komt in de richting van een middelgroot effect (0,44) op het laatste meetmoment.

Tabel 4 Gemiddelde vaardigheid selectieve aandacht

Leeftijd	Experimentele groep			Controlegroep			t	P waarde*	Effect-grootte ≥ 0.20
	N	Mean	Sd	N	Mean	Sd			
2.30				161	-2.68	1.21			
2.72	262	-1.64	1.05	184	-1.69	1.02	0.53	N.S.	
3.26	274	-0.19	0.95	206	-0.42	0.84	2.79	0.003	0.26
3.27				206	-0.39	0.84			
4.31	262	1.15	0.77	142	0.98	0.74	2.14	0.017	0.22
4.73				78	1.54	0.63			
5.27	125	2.50	0.62	82	2.23	0.60	3.13	0.001	0.44
5.80				85	2.91	0.57			

* De p-waarde (significantieniveau) bij eenzijdige toetsing

Resultaten subvragen

De subvragen zijn gericht op het analyseren van de verschillen tussen startgroepen en de rol die verscheidene kenmerken van startgroepen hebben op de ontwikkeling van kinderen. Om een beeld te krijgen van de verschillen tussen startgroepen is hiervan een overzicht gemaakt welke wordt weergegeven in tabel 11 (zie bijlage). In de tabel is te zien dat iets minder dan de helft van de startgroepen meer uren dan het minimum aantal uren van 12,5 uur aanbiedt per week. Verder is er op een derde van de startgroepen geen stagiaire, terwijl bij een kwart van de startgroepen meer dan de helft van de week een stagiaire is. De groepssamenstelling varieert sterk: er zijn gemengde groepen[22], in sommige startgroepen zitten voornamelijk doelgroepkinderen en in andere groepen zitten bijna geen doelgroepkinderen. De variatie in groepssamenstelling tussen startgroepen wordt verder duidelijk door de verschillen in sociaal economische groepskenmerken (op deze kenmerken wordt in vrijwel alle gemeentes in Nederland de doelgroep- of VVE-indicatie gebaseerd). Met betrekking tot educatief partnerschap blijkt dat groepsbijeenkomsten voor ouders in bijna alle groepen worden georganiseerd. Toch worden in bijna een derde van de startgroepen geen thuisopdrachten aan ouders gegeven. Tot slot blijkt dat ruim twee derde van de startgroepen een algeheel succesvolle implementatie kent.

[22] In gemengde groepen komen doelgroepkinderen alle dagdelen en de kinderen zonder VVE-indicatie komen maar twee dagdelen per week, waardoor de groepssamenstelling varieert van dag tot dag.

Hoewel de variatie op de variabelen dus uiteenloopt, zijn er 390 multilevel analyses uitgevoerd om de invloed van 39 onafhankelijke variabelen (kenmerken van startgroepen) op 10 afhankelijke variabelen (taal, rekenen en selectieve aandacht) te onderzoeken. Uit 390 analyses komen 18 significante (P<,05, eenzijdige toetsing) resultaten naar voren (waarbij gecorrigeerd is voor covariaten, zie kopje analytische strategie). De resultaten zijn niet robuust, oftewel er hangt steeds een andere onafhankelijke variabele samen met scores op taken op meetmomenten 3 en 4 en CITO scores uit groep 1. Bovendien tonen 11 van de 18 significante toetsen een relatie in de verkeerde richting aan, m.a.w. tegengesteld aan onze hypothesen. Op basis van het significantieniveau kan wel gesteld worden dat het aantal significante relaties aardig overeen komt met wat men op basis van toeval zou verwachten (ongeveer 5%) en niet duidt op betekenisvolle verbanden. In de discussie zullen hiervoor mogelijke verklaringen gegeven worden.

Discussie

Proceskwaliteit

Uit de beschrijving van de proceskwaliteit binnen de startgroepen komt naar voren dat de emotionele regulatie van midden-hoge kwaliteit is. Vooral de sfeer op de groep en sensitieve responsiviteit is van hoge kwaliteit. De ruimte die op de groep geboden wordt voor het perspectief van het kind laat een iets wisselender beeld zien. Met betrekking tot groepsorganisatie is er eveneens veel variatie, maar gemiddeld is de groepsorganisatie van midden-hoge kwaliteit op de startgroepen. De midden tot hoge kwaliteit binnen de startgroepen op het gebied van emotionele regulatie en groepsorganisatie, demonstreren de aanwezigheid van warme en ondersteunende interacties binnen een gestructureerde setting. Beide aspecten dragen bij aan de ontwikkeling van zelfregulatie van kinderen. Het bevorderen van zelfregulatie is van belang om schoolrijpheid en later schoolsucces te bevorderen (Blair, & Diamond, 2008). Er is echter ruimte voor verbetering op het gebied van educatieve ondersteuning. Er zijn activiteiten geobserveerd waarbij leren en ontwikkeling in grote mate gefaciliteerd werd, de kwaliteit van feedback aan kinderen van hoge kwaliteit was en de taalontwikkeling op adequate

wijze werd gestimuleerd[23]. Echter, dit gebeurt niet systematisch, dat wil zeggen: niet gedurende de hele ochtend, waardoor de gemiddelde scores een kwaliteit in de klasse laag-midden weergeven.

Dit beeld komt overeen met wat er op het gebied van geletterdheid en rekenen (ECERS-E) gemiddeld is geobserveerd. Ook hier scoren de startgroepen op sommige items in de range 'goed tot uitstekend'. Gemiddeld genomen is er echter ruimte voor verbetering op geletterdheid en rekenen bij alle startgroepen.

De nauwkeurige uitsplitsing naar items en dimensies hebben we alleen kunnen maken voor de startgroepen. Voor de controlegroep waren in D.A.N.S. alleen de domeinscores van de CLASS-toddler en ECERS-E aanwezig. Vergelijking van deze gemiddelde domeinscores laat significante verschillen zien tussen de experimentele groep en controlegroep op alle domeinen van de CLASS-toddler. De verschillen zijn echter klein: zowel binnen de controlegroep als op de startgroepen zijn pedagogisch handelen en groepsmanagement van midden-hoge kwaliteit en wordt laag (laag-midden) gescoord op educatieve ondersteuning. Hetzelfde geldt voor geletterdheid en rekenen (ECERS-E) waarop beide groepen gemiddeld laag scoren. Dit laat zien dat er ruimte is voor verbetering op het gebied van faciliteren van leren en ontwikkeling, kwaliteit van feedback op uitlatingen van kinderen, alsmede de manier waarop taalontwikkeling, geletterdheid en rekenen worden gestimuleerd. Zowel op reguliere VVE-groepen als op groepen waar een hbo-geschoolde begeleider met onderwijsbevoegdheid werkzaam is. De invloed van de inzet van een hbo-geschoolde begeleider met onderwijsbevoegdheid (subvraag 2) lijkt daarmee op het eerste gezicht geen verschil te maken voor de proceskwaliteit. Wel is in de startgroepen geobserveerd dat het optreden van leerkrachten bij activiteiten vier keer zo vaak als sterker is gecategoriseerd op het educatieve domein van de CLASS-toddler dan bij pedagogisch medewerkers (zie tabel 10 bijlage).

Effecten van startgroepen (hoofdvraag)

Kinderen uit de experimentele- en controlegroep blijken op driejarige leeftijd niet significant van elkaar te verschillen als het hun receptieve woordenschat betreft. Op de lange termijn zijn er echter wel, weliswaar kleine, effecten gevonden voor receptieve woordenschat op vier- en vijfjarige leeftijd en voor woordenschat en kritisch luisteren in groep 1 (CITO toets taal voor kleuters).

[23] Zie minimum en maximum scores op educatieve ondersteuning in tabel 8 (bijlage), alsmede de maximum scores op educatieve ondersteuning in tabel 9 (bijlage).

Met betrekking tot rekenen is een ander beeld te zien. De kinderen uit de startgroepen scoren ongeveer een half jaar na de eerste meting hoger op rekenen dan vergelijkbare kinderen uit de controlegroep. Na de doorstroom op de basisschool, zo rond de leeftijd van 5 jaar, lijken deze effecten echter uitgedoofd. Het kleine gevonden effect voor rekenen op 3 jaar kan mogelijk verklaard worden doordat beroepskrachten van de startgroepen tijdens de observatie tussen minimaal en goed scoorden op het item[24] sorteren, vergelijken en matchen, waarbij bijvoorbeeld wordt aangemoedigd dat kinderen vergelijkingen gebruiken bij meten (groot, groter, grootst). Naast getalbegrip, zijn meten en meetkunde belangrijke onderdelen in de CITO-toetsen voor peuters en kleuters. Mogelijk is hier dus sprake van een effect van proceskwaliteit op de ontwikkeling van kinderen.
De resultaten voor rekenen op de lange termijn sluiten aan bij wat er in de kleutergroepen in de praktijk aangeboden wordt op het gebied van rekenen. Uit onderzoek van de Haan et al. (2011) blijkt namelijk dat minder dan 5% van de tijd in de kleuterklassen wordt besteed aan leerkrachtgestuurde rekenactiviteiten.
Voor selectieve aandacht wordt er een klein effect gevonden. De vaardigheid van selectieve aandacht van kinderen uit de startgroepen neemt het eerste half jaar iets sneller toe dan de vaardigheid van vergelijkbare kinderen in reguliere voorschoolse voorzieningen. Daarna houdt deze voorsprong aan. Uit onderzoek van Garon, Bryson, & Smith (2008) en Veer, Luyten, Mulder, van Tuijl, & Sleegers (2015) blijkt dat selectieve aandacht centraal staat in de ontwikkeling van executieve functies: werkgeheugen, inhibitie en cognitieve flexibiliteit. Zoals eerder beschreven zijn executieve functies gerelateerd aan leren (Blair, & Razza, 2007). De bevinding dat de vaardigheid van kinderen uit de startgroepen met betrekking tot selectieve aandacht sneller toeneemt dan bij kinderen uit de controlegroep lijkt dan ook positief te zijn, gezien het belang van selectieve aandacht voor de ontwikkeling van executieve functies en daarmee ook voor taal- en rekenvaardigheden. Het gevonden kleine effect voor selectieve aandacht zou tevens een verklaring kunnen bieden voor de gevonden verschillen tussen kinderen uit de startgroepen en de controlegroep op het gebied van hun ontwikkeling van taalvaardigheden op lange termijn.

Kenmerken van startgroepen (subvragen)

Ten einde de subvragen te kunnen beantwoorden, is de variatie tussen de startgroepen gerelateerd aan scores van kinderen op taal, rekenen en selectieve aandacht. De startgroepen

[24] Bij 30% van de startgroepen is dit item gescoord.

verschillen onderling in structurele kwaliteit (o.a. de intensiteit), proceskwaliteit, leiderschap, opbrengstgericht werken, de doorgaande lijn, succes van het team van beroepskrachten, groepssamenstelling, educatief partnerschap, en succes van de implementatie. Toch hebben we niet kunnen vaststellen dat deze verschillen in kenmerken ook leiden tot verschillen in ontwikkelingsuitkomsten van kinderen. Dit terwijl eerder onderzoek laat zien dat zowel structurele kwaliteit als proceskwaliteit samenhangt met de ontwikkeling van kinderen (Slot, 2014; Sylva et al., 2011).

De meest aannemelijke verklaring voor het niet vinden van verbanden lijkt het beperkte aantal groepen (30), en de geringe variatie tussen de startgroepen te zijn. Een voorbeeld van de geringe variatie is groepssamenstelling: er zijn slechts vijf gemengde groepen tegenover 25 homogene groepen. De variatie in intensiteit is eveneens beperkt: van 12,5 tot 15 uur per week. De kleine variatie op proceskwaliteit is te zien aan de standaarddeviaties in tabel 10. Ter illustratie: de scores op het educatieve domein van de CLASS-toddler variëren tussen 2,2 en 4,3 terwijl de mogelijke range van 1 tot 7 loopt.

Samenvattend kunnen we stellen dat we in dit onderzoek geen bewijs vinden voor verschillen in ontwikkelingsuitkomsten naar aanleiding van verschillen op groepsniveau binnen de startgroepen. De groepssteekproefgrootte lijkt te klein te zijn om duidelijke uitspraken te kunnen doen. Zodoende is het ook niet mogelijk om de gevonden effecten van startgroepen (hoofdvraag) uit te splitsen naar afzonderlijke elementen van startgroepen.

Beperkingen van het onderzoek

Door middel van de propensity score matching is er gestreefd naar de benadering van een gerandomiseerde studie. De propensity score is eigenlijk een soort samengestelde covariaat en kan alleen maar worden bepaald op grond van gemeten variabelen en niet op grond van ongemeten variabelen. Dit betekent dat zogenaamde residuele systematische bias niet volledig kan worden uitgesloten.

Een andere beperking heeft betrekking op het grote aantal missende waarden binnen het Pre-COOL bestand. Hierdoor hebben we ons moeten beperken tot die respondenten van wie we over complete informatie konden beschikken. De voor dit onderzoek belangrijkste variabelen die veel missende waarden hadden zijn: opleidingsniveau ouder(s), geboorteland ouder(s), VVE (0,1) en groepsidentificatienummer[25],. Dit heeft er dan ook toe geleid dat respondenten van wie

[25] Een instelling uit het Pre-COOL bestand kan bestaan uit meerdere groepen, ieder met een unieke proceskwaliteit. Van de meeste kinderen is niet bekend tot welke groep zij behoorden. Daartoe is besloten de gemiddelde proceskwaliteit van een instelling te gebruiken. Bij de startgroepen is er slechts één groep per instelling.

de informatie wel bekend was, veel gewicht in de schaal hebben gelegd. Het is daardoor mogelijk dat alleen de kinderen van welwillende laagopgeleide ouders uit Pre-COOL, door ons zijn gebruikt voor de vergelijking. Mogelijk zijn dit precies de ouders die eveneens actiever zijn in het stimuleren van hun kinderen. Dit betekent dat de kans groot is dat er selectie-bias heeft kunnen optreden. Echter, door de aanvullende koppeling die we via het CBS hebben laten maken, hebben we deze bias enigszins kunnen bestrijden.
Door het grote aantal missende data omtrent groepsidentificatienummer, is de data op instellingsniveau gebruikt voor de gegevens over proceskwaliteit. Het moge duidelijk zijn dat de mogelijke variatie in proceskwaliteit tussen groepen binnen een instelling hierdoor niet is meegenomen en de resultaten minder precies zijn. Bovendien is daardoor niet helemaal duidelijk wat het opleidingsniveau was van de leidsters op de groepen waar de kinderen daadwerkelijk zaten.
Verder is er vanwege het enorme aantal missings op de VVE variabele[26] de keuze gemaakt om grotendeels, maar niet uitsluitend, kinderen die een VVE-programma hebben gevolgd mee te nemen voor de vergelijking.

Implicaties voor de praktijk

De gevonden effecten op taal, rekenen en selectieve aandacht hangen mogelijk samen met de hoge proceskwaliteit op pedagogisch gebied (emotionele regulatie en groepsorganisatie). Een hoge proceskwaliteit op pedagogisch gebied – waarbij warme, ondersteunende interacties binnen een gestructureerde setting plaats vinden – kan de zelfregulatie van kinderen stimuleren en zo een belangrijke bijdrage leveren aan later schoolsucces (Blair, & Diamond, 2008). Daar staat tegenover dat de effecten zich voordoen ondanks een relatief lage proceskwaliteit op educatief gebied en de beperkte implementatie van de kenmerken van startgroepen op een derde van de startgroepen (Oberon, 2015). De effecten zouden bovendien mogelijk groter zijn geweest wanneer er een vergelijking gemaakt zou worden met kinderen die niet op een voorschoolse voorziening zaten. Het is ook belangrijk om in ogenschouw te nemen dat de kinderen na de startgroep ingestroomd zijn in het basisonderwijs. De proceskwaliteit op de basisscholen hebben we niet kunnen meten, maar is mede van invloed geweest op de ontwikkeling van taal, rekenen en selectieve aandacht op de langere termijn.
Uit deze beschouwingen volgen een aantal belangrijke implicaties voor de praktijk.

[26] De VVE variabele verwijst naar de informatie uit de vragenlijsten van het Pre-COOL onderzoek (verkregen uit D.A.N.S.) waaruit afgeleid is of voorschoolse voorzieningen ten tijde van het onderzoek met een VVE-programma werkten.

Een eerste implicatie betreft de ruimte voor verbetering op het educatieve domein. De aanstelling van een onderwijsbevoegde begeleider leidt niet één op één tot een continue hoge educatieve kwaliteit op de startgroepen. Echter, de ingrediënten van de startgroepen tezamen (meer uren, opbrengstgericht werken, hbo-er op de groep, regie schoolleider basisschool, doorgaande lijn VVE van peutergroep naar kleutergroep) lijken wel tot een klein effect op de ontwikkeling van taal, rekenen en selectieve aandacht te leiden. Het verhogen van de educatieve kwaliteit op de groepen zou mogelijk tot grotere effecten op de ontwikkeling (op de genoemde domeinen) bij kinderen kunnen leiden (Howes et al., 2008; Sylva et al., 2011). Het is dan ook van belang dat er niet alleen aandacht is voor structurele kenmerken (hbo- versus mbo geschoolde beroepskracht), maar dat er tevens continu gewerkt wordt aan het verhogen van de proceskwaliteit, voornamelijk op het educatieve domein. Voor meer informatie verwijzen wij naar het rapport van het Pre-COOL consortium (2016) waar (onder andere) op dit punt uitgebreide toelichting gegeven wordt.

Hetzelfde geldt voor de aanbod en de kwaliteit van het onderwijs op de basisschool. Het is van belang dat er voldoende aandacht wordt besteed aan vroege vaardigheden die gerelateerd zijn aan rekenen.[27] Ook dient de proceskwaliteit op het educatieve domein op orde te zijn: stimuleren van de taalontwikkeling (o.a. het stellen van open vragen, het gebruik van een rijk en gevarieerd arsenaal aan woorden), actief faciliteren en stimuleren van leren en ontwikkeling (o.a. kinderen cognitieve uitdagingen bieden) en kwalitatieve feedback (o.a. scaffolding: aansluiten bij wat kinderen al weten en dit uitbreiden). Verder biedt de kleine groep (de Haan et al., 2011) meer mogelijkheden om aan te sluiten bij individuele kinderen. In een dergelijke situatie is het dan met name van belang dat de vaardigheden van de leerkracht op het gebied van groepsmanagement van voldoende niveau zijn.

Een tweede implicatie betreft het beleid omtrent het invoeren van het idee van de startgroepen in de dagelijkse praktijk. Het is van belang dat vóór de start van een ‘startgroep’ de basiskenmerken op orde zijn: de schoolleider dient nauw betrokken te zijn bij de startgroep zodat de regierol goed uitgevoerd kan worden, het VVE-programma in de startgroep dient hetzelfde programma te zijn als het programma waarmee in de kleutergroepen wordt gewerkt en de beroepskrachten dienen getraind te zijn in opbrengstgericht werken en het werken met het VVE-programma. Daarnaast is het van belang om aandacht te besteden aan de match tussen de hbo-er en de mbo-er. Ook het beleid met betrekking tot ouderbetrokkenheid – een

[27] Dit dient wel op een voor de leeftijd van de kinderen passende wijze te geschieden, zoals in de vorm van begeleide spelactiviteiten.

belangrijke component in VVE programma's – dient vóór de start van een startgroep duidelijk geformuleerd te zijn (Oberon, 2015).
Een laatste implicatie van dit onderzoek heeft betrekking op de discussie rondom het testen en toetsen bij jonge kinderen. Zoals beschreven in de instrumentensectie, en in de bijlage (tabellen 13 tot en met 23), is het mogelijk om met goed getrainde testleiders en door maatwerk (individuele afnames) een vaardigheid op het gebied van taal, rekenen en selectieve aandacht – op een voor jonge kinderen prettige wijze – betrouwbaar te meten. Bovendien laat dit onderzoek zien dat het heel belangrijk is om achterstanden op basis van testen zo vroeg mogelijk vast te stellen. De verkregen testscores van deze vroege afnames blijken namelijk in grotere mate voorspellend te zijn voor latere vaardigheid (betrouwbaarder) dan de opleiding of thuistaal: de variabelen waarop in het huidige systeem vastgesteld wordt of een kind in aanmerking komt voor VVE. Ook blijkt uit dit onderzoek dat het afnemen van toetsen of testen bij jonge kinderen leer- en ontwikkelingsstagnatie bij kinderen kan voorkomen, omdat achterstanden vroegtijdig op een betrouwbare en valide manier vastgesteld kunnen worden. Belangrijker nog, op jonge leeftijd kan er een goed afgestemd en beproefd (evidence-based) aanbod gegeven worden, waarmee de ontwikkeling van kinderen, die er het meest behoefte aan hebben, zo goed mogelijk gestimuleerd kan worden. Voorwaarde is wel dat het testen op een goede manier wordt gedaan, namelijk met valide taken die leuk zijn voor kinderen en waarbij ze veel complimentjes ontvangen. Een tweede voorwaarde is dat de gegevens worden gebruikt om vervolgstappen te bepalen en niet om een stempel (slim, minder slim) op een kind te plakken. Het afnemen van toetsen is geen eindstation, maar dient juist gezien te worden als een diagnostische werkwijze om een passend aanbod naar het kind toe te realiseren.

Conclusie

Dit onderzoek laat korte- en lange termijn effecten van startgroepen zien ten opzichte van reguliere voorschoolse voorzieningen. Het betreft effecten op rekenen, taal en selectieve aandacht, onderdelen die belangrijk zijn voor de schoolloopbaan. Voor de ontwikkeling van doelgroepkinderen – kinderen met een risico op vroege en blijvende leer- en ontwikkelingsachterstand – lijkt vroeg interveniëren (startgroepen) en het continu streven naar hoge proceskwaliteit van belang te zijn. Hoewel binnen de startgroepen de proceskwaliteit op pedagogisch gebied (emotionele regulatie en groepsorganisatie) redelijk hoog is, is er op het educatieve domein nog ruimte voor verbetering. Om de startgroepen een succesvol vervolg te kunnen geven, is het bovendien belangrijk dat de basiselementen van de startgroepen op orde

zijn, zodat direct gestart kan worden met het werken volgens het VVE-programma en het opbrengstgericht werken. De verwachting is dat deze manier van werken ten goede komt aan de proceskwaliteit en zo de ontwikkelingskansen van doelgroepkinderen worden vergroot.

Referenties

Blair, C. & Diamond, A. (2008). Biological processes in prevention and intervention: The promotion of self-regulation as a means of preventing school failure, *Development and Psychopathology,* 20, 899–911.

Blair, C. & Razza, R.P. (2007). Relating Effortful Control, Executive Function, and False Belief Understanding to Emerging Math and Literacy Ability in Kindergarten, *Child Development, 78:2,* 647-663.

Broekhuizen, M.L. (2015). Differential effects of early child care quality on children's socio-emotional development (Doctoral thesis, Universiteit Utrecht). Ridderkerk: Ridderprint BV.

Cohen, J. (1992). A Power Primer, *Psychological Bulletin, 112*(1), 155-159.

Dunn, M., & Dunn, L. M. (2005). *Peabody Picture Vocabulary Test-III-NL*. Amsterdam: Harcourt Test Publishers.

Garon, N., Bryson, S.E., & Smith, I.M. (2008). Executive Function in Preschoolers: A Review Using an Integrative Framework, *Psychological Bulletin, 134: 1,* 31-60.

Gerhardstein, P., & Rovee-Collier, C. (2002). The Development of Visual Search in Infants and Very Young Children, *Journal of Experimental Child Psychology, 81,* 194-215.

Haan, A.K.E. de & van Tuijl, C. (2011). Besteding onderwijstijd voor- en vroegschool. De leidster/leerkracht doet ertoe. *De Wereld van het Jonge Kind*, februari/2, 22-25.

Haan, A. de, Leseman, P., & Elbers, E. (2011). Pilot gemengde groepen 2007-2010, Onderzoeksrapportage oktober 2011. Utrecht: Universiteit Utrecht.

Haan, A. de, Elbers, E., Hoofs, H., & Leseman, P. (2012). Targeted versus mixed preschools and kindergartens: effects of class composition and teacher-managed activities on disadvantaged children's emergent academic skills, *School Effectiveness and School Improvement: An International Journal of Research, Policy and Practice,* DOI:10.1080/09243453.2012.749792.

Howes, C., Burchinal, M., Pianta, R., Bryant, D., Early, D., Clifford, R., & Barbarin, O. (2008).

Ready to learn? Children's pre-academic achievement in pre-Kindergarten programs, *Early Childhood Research Quarterly,* 23, 27–50.

Koerhuis, I. (2010). Rekenen voor kleuters. Arnhem: Cito.

Lansink, N. (2009). Taal voor kleuters. Arnhem: Cito.

Le Paro, M., & Hamre, B.K., & Pianta, R.C. (2011). Classroom Assessment Scoring System. Toddler Manual. Teachstone, Charlottsville.

McClelland, M. M., Cameron, C.E., McDonald Connor, C., Farris, C.L., Jewkes, A.M., & Morrison, F.J. (2007). Links between behavioral regulation and preschoolers' literacy, vocabulary, and math skills. *Developmental Psychology*. 43(4), 947–959.

Mulder, H., Hoofs, H., Verhagen, J., van der Veen, I., & Leseman, P.P.M. (2014). Psychometric properties and convergent and predictive validity of an executive function test battery for two-year-olds, *Frontiers in Psychology, 5,* 733-768.

Oberon (2015). Implementatie-onderzoek Startgroepen peuters. Eindrapportage: verslag van de derde en laatste meting. Utrecht: Oberon.

Onderwijsraad (2008). Een rijk programma voor ieder kind. Den Haag: Onderwijsraad.

Onderwijsraad (2010). Naar een nieuwe kleuterperiode in de basisschool. Den Haag: Onderwijsraad.

Op den Kamp, M. (2010). Rekenen voor peuters. Arnhem: Cito.

Pre-COOL consortium. (2016). Pre-COOL cohortonderzoek. Effecten van kwaliteit van voorschoolse instellingen. Amsterdam: Kohnstamm Instituut.

Rosenbaum, P.R., & Rubin, D.B. (1983). The central role of propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.

Rubin D.B., & Thomas, N. (1996). Matching using estimated propensity scores: relating theory to practice. *Biometrics*, 52, 249-264.

Scerif, G., Cornish, K., Wilding, J., Driver, J., and Karmiloff-Smith, A. (2004). Visual search in typically developing toddlers and toddlers with Fragile X or Williams syndrome. *Developmental Science, 7*(1), 116-130.

Shonkoff, J. P. (2010). Building a new biodevelopmental framework to guide the future of early childhood policy. *Child Development, 81*(1), 357–367. doi:10.1111/j.1467-8624.2009.01399.

Slot, P.L. (2014). Early Childhood Education and Care in the Netherlands. Quality, Curriculum, and Relations with Child Development (Doctoral thesis, Universiteit Utrecht). Ridderkerk: Ridderprint BV.

Stürmer, T., Joshi, M., Glynn, R.J., Avorn, J., Rothman, K.J., & Schneeweiss, S. (2006). A

review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *Journal of Clinical Epidemiology*, 59(5), 437-447. doi: 10.1016/j.jclinepi.2005.07.004.

Sylva, K.S, Sammons, P., Siraj-Blatchford, I., Taggart, B. (2008). Assessing quality in early years. Early Childhood Rating Scale Extension (ECERS-E) (four curricular subscales)– revised edition. Stoke on Trent: Trentham Books.

Sylva, K., Melhuish, E., Sammons, P., Siraj-Blatchford, I., & Taggart, B. (2011). Pre-school quality and educational outcomes at age 11: Low quality has little benefit, *Journal of Early Childhood Research*, 9(2), 109-124.

Veen, A., van der Veen, I., Heurter, A. M. H., Ledoux, G., Mulder, L., Paas, T., Leseman, P., Mulder, H., Verhagen, J., & Slot, P. (2012). Pre-COOL cohortonderzoek. Technisch rapport tweejarigencohort, eerste meting 2010/2011. Amsterdam: Kohnstamm Instituut.

Veen, A., van der Veen, I., Heurter, A.M.H., Paas, T., m.m.v. Karssen, A.M. (2013). Pre-COOL cohortonderzoek. Technisch rapport vierjarigencohort, tweede en derde meting, 2010-2011 en 2011-2012. Amsterdam: Kohnstamm Instituut.

Veen, A., van der Veen, I., Heurter, A. M. H., Ledoux, G., Mulder, L., Paas, T., Leseman, P., Mulder, H., Verhagen, J., & Slot, P. (2014). Pre-COOL cohortonderzoek. Technisch rapport tweejarigencohort, tweede meting 2011/2012. Amsterdam: Kohnstamm Instituut.

Veen, A., & Leseman, P. (2015). Pre-COOL cohortonderzoek. Resultaten over de voorschoolse periode. Amsterdam: Kohnstamm Instituut.

Veer, I.M., Luyten, J.W., Mulder, H., van Tuijl, C., & Sleegers, P.J.C. (2015). *Early development of executive functions in 2 to 3 year olds: a longitudinal study.* In: EARLI: 16th European conference for Research on Learning and Instruction, 25-08-2015 - 29-08-2015, Limassol, Cyprus.

Verhagen, J., Mulder, H., & Leseman, P.P.M. (2015). Effects of home language environment on Inhibitory control in bilingual three-year-old children, *Bilingualism: Language and Cognition*, DOI:10.1017/S1366728915000590.

