GEON: Summatieve evaluatie (II)

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

Dit is het tweede gedeelte van het afsluitende artikel in de reeks in dit tijdschrift over GEON. Deze reeks startte in maart 1980. Dit laatste artikel gaat over de summatieve evaluatie. Het eerste deel (waaraan een samenvatting van het geheel voorafging) verscheen in het vorige nummer (januari 1981). Een inhoudelijke splitsing was, gezien de opbouw van de tekst, niet mogelijk. Aanbevolen wordt dan ook beide deel-artikelen als één geheel te lezen.

14. *Is er eigenlijk niet gewoon testgericht getraind?*

Onderzoek als dat van Sherrets en Quattrocchi lijkt te tonen dat dat niet zo'n vaart loopt (1979). Sandbergen besteedde er grondig aandacht aan in 1968. Hij stelt terecht dat soms niet uit te maken is 'in hoeverre men bij "opvoedkundige experimenten", waarbij men bijvoorbeeld de ontwikkeling van intellectuele functies tracht te stimuleren en het effect van zo'n programma evalueert met een test, in feite bezig is met coaching voor die specifieke test' (p. 503). Daarbij moet bedacht worden dat alleen problematisch is de invloed van factoren als 'coaching, oefening en "test sophistication" ', voor zover die niet 'worden gedekt door het begrip dat het instrument pretendeert te meten' (p. 502). Immers: onderwijs-effecten op concentratie, begrip e.d. zijn zonder meer wenselijk, zodat een hogere score als gevolg van zulke effecten relevante informatie geeft (cf p. 509). Verder koppelt Sandbergen 'test-sophistication' en 'coaching' duidelijk aan het doen van tests zelf, onder reële condities (pp. 505-6, 522).

We denken niet dat deze valkuil in het geval van GEON een bedreiging van de effectconclusie inhoudt, en wel omdat er tussen projectschoolgroepen en vergelijkingsschoolgroepen geen verschil in testervaring was: in beide experimentele condities werd hetzelfde test-programma afgewerkt. Bovendien werden de effecten waarover Sandbergen rapporteert bij oudere kinderen geconstateerd (cf pp. 506-7, 512), werden perioden van 3 tot 12 maanden bezien (p. 510) terwijl het bij de WPPSI-afnames in GEON om 19 maanden ging, bleek oefening het meest te helpen bij testen met een tijdslimiet (p. 511) wat bij de WPPSI amper voorkomt en bleken 'slimme' kinderen duidelijk meer dan 'domme' kinderen te profiteren (p. 521), terwijl we in GEON met achterstandssituaties te maken hebben. Tenslotte vermelden we dat Sandbergen rapporteert (p. 520) dat door oefening de spreiding toeneemt, terwijl in GEON over het algemeen (zeker op de projectscholen) de spreiding afneemt.

Wellicht blijft er ruimte voor twijfel, gezien het gebruik van de zgn. Takenreeksen in het project. Men zou dat testgerichte training kunnen noemen. Argwaan in die richting menen we echter te kunnen ontzenuwen door middel van de volgende argumenten.

- Het effect van de takenreekslesjes, afgemeten aan de vooruitgang op de bij die reeksen behorende toetsen, was weinig spectaculair. Het viel ronduit tegen.
- De meeste kinderen waarvan we de WPPSI-gegevens verwerken hebben *nooit* in een takenreeksgroepje gezeten.
- Subgroepen kinderen die aan twee of meer takenreeksen hebben meegedaan gingen op de WPPSI niet meer vooruit dan subgroepen kinderen die aan geen enkele takenreeks meededen.
- De relatie tussen de takenreekstoetsen en de WPPSI is, uitgedrukt in een produktmomentcorrelatiecoëfficiënt, beperkt: zo rond de .50.
- De ervaring met de takenreekstoetsen was op projectscholen en vergelijkingsscholen identiek, door de vergelijkingstakenreekstoetsafnames.
- Analyses van de relaties tussen projectdeelname en testresultaat laten zien dat het effect van het zitten in een klas waar de juf een takenreeks doorwerkt (ook al zit het kind niet in het groepje dat de lesjes krijgt) gemiddeld geringer is dan waar het een cursus betreft.

Het laatste is niet zo verwonderlijk. De cursussen waren in het project essentieel, dáár ging ook de meeste begeleidingstijd in zitten.

15. *Daar zegt U zowat: U heeft gewoon één groot Hawthorne-effect geboekt?*

De situatie die bij uitstek aanleiding geeft tot een 'Hawthorne'-interpretatie is die waarin ondanks forse programmavariatie effectiviteit alom wordt geconstateerd. Weikart (1975, p. 502) vond iets dergelijks. Er zij opgemerkt dat dergelijke algemene verklaringspogingen elkaar logisch gesproken vaak uitsluiten. Zo hoor je ook wel dat (bijna) alles in dit soort werk uiteindelijk teruggevoerd kan worden op een 'school-' of 'persoons'-factor. Maar hoe dan te verklaren dat op de 9 scholen die in de eerste experimentele ronde vergelijkingsschool waren, en in de tweede ronde projectschool, de testresultaten in beide perioden waren als in Tabel 8?

Tabel 8 *Testresultaten op 9 scholen, in beide rondes*

	1e ronde, als vergelijkingsschool	2e ronde, als projectschool
d	+ 5	+ 11
s	+ 0	+ 8

Toelichting:
d = doelgroep s = steekproef

Dit lijkt eerder op een Hawthorneeffect dan op een Schooleffect, zou men zeggen. Echter: de testresultaten, zoals eerder gepresenteerd in dit artikel, zijn niet overal even groot. Men zou dan kunnen veronderstellen dat er toch een 'School'-effect opgetreden is, ware het niet

- dat je daarmee niets verklaart;
- dat het lastig is om van een zo algemeen bedoeld mechanisme als 'Hawthorneeffect' te stellen dat het op de ene school wel optreedt, op de andere niet;
- dat er nogal wat verschil zit in preciese deelname aan het programma, van school tot school bekeken;
- dat 'testresultaat' en 'deelname' aan elkaar gerelateerd blijken te zijn, óók indien 'gecorrigeerd' voor een scholenfactor.

Het laatste moeten we toelichten. De WPPSI was niet bedoeld om de effectiviteit van afzonderlijke programma-onderdelen te evalueren. Maar nu er flink wat programma-variatie optrad èn geregistreerd werd, konden we niettemin vaststellen, dat niet elk programma-onderdeel even sterk gerelateerd is aan meetbare effecten op de test. De verwerkingsmethode staat beschreven in Stokking (1980[b]). We zien dan het volgende. (Zie voor een iets uitgebreidere weergave De Vries et al. (1980)). Op regioniveau analyserend zijn er indicaties dat 'Zelfstandig Werken' nauwer (positief) aan de testresultaten is gerelateerd dan 'Verwoorden van Probleemgedrag' en de 'Takenreeksen'. Op kindniveau analyserend is de samenhang tussen projectdeelname en testresultaat relatief het duidelijkst ten aanzien van de cursussen 'Zelfstandig Werken' en 'Voorspellen en Reageren' en de eerste takenreeks. De tweede takenreeks vertoont zo te bezien de minste relatie met testresultaat. Op kindniveau analyserend gaat 'méér onderdelen doorwerken' niet per se samen met grotere testresultaten.

Resumerend: op projectscholen wordt een gemiddeld duidelijk positief testresultaat geconstateerd, vergeleken met vergelijkingsscholen; binnen de projectscholengroep verschillen de resultaten al naar gelang de preciese deelname, ook indien gecorrigeerd voor de factor 'school'; het is niet zo dat hoe intensiever begeleid is hoe meer effect optreedt, dat hangt duidelijk af van wat er is gedaan. (N.B.: deelname en *voor*testgemiddelde zijn *niet* gerelateerd!)

Op schoolniveau analyserend zijn er indicaties dat waar weinig deelname plaatsvindt aan schoolprogramma-onderdelen, intensieve oudercontacten desalniettemin gepaard kunnen gaan met goede testresultaten, en dat waar wel veel deelname plaatsvindt de factor oudercontact er voor wat betreft de testresultaten minder toe doet. Dit laatste zou kloppen met wat Groenendaal (1978, p. 27) in zijn literatuuroverzicht stelt: 'Gezinsbeïnvloeding of ouderactivering droeg weinig extra bij tot de intellectuele ontwikkeling van de kinderen wanneer op school reeds een speciaal programma draaide'.

Bij wijze van illustratie geven we in Tabel 9 de gemiddelde vooruitgang op de WPPSI van de groepen kinderen op de projectscholen die respectievelijk 0, 1, 2, 3, en 4 semesters van hun kleuterschoolperiode bij een leidster zaten die op dàt moment de minicursus Zelfstandig Werken doorwerkte met haar klas. We zien dan dat het *per* programma-onderdeel kan voorkomen dat een intensievere confrontatie ermee gepaard gaat met een groter testresultaat. Maar dat gaat niet altijd op (in dit geval met name niet in de eerste regio). En waar het wel optreedt kan het moeilijk een Hawthorne effect genoemd worden, omdat 'méér van de ene cursus' meestal 'minder van de andere' impliceerde.

Het wordt tijd om wat nader te kijken naar de merites van een 'Hawthorne-interpretatie'. We denken dat wat McDill et al. schrijven (1969, p. 49-50) representatief is voor de wijze waarop over het alge-

Tabel 9 *Gemiddelde WPPSI-vooruitgang per telcategorie voor de minicursus Zelfstandig Werken*

aantal	1e ronde				2e ronde		
semesters	1e regio	2e regio	3e regio	4e regio	1e regio	2e regio	3e regio
0	18 (7)	−6 (3)	2 (3)	6 (21)	9 (5)	−2 (26)	
1	3 (41)	16 (58)	6 (51)	15 (18)	2 (27)	12 (20)	15 (74)
2	7 (6)	22 (12)	13 (26)	10 (25)	5 (32)	12 (27)	31 (7)
3	1 (14)						
4							

Toelichting:
0, 1, 2, 3, 4 is het aantal semesters dat kinderen bij een leerkracht in de klas hebben gezeten die op dàt moment die cursus deed. Tussen haakjes staan de aantallen kinderen waarover gemiddeld is.

meen aan een Hawthorne-effect wordt gedacht: 'Singling out people for special attention has been shown to affect their behavior. It is the well-known Hawthorne-effect. If compensatory education programs achieve improvement initially, it may be that the treatment itself has little to do with the gain and that any special attention to the enrollees would have achieved the same results. The potential presence of a Hawthorne-effect makes it hazardous to generalize results or to try to replicate them in another setting.'

Ons inziens moet men een 'Hawthorne-effect' niet als een bedreigende interpretatie zien. Inderdaad gaat het vooral om de generaliseerbaarheid. Maar als bepaalde ('experimentele') condities bevordelijk werken, waarom die dan niet in de voorwaarden voor een programma opgenomen? Voorzover we in GEON geconstateerd zouden hebben dat een bepaalde, zeer school-nabije, begeleiding en training, resultaat heeft, is dat op zich een waardevolle conclusie.

Toch wordt een Hawthorne-effect meestal als een artefact gezien, waarbij echter niet duidelijk is wat zo'n interpretatie nu precies inhoudt. Bracht en Glass (1968, pp. 457-459) brengen een handvol zeer verschillende zaken onder deze noemer. Strikt methodologisch bezien gaat het om een interpretatie waarin 'het in een experimentele situatie verkeren van de onderzochten' de feitelijke beïnvloedende factor is. In het historische onderzoek in de Hawthorne fabrieken was daarbij de afhankelijke variabele, de effectmaat: toename van produktie. Er zijn verschillende redenen waarom een 'Hawthorne'-interpretatie in het geval van GEON niet zo aannemelijk is.

1. Onze belangrijkste effectmaat: invloed op de ontwikkeling van kinderen, is niet simpelweg een kwestie van méér, onderwijzen is geen fabriceren.
2. We hebben meerdere effectmaten tot onze beschikking, waarbij het opvallende is dat ze allemaal dezelfde kant uitwijzen (zo blijkt bijvoorbeeld ook uit vragenlijstgegevens dat de cursus Zelfstandig Werken relatief beter aanslaat dan de cursus Verwoorden van Probleemgedrag; zo blijkt ook uit contacttellingen het belang van werken aan dit soort training in teamverband; e.d.).
3. Specifieke conclusies als bijvoorbeeld de positieve relatie tussen gebruik van de cursus Zelfstandig Werken en de testresultaten (die op elk analyseniveau weer terugkomen) vallen moeilijