Appels en peren zijn allebei fruit

Reactie op André Aleman en Cathy van Tuijl, Meta-analyse, heterogeniteit, en de effecten van voorschoolse educatieve programma's: een kritische beschouwing

P.P.M. Leseman & H. Blok

Paul Leseman is verbonden aan bij de afdeling Pedagogische en Onderwijskundige Wetenschappen van de Universiteit van Amsterdam.

Henk Blok is onderzoeker bij het SCO-Kohnstamm instituut van de Universiteit van Amsterdam.

Correspondentieadres: Paul Leseman, Pedagogische en Onderwijskundige Wetenschappen, Universiteit van Amsterdam, Postbus 94208, 1090 GE Amsterdam.

De kritische beschouwing van Aleman en Van Tuijl (in dit nummer) over een meta-analyse van voorschoolse educatieve programma's van onze hand (Leseman, Otter, Blok & Deckers, 1998, 1999) leidt tot enkele slotconclusies die wij van harte kunnen onderschrijven. Dat geldt niet voor de met stelligheid gedebiteerde opinies over twee methodologische kwesties die aan statistische metaanalyses in het algemeen en aan die van ons in het bijzonder kleven: het probleem van de statistische afhankelijkheid van effecten die uit dezelfde studie voortkomen en de vraag hoe ernstig het is als de uiteindelijke verzameling effecten heterogeen is. Aleman en Van Tuijl nemen bij deze kwesties een rigide standpunt in en kritiseren vanuit die positie onze aanpak. Wij zullen hierna betogen dat de zeggingskracht van de statistische meta-analyse, als hulpmiddel om greep te krijgen op een grote hoeveelheid onderzoeksgegevens, niet gebaat is bij zo'n opstelling.

Hoe ernstig is statistische afhankelijkheid?

In de methodologische literatuur over meta-analyses staat minder vast dan Aleman en Van Tuijl suggereren. Zoals in vele andere professionele gemeenschappen vindt men onder de deskundigen op dit terrein rekkelijken en preciezen. Dat geldt bijvoorbeeld voor het standpunt over het gevaar van statistische afhankelijkheid van effecten als er meer effecten per studie worden geanalyseerd. Er zijn auteurs die heel ver gaan in het bestrijden van de afhankelijkheid van effecten. Zij suggereren dat eigenlijk ook effecten van hetzelfde programma maar in een andere populatie uitgevoerd, en effecten van verschillende studies maar van dezelfde onderzoeksgroep afhankelijk zijn en eigenlijk dus geïntegreerd zouden moeten worden tot één effect voordat ze in een meta-analyse opgenomen kunnen worden. Er zijn ook rekkelijken die, misschien met het idee in het achterhoofd dat de afhankelijkheidskwestie niet bevredigend op te lossen is, veel ruimere oplossingen voorstellen. Zo stellen Glass, McGaw en Smith (1981) de '...simple (but risky) solution...' voor '...to regard each finding as independent of the others. The assumption is untrue but practical' (p.200).

Het grote bezwaar van een heel precieze opstelling is dat er veel informatie verloren gaat. Vooral wanneer de onderzoeksvraag zich richt op differentiële effecten van bepaalde programma-variaties (zoals bijv. aanbiedingsduur, leeftijd waarop het programma start) of op interactie-effecten van programma en subgroepen, moet men idealiter juist kunnen beschikken over verschillende effecten die binnen één studie zijn gevonden. Daarmee is het statistische probleem - afhankelijkheid van de waarnemingen - niet van de baan. Toch kan misschien beter gezocht worden naar pragmatische

oplossingen zoals het achteraf corrigeren van het significantieniveau, en het betrachten van de nodige voorzichtigheid bij het trekken van conclusies en generaliseren van de bevindingen (Hedges & Olkin, 1985, p. 256). Ook multilevel meta-analyse technieken bieden een mogelijke oplossing. In een multilevel-analyse kan de afhankelijkheid van effecten expliciet onderzocht worden door in de analyse effecten onder studies te nesten.

Waar behoefte aan bestaat is onderzoek naar de mate van vertekening die afhankelijkheid van effecten oplevert. Wat dit laatste betreft zijn er enkele gegevens die de ernst van de validiteitsbedreiging die aan afhankelijkheid verbonden is relativeren. Hunter en Schmidt (1990) concluderen dat schending van de onafhankelijkheidseis geen systematische invloed heeft op de gemiddelde effectgrootte die in de meta-analyse wordt geschat. Wel leidt afhankelijkheid tot *onderschatting* van de fouten-variantie die het gevolg is van steekproeffluctuaties en daarmee tot overschatting van de systematische variantie van de geschatte populatie- effecten.
De *verschillen*tussen de populatie-effecten (heterogeniteit) worden dan eerder als systematisch dan als toevallig beschouwd. Het gevolg zou daarom kunnen zijn dat de alternatieve hypothese ('er is een effect') te snel wordt verworpen ten gunste van de nulhypothese (er is *geen* effect). De 'power' van de analyse is dus verminderd en afhankelijkheid zou dus een 'conservatieve' invloed hebben op de uiteindelijke conclusie.

De vertekenende invloed hangt ook af van het aantal afhankelijke effecten ten opzichte van het totaal aantal effecten. Wanneer het aantal onderling afhankelijke effecten klein is ten opzichte van het totale aantal effecten dat in de metaanalyse wordt betrokken, is er volgens Hunter en Schmidt weinig aan de hand. Een praktisch voorschrift zou kunnen zijn het aantal effecten van één studie te beperken, maar niet op voorhand tot één of enkele (Wolf, 1986).

Nog afgezien van het statistische afhankelijkheidsprobleem, blijft staan dat één studie soms tientallen effecten oplevert en een andere slechts één. Daardoor wordt het totaalbeeld, de gewogen gemiddelde effectgrootte die berekend wordt, te veel bepaald door deze ene studie, terwijl die studie, beter gezegd het programma dat geëvalueerd is, maar één greep is uit de verschillende mogelijkheden die er zijn, net als dat andere programma waarover slechts één effect werd gerapporteerd. Naar onze mening is hierbij primair een conceptuele benadering aangewezen.

Verschillende effecten van één studie kunnen weliswaar binnen hetzelfde, ruim gedefinieerde ontwikkelingsdomein vallen, maar toch duidelijk conceptueel te onderscheiden zijn. Effecten kunnen bijvoorbeeld betrekking hebben op zogenaamde *fluid abilities* die, zo is toch een goed ondersteund standpunt, minder beïnvloedbaar zijn door interventies, en *crystalized abilities* die dat juist meer zijn. In onze meta-analyses hebben we verondersteld dat uitkomstmaten zoals woordenschat, semantische begripskennis en (pre)geletterdheid een andere ontwikkelingsgeschiedenis kennen en meer door de culturele en sociolinguïstische context beïnvloed worden dan performale intelligentie en basaal rekeninzicht. Om die reden zijn van verschillende studies meer effecten in onze meta-analyses opgenomen, betrekking hebbend op deze conceptueel onderscheiden uitkomstmaten. Achteraf kan uiteraard blijken dat de gemiddelde effectgrootten toch niet wezenlijk verschillen.

Zoals al gesteld kunnen, eveneens vanuit een conceptueel kader, interessante programmavariaties geïdentificeerd worden en eveneens rechtvaardigen dat meer effecten van één studie worden

opgenomen. Dit geldt naar onze mening ook voor de kwestie of effecten onmiddellijk na afloop van een programma dan wel op langere termijn in een follow-up studie zijn gemeten. Gelet op de relevantie van de vraag naar lange-termijneffecten is het gerechtvaardigd van een studie die korte zowel als lange-termijneffecten rapporteert, beide effecten op te nemen.

Wat de studie van Campbell en Ramey (1994) betreft naar het Abecedarian Project aarzelen wij. Door onze selectie te richten op studies die na 1984 zijn gepubliceerd, kwamen wij uit bij het artikel uit 1994, waarin sprake is van drie experimentele groepen, die te beschouwen zijn als drie onafhankelijke, aselect gevormde steekproeven uit dezelfde populatie, naast één aselecte no-treatment controlegroep. In onze eerste meta-analyse beperkten we ons niet tot de strikt voorschoolse periode en bleek de Abecedarianstudie interessant omdat met deze drie experimentele groepen verschillende vormen van continuering (programmaverlenging, intensiteitsvergroting) werden getoetst. Onze veronderstelling was dat de afhankelijkheid van de drie vergelijkingen die per ontwikkelingsdomein (respectievelijk intelligentie en taal-geletterdheid) gemaakt werden met steeds dezelfde controlegroep geen ernstige bedreiging vormde voor de statische conclusievaliditeit en eerder een 'conservatieve' dan hypothese-bevestigende invloed zou hebben. Per ontwikkelingsdomein werden uiteindelijk maar *drie* (min of meer) afhankelijke effecten in onze analyses opgenomen. Het bezwaar van Aleman en Van Tuijl tegen het opnemen van meer effecten van deze studie is technisch gezien, vanuit een precies standpunt, correct. Of het tot ernstige vertekeningen heeft geleid, betwijfelen wij.

Het verbaast ons trouwens dat de auteurs zelf zonder blikken of blozen nog in hetzelfde artikel zondigen tegen de door hen verdedigde strikte regels. In Tabel 1 treffen we namelijk meteen al een *d*-score en bijbehorende betrouwbaarheidsindicatoren aan van de*samengevoegde* domeinen Intelligentie en Taal/geletterdheid, gebaseerd op *8* effecten, waarvan er - inderdaad! - twee uit dezelfde studie, die van Campbell en Ramey, voortkomen.

Hoe ernstig is heterogeniteit?

De vraag hoe zwaar heterogeniteit moet wegen is een andere bron van meningsverschil tussen de experts. Strikt genomen moeten effecten volstrekt homogeen zijn (dus niet meer dan op toevallige gronden van elkaar afwijken) om zinvol en statistisch betrouwbaar een gemiddelde effectgrootte te kunnen berekenen; *appels* en *peren* kun je niet optellen, luidt de volkswijsheid. Maar, om die analogie door te trekken, op het terrein van voorschoolse educatie gaat het niet meteen om de vraag of appels effectief zijn of peren, maar of *fruit* gezond is en of er eventueel kenmerken te verzinnen zijn (bijv. het appelachtig zijn, het vitamine C-gehalte, de kleur) die met de gezondheid van fruit samenhangen. Op het gebied van voorschoolse educatie bestaat evident nog geen eenvoudige receptuur, zo die er al ooit zal komen; meta-analyses zoals de onze kunnen bijdragen aan een scherper inzicht in de kenmerken waarom het gaat en kunnen ook tot zinvol afgebakende nieuwe meta-analyses leiden.

Voorschoolse programma's zijn georganiseerde omvattende en intern complexe sociale praktijken van mensen, afgestemd op specifieke culturele contexten en doelgroepen. Streven naar homogeniteit draagt het gevaar in zich dat van tevoren al zoveel van die variabele, heterogene sociale praktijken worden uitgesloten, dat er een heel specifieke selectie resteert. De gemiddelde effectgrootte die berekend kan worden mag dan te vertrouwen zijn binnen de aangegeven marges

omdat de homogeniteit groot is, het resultaat is nog maar zeer beperkt te generaliseren - niet alleen omdat er doorgaans weinig studies overblijven, maar ook en vooral omdat deze studies programma's evalueren die heel nauw overeenkomen in opzet, meetinstrumenten, context en doelgroep. De moeilijkheid is een goede balans te zoeken tussen zeggingskracht en precisie, tussen externe validiteit en statistische conclusievaliditeit. Welke balans precies gevonden wordt hangt in laatste instantie af van de onderzoeksvraag.

Het is in dit licht niet zo verwonderlijk, en naar onze mening niet verwerpelijk, dat in verschillende recente meta-analyses op verwante terreinen (bijv. voorlezen door ouders thuis; cf. Bus, Van IJzendoorn & Pellegrini, 1995) het met de eis van homogeniteit niet zo nauw wordt genomen. Onze interesse was niet alleen hoe groot de gemiddelde *d*-score van een bepaalde specifieke voorschoolse aanpak in de populatie van voorschoolse programma's is, maar evenzeer of we effectiviteitskenmerken konden identificeren (zoals periode van uitvoering, intensiteit, globale pedagogisch-didactische aanpak, brede dan wel smalle doelen enz.). Heterogeniteit van programma's, contexten en doelgroepen is voor dit doel geen obstakel, maar juist informatief. Het kan uiteraard blijken dat zelfs bij een vrij heterogene selectie van studies, de effecten toch homogeen zijn. Dit is wat in onze tweede meta-analyse bleek: beperking tot strikt voorschoolse programma's levert binnen de beschikbare studies - die onderling nog steeds verschilden - een voldoende homogene subset op (los van programmavariaties) met gemiddeld grotere effecten in het cognitieve en talig-geletterde domein dan wanneer de hele range van programma's werd geanalyseerd. Dit bevestigde de bevinding in de eerste meta-analyse dat timing een belangrijk programmakenmerk is: beginnen met het programma in de voorschoolse periode, wanneer het kind 2 à 3 jaar is, is waarschijnlijk effectiever dan veel later of veel eerder.

Er moet natuurlijk afgebakend worden, van tevoren al. Die afbakening moet primair gestuurd zijn door de onderzoeksvraag en natuurlijk is niet elke vraag te beantwoorden met de beschikbare gegevens. Vanwege onze interesse in voor- en vroegschoolse strategieën van voorkómen van onderwijsachterstand van kinderen in sociaal-economische achterstandssituaties hebben wij gekozen voor een ruime definitie van voorschoolse programma's. De beperking hierbij tot *centerbased* programma's werd ingegeven door het geringe aantal geschikte studies met *home-based* programma's dat we konden vinden, en door onze verwachting dat bij home-based programma's andere kenmerken van effectiviteit relevant zouden zijn, zodat beide benaderingen niet zomaar op één hoop geveegd konden worden. Binnen deze beperking mochten, gelet op onze onderzoeksvraag, juist vele varianten meedoen.

De studie van Burchinal, Lee en Ramey (1989) beschrijft een kinderopvang (daycare) programma voor kinderen in achterstandssituaties, in de woorden van de auteurs '...children determined to be 'at risk' for school failure due to socioeconomic factors...' (p.129), en rapporteert effecten in het cognitieve domein. Het programma voldoet om die redenen geheel en al aan de definitie, en mag dus *niet*worden uitgesloten. Andere kinderopvang-programma's die we in de literatuur tegenkwamen hadden echter vrijwel altijd betrekking op een normale, gemiddelde doelgroep of rapporteerden niet de gewenste uitkomstmaten.

Slot

De conclusie van Aleman en Van Tuijl, dat er meer (gedegen) studies nodig zijn alvorens we met meer zekerheid kunnen concluderen dat voorschoolse programma's effectief zijn, is een open deur. Dat wij zouden menen dat de kwestie van de effectiviteit bevredigend is opgelost, is onjuist. Elders hebben wij met klem gewezen op het belang van studies in een Nederlandse context, alvorens we met enige zekerheid iets over mogelijk succes van de voorschoolse aanpak *in Nederland* zouden kunnen concluderen (Blok & Leseman, 1996). Het is wel juist dat we menen van onze meta-analyses iets geleerd te hebben, en dat we onze eerdere veronderstellingen over effectieve kenmerken verder hebben kunnen aanscherpen, ten nutte van nieuwe programma's en meta-analyses.

Literatuur

Blok, H. & Leseman, P.P.M. (1996). Effecten van voorschoolse stimuleringsprogramma's: een review van reviews. *Pedagogische Studiën, 73*, 184-197.

Burchinal, Lee & Ramey, C.T. (1989). Type of day-care and preschool intellectual development in disadvantaged children. *Child Development, 60,* 128-137.

Bus, A.G., IJzendoorn, M.H. van & Pellegrini, A.D. (1995). Joint book reading makes for success in learning to read. A meta-analysis on intergenerational transmission of literacy. *Review of Educational Research, 65,* 1-21.

Campbell, F.A. & Ramey, C.T. (1994). Effects of early intervention on intellectual and academic achievement: a follow-up study of children from low-income families. *Child Development, 65,* 684-698.

Glass, G.V., McGaw, B. & Smith, M.L. (1981). *Meta-analysis in social research*. Beverly Hills, California: Sage.

Hedges, L.V. & Olkin, I. (1985). Statistical methods for meta-analysis. Orlando, Florida: Academic Press.

Hunter, J.E. & Schmidt, F.L. (1990). *Methods of meta-analysis. Correcting error and bias in research findings*. Newbury Park: Sage.

Leseman, P.P.M., Otter, M.E., Blok, H. & Deckers, P. (1998). Effecten van voor- en vroegschoolse stimuleringsprogramma's: een meta-analyse van evaluatiestudies 1985-1996. *Nederlands Tijdschrift voor Opvoeding, Vorming en Onderwijs, 14,* 134-154.

Leseman, P.P.M., Otter, M.E., Blok, H. & Deckers, P. (1999). Effecten van voorschoolse educatieve centrumprogramma's: Een aanvullende meta-analyse van studies 1985-1996. *Nederlands Tijdschrift voor Opvoeding, Vorming, en Onderwijs, 15,* 28-37.

Wolf, F.M. (1986). *Meta-analysis, quantitative methods for research synthesis*. Beverly Hills, California: Sage.