Bijlage

Tabel 5 Beschrijvende gegevens respondenten startgroepen (N=306)

		Frequentie	*Percentage*
Geslacht	Man	144	47%
	Vrouw	162	53%
Thuistaal	alleen Nederlands	135	44%
	Nederlands en anders	66	22%
	Alleen anders	105	34%
Doelgroepkinderen (volgens beroepskrachten)		227	74%
Geboorteland niet-Westers (één of beide ouders)		168	55%
Netto inkomen hoofdverdiener	€ 0-1000 p. mnd	45	18%
	€1000-€1500 p. mnd	95	38%
	> €1500 p. mnd	109	44%
	Onbekend	57	n.v.t.
Opleiding moeder	Lager onderwijs	43	15%
	LBO	47	16%
	MBO of Alg. voortg. onderw	137	48%
	Hoger onderwijs	59	21%
	Onbekend	20	n.v.t.
Opleiding ouders (hoogste)	Lager onderwijs	26	9%
	LBO	44	15%
	MBO of Alg. voortg. onderw.	141	49%
	Hoger onderwijs	80	28%
	Onbekend	15	n.v.t.
Doorgaande lijn	Ja	200	74%

(Startgroep basisschool)	Nee	69	26%
	n.v.t. (uitgevallen)	37	n.v.t.

Tabel 6 Meetinstrumenten individuele testafnames startgroepen en respons

Soort test: vaardigheid	**Respons MM1** (2 jr)		**Respons MM2** (3jr)		**Respons MM3** (4jr)		**Respons MM4** (5 jr)	
	N	**%**	**N**	**%**	**N**	**%**	**N**	**%**
	N tot. = 307		N tot. = 284		N tot. = 271		N tot. = 128	
Taal								
Peabody Picture Vocabulary Task (PPVT-III-NL): receptieve woordenschat (Verhagen, Mulder, & Leseman, 2015). Gebaseerd op werk van Dunn & Dunn (2005).	288	94%	280	99%	270	100%	127	99%
Rekenen								
CITO rekenen voor peuters: getalbegrip, meten en meetkunde (Op den Kamp, 2010).	272	89%	274	97%	n.v.t.		n.v.t.	
Aandacht								
Aandachtstaak: selectieve aandacht (Mulder et al., 2014). Gebaseerd op werk van: Gerhardstein & Rovee-Collier (2002); Scerif, Cornish, Wilding, Driver, & Karmiloff-Smith (2004).	286	93%	282	99%	267	99%	127	99%

Tabel 7 Respons op CITO-toetsen groep 1

	CITO Taal voor kleuters M1		CITO Taal voor kleuters E1		CITO rekenen voor kleuters M1		CITO rekenen voor kleuters E1	
	N	Perc.	N	Perc.	N	Perc.	N	Perc.
Startgroepen								
Afgenomen	102	38%	151	56%	101	38%	157	58%
Niet afgenomen	108	40%	53	23%	108	40%	50	19%
Wordt in 2016 afgenomen	55	20%	61	20%	56	21%	58	22%
Non-respons	4	2%	4	2%	4	2%	4	2%

Tabel 8 Respons op oudervragenlijst

Startgroep	**Respons Ouder-vragenlijst**	**Startgroep**	**Respons Ouder-vragenlijst**
Amsterdam W	100%	Nijmegen	100%

Amsterdam ZO	91%	Tilburg	100%
Den-Haag NO	100%	Assen	100%
Den-Haag ZW	93%	Beverwijk	100%
Rotterdam	100%	Borssele	100%
Utrecht N	90%	Den-Bosch	92%
Utrecht ZW	91%	Dordrecht	92%
Almere Buiten	100%	Koewacht	100%
Almere Haven	100%	Nieuwegein	100%
Amersfoort	100%	Ruurlo	80%
Apeldoorn	88%	Ter Aar	100%
Arnhem	100%	Twello	100%
Delft	100%	Veendam	100%
Emmen	100%	Wijchen	100%
Leiden	71%		
		Totaal	**96%**

Tabel 9 Tijdspad individuele testafnames

Tijdstip	Meetmoment	Cohort	Testleider
Najaar 2012	1	1	Eerste auteur
Voor- en najaar 2013	2	1	Eerste auteur
Najaar 2013	1	2	Eerste auteur
Voor- en najaar 2014	2	2	Eerste auteur
	3	1	Eerste auteur, OA1
Voor- en najaar 2015	3	2	Eerste auteur, OA2, OA3
	4	1	Eerste auteur, OA2, OA3

OA: onderzoeksassistent.

Tabel 10 Proceskwaliteit startgroepen (observaties meetmoment 2)

CLASS-toddler *1=laag* *4=midden* *7=hoog*	**N**	**Min.**	**Max.**	**Gem** **.**	**sd**	**Aantal keer LK sterker**	**Aantal keer PM sterker**
Emotionele regulatie				**5.42**	**.45**		
Positieve sfeer	30	3	7	5.37	.56	15	3

Negatieve sfeer *(omgescoord)*	30	3	7	6.79	.38	3	3*
Sensitiviteit	30	3	7	5.27	.66	9	4
Aandacht voor kindperspectief	30	2	7	4.18	.62	14	3
Groepsorganisatie				**5.04**	**.63**		
Gedragsregulering	30	3	7	5.04	.63	20	6
Educatieve ondersteuning				**2.87**	**.54**		
Faciliteren leren & ontwikkeling	30	1	6	3.03	.57	37	6
Kwaliteit feedback	30	1	6	2.28	.65	35	9
Stimuleren taalontwikkeling	30	1	6	3.11	.50	34	7
ECERS-E (*1=onvoldoende, 3=minimaal, 5=goed, 7=uitstekend)*						Aantal groepen score 4	Aantal groepen score 5 of hoger
Geletterdheid				**2.67**	**0.47**		
Aanwezigheid geschreven letters en woorden	30	1	5	2.20	1.00	1	1
Boeken en leeshoek	30	1	6	2.50	1.38	3	3
Voorlezen	30	1	5	2.77	1.41	4	5
Klanken in woorden	30	1	3	2.43	0.57	0	0
Ontluikende geletterdheid	30	1	4	1.90	1.03	3	0
Praten en luisteren	30	1	7	4.07	1.70	14	9
Rekenen				**1.88**	**0.77**		
Tellen en het toepassen van tellen	30	1	7	2.43	1.28	4	1
Lezen en schrijven van eenvoudige cijfers	30	1	2	1.03	0.18	0	0
Vormen en ruimte**	20	1	5	1.60	1.23	2	1
Sorteren, vergelijken en matchen**	9	2	6	3.56	1.42	3	2

* Sterker zijn op negatieve sfeer betekent het meest bepalend zijn voor de score op dit item.

** Alleen het item dat het meest zichtbaar was tijdens de observatie is gescoord door de observator.

Tabel 11 Beschrijvende gegevens startgroepen

		Frequentie	*Percentage*
Uren per week (Oberon)	12,5	17	57%
	13 – 15	13	43%
Middaguren (per week)	0	14	47%
	2 – 5,5	16	53%

Stagiaire (dagdelen per week)		0	9	30%
		1 – 2	12	40%
		3 – 4	7	23%
		5	2	7%
Groepssamenstelling	(bijna) alleen doelgroepkinderen		15	50%
	Gemengde groep		5	17%
	(bijna) alleen niet-doelgroep kinderen		10	33%
Aantal kinderen per groep *		10-13	4	14%
		14-16	24	80%
		missing	2	6%
Sociaal economische groepskenmerken:				
Thuistaal niet Nederlands		0 %	8	27%
		7 – 20%	7	23%
		21 – 40%	11	37%
		41 – 86%	4	13%
Leerlinggewicht (0.3 of 1.2)		5 – 20%	7	23%
		21 – 40%	7	23%
		41 – 60%	10	33%
		61 – 80%	4	13%
		> 80%	2	7%
Pedagogische ondersteuning op de werkvloer (gemiddeld en daarna afgerond)		Enkele keren per jaar	10	33%
		Maandelijks	16	53%
		Wekelijks	4	13%
Splitsen van de groep (vragenlijst beroepskrachten)		Nooit/soms	12	40%
		Vaak/altijd	18	60%
CLASS-toddler emotionele regulatie (max.) **		4,25 – 5,75	11	37%
		6 - 7	19	63%
CLASS-toddler educatieve ondersteuning (max.) **		2,33 – 3,67	19	63%
		4	3	10%

	4,33 – 5,33	8	27%
Educatief partnerschap (Oberon)			
Thuisopdrachten voor ouders	Ja	21	70%
Individuele oudergesprekken (10 min. gespr.)	Ja	24	80%
Thematische groepsbijeenkomsten voor ouders	Ja	28	93%
Ouderparticipatie (Oberon): koffie-ochtenden,	Nee	1	3%
deelname activiteiten op de groep	Ja	29	97%
Effectief leiderschap (Oberon)			
Schoolleider is inhoudelijk betrokken	Beperkt	6	20%
	Enige mate	10	33%
	Sterk	14	47%
Faciliterend leiderschap	Beperkt	2	6%
	Enige mate	7	23%
	Sterk	21	70%
Delegerend leiderschap	Beperkt	1	3%
	Enige mate	8	27%
	Sterk	21	70%
Democratisch leiderschap	Beperkt	2	7%
	Enige mate	9	30%
	Sterk	19	63%
Aantal wisselingen in schoolleider	Geen	25	83%
	eenmalig	4	13%
	meermalig	1	3%
Opbrengstgericht werken (Oberon)			
Gegevens verzamelen (leerlingvolgsysteem)	Deels	7	23%
	Ja	23	77%
Doelen formuleren	Deels	6	20%
	Ja	24	80%
Groepsplannen opstellen	Deels	9	30%
	Ja	21	70%
Evalueren	Nee	1	3%
	Deels	11	37%
	Ja	17	57%

	missing	1	3%
Mate van implementatie opbrengstgericht	Deels	13	43%
werken	Ja	17	57%
Doorgaande lijn (Oberon)			
Leerkracht intern aangetrokken	Nee	12	40%
	Ja	18	60%
Zelfde VVE-programma in startgroep en groep	Nee	5	17%
1/2	Ja	25	83%
Overkoepelend succes implementatie	Ja	21	70%
(Oberon)	Deels / Nee	9	30%

* op het gemeten tijdstip.

****** Gemiddelde van de maximum scores op de dimensies: positieve sfeer, negatieve sfeer (omgescoord), sensitiviteit, aandacht voor kindperspectief.

******* Gemiddelde van de maximum scores op de dimensies: faciliteren van leren en ontwikkeling, kwaliteit feedback, stimuleren taalontwikkeling.

Tabel 12 Beschrijvende gegevens gematchte respondenten (PSM variabelen)

		Startgroepen (N=276)	**Pre-COOL** (gematchte en gewogen groep)
		Percentage	*Percentage*
Geslacht	Man	46%	41%
	Vrouw	54%	59%
Thuistaal	alleen Nederlands	44%	42%
	Nederlands en anders	21%	21%
	Alleen anders	34%	37%
Geboorteland moeder	Oude immigrant[28]	24%	32%
	Nieuwe immigrant[29]	17%	15%

[28] Oude immigrant is geboren in: Turkije, Marokko, Suriname, Antillen of Aruba
[29] Nieuwe immigrant is geboren in: Irak, Afghanistan, Somalië, of overig niet-Westers land.

Geboorteland vader	Oude immigrant	28%	32%
	Nieuwe immigrant	18%	17%
Eenoudergezin	Ja	17%	19%
	Nee	83%	81%
Belangrijkste inkomensbron van het huishouden	Loon	72%	78%
	Bijstand	18%	15%
	Arbeidsongeschiktheidsuitkering	6%	4%
	Overig geen loon	4%	3%
		Gem. (sd)	*Gem. (sd)*
Opleiding moeder[30]		3.47 (1.45)	3.46 (1.38)
Leeftijd mm2		3.26 (0.17)	3.27 (0.18)

Tabel 13 Receptieve woordenschat mm1* en mm3

Onafhankelijke variabelen	Bèta	t	Sig.	Afhankelijke variabele
Geslacht vrouw	.075	1.57	.117	*Receptieve woordenschat*
Geboorteland niet-Westers	.088	1.50	.134	*mm3*
Opleiding ouders (hoogste)	.039	.82	.412	
Thuistaal niet NL	-.312	-4.62	.000	
Leeftijd mm1	-.106	-1.97	.050	
Leeftijd mm3	.092	1.75	.081	
Receptieve woordenschat mm1	.485	7.78	.000	

* mm: meetmoment.

Tabel 14 Receptieve woordenschat mm2 en mm3

Onafhankelijke variabelen	Bèta	t	Sig.	Afhankelijke variabele
Geslacht vrouw	.110	2.35	.019	*Receptieve woordenschat*
Geboorteland niet-Westers	.057	.98	.328	*mm3*
Opleiding ouders (hoogste)	.049	1.05	.296	
Thuistaal niet NL	-.335	-5.14	.000	
Leeftijd mm2	.003	.08	.935	

[30] 1: geen opleiding – 6: WO en hoger. Indien onbekend: opleiding vader.

Leeftijd mm3	-.009	-.12	.906
Receptieve woordenschat mm2	.441	7.95	.000

Tabel 15 Receptieve woordenschat mm1 en mm4

Onafhankelijke variabelen	Bèta	t	Sig.	Afhankelijke variabele
Geslacht vrouw	.157	2.12	.036	*Receptieve woordenschat*
Geboorteland niet-Westers	-.119	-1.24	.217	*mm4*
Opleiding ouders (hoogste)	.116	1.57	.120	
Thuistaal niet NL	.146	1.24	.217	
Leeftijd mm1	.053	.53	.597	
Leeftijd mm4	.091	1.00	.319	
Receptieve woordenschat mm1	.620	5.81	.000	

Tabel 16 Receptieve woordenschat mm2 en mm4

Onafhankelijke variabelen	Bèta	t	Sig.	Afhankelijke variabele
Geslacht vrouw	.202	2.69	.008	*Receptieve woordenschat*
Geboorteland niet-Westers	-.065	-.65	.518	*mm4*
Opleiding ouders (hoogste)	.188	2.50	.014	
Thuistaal niet NL	.058	.51	.611	
Leeftijd mm2	-.022	-.36	.716	
Leeftijd mm4	.087	.84	.405	
Receptieve woordenschat mm2	.503	5.76	.000	

Tabel 17 Receptieve woordenschat mm1 en taal groep 1

Onafhankelijke variabelen	Bèta	t	Sig.	Afhankelijke variabele
Geslacht vrouw	.054	.83	.406	*CITO taal voor kleuters E1*
Geboorteland niet-Westers	-.122	-1.46	.148	
Opleiding ouders (hoogste)	.093	1.48	.142	
Thuistaal niet NL	-.024	-.25	.806	
Leeftijd mm1	-.169	-2.23	.028	
Leeftijd CITO Taal E1	.327	5.31	.000	
Receptieve woordenschat mm1	.587	6.67	.000	

Tabel 18 Receptieve woordenschat mm2 en taal groep 1

Onafhankelijke variabelen	Bèta	t	Sig.	Afhankelijke variabele
Geslacht vrouw	.067	1.03	.303	*CITO taal voor kleuters E1*
Geboorteland niet-Westers	-.140	-1.69	.094	
Opleiding ouders (hoogste)	.078	1.25	.212	
Thuistaal niet NL	-.114	-1.24	.218	
Leeftijd mm2	-.060	-1.83	.070	
Leeftijd CITO Taal E1	.286	4.55	.000	
Receptieve woordenschat mm2	.465	6.45	.000	

Tabel 19 Rekenen mm2 en rekenen groep 1

Onafhankelijke variabelen	Bèta	t	Sig.	Afhankelijke variabele
Geslacht vrouw	-.092	-1.40	.164	*CITO rekenen voor kleuters*
Geboorteland niet-Westers	-.201	-2.43	.017	*E1*
Opleiding ouders (hoogste)	.156	2.58	.011	
Thuistaal niet NL	.235	2.59	.011	
Leeftijd mm2	-.050	-1.54	.126	
PPVT mm2	.197	2.31	.023	
Leeftijd CITO Rekenen E1	.303	5.10	.000	
CITO rekenen voor peuters mm2	.439	5.13	.000	

Tabel 20 Selectieve aandacht mm1 en selectieve aandacht mm3

Onafhankelijke variabelen	Bèta	t	Sig.	Afhankelijke variabele
Geslacht vrouw	.152	2.54	.012	*Selectieve aandacht mm3*
Geboorteland niet-Westers	.001	.01	.994	
Opleiding ouders (hoogste)	.049	.82	.415	
Thuistaal niet NL	-.102	-1.34	.181	
Leeftijd mm1	-.008	-.11	.909	
Leeftijd mm3	.160	2.42	.016	
Selectieve aandacht mm1	.273	4.26	.000	

Tabel 21 Selectieve aandacht mm2 en selectieve aandacht mm3

Onafhankelijke variabelen	Bèta	t	Sig.	Afhankelijke variabele
Geslacht vrouw	.156	2.85	.005	*Selectieve aandacht mm3*
Geboorteland niet-Westers	-.006	-.09	.929	
Opleiding ouders (hoogste)	.061	1.13	.262	
Thuistaal niet NL	-.080	-1.13	.259	
Leeftijd mm2	.164	1.82	.071	
Leeftijd mm3	-.021	-.24	.813	
Selectieve aandacht mm2	.437	7.32	.000	

Tabel 22 Selectieve aandacht mm1 en selectieve aandacht mm4

Onafhankelijke variabelen	Bèta	t	Sig.	Afhankelijke variabele
Geslacht vrouw	.308	3.63	.000	*Selectieve aandacht mm4*
Geboorteland niet-Westers	.024	.21	.832	
Opleiding ouders (hoogste)	-.001	-.02	.988	
Thuistaal niet NL	-.077	-.66	.508	
Leeftijd mm1	.108	.90	.369	
Leeftijd mm4	.192	1.82	.072	
Selectieve aandacht mm1	.170	1.70	.092	

Tabel 23 Selectieve aandacht mm2 en selectieve aandacht mm4

Onafhankelijke variabelen	Bèta	t	Sig.	Afhankelijke variabele
Geslacht vrouw	.267	3.59	.001	*Selectieve aandacht mm4*
Geboorteland niet-Westers	.011	.11	.910	
Opleiding ouders (hoogste)	-.005	-.07	.948	
Thuistaal niet NL	-.074	-.72	.475	
Leeftijd mm2	-.072	-.61	.546	
Leeftijd mm4	.215	2.09	.039	
Selectieve aandacht mm2	.419	5.22	.000	