GEON: Summatieve evaluatie (I)*

K. M. STOKKING
Instituut voor Pedagogische en Andragogische Wetenschappen, Rijksuniversiteit Utrecht

Samenvatting

Dit artikel is het laatste uit een reeks van vijf over het GEON-project (1973-1979) en de evaluatie ervan. Het behandelt de summatieve evaluatie op basis van testafnames bij kinderen. De resultaten op de WPPSI, afgenomen aan het begin en aan het einde van de kleuterschool, zijn positief, ook indien vergeleken met gegevens van scholen waar wel getest werd, maar het project (nog) niet werd uitgevoerd. Testafnames in de periode nadat scholen intensief begeleid werden tonen resultaten die wijzen op een zekere beklijving van het effect van de inservice trainingsactiviteiten van leerkrachten. Testafnames op 8-jarige leeftijd (WISC-R) tonen voor de kinderen een zekere terugval, maar het experimentele resultaat blijft duidelijk zichtbaar.

Kinderen van geschoolde en ongeschoolde arbeiders bleken evenzeer te profiteren van de activiteiten als overige kinderen. Jongens lijken wat meer te profiteren dan meisjes. De resultaten zijn niet te danken aan werken in kleine klassen of met ervaren leerkrachten. De betrouwbaarheid van de testgegevens is goed. Getracht wordt achtereenvolgens de volgende typen kritiek te weerleggen: proefleidereffecten, statistische regressie, selectief experimenteel verlies, differentiële groei, ongeldige normering, testgerichte training, Hawthorne-effect.

Er is zuinig omgesprongen met statistische analyse, in het bijzonder statistische toetsing. Dit uit geldigheids- en bruikbaarheidsoverwegingen. Aandacht wordt gegeven aan verklaringsmogelijkheden bij dit type werk, evaluatiecriteria en evaluatie-opvattingen.

Slotconclusie is dat, aangezien gewerkt is op scholen in achterstandssituaties een meetbaar effectief stimuleringsbeleid mogelijk is, via positieve discriminatie in de klas, waarbij de rest van de klas er niet onder hoeft te lijden, de leerkrachten zelf de veranderingen doorvoeren en de resultaten relatief blijvend zijn. Dit neemt niet weg dat we niet weten wat voor de betrokken kinderen uiteindelijk aan positieve gevolgen (in termen van cognitieve ontwikkeling, zelfstandigheid, of schoolcarrière?) resteert. Laat staan dat we zouden kunnen aangeven wat van het type ondernomen activiteiten de maatschappelijke betekenis zou kunnen zijn. Desalniettemin lijkt het verantwoord om de activiteiten voor herhaling en navolging aan te raden. Aan een verdere implementatie wordt ook gewerkt.

Om redenen van omvang diende het artikel in twee delen te worden gesplitst. Het tweede deel verschijnt in het eerstvolgende nummer (februari 1981). Gezien de mate waarin de verschillende paragrafen één geheel vormen was het niet mogelijk tot twee relatief op zich staande deelartikelen te komen. Aanbevolen wordt beide delen als één geheel te lezen.

1. *Wat mogen we in dit artikel verwachten?*

Dit is het eerste gedeelte van het laatste artikel in de serie over het GEON-project die we in maart 1980 in Pedagogische Studiën begonnen. Dit vijfde artikel sluit nauw aan bij Stokking (1980[b]) waarin de evaluatie-aanpak werd beschreven en De Vries, Kramer-van Walderveen, Stokking en Thierens (1980) waarin we ervaringen en evaluatiegegevens per cursus vermeldden. Moesten we in laatstgenoemd artikel al flink selecteren uit het beschikbare vragenlijst–, observatie– en registratiemateriaal, in dit artikel hebben we onszelf zo mogelijk nog meer moeten beperken. Van de beschikbare kwantitatieve evaluatiegegevens beschrijven we hieronder slechts de test- en toetsgegevens; overige informatie, met name ook beschikbaar op basis van de gehanteerde administratieve aanpak, vermelden we slechts hier en daar bij wijze van illustratie en ondersteuning. We veronderstellen de evaluatie-opzet bekend en herhalen daarom niets uit het tweede artikel (Stokking, 1980[b]). Noodgedwongen laten we ook van de testgegevens het meeste cijfermateriaal weg. We geven in dit artikel de uitkomsten per regio per ronde per experimentele conditie. Het gaat te ver om de gegevens per school te presenteren. Anderzijds geven we ook geen totaaltabel. Regio's en rondes worden door

* Met dank aan M. Zwarts, W. Molenaar, A. K. de Vries en G. A. Kohnstamm.

ons als replicaties behandeld, in de zin van replicatie-onderzoek. Wat betreft bijkomende factoren als: geslacht van het kind, leeftijd, beroep van de vader, bij welke leerkrachten zat het kind in de klas en wanneer, leeftijd en ervaring van die leerkrachten, deelname van die leerkrachten aan het project (per programma-onderdeel), klasse-grootte e.a.: hierbij zullen we voornamelijk de conclusies vermelden. We geven dan wel aan hoe we analyseerden, maar zullen slechts bij wijze van illustratie cijfermateriaal opnemen in dit artikel.

De bovengenoemde beperkingen zijn temeer nodig, omdat in een artikel als dit niet volstaan kan worden met een presentatie van de resultaten op zich. We willen de uitkomsten bespreken tegen het licht van relevante onderzoeksliteratuur, en bovendien niet voorbij gaan aan de discussie die gevoerd wordt aangaande keuzen en problemen bij evaluatie en evaluatie-onderzoek zelf. Omdat verschillende delen van de te melden uitkomsten een eigen, aparte bespreking nodig hebben, lijkt het ons niet zo overzichtelijk om de gebruikelijk opbouw (eerst alle resultaten, dan de discussie) aan te houden. Een dialoog is waarschijnlijk leesbaarder. Reden voor ons om een criticus als gesprekspartner ten tonele te voeren. Op diens eerste vraag hebben we hierboven al geanticipeerd. Het gesprek gaat als volgt verder.

2. *Wat waren de testresultaten van het GEON-project?*

De summatieve evaluatie is gecentreerd rond de afname van de WPPSI op de kleuterscholen. Op elke door het project begeleide school werd in de maand oktober van het eerste jaar dat de school werd begeleid deze test afgenomen bij een tiental kleuters, aselect getrokken uit de groep met jongste kleuters. De steekproef was gestratificeerd naar sekse en naar klas indien op de betreffende school een heterogene leeftijdsopbouw per klas werd toegepast. Verder werd de test afgenomen bij die kleuters die naar de mening van de kleuterleidster risico's liepen achterstand op te lopen dan wel die reeds hadden. Achterstand was gedefinieerd als het niet mee kunnen komen met de rest van de groep waar het gaat om het begrijpen van de dagelijkse opdrachtjes, risico sloeg ook op eventuele verwachtingen zoals b.v. dat een kind een gerede kans liep een jaar langer op de kleuterschool te zullen blijven, in de eerste klas G.L.O. te blijven zitten, dan wel verwezen te worden naar een vorm van buitengewoon onderwijs. De aldus aangegeven 'doelgroep' kon uiteraard overlappen met de steekproef. Op deze wijze werden per school in totaal 12-15 kinderen getest. Na ruim anderhalf jaar, in de maand mei, werd dezelfde groep

Tabel 1 *Testresultaten Kleuterschoolperiode, Projectscholen*

regio	ronde	exp. groep	aantal kinderen	voortest x	s_x	natest x	s_x	verschil	S_{ed}
1	1	pd	41	89	21	95	17	+ 6	6,7
		ps	50	104	14	105	14	+ 1	5,1
2	1	pd	37	88	23	107	23	+19	8,3
		ps	48	105	18	121	19	+16	6,8
3	1	pd	29	83	20	95	18	+12	7,5
		ps	63	94	21	102	19	+ 8	7,9
4	1	pd	–						
		ps	65	88	21	99	17	+11	5,5
1	2	pd	20	79	21	89	19	+10	7,3
		ps	59	99	18	104	15	+ 5	5,9
2	2	pd	30	85	24	98	21	+13	6,9
		ps	70	105	16	113	14	+ 8	4,6
3	2	pd	41	92	20	107	19	+15	4,8
		ps	58	101	18	114	18	+13	4,4

Toelichting: x = gemiddelde s_x = standaardafwijking
pd = doelgroep ps = steekproef verschil = (x na − x voor)
$S_{ed} = (S^2_{voor}\,(1-r_{voor,\,voor}) + S^2_{na}\,(1-r_{na,\,na}))^{1/2}$

Tabel 2 *Testresultaten per ronde samengevat, Projectscholen*

	vier 1e rondes			drie 2e rondes		
	vooruitgang	aantal scholen	aantal kinderen	vooruitgang	aantal scholen	aantal kinderen
doelgroep	+12	18	107	+13	20	91
steekproef	+ 9	25	266	+ 9	20	187

weer de WPPSI afgenomen. Dat was dus tegen het einde van de periode van twee jaar dat de school intensief werd begeleid en tegen de tijd dat de aanvankelijk 4-jarigen als 6-jarigen zouden overstappen naar het G.L.O. Per regio per ronde zijn meestal zes scholen begeleid. De gegevens van de zes 'steekproeven' kunnen worden samengenomen; evenzo van de 'doelgroepen'. Op die manier kunnen de testresultaten worden samengevat als in Tabel 1.

In deze tabel zijn de testresultaten uitgedrukt in de gewogen scores, dat zijn ruwe scores, omgerekend op basis van de Amerikaanse ijkingssteekproef. Voor details zie de WPPSI-manual (Wechsler, 1967). Bij het bezien van de vooruitgang in de testgemiddelden moet de gevonden standaardafwijking in aanmerking worden genomen. Een van de meer geformaliseerde manieren om dat te doen is via de standaardfout van de verschilscores (S_{ed}). Afgemeten daaraan was het resultaat in de tweede ronde van de derde regio (1977-1979) het indrukwekkendst (ruim driemaal de standaardfout), dat in de eerste ronde van de eerste regio (1974-1976) het minst (ongeveer éénmaal de standaardfout).

De voortestgemiddelden van de onderscheiden 'doelgroepen' liggen steeds lager (9 à 20 punten) dan die van de steekproeven. Dit betekent dat het oordeel van de kleuterleidster (gevraagd ± 1½ maand na het begin van het schooljaar) een zinvolle maat is voor het opsporen van ontwikkelingsachterstanden. Er zijn duidelijk verschillen tussen de zeven experimentele rondes, zowel qua voortestgemiddelde (de problematiek is niet overal even zwaar) als wat betreft experimenteel 'effect'. Bij het laatste vallen vooral de beide extremen op: eerste ronde eerste regio versus eerste ronde tweede regio. Het verhoudingsgewijs geringe resultaat in de eerste ronde eerste regio valt terug te voeren op het aanloopkarakter ervan; behalve dat er sprake was van startproblemen was het ook zo dat bewust een minder intensieve begeleiding werd gegeven dan in de overige rondes, dit om eerst de nodige ervaring op te doen als project. Het resultaat in de eerste ronde van de tweede regio is enigszins geflatteerd; er zijn diverse aanwijzingen dat we de testgegevens van die ronde met enkele korrels zout moeten nemen (zo was ondermeer de betrouwbaarheid van de natestafname op een van de scholen beneden de maat). Tenslotte valt op dat in de vierde regio geen doelgroepgegevens vermeld staan. Dat komt omdat men daar geen doelgroepen heeft onderscheiden, deels omdat men van mening was dat de hele klas 'doelgroep' was, deels omdat men op ideologische gronden geen onderscheid binnen een klas wenste aan te brengen en deels omdat men dermate vroeg in het schooljaar met testen begon (eerste helft september) dat de kleuterleidsters de kinderen nog niet voldoende kenden om een doelgroep te kunnen aanwijzen.

Als we, ondanks bovengemaakte kanttekeningen, de gegevens nog wat verder samenvoegen, dan komen we voor de steekproefgegevens op een vooruitgang van 9 punten, voor de doelgroep-gegevens op 12 à 13 punten (Tabel 2). Resumerend menen we te mogen spreken van een positief experimenteel resultaat. Op de betekenis ervan komen we nog terug.

3. *Alles goed en wel, maar zouden de testresultaten niet net zo zijn uitgevallen zonder GEON?*

Rekening houdende met een aantal mogelijke valkuilen bij het trekken van conclusies (zie Stokking, 1980[b]) hebben ook testafnames plaatsgevonden op een aantal zogenoemde vergelijkingsscholen. De afnameprocedure (keuze van kinderen, tijdstippen) was identiek aan wat boven werd beschreven met betrekking tot de projectscholen. De vergelijkingsscholen zelf waren die zes kleuterscholen per regio waarmee tijdens de schoolkeuzeprocedure, voorafgaande aan de start van het project, in principe was overeengekomen dat zij in de tweede experimentele ronde zouden worden begeleid. Tijdens de eerste ronde was er in elke regio dus sprake van zes projecten zes vergelijkingsscholen. De testresultaten op de vergelijkingsscholen staan in Tabel 3.

Het blijkt zo te zijn dat, per regio bezien, de voortestgemiddelden op de vergelijkingsscholen steeds hoger liggen (5 à 13 punten) dan op de projectscholen. Dat komt doordat in de schoolkeuzeprocedure

Tabel 3 *Testresultaten Kleuterschoolperiode Vergelijkingsscholen*

regio	ronde	exp groep	aantal kinderen	voortest x	s_x	natest x	s_x	verschil	S_{ed}
1	1	vd	18	98	19	92	17	− 6	6,4
		vs	–						
2	1	vd	21	93	23	105	19	+12	7,2
		vs	40	110	21	109	15	− 1	5,9
3	1	vd	14	90	11	91	12	+ 1	4,7
		vs	48	103	20	107	14	+ 4	6,6
4	1	vd	–						
		vs	60	101	20	108	18	+ 7	5,6

Toelichting: vd=doelgroep vs=steekproef Zie verder Tabel 1

bij de uiteindelijke verdeling over eerste en tweede experimentele ronde de scholen met de zwaarste achterstandsproblematiek 'voorrang' kregen voor wat betreft deelname aan het project. Over het niet onderscheiden hebben van een doelgroep in de vierde regio schreven we reeds. In de eerste regio zijn geen steekproefgegevens van de vergelijkingsscholen beschikbaar, doordat de hele testprocedure eerst nà die afnameperiode (najaar '74) werd ontworpen. Dit is een schoolvoorbeeld overigens van het nadeel van het starten van zo'n project alvorens aan de evaluatie te laten werken (zoals noodgedwongen feitelijk geschiedde). Vergelijking met Tabel 1 toont verder dat de gemiddelde vooruitgang op de projectscholen steeds groter is dan die op de vergelijkingsscholen. Ervan uitgaande dat gewogen (dat is: 'voor leeftijd gecorrigeerde') gemiddelden min of meer constant zullen blijven, vallen de resultaten van enkele vergelijkingsgroepen op. De doelgroep in de eerste regio gaat 6 punten achteruit. Hoewel hier sprake was van afname van de natest door anderen dan bij de voortest is de achteruitgang waarschijnlijk wel reëel, gezien de latere WISC-cijfers. Wel is de groep klein (n=18), wat alweer komt door de aanvankelijk onvolkomen testprocedure. De doelgroep in de tweede regio en, in mindere mate, de steekproef in de vierde regio, gaan duidelijk vooruit. Dat bleek echter terug te voeren op gerichte trainingsactiviteiten van kleuterleidsters: in de tweede regio is op enkele vergelijkingsscholen met takenreeksachtige lessen gewerkt (dit is precies de in het tweede artikel met 'diffusie' aangeduide valkuil); in de vierde regio werd op enkele scholen intensief gewerkt met een schoolvoorbereidingsprogramma.

Als we ondanks de gemaakte kanttekeningen (enerzijds bedenkingen; anderzijds juist extra's op de vergelijkingsscholen, dus een 'inflatie' van het controlegroepkarakter) de resultaten samenvatten (Tabel 4) zien we dat de eerder (Tabel 2) van de projectscholen gemelde vooruitgang duidelijk gunstig afsteekt bij die op de vergelijkingsscholen. We wijzen bovendien nog even op het verschijnsel dat op de vergelijkingsscholen de steekproef meer wint dan de doelgroep, op de projectscholen juist de doelgroep het meeste vooruitgaat. Het belang van deze constatering zit vooral ook daarin, dat dit patroon ook per regio bekeken steeds terugkomt en zelfs per school bekeken vrijwel zonder uitzondering opgaat.

Tabel 4 *Testresultaten samengevat, Vergelijkingsscholen*

	vooruitgang	aantal scholen	aantal kinderen
doelgroep	+ 2	16	53
steekproef	+ 4	16	148

4. *Blijft er van de experimentele winst wel iets over na verloop van tijd?*

Juist omdat er in vergelijkbaar onderzoek vaak werd geconstateerd dat effecten na enkele jaren weer min of meer waren uitgewist, hebben we bij alle kinderen waarvan we WPPSI-gegevens verzamelden op 8-jarige leeftijd de WISC-R laten afnemen. De uitkomsten staan opgenomen in Tabel 5.

We waren binnen het bestek van het project alleen in staat om deze follow-up te doen bij de kleuters die in de eerste ronde in de eerste, tweede en derde regio op de door het project begeleide kleuterscholen zaten. Tabel 5 kan vergeleken worden met de Tabellen 1 en 3. We zien in Tabel 5 een bekend patroon: hoe groter de winst aanvankelijk was (tussen 4- en 6-jarige leeftijd), hoe groter het verlies daarna (tussen 6

Tabel 5 *WISC-R-gegevens, 8-jarige leeftijd*

regio	groep	aantal	WISC-R x	WISC-R s	WISC-R minus WPPSI na	WPPSI na – WPPSI voor	WISC x	relatieve positie 1	relatieve positie 2	relatieve positie 3	WISC minus WPPSI voor
1	pd	39	86	15	– 8	+ 6		–0,44	–0,25	–0,28	– 2
	ps	37	99	17	– 6	+ 1	91	–0,39	+0,38	+0,44	– 5
	vd	14	86	21	– 6	– 6		+0,06	–0,44	–0,50	–12
2	pd	36	96	23	–11	+19		–0,57	–0,26	–0,29	+ 8
	ps	42	106	15	–14	+16	101	+0,24	+0,47	+0,29	+ 3
	vd	21	95	17	–10	+12		–0,33	–0,37	–0,35	+ 2
	vs	39	105	14	– 4	– 1		+0,48	–0,16	+0,24	– 5
3	pd	24	94	16	– 4	+13		–0,66	–0,38	–0,50	+ 9
	ps	56	103	17	– 0	+ 8	103	–0,11	–0,06	+0,00	+ 8
	vd	8	97	14	– 3	– 8		+0,61	–0,25	–0,36	–11
	vs	46	108	19	– 1	+ 4		+0,44	+0,31	+0,30	+ 3

Toelichting: pd = projectscholen doelgroep
ps = projectscholen steekproef
vd = vergelijkingsscholen doelgroep
vs = vergelijkingsscholen steekproef
1, 2, 3 = de meetmomenten (WPPSI voor, WPPSI na, WISC-R)
relatieve positie = z-score per regio

en 8). Maar dat klopt alleen goed als je de eerste en tweede regio vergelijkt. Daar komt bij dat de teruggang in de tweede regio als een bevestiging kan worden gezien van de eerder (en vóór de WISC-afnames!) geuite bedenkingen tegen de wel erg fors uitgevallen winst daar van 4 op 6. En de gegevens van WPPSI-voor en WISC-R vergelijkend, zien we dat de vooruitgang van alle projectgroepen in elke regio boven die van alle vergelijkingsgroepen ligt. Opvallend, vooral ook bij het WISC-R gemiddelde per regio, zijn de niveauverschillen tussen de regio's. Tegen de achtergrond van de WPPSI-voortestgegevens springt vooral regio drie eruit. We hebben daarvoor twee verklaringen. Ten eerste is er juist in die regio het meest gewerkt met die programma-onderdelen waarvan een meer blijvend effect verwacht mag worden (ook dit was al vóór de WISC-afnames gespecificeerd). En vervolgens kent de OBD in die regio een jarenlange traditie op het terrein van de begeleiding van de onderbouw van het G.L.O. Dat zal er ook wel mee te maken hebben.

Uiteraard zouden we kunnen praten over het experimentele verlies, of over verschillen in moeilijkheid tussen WPPSI en WISC, maar zeker is toch wel (zie de standaardscores in Tabel 5), dat er sprake is van duidelijke, voor de projectgroepen positieve verschuivingen tussen 4- en 6-jarige leeftijd, die zich tussen 6 en 8 in essentie grotendeels stabiliseren.

5. *Maar het zullen wel bepaalde (groepen) kinderen zijn die vooral geprofiteerd hebben van het project?*

5.1. *De factor sociaal milieu*

Twee van de meest dominerende kenmerken waarop in samenlevingen als de onze gediscrimineerd wordt zijn sociaal milieu en sekse. Het laatstgenoemde zullen we dadelijk bespreken. Eerst willen we ingaan op de 'factor' sociaal milieu. We hebben geprobeerd om zoveel mogelijk gegevens te verzamelen aangaande beroep en opleiding van de ouders. Alleen het gegeven 'beroep vader' is voldoende compleet beschikbaar om er in dit artikel over te kunnen rapporteren. Gehanteerd is de beroepenindeling van het CBS die door het RITP tot een alfabetische beroepenklapper is omgewerkt. De categorieën zijn: 1. boeren en tuinders, 2. academische- of leidinggevende beroepen, 3. middenkader, 4. administratief- en dienstverlenend personeel, 5. geschoolde arbeiders, 6. ongeschoolde arbeiders, 7. landarbeiders, 8. zelfstandige beroepen, 9. geen beroep/onbekend. Wij hebben daar nog aan toegevoegd de categorie: 0. werkloos.

Ten aanzien van gebruik van het gegeven 'beroep vader' kan worden opgemerkt dat daarmee subgroepen gevormd worden waarvan het belangrijk is om ze separaat te bezien, maar dat het een zeer grove maat betreft waar het gaat om het 'vangen' van zoiets als

'gezinsmilieu'. Het is een 'status'-variabele, geen 'proces'-variabele, zie bijvoorbeeld Shipman (1973, p. 169), Groenendaal (1978, p. 9) en Bronfenbrenner (1979, p. 419). Shipman stelt (p. 170): 'Precious research (..) suggests that it is the proces variables which have the greater impact on a child's life, and they certainly have greater theoretical utility than demographic indices for explaining how the environment mediates experience in critical ways. (..)". En inderdaad zijn er grotere verschillen aangetroffen in opvoedingspraktijk binnen door statusvariabelen gedefinieerde groepen dan ertussen (Shipman, p. 172; zie ook Majoribanks, 1979).

Shipman laat de mogelijkheid open dat 'situational and status characteristics (..) define important aspects of the child's psychological environment' (p. 170). Vergelijk ook Groenendaal (1978, p. 19): 'Het sociaal-economisch niveau van het gezin (beroep/opleiding ouders) vormt weliswaar een globale maar goede index voor het opsporen van kinderen op scholen in achterstandssituaties en daarmee voor het opsporen van onze doelgroep: zwakfunctionerende kleuters.'

Het 'sociaal milieu' waaruit de school de leerlingen rekruteert, was een belangrijk criterium bij de schoolkeuzeprocedure. Daarnaast zijn andere criteria gehanteerd, bijvoorbeeld m.b.t. het functioneren van het schoolteam. Het percentage kinderen van arbeiders lag meestal zo tussen de 50 en 60%, gemiddeld per regio/ronde genomen.

De factor 'beroep vader' bleek niet consistent en substantieel gecorreleerd te zijn aan variabelen als 'ervaring leerkracht', 'deelname aan het project', 'klassegrootte'. Opvallend is verder, dat in de vier eerste rondes géén verband kon worden geconstateerd, ten aanzien van de kinderen waarvan we het beroep van de vader kennen, tussen dat beroep, en het al of niet aangewezen worden door de kleuterleidster als 'doelgroepkind' vóór de voortest WPPSI. In de drie tweede rondes echter waren kinderen van ongeschoolde arbeiders duidelijk 'oververtegenwoordigd' in de doelgroep.

Afgezet tegen de testgegevens ontstaat voor 'beroep vader' Tabel 6. We hebben ons in deze figuur beperkt tot die beroepscategorieën waarin wat grotere aantallen kinderen vielen. Van de eerste ronde in de eerste regio ontbreekt het gegeven van de vergelijkingsscholen. In de vierde regio was er zoals al gezegd geen follow-up met de WISC-R meer mogelijk. We zien dat er sprake is van een duidelijke relatie tussen 'voortestgemiddelde' en 'beroep vader': kinderen van vaders met maatschappelijk minder gewaardeerde beroepen scoren gemiddeld lager. Er is echter géén duidelijke relatie met de scorevooruitgang: kinderen van (on)geschoolde arbeiders gaan meestal gemiddeld of zelfs meer dan gemiddeld vooruit op de WPPSI. Alleen in de eerste regio blijven kinderen van ongeschoolde arbeiders als groep

Tabel 6 *Relatie 'beroep vader' – testgegevens, projectscholen en vergelijkingsscholen*

	regio	ronde	WPPSI-voortest				WPPSI na minus WPPSI voor				WISC-R minus WPPSI na			
			3	4	5	6	3	4	5	6	3	4	5	6
	1	1	108	98	96	87	7	1	8	−2	−5	−5	−10	−7
project-	2	1	104	110	96	82	(12)	11	16	18	−6	−10	−5	−6
scholen	3	1	126	98	90	80	−6	8	12	18	−2	−4	−4	−7
	4	1	(122)	102	91	83	(−12)	6	7	16				
	1	2	93	102	94	87	8	8	5	0				
	2	2	120	98	105	93	1	10	7	6				
	3	2	121	104	93	85	5	17	20	18				
	1	1												
vergelij-	2	1	100	106	96	(87)	11	−4	10	(−1)				
kings-	3	1	(131)	109	91	92	(−11)	2	13	2				
scholen	4	1	106	102	94	86	16	15	9	13				

Toelichting: 3 = middenkader
4 = adm. en dienstverlenend personeel
5 = geschoolde arbeiders
6 = ongeschoolde arbeiders
De tussen haakjes staande getallen hebben betrekking op kleine aantallen kinderen.

achter, zowel in de eerste als in de tweede ronde. We hebben daarvoor geen duidelijke verklaring; de enige niet geheel onplausibele veronderstelling is dat genoemd verschijnsel gerelateerd is aan de in de eerste regio geringere heterogeniteit in samenstelling van de schoolbevolking op dit punt (beroep vader). Er zijn namelijk in ons gegevens-bestand indicaties dat per school bezien een heterogenere samenstelling van de schoolbevolking naar de factor 'beroep vader' gemiddeld met hogere testresultaten gepaard gaat; op scholen met een groot percentage kinderen van arbeiders is de kans op een relatief achterblijvend testresultaat groter. Dit is vaker geconstateerd, zie bijvoorbeeld Smith en Bissell (1970, p. 99).

In Tabel 6 is te zien dat de WISC-R-resultaten nogal uiteenliepen van regio tot regio, kijkend naar de verschillende opgenomen categorieën. Er is geen groep waar de beklijving consistent het grootst is. De conclusie dat er geen duidelijke samenhang is tussen 'testresultaat' en 'beroep vader' is niet nieuw, zie bijvoorbeeld Madaus et al. (1979, pp. 221-2) en Hindley and Owen (1978, p. 345).

Evenals Hermanns (1979, p. 355) was onze ingang bij het definiëren van de 'risico-groep' een ontwikkelingspsychologische, waarbij met relatieve maatstaven wordt gewerkt. Hij stelt verder: 'De gegevens wijzen er aldus op dat kleuters uit lager socio-economische niveau's niet alleen het onderwijs beginnen met een grotere kans op achterstand in de verstandelijke ontwikkeling, maar dat ze indien deze achterstand er is, bovendien minder kans hebben om in de loop van het kleuteronderwijs deze achterstand weg te werken dan kinderen uit andere milieu's'. Relatieve criteria hanterend moesten wij hetzelfde constateren, als we per experimentele groep kijken, echter *niet* als we projectscholen en vergelijkingsscholen vergelijken.

5.2. *De factor geslacht*

Het percentage jongens per regio per ronde per experimentele groep fluctueerde van 44% in de eerste ronde van de vierde regio op de projectscholen tot 58% in de tweede ronde van de tweede regio op de groep scholen die daar in de eerste ronde begeleid was (en waar tijdens de tweede ronde voortgezette evaluatie plaatsvond). Het percentage jongens dat de kleuterleidsters in de WPSSI-'doelgroep' kozen lag meestal ongeveer 10% hoger. Die 'doelgroep' bestond derhalve uit gemiddeld ½ keer zoveel jongens als meisjes. Dit is geen onbekend verschijnsel. Zie bijvoorbeeld Stevens (1973, pp. 71-72), waar hij stelt: 'De jongens blijken kwetsbaarder dan de meisjes'. We komen hier straks op terug. Variantie-analyses, op verschillende wijzen uitgevoerd, gaven voor 'sekse' géén statistisch significante relatie met de WPPSI-totaalscore te zien. Hierbij was onder meer gecorrigeerd voor leeftijd en experimentele conditie. Daaruit zou je kunnen besluiten tot het ontbreken van een relatie tussen de testgegevens en de factor geslacht. Dat wordt vaker gevonden (Stevens, 1973, p. 92; Hermanns, 1979, p. 354), en Kohnstamm (1968, p. 280) stelt ten aanzien van de normering van de WPPSI zelfs: 'Zo is de variabele geslacht op deze leeftijd dikwijls irrelevant gebleken'. Bedacht moet echter worden dat, zeker in geval van de WPPSI, de test daar ook op is gemaakt. Aan de andere kant worden soms wel degelijk samenhangen gevonden. Zo vonden Hindley en Owen (1978, pp. 336-339) op de Stanford Binet bij vergelijking van testuitkomsten, ook op 3-, 5- en 8-jarige leeftijd, dat de jongens vaak hoger scoorden, en dat 'there is a tendency for a greater spread in the change scores of boys'. Wij vonden een tendens dat de jongens op 4-jarige leeftijd gemiddeld lager 'begonnen' dan de meisjes, om dat op 8-jarige leeftijd helemaal te hebben 'ingehaald'. Overigens doorbreekt de eerste regio hier weer het algemene beeld (evenals dat in 5.1. het geval was).

De gegevens zijn genomen over beide experimentele condities tesamen. De trend is sterker dan het gegeven dat gemiddeld ongeveer 60% van de om evaluatieredenen aangewezen doelgroep-WPPSI uit jongens bestond, in combinatie met de testresultaten van die 'doelgroep' (zie Tabellen 1-5), kan verklaren. We vonden géén consistente relatie tussen geslacht en vooruitgang op bepaalde *sub*tests van de WPPSI, evenmin als duidelijke, zich replicerende relaties tussen geslacht en de resultaten bij de takenreeksen. Wat het laatste betreft lijkt het er alleen een beetje op dat van de kinderen die aan de lesjes meededen de meisjes gemiddeld meer profiteren, maar groot zijn de verschillen niet. Tenslotte: op 8-jarige leeftijd scoren de jongens in alle drie de regio's gemiddeld duidelijk hoger op Kennis, Inzicht en Legpuzzels, de meisjes op Codes.

Meer jongens in de 'doelgroep' dus en gemiddeld meer vooruitgang in score bij hen. We veronderstellen dat hier in ieder geval ook een, waarschijnlijk niet bewust gehanteerd, positief discriminatieproces een rol speelt. Maar dat is niet de positieve discriminatie die we bedoelen (behalve dan voor zover gekoppeld aan achterstand), en eerder als een negatieve discriminatie naar de andere kant te beschouwen.

6. *Was u eigenlijk wel op achterstandsituaties gericht?*

Omdat niet gewerkt werd met aselecte steekproef-

trekking uit een duidelijk omschreven populatie in Nederland is het voor generalisatie van de resultaten nodig om nauwkeurig te omschrijven met welke groepen het project heeft gewerkt. Ten aanzien van de kinderen zijn er een aantal momenten (geweest) waarop doelgroepkeuzen plaatsvonden. Het eerste moment was de schoolkeuzeprocedure. Daaruit kwam in elke regio een keuze voort van buurten, scholen, leidsters en kinderen. De procedure en de daarbij gehanteerde criteria zijn weergegeven in een afzonderlijk daaraan gewijd rapport. Het resultaat van de procedure, de buurten en de scholen waar het project werkte, zal in het eindverslag op een aantal punten worden beschreven. Als de keuze van scholen vaststaat vinden tijdens de projectperiode nadere keuzen plaats. Naar de leidsters toe gebeurt dit doordat niet elke leidster even intensief meedoet, hetgeen samen met de schoolteams van moment tot moment wordt bepaald. Naar de kinderen toe zijn er een aantal momenten waarop nadere keuzen plaatsvinden (een rechtstreeks gevolg van deelname aan het project, immers GEON = Gedifferentieerd Onderwijs). Voorbeelden zijn:

- Voor de voortest WPPSI geeft de leidster een oordeel door aan te geven bij welke kinderen zij een achterstand verwacht (de 'doelgroep' kinderen, zoals die aanduiding bij de verwerking van de WPPSI-gegevens werd gehanteerd).
- Bij de voortoetsafname van de takenreekstoetsen wordt, voor elke takenreeks weer opnieuw, toegespitst op de daarin verwerkte taal-denkprocessen, een groepje in de klas uitgekozen voor extra activiteiten; dat gebeurt door een combinatie van het oordeel van de leidster en de uitslag van de voortoets.
- In het kader van de cursus 'Verwoorden van Probleemgedrag' worden door elke leidster één of enkele kinderen voor observatie (eventueel gevolgd door functionele analyse) uitgekozen.
- In het kader van de cursus 'Voorspellen en Reageren' worden door elke leidster één of enkele kinderen uitgekozen om de voorspellingen en de controles daarop te doen aan de hand van de werklijst in de cursus.

De verandering die het project wilde bevorderen betreft de kansen van kinderen die, met name vanuit sociaal-economische achtergronden, in het onderwijs het minst aan hun trekken komen. Het gaat niet om het exemplarische bereiken van 'onderwijswonderen', maar om het wegnemen van belemmeringen die een normale onderwijsparticipatie door, respectievelijk spreiding van aandacht aan kinderen in de weg staan. In de wandeling werd dit wel de 'achterhoedefilosofie' van het project genoemd[1].

Met name kinderen uit arbeidersmilieus trekken onvoldoende profijt van het onderwijs. Dat is voldoende reden om iets aan dat onderwijs te gaan doen. Alleen: die reden is een optelling en bij de te ondernemen actie moet hij weer afgebroken worden naar zijn samenstellende delen. Oorzaken binnen de school voor de ongelijkheid van kansen zijn immers, ook uit onderzoek, wel bekend:

- de afstand tussen leerstofinhouden en presentatie daarvan en de belevingswereld van het kind is niet voor ieder kind gelijk;
- waar de afstand groter is, is het vaak niet mogelijk extra overbrugging aan te reiken, wegens ontbreken van vaardigheden, gelegenheid of hulpmiddel;
- de school kan door interne zwakte en externe overspoeling (waaronder begrepen die van innovatoren) niet zelf die problemen structureel oplossen; en
- de aanwezige ondersteuningsfaciliteiten (nascholing, begeleiding) zijn niet voor iedere school even adequaat.

Deze oorzaken versterken elkaar. Scholen d.w.z. leerkrachten, kunnen zich dat heel goed bewust zijn. Men geeft dan soms, althans schijnbaar, de moed op. Het begrip 'achterhoede' is dan niet meer alleen toepasbaar op de situatie van het kind, maar ook op het schoolgebeuren en de relatie van de school met zijn omgeving.

De factoren die binnen de school aan de orde zijn, zijn hoofdzakelijk van belang vanwege datgene wat zich weliswaar buiten de beïnvloeding van de school bevindt maar voor de opvoeding van cruciaal belang is.

Van directe invloed zijn:

- de behuizing geeft allerlei beperking aan activiteiten (kleine huiskamer met televisie, keuken die te klein is om kinderen te laten helpen, gemeenschappelijke slaapkamers);
- de inkomenspositie laat het aanschaffen van leerstimulerende middelen (speelgoed) of ontplooiing van zulke activiteiten, slechts beperkt toe;
- de leefsituatie van het gezin is gereguleerd door beperkingen (b.v. ploegendienst, 's avonds afwezige moeder);
- het is praktisch onmogelijk om uit te vinden hoe men het kind kan stimuleren bij het leren van iets wat men zelf niet gekend heeft;

terwijl van indirecte invloed is:

- de sociale situatie van het gezin waarin de genoemde problemen zich voordoen is zodanig dat het nauwelijks stimulerend, zo niet ontmoedigend is om er iets aan te doen.

Het is voor de hand liggend dat men in een confrontatie met al deze factoren tot hopeloosheid kan vervallen. Op zulke scholen is het GEON-project gericht geweest. Ruimtegebrek belet ons hier om gedetailleerder te zijn.

7. *Het effect zal echter wel vooral op de goed lopende scholen bereikt zijn, in kleine klassen, bij ervaren leerkrachten?*

Criterium in de schoolkeuzeprocedure was dat het project moest worden uitgevoerd waar het het meest nodig was. Daar werd bij gezegd dat dat niet samen hoefde te vallen met 'het meest willen'. De schoolteams pasten dit criterium in gezamenlijk overleg toe. Het project werkte aldus met scholen die met meer problemen geconfronteerd worden dan je 'gemiddeld' kunt noemen. Dat blijkt ook wel, al was het alleen maar uit de voortest-gemiddelden per school op de WPPSI. Andere indicaties zijn: veel personeelsverloop, stagnerende contacten met de ouders, strubbelingen in het team, ziekte en werken met invalsters, weinig contact met de OBD, problemen in de buurt waar de school staat. Een aantal factoren zijn afgezet tegen de testgegevens. Dan blijkt het volgende.

Analyse op verschillende analyseniveaus toont dat er geen duidelijke relatie is tussen de testgegevens en de omvang van het team, respectievelijk de gemiddelde klassegrootte. Op schoolniveau analyserend is er geen relatie constateerbaar tussen veranderingen in het leerlingenaantal (met name: terugloop) en testresultaat. Bedacht moet echter worden dat deze variabele is gerelateerd aan schoolgrootte en klassegrootte. Het kan zijn dat een eventueel positief effect van een kleine klas wordt tenietgedaan, doordat er klasopheffing en dus baanverlies dreigt bij teruglopend leerlingenaantal (zie ook Colthof, 1979, p. 190, die terugval in functioneren door onzekere rechtspositie kan verklaren). Op regioniveau analyserend zijn er inderdaad indicaties dat dreigende klasopheffing de resultaten kan drukken. Er is geen relatie tussen het aantal malen dat een kind van leerkracht wisselt (anders gezegd: hoeveel leidsters het heeft gehad op de kleuterschool) en testresultaat.

Waar een groot team of personeelsverloop gepaard ging aan relatief tegenvallend testresultaat, was er tevens minder deelname aan het project. Dit deed zich vooral in de eerste rondes vóór 1976 voor. Anders gezegd: het project heeft kennelijk geleerd hoe ook grotere en van samenstelling veranderende teams intensief bij de training kunnen worden betrokken. Wél blijft de factor 'strubbelingen in het team' enigszins samengaan met relatief achterblijvende resultaten (als we ook het weinig vooruitgaan in een ronde waar de andere scholen flink vooruitgaan als zodanig meetellen). Er blijken geen relaties van betekenis gevonden te kunnen worden tussen leeftijd en ervaring van de leerkracht enerzijds, testgegevens anderzijds. Op schoolniveau analyserend zijn er aanwijzigingen dat er weinig relatie bestaat tussen extra personele ondersteuning (in de vorm van een stimuleringskracht, een leerkracht voor buitenlandse kinderen, een taakverlichtster, een logopediste) en testresultaat. Bedacht moet echter worden dat waar zulke ondersteuning bestaat de problematiek ook groter zal zijn. En zulke ondersteuning is veelal slechts op enkele individuele kinderen gericht geweest. Verder zijn er voor zover de gegevens beschikbaar zijn geen relaties te ontdekken tussen mate van nevenactiviteiten van leerkrachten (studie, het volgen van een cursus op de plaatselijke OBD e.d.) en testresultaat.

Bij dit alles moet bedacht worden dat de binnen het project optredende variatiebreedte op de verschillende besproken factoren aanzienlijk was. De eerder gepresenteerde testresultaten zijn dus niet behaald op scholen in bevoorrechte situaties.

8. *Kloppen de gegevens dan wel? Wie namen eigenlijk de tests af?*

De testafnames werden aanvankelijk in principe gedaan door de intermediaire krachten. Daartoe was besloten omdat afname door een vertrouwd persoon juist voor onze doelgroep van belang is en omdat een externe testassistent naar de leerkracht toe het testen in een ongewenst, bijvoorbeeld controlerend of mystificerend daglicht zou kunnen plaatsen.

Methodologisch gezien was deze opzet zwak, gezien de mogelijkheid van *rolvermenging*. Immers: de intermediaire kracht kent het kind en laat wellicht haar verwachting en/of hoop doorklinken in de wijze waarop ze de test afneemt. (We laten bewuste beïnvloeding van uitkomsten als mogelijkheid achterwege. Ten eerste wisten veel intermediaire krachten niet precies hoe de gegevens verwerkt zouden worden. En vervolgens waren velen écht benieuwd of het projectwerk resultaat zou hebben of niet – en ze waren daarover nogal eens sceptisch). Er zou zelfs sprake kunnen zijn van interactie tussen proefleider en plaatsgevonden begeleiding.

Genoemde 'rolvermenging' is een voorbeeld van een 'proefleidereffect'. Een ander type daarvan doet zich voor als *tussen* proefleiders systematische scoreverschillen bestaan. (Waarschijnlijk is de uitdrukking 'proefleidereffect' vooral in die betekenis ingebur-

gerd, als hoofdeffect in een variantie-analytische toetsing op zulke verschillen). Als zulke verschillen zich voordoen, zou *proefleiderwisseling* (de natest en de voortest worden ieder door een ander afgenomen) voor een vertekening van de resultaten kunnen zorgen.

We hebben ons steeds heen en weer geslingerd gevoeld tussen de gedachte dat vermijden van proefleider-wisseling gunstig was, daar het een systematisch strenger/minder streng scoren zou kunnen voorkomen en de wens tot vermijden van de valkuil dat een proefleider het kind zo goed kent dat ze, willens of niet, er meer 'uithaalt' op de natest dan een andere proefleider zou doen.

Het beleid was dus halfslachtig: pleiten voor voorkomen van proefleiderwisseling, maar alle organisatorische factoren die dat veroorzaakten accepteren. Wat daardoor wel gezegd kan worden is dat er geen invloed is uitgeoefend op die wisseling, er zat dus geen systematiek in die de testresultaten zou kunnen vertroebelen.

In feite is de natest WPPSI vrij vaak afgenomen door een andere persoon (intermediaire kracht, begeleidster, medewerk(st)er uit een andere regio of van de centrale projectstaf) dan degene die de voortest afnam. Als je het zo negatief mogelijk wilt afschilderen zou je kunnen zeggen dat we dus met beide typen proefleidereffecten te maken kunnen hebben. We zagen dat in 40% van de gevallen gewisseld werd, dat wisseling meer voorkwam in de eerste en derde regio dan in de tweede en vierde, en dat het meer voorkwam op scholen die op dàt moment geen 'projectschool' waren, dan op projectscholen. Tenslotte zagen we, dat wisseling vaker met meer dan gemiddelde testresultaten (gemiddelde scorevooruitgang per school) gepaard ging. Beide laatste beweringen tesamen genomen betekent dit dat op scholen die op dat moment projectschool waren geprobeerd is om de natest zoveel mogelijk door dezelfde persoon te laten afnemen als de voortest en dat dit de testresultaten wel eens juist gedrukt kan hebben.

Verder geldt dat de intermediaire krachten de kinderen op de vergelijkingsscholen ook al kenden, met name tijdens de eerste ronde toen naast de WPPSI-afnames ook vergelijkingstakenreekstoetsafnames plaatsvonden, zodat een argument dat kinderen zich meer op hun gemak voelen bij iemand die ze al kennen slechts ten dele kan worden opgevoerd. Bovendien constateerden we dat, waar een ander de voortest afnam en de intermediaire kracht van de school de natest, de resultaten vaker verhoudingsgewijs achterbleven, en dat waar juist bij de *natest* werd ingesprongen, de vooruitgang vaker méér dan gemiddeld was. Anders gezegd: onze intermediaire krachten scoorden gemiddeld nogal 'streng'. (Dit komt overeen met een persoonlijke mededeling van de wetenschappelijke projectleider, gedaan ongeveer halverwege de projectperiode, en gebaseerd op zijn gesprekken met de intermediaire krachten over de afgenomen tests.)

Nu kan er van dergelijke 'effecten' alleen sprake zijn waar de standaardisatie van de test tekort schiet. Bij tests als de gebruikte is die standaardisatie vrij ver doorgevoerd en in zoverre zal de bedreiging 'proefleidereffecten' niet zo groot zijn. Feit blijft ondertussen wel dat tegen een dergelijke kritiek geen echt verweer mogelijk is. Er is slechts één alternatief en dat is gebruik maken van testassistenten.

Vanaf 1978 heeft het project dan ook gewerkt met externe testassistenten. Deze zijn voornamelijk ingezet bij de afnames van de WISC-R. In de eerste regio hebben twee assistenten ongeveer de helft van alle WISC-afnames gedaan, in de tweede en derde regio is door 5 respectievelijk 7 assistenten (waarvan in het laatste geval 3 slechts beperkt zijn ingezet) het geheel aan WISC-afnames verzorgd. Deze assistenten waren willekeurig verdeeld over de lagere scholen en bleven onkundig van de experimentele status van de afnames. Ook is er voor gewaakt om een bepaalde experimentele groep niet geheel door (een) bepáálde proefleider(s) de WISC-R af te laten nemen.

Samengevat: door proefleider-wisselingen en door het inzetten van testassistenten van buiten blijft er van het methodologische bezwaar van 'rolvermenging' niet zoveel over; waar dat wel een rol zou kunnen hebben gespeeld, zijn er indicaties dat het juist 'conservatief' uitpakte.

9. *Als de testafnames zo zijn verlopen, zijn de uitkomsten dan nog wel betrouwbaar?*

Steeds is er een split-half betrouwbaarheidscoëffici-

Tabel 7 *Split-half betrouwbaarheidscoëfficiënten Testafnames*

		WPPSI			WISC-R
regio	ronde	voortest	natest	totaal	totaal
1	1	.95	.92	.96	.91
2	1	.90	.96	.97	.92
3	1	.89	.94	.95	.89
4	1	.97	.94	.98	
1	2	.96	.90	.96	
2	2	.97	.93	.98	
3	2	.97	.97	.99	

ënt berekend. Bij de WPPSI gebeurde dat zonder de subtest transponeren, bij de WISC-R zonder legpuzzels en codes; dit om redenen van lengte en/of berekeningswijze van die subtests. In Tabel 7 staan de uitkomsten per regio per ronde weergegeven. De coëfficiënten zijn bevredigend en komen voor wat betreft de WPPSI overeen met wat de manual meldt (.96). Bij de WISC-R komen we wat lager uit dan Wechsler voor 8½-jarige leeftijd meldt (.95).

Per subtest bekeken blijkt de coëfficiënt voor de WPPSI vrijwel steeds rond de .90 te liggen. Het betrouwbaarst komen uit de bus: Kopiëren, Doolhoven, Algemene Kennis. Het lastigst betrouwbaar te krijgen bleken Analogieën (meestal tussen .80 en .85, met enkele uitschieters naar beneden: .66, .73) en Woordenschat (meestal tussen .80 en .90). Vergeleken met de opgave uit de Manual vallen Kopiëren en Algemene Kennis bij ons betrouwbaarder uit en Analogieën juist minder. Wat de WISC-R aangaat komen onze gegevens qua patroon overeen met de Amerikaanse: Inzicht valt laag uit (rond .60), Blokpatronen en Woorden het hoogst (ruwweg .85 en .79). Opvallend is dat onze WPPSI-uitkomsten over het geheel genomen hoger uitvallen dan de Amerikaanse, onze WISC-R duidelijk lager. Het vaakst werd een teleurstellende interne consistentie per subtest aangetroffen in de derde regio (dat geldt zowel voor WPPSI als voor WISC-R).

Per proefleidster bekeken komt de coëfficiënt bijna altijd boven de .90 uit, vaak boven de .95 (WPPSI), bij de WISC-R rond de .90, met enkele uitschieters naar boven (eerste regio PL 12, tweede regio PL 15: .99) en beneden (tweede regio PL 16: .80).

Per school bezien komt de WPPSI-coëfficiënt vaak zeer hoog uit (.97-.99) en hier en daar tussen de .80 en de .90. Voor de WISC-R geldt hierbij: rond de .90, maar wat meer fluctuerend; vooral in de eerste en tweede regio enkele uitschieters naar beneden.

Al met al laat de interne consistentie weinig te wensen over. Incidentele dubbel-afnames van de WPPSI (bijvoorbeeld: in februari door een psycholoog in verband met een onderzoek, in mei voor GEON) kwamen binnen een marge van 1 punt op hetzelfde gewogen totaal uit. Ook dat geeft moed.

Een andere manier om de betrouwbaarheid van de testgegevens te benaderen is een indruk geven van training, afname en controle. Training vond plaats in de inwerkperiode van de intermediaire krachten, in een drietal bijeenkomsten, door een klinisch pedagoge. Daarna werden bij 3 of 4 kinderen proefafnames gedaan, waarna over afname en scoring weer discussie plaatsvond. Deze procedure is per regio gehanteerd. Er werd aandacht besteed aan testen in het algemeen, het gebruik in het project, de keuze van de WPPSI. Benadrukt werd het voorkómen dat de kleuterleidster er teveel waarde aan zou hechten; het was meer 'voor wetenschappelijk publiek.' Men observeerde elkaar en zichzelf, met behulp van audio- en video-opnamen, oefende het invullen, etc. Afname vond meestal plaats in het kamertje van de school of anders in een leegstaand lokaal. Er werd aandacht besteed aan de nodige rust en door hooguit drie kinderen per dag een test af te nemen was er geen sprake van gejacht. Geluidsoverlast was echter soms niet te vermijden.

De intermediaire krachten vonden het steeds terugkerende testen over het algemeen niet al te onaangenaam, maar het was wel iets 'dat er nu eenmaal bijhoorde'.

Controle op de gegevens vond op de volgende manieren plaats. Voor het gehele bestand zijn de ruwe gegevens per item, die apart werden ingevoerd voor de betrouwbaarheidsberekeningen, per subtest getotaliseerd; de uitkomsten daarvan zijn vergeleken met de door de intermediaire krachten opgegeven ruwe subtesttotalen, zoals die gebruikt werden bij de effectanalyses. Afwijkingen bleken gering, in aantal zowel als in grootte en bovendien niet systematisch. Verder is steekproefsgewijs (1 à 2 scholen per regio per ronde) nagegaan of de leeftijd op het moment van testafname correct was berekend en of de omzetting van ruwe in gewogen subtesttotalen (welke óók werden gebruikt bij de effect-analyses) juist had plaatsgevonden. In het eerste werd slechts hoogst zelden een fout aangetroffen, in het tweede af en toe, maar niet systematisch. Incidenteel is de scoring langsgelopen, wat soms de mogelijkheid tot interpretatieverschillen aan het licht bracht. Tenslotte zijn de testafnames steevast doorgesproken tussen intermediaire kracht en wetenschappelijk projectleider. In die gesprekken kwamen geen onregelmatigheden aan het licht.

De testassistenten bij de WISC-R waren in de eerste regio ex-kleuterleidsters, die een training ontvingen als boven beschreven. In de tweede regio waren het studentes orthopedagogiek met testervaring, in de derde regio ervaren testassistenten. Steeds werd discussie ingelast over eventuele afname- of scoringsproblemen.

10. Gezien uw doelgroepkeuzeprocedure en de waarden van betrouwbaarheidscoëfficiënten zal het effect dat u rapporteert wel een kwestie zijn van statistische regressie.

Dat is voor een deel inderdaad het geval. Echter meer 'regressie' als *verschijnsel* dan als oorzaak van een

artefact. Bovendien was ons onderzoeksdesign bedoeld mede juist om dit type kritiek op te kunnen vangen en we hebben dan ook diverse contra-indicaties.

We hebben aan de kwestie van statistische regressie naar het gemiddelde een apart artikel gewijd (Stokking, 1980c). Daar gaan we ook in op de discussie hieromtrent in de eerste jaargang van het Tijdschrift voor Onderwijsresearch (Groen, 1975; De Groot en Van Peet, 1975, 1976; Peschar 1976a+b). Zie ook Peschar, 1978. Op deze plaats volstaan we met het volgende.

Vroon (1978) onderscheidt verschillende typen van regressie. Waar het in ons geval om gaat is regressie bij herhaalde meting binnen één groep. Wat optreedt is dat aanvankelijk extremere scores in tweede instantie dichter bij het gemiddelde blijken te liggen. In het bijzonder dus: aanvankelijk lage scores stijgen. Het zal duidelijk zijn dat als zoiets door andere factoren dan experimentele interventie kan zijn veroorzaakt, daarmee voor evaluatie van programma's als GEON een lastig probleem ontstaat. Thorndike (1942) wijdde reeds een grondig artikel aan regressie naar het gemiddelde.

Samengevat komt een en ander op het volgende neer: Als scores bij voormeting extreem laag liggen ten opzichte van het gemiddelde en men kiest de 'cases', in dit geval kinderen, die zulke scores behaalden op grond van die extremiteit uit, in dit geval voor experimentele beïnvloeding, terwijl de meetbetrouwbaarheid van de voormeting niet perfect was, dan mag men verwachten dat hun scores bij een tweede meting sowieso hoger uitvallen (omdat er wordt gekapitaliseerd op negatieve meetfouten), zodat een eventuele vooruitgang niet (geheel) aan de interventie mag worden toegeschreven. Vooruitgang kan dan een *artefact* zijn, 'veroorzaakt door' regressie. Als aan één van de voorwaarden niet of slechts in geringe mate is voldaan, bijvoorbeeld: de meetbetrouwbaarheid is hoog, of: extremiteit was geen selectiecriterium, dan kan er niet of amper sprake zijn van een regressie-artefact (zie ook Roskam en Van der Sanden (1974); Bereiter (1976); Hindley en Owen (1978); Cambell en Erlebacher (1970); Molenaar en Thomas (1978)).

Als aan de genoemde condities wél is voldaan, is de correlatie tussen beide meetmomenten noodzakelijkerwijs kleiner dan 1. Het omgekeerde geldt niet: alleen al 'op grond van' het laatste zal regressie optreden, maar dan betreft het een mathematisch verschijnsel, dat kan optreden op basis van verschillende substantiële oorzaken. Behalve aan meetonbetrouwbaarheid gerelateerde oorzaken (a) kan er namelijk ook sprake zijn van een verschuivende samenstelling van de test, anders gezegd: men meet op latere leeftijd iets anders (b). En uiteraard ook: werkelijke veranderingen in ontwikkeling (c). (Zie ook: Vroon, 1978, p. 289; Hindley en Owen, 1978, p. 347.) Men kan deze hele discussie willen omzeilen door een betrouwbaarheidscorrectie toe te passen. Motto: dan zit het altijd goed. Bereiter (1976, pp. 3, 9) spreekt van een 'over-correction-undercorrection dilemma' en zegt o.i. terecht dat men het probeem dan verschuift naar de keuze van een betrouwbaarheidscoëfficiënt (respectievelijk correctieformule). Cronbach (1976, p. 6, 8) zegt naar aanleiding van zes verschillende coëfficiënten voor correctie voor attenuatie dat normaliter een standaardmeetfout veel meer van belang is dan een (generaliseerbaarheids)coëfficiënt. Naar aanleiding van genoemd keuzedilemma en de mogelijkheid in ieder geval de standaardmeetfout te berekenen hebben we ons tot het laatste beperkt en hebben we niet een of andere correctie aangebracht. Als men niet oppast corrigeert men het experimentele effect weg. (Zie ook Hindley en Owen, 1978, p. 343, en Bereiters verwijzing naar een artikel van Garside uit 1956.) Hindley en Owen, (1978, p. 347): 'Correction for test unreliability substantially reduces regression effects only when they are in any case relatively small because the correlation between scores at two ages is high. This tends to occur over shorter intervals. Most of the larger regression effects are attributable predominantly to a mixture of (b) and (c), which cannot be disentangled here.' We denken dat onvoldoende wordt geëxpliciteerd dat 'regressie' die uitsluitend gepaard gaat met een niet perfecte correlatie niet meer is dan een etiket. Dan is namelijk, ook bij een perfect betrouwbare eerste meting, de voorspelbaarheid van de tweede niet perfect. Regressie is dan een naam voor een gevolg van wat er gebeurd is (bijvoorbeeld: 'true changes in relative ability') en geen artefact in de zin van een bedreiging van een effectconclusie.

Contra-indicaties voor een regressie-artefact in GEON zijn de volgende:

1. Er is niet geselecteerd op basis van extreme scores. Onze projectscholen (en overigens ook vergelijkingsscholen) werden gekozen via een schoolkeuzeprocedure waarin allerlei criteria een rol spelen, maar geen testscores. En op de scholen zijn steekproeven getrokken. De 'doelgroepen', die daarnaast worden gehanteerd werden gevormd op aanwijzing van de kleuterleidster.
2. We zien dat de doelgroep op de projectscholen qua scores gemiddeld dichter bij de steekproef (klasgemiddelde) komt te liggen, terwijl het verschil tussen doelgroep en steekproef op de vergelijkingsscholen juist toeneemt. (Uitgaande van de

'fan-spread' hypothese van Cambell en Erlebacher (1970) zou je het omgekeerde verwachten!)
3. Hoewel kinderen van vaders met diverse beroepen gemiddeld per categorie qua voortestscores verschillen (kinderen van arbeiders scoren lager), zien we geen duidelijke verschillen tussen de categorieën in verschil tussen voor- en natest.
4. Als Lord stelt: 'In general, the analysis of observed gains results in a built-in bias in favor of whatever treatments happens to be assigned to initially low-scoring groups' (1967, p. 37), dan geldt in ieder geval in GEON dat binnen de experimentele conditie 'projectscholen' kinderen niet meer of minder bij leidsters zaten die cursussen doorwerkten, al naar gelang hun voortestscore. (Zie ook Molenaar en Thomas, 1978, p. 154).
5. We vinden van veel relaties (bijvoorbeeld: ontbreken van een verband tussen klassegrootte en vooruitgang; juist wel verband tussen deelname en vooruitgang), dat ze zich van regio tot regio/van school tot school herhalen. Voor zover al van toevallige meetfouten met een zekere omvang sprake is geweest zouden ze dus steeds dezelfde kant uit hebben moeten vallen, wat erg onwaarschijnlijk is. Ook de meta-evaluatie van de evaluatie van het GEON-project komt tot de conclusie dat 'regressie' geen groot probleem is (Zwarts, 1979).

11. *O.K., Regressie 'verklaart' niet zo veel, maar is jullie design wel gerealiseerd? Hoe zit het met het experimentele verlies, met de mogelijkheid van differentiële groei, e.d.?*

In een eerder artikel (Stokking, 1980[b]) is uitgebreid ingegaan op problemen bij het trekken van conclusies. Er werden ruim 20 mogelijke valkuilen onderscheiden waarmee zo goed mogelijk rekening werd gehouden bij de opstelling van het design. Zo werd besloten om in ieder geval ook vergelijkingsgegevens te verzamelen vanwege de bedreigingen 'reactieve metingen', 'niet representatief verlies', 'autonome ontwikkeling' en 'unieke gebeurtenissen'. Als gewezen wordt op de mogelijkheid van acceleratie, plotselinge groei zoals die in de ontwikkelingspsychologisch gezien minder stabiele periode van 5-7 jaar wel voorkomt (zie Stevens, 1973, p. 27), dan menen we dat er geen redenen zijn om aan te nemen dat dat verschijnsel in de verschillende experimentele groepen in ongelijke mate zich zou hebben voorgedaan. Een extra controle daarop is het gegeven dat de gemiddelde leeftijden op project- en vergelijkingsscholen elkaar zelden veel ontliepen en dat het interval tussen voor- en natest ook vrijwel identiek kon worden gehouden. Over alle vier eerste rondes samen was de leeftijd bij voortestafname op de projectscholen gemiddeld 55 maanden, op de vergelijkingsscholen 56 maanden. Het interval was gemiddeld respectievelijk 19 en 18 maanden. (De leeftijd bij natestafname lag dus gelijk.)

Toch vallen bij het design wel enkele kanttekeningen te plaatsen. Doordat de steekproeven uit hele groepen worden getrokken vallen er meestal al enkele doelgroepkinderen in. De kinderen in deze 'overlap' worden uiteraard niet tweemaal getest. Hun gegevens worden wel tweemaal gebruikt (bij berekening van zowel doelgroepgemiddelden als steekproefgemiddelden). Dit betekent dat eventuele verschillen tussen doelgroep en steekproef onderschat zullen worden. De bepaling van de resultaten bij de doelgroepen wordt hierdoor uiteraard niet beïnvloed. Wat wel beïnvloed wordt is de mogelijkheid om het experimentele effect te zuiveren volgens de elders (Stokking, 1980[b], par. 7) voorgestelde optel-aftrek procedure (vd − vs = regressie, ps − vs = algemene GEON-begeleiding, etcetera). We hebben in de deelrapportages deze procedure wèl gevolgd voor de takenreekstoetsgegevens uit de 1e rondes van de eerste, tweede en derde regio. Dat bracht soms een flinke relativering van het effect van de takenreekslesjes met zich mee (zie ook De Vries et al., 1980). Dit artikel is echter gebaseerd op de argumenteermogelijkheden die het design als geheel ons biedt. Voor de testgegevens zullen we de optel-aftrekprocedure niet hanteren.

Een tweede kanttekening betreft de WISC-R. Het design is geënt op de kleuterscholen. Daarna waaieren de kinderen uit over verschillende lagere scholen. Dat zou de experimentele opzet kunnen vertroebelen als 'experimentele' en 'controle'-kleuters bij elkaar in de (eerste en tweede) klas zouden komen te zitten. We stellen met opzet: 'kunnen' vertroebelen, want het hangt een beetje van de analysevraagstelling af wat het meest gunstig is. Wij zijn er van uitgegaan dat het mengen van experimentele groepen een effectafvlakkend effect zou hebben (cf. McDill et al. 1969, p. 23). Daartegenover staat dat de omgevingscondities van lagere school tot lagere school nogal kunnen verschillen, wat de vergelijkbaarheid niet ten goede komt als de experimentele groepen gescheiden blijven optrekken tussen hun 6e en 8e jaar. Welnu, wat zich feitelijk heeft voorgedaan is dat de groepen grotendeels (voor meer dan 80%) gescheiden bleven optrekken. Wellicht dat dit mede onze redelijke gunstige beklijvingsgegevens verklaart. Daarnaast bleken de verschillende groepen lagere scholen redelijk vergelijkbaar waar het ging om gemiddelde klasse-

grootte, ervaring en leeftijd der leerkrachten, deelname aan bijzondere activiteiten, e.d.

Van de testprocedure maakte deel uit dat per school steekproeven van 10 werden getrokken (uit de leeftijdsgroep). Er werd van afgezien om de hele groep te testen, omdat daarmee teveel tijd gemoeid zou zijn, in relatie tot het begeleidende werk. Reductie van het aantal te testen kinderen kon gewaagd worden, gezien de uiteindelijk te verwachten grote aantallen. Het aantal van 10 werd bepaald op inferentieel-statistische gronden (bij een experimenteel verlies van 50% nog voldoende aantallen per experimentele conditie overhouden om een onderscheidingsvermogen van ongeveer 2/3 te realiseren bij een verwachte behoorlijke effectgrootte. Zie Cohen, 1969). Toch is een aselecte steekproef van 10 per school niet altijd gerealiseerd. Zowel in de tweede als derde regio viel één vergelijkingsschool af. Een reeds geselecteerd kind kon bij voortestafname afwezig zijn, bijvoorbeeld door ziekte (waarbij dan soms wel, soms niet willekeurig een ander werd uitgenodigd). Er waren uiteraard nogal eens kinderen waaraan wel een voortest was afgenomen, niet voor de natest beschikbaar (ruwweg 5%). In de vierde regio werden geen doelgroepen aangewezen, maar wel extra kinderen (buiten de 10 per school) getest, echter zonder de steekproefkeuze zorgvuldig vast te leggen. Deze was daardoor niet meer reconstrueerbaar..

Door dergelijke factoren zijn de aantallen vaak niet 'mooi 10 per school', en kan men zich afvragen in hoeverre van aselectheid sprake is. Wat bijvoorbeeld te denken van de situatie dat er 16 kinderen in een klas zitten, waarvan er drie ziek zijn, waarna besloten wordt om de overige 13 'dan maar allemaal' te testen? We gaan er vooralsnog van uit dat het willekeurige karakter niet is aangetast, zodat generalisatie (naar de scholen!) mogelijk blijft. Checks ten aanzien van variabelen als leeftijd en beroep vader hebben ook geen substantiële vertekeningen aan het licht gebracht. Voor de groep waarvan we gegevens tot en met het achtste jaar konden verzamelen was het experimentele verlies van 4 op 8 jaar als volgt. De kinderen van 18 'projectkleuterscholen' en 16 'vergelijkingskleuterscholen' uit de eerste rondes in de eerste, tweede en derde regio, waaierden, voorzover ze binnen de regio's bleven wonen, uit over 103 lagere scholen. Van de 344 betrokkenen met WPPSI geteste kleuters verhuisden 38 naar buiten de regio's, dat is 11%. Dat is, de periode in aanmerking genomen, niet veel. Cicirelli et al. (1970, p. 114) meldden dat 'the principle findings of that (migration) study were that a total of only 12% of the original Head Start population (. .) had left the sample target areas'.

Een laatste kanttekening betreft de analyse van de gegevens. Kenmerkend voor de gebruikte geautomatiseerde analysetechnieken is dat berekeningen zoveel mogelijk gemaakt worden op dàt deel van de proefpersonen waarbij de waarden op alle variabelen in kwestie bekend zijn. Met name bij meer overkoepelende berekeningen kan het voorkomen dat er enkele 'cases' buiten beschouwing worden gelaten, omdat ze niet meer 'volledig' zijn, als een groter aantal variabelen (i.c. scores) tegelijkertijd worden beschouwd. Doordat niet van àlle kinderen àlle gegevens beschikbaar waren, zijn sommige analyses daardoor uitgevoerd op een iets uitgedunde onderzoeksgroep. We hebben dat niet uitgebreid op mogelijke vertekenende invloed onderzocht omdat we het zich al dan niet herhalen van verbanden in andere rondes/regio's doorslaggevend achten. Bovendien bleef dit 'analyseverlies' beperkt tot 1 à 2%. We menen dat op grond van bovenstaande vertrouwen in de gegevens zoals gepresenteerd in de Tabellen 1-5 gerechtvaardigd blijft.

12. *Maar u presenteert steeds de 'gewogen' scores. Mag dat zomaar?*

De gewogen scores zijn uiteraard niet 'zomaar' te gebruiken. In feite zijn er 3 typen scores. De scores die de kinderen op de verschillende subtests behalen, in de zin van 'aantal items goed' worden *ruwe* scores genoemd. Opgeteld geeft dit het ruwe totaal (opgebouwd uit een verbaal en een performaal subtotaal). Gebaseerd op een Amerikaanse ijkingssteekproef geeft Wechsler omrekentabellen waarbij, per leeftijdsrange van 3 maanden, het gemiddelde per subtest (ongeveer) 10 wordt, de standaardafwijking (ongeveer) 3. We spreken dan van *gewogen* scores, die weer opgeteld kunnen worden tot een gewogen totaalscore. Op deze manier wordt 'gecorrigeerd voor leeftijd', komt het gemiddelde op (ongeveer) 100 uit, bij een standaardafwijking van ± 20. Tenslotte zijn de gewogen totaalscores omzetbaar in *IQ*-scores, wat er eenvoudigweg op neerkomt, dat de standaardafwijking wordt teruggebracht op 15 (uitgaande van 21 voor het gewogen totaal).
Het project heeft de IQ-scores nooit gehanteerd. We hebben altijd gewerkt met de ruwe en de gewogen scores en gepresenteerd werden steeds (alleen) de gewogen scores, omdat de ruwe alleen hun die de test goed kennen iets zeggen. Een en ander geldt zowel voor WPPSI als WISC-R.

Het is op voorhand niet aan te nemen dat de Amerikaanse ijkingsgegevens geldig zijn in de Nederlandse situatie. De ervaringen, met name voor de

WPPSI, zijn dusdanig dat inmiddels wel duidelijk is dat dat niet zo is. De WPPSI valt in ons land relatief 'gemakkelijk' uit, als je redeneert dat een gewogen gemiddelde van 100 richtpunt is. Maar dat is voor ons geen probleem, aangezien we de gegevens relatief waarderen (vergelijkenderwijs, 'projectscholen' en 'vergelijkingsscholen') en aangezien we statistische analyses ook steeds op basis van de ruwe scores uitvoerden, zonder andere resultaten. Wat anders is dat de structuur van de test anders kan uitvallen. Als we naar de daadwerkelijk gevonden scores kijken dan zien we dat bepaalde subtests (met name Woordenschat) laag uitkomen, andere (met name Blokpatronen) hoog. Is dat echter te wijten aan daadwerkelijke verschillen met de Amerikaanse situatie (b.v. door een andere kleuterschoolpraktijk?), of aan onze specifieke onderzoeksgroep? Belangrijker argument nog: men kan zo'n vraag niet beantwoorden aan de hand van de gegevens zèlf, zonder in ogenschouw te nemen hoe ze verzameld zijn. Immers: al hadden we op alle subtests gemiddelden van 10 gevonden, dan nog kan onze onderzoeksgroep specifiek van samenstelling zijn. Een stap verder komen we dus misschien als we kijken hoe onze groep is samengesteld. Van de Amerikaanse ijkingssteekproef is bijvoorbeeld de verdeling bekend op de variabelen 'beroep van de vader' en 'genoten onderwijs van de vader'. We kunnen daar echter geen gebruik van maken, want:

- vergelijking van dergelijke variabelen tussen twee verschillende landen (en dus economieën en onderwijssystemen) is een hachelijke zaak;
- ook al vonden we identieke verdelingen in onze groep, de groep kan dan toch specifiek van samenstelling zijn. (De verklarende waarde van dergelijke stratificatievariabelen is nu eenmaal nooit 100%.)

Rest, ten derde, als enige mogelijkheid: het ijken van de WPPSI in Nederland via een aselekte steekproef uit de relevante populatie. Dat viel buiten onze mogelijkheden (wel zullen we onze gegevens publiceren).

Zoals gezegd: het ontbreken van Nederlandse normeringsgegevens maakt gebruik door ons van de gewogen scores nog niet zinloos. Wel is het zo dat een waardering van vooruitgangen *op zich* een hachelijke zaak is. En dat op een aantal gronden:

- de standaardafwijking is groter dan voor IQ's gebruikelijk en per gegevensgroep verschillend; dit is opgevangen door in de Tabellen 1, 3 en 5 over te stappen op de standaardfout van verschilscores, en op de standaardnormale verdeling;
- de betekenis van een bepaalde vooruitgang kan afhankelijk zijn van het beginniveau; het is gebruikelijk om intervalschaalkarakter te veronderstellen, maar een toename van 85 naar 95 betekent toch wel iets anders dan een toename van 105 naar 115, al was het alleen maar gezien procedures voor toelating tot het buitengewoon onderwijs;
- de betekenis van een bepaalde toename kan afhankelijk zijn van de ontwikkelings*snelheid*. Dat betekent, dat je eigenlijk minimaal twee pretestafnames (en ook twee posttestafnames) nodig zou hebben, om de groepen die je wilt vergelijken, te kunnen beoordelen op èn verschil in prestatie c.q. verschil in toename van prestatie, èn interval, èn leeftijd. Dan zijn we echter in een andere researchsituatie aangeland. De gegeven verantwoording gaat zover als onze opzet mogelijk maakt.
- De WPPSI- en de WISC-R-gegevens zijn niet zonder meer vergelijkbaar. Het kan niet anders of de WISC-R moet als een 'moeilijker' test gezien worden dan de WPPSI: De Amerikaanse normeringen gebruikend ligt het WISC-gemiddelde van de diverse experimentele groepen 0-14 punten onder het WPPSI-gemiddelde op 6-jarige leeftijd. Alternatieve overwegingen zijn verder: is het intervalschaalkarakter gelijk? Meet men wel hetzelfde op verschillende leeftijden?

Voor de effectconclusie beschouwen we als afdoende, dat analyses op de ruwe scores tot dezelfde conclusies leiden. Uiteraard worden daarbij 'leeftijd bij voortestafname' en 'interval tussen voor- en natest' als aparte variabelen ingebracht. Het leek ons weer te ver gaan om in plaats van de gewogen scores, omgezet volgens de manuals, 'eigen' gewogen scores te gebruiken in de presentatie.

13. *Over presentatie gesproken, heeft u eigenlijk wel statistisch getoetst?*

In het artikel over de evaluatie-aanpak in het GEON-project (Stokking, 1980b) zijn al enkele relativerende opmerkingen gemaakt over deze vorm van statistische analyse. Het komt er op neer dat je kunt twijfelen aan de geldigheid van toetsingsuitkomsten zowel als aan de bruikbaarheid ervan (ook al zijn ze statistisch gesproken geldig). De reden om statistisch te gaan toetsen is veelal dat men niet steeds weer opnieuw onderzoek kan doen. Er moet een conclusie of beslissing komen en een toets is daarbij een hulpmiddel. Nu blijkt een dergelijk 'knopen doorhakken' eerder adequaat bij het nemen van beslissingen dan bij het (wetenschappelijke proces van) trekken van conclusies. In die discussie willen we nu niet treden. Wat we wel naar voren willen halen is dat in het GEON-project sprake was van replicaties. Daarmee stellen we een traditie aan de orde die zegt dat niets zo overtuigend is als een zich (steeds weer) replicerend

verband (respectievelijk effect). Men zou kunnen redeneren dat men dan die met zoveel onzekerheden omgeven statistische toetsing niet nodig heeft. Maar zo simpel ligt replicatie-onderzoek ook weer niet. Twee levensgrote vragen zijn immers: welke (deel)onderzoeken zijn replicaties van elkaar? En: hoe de uitkomsten van meerdere onderzoekingen te waarderen? Bij de laatste vraag is men geneigd om te denken aan... statistisch toetsen. We stappen nu maar uit deze cirkel en sommen even op wat we daadwerkelijk aan toetsing deden.

Centraal stond steeds een twee-factoren variantieanalyse (meetmoment, experimentele conditie), die afzonderlijk werd gedaan voor de subtests en de totalen, equivalent aan een enkelvoudige analyse op verschilscores met als factor experimentele conditie. Daarbij kwamen covariantieanalyses, met als extra factoren/variabelen: 'school', 'sekse', 'leeftijd bij voortest' van het kind en 'interval tussen voor- en natest'. Met behulp van beide laatste covariabelen waren ook de ruwe gegevens analyseerbaar (zie boven). Tussen een aantal bijkomende variabelen (zoals 'beroep vader', 'klassegrootte', 'ervaring leerkracht', 'deelname aan het project') werden correlaties berekend. Weliswaar werden die ook op significantie getoetst, maar door het redelijk gelijk blijven van het aantal cases per ronde per experimentele groep (± 70) kwam dat er op neer dat een grens werd getrokken bij een r van ongeveer .20. Tenslotte werden (co)variantieanalyses gedaan met als factoren/variabelen: òf 'school' òf 'experimentele conditie'; en 'beroep vader' en/of 'klassegrootte' etc. (de bijkomende variabelen).

Het is hier even nodig er op te wijzen dat er steeds gebruik is gemaakt van het SPSS-pakket (Nie et al, 1975. Zie ook Berk en Francis, 1978). Dat had zo zijn beperkingen, zeker in de eerste periode van het project: géén genestelde structuur, géén repeated measurement design, géén multivariatie mogelijk bij de variantieanalyses (Het systeem van opties bij dat programma bood overigens ook voordelen; (zie ook Cohen (1968); Overall c.s. (1969, 1972, 1975); Woodward en Overall (1975)). Inmiddels zijn de mogelijkheden ruimer en overigens zijn er meer mogelijkheden dan SPSS. Maar de bovengeschetste relativering van het belang van significantietoetsing bracht met zich mee dat er bij afronding van de evaluatie niet meer veel tijd in is geïnvesteerd. Mede gezien de omvang van het gegevensbestand ligt er dus nog een heel terrein aan analysewerk braak. Met name ook: effect-grootteschattingen (Verberk, 1970) en modelvorming[2].

Er is tenminste één technisch argument voor onze keus te blijven bij de aanvankelijke beperkte SPSS-analysemogelijkheden: de door ons gehanteerde analyses waren 'conservatief', dat wil zeggen: ze pakken minder snel significant uit (dan bijvoorbeeld bij een repeated measurement design te verwachten valt). En dat is wetenschappelijk gesproken toch een goed gebruik. De kwestie van de keuze van een analysetechniek is een terrein vol discussie (zie Campbell en Erlebacher, 1970; Lord, 1967, 1969; Feldt, 1958; Edwards, 1974; Verberk, 1970; Kenny, 1975; Huck en McLean, 1975; Porter en Chibucos, 1974; Evans en Anastasio, 1968). In die discussie worden tal van technieken vergeleken op punten als assumpties, onderscheidingsvermogen, robuustheid, feitelijke nulhypothese, aard van de covariabele, interpretatiemogelijkheden. Een eenduidige uitkomst levert dat vooralsnog niet op of het zou moeten zijn dat men in onze researchsituatie (ontbreken van randomisering) het beste maar helemaal kan afzien van statistische toetsing. Immers: aan de voor het hele lineaire model essentiële voorwaarde van onafhankelijkheid tussen de componenten (factoren en covariaten en 'fout') zal dan meestal niet zijn voldaan. En: robuustheidsonderzoek strekt zich tot die fundamentele assumptie niet uit.

Al deze treurnis ten spijt hebben we toch getoetst, echter niet na een aantal assumpties gecheckt te hebben. Kruistabelleringen lieten zien dat over het algemeen variabelen als leeftijd, geslacht, klassegrootte, ervaring leerkracht, beroep vader, keurig min of meer gelijk verdeeld waren per meetmoment en experimentele groep (wat dus ook betekent dat er geen sprake was van selectieve uitval). Ook bleken voor de hoofdtoetsing (conditie x meetmoment) de varianties redelijk homogeen; bovendien was er sprake van grote en niet sterk verschillende steekproefomvangen. De homogeniteit van regressies hebben we niet gecheckt, wat niet zo'n bezwaar is omdat voorzover bekend heterogeniteit weer een conservatieve toets oplevert.
Helaas is niet bekend wat heterogene regressies doen in combinatie met heteroscedasticiteit en statistische afhankelijkheid. En met een dergelijke wirwar moet, zo is gebleken, rekening worden gehouden als de factor 'school' wordt ingebracht. Want 'beroep vader', 'klassegrootte', 'ervaring leerkracht', e.d. verschillen van school tot school en spreiden daarbinnen soms amper of helemaal niet.

Tegen de achtergrond van bovenstaande opsomming rapporteren we nu, dat het experimentele effect zelden significant uitviel en dat bij een hoog geschat onderscheidingsvermogen. Dit betekent dat er nogal wat verschillen waren van school tot school, binnen eenzelfde experimentele conditie en nogal wat verschillen van kind tot kind, ook binnen eenzelfde

school.

Dit brengt ons tot een laatste punt, namelijk het thema van de keuze van het analyseniveau. Cronbach (1976) schreef daar belangrijke dingen over. Zie ook Bronfenbrenner (1979). Ruimtegebrek verbiedt ons daar nader op in te gaan. Zie ons paper voor de ORD '80 (Stokking, 1980a). Alleen dit: Cronbach volgend is het juist in projecten als GEON essentieel of gemiddelde effecten substantieel zijn. Dat is duidelijk beleidsmatig gedacht. Dat de voorspelbaarheid naar individuele kinderen toe tegenvalt vinden wij ernstig. Het betrouwbaar zijn in de zin van steeds weer optreden van de (gemiddelde) effecten doet ons daar overheen stappen, waar het om verspreiding gaat. Er kan zinvol gewerkt worden aan sociale problemen, ook al zijn de oplossingen niet perfect.[3]

Noten

1. De rest van deze paragraaf is gebaseerd op de projectnota over verspreidingsaspecten (GEON, 1977).
2. Als er bijvoorbeeld iemand voor voelt om daarop een dissertatie te baseren zijn wij graag tot medewerking bereid.
3. Zie voor een 'Demystifying social statistics' verder Irvine et al. (1979).

Literatuur

De literatuurlijst behorende bij dit artikel is opgenomen bij het tweede, afsluitende deel, dat in het februarinummer verschijnt.

Curriculum vitae: zie Pedagogische Studiën, 1980 (57) p. 194.

Manuscript aanvaard 22-8-'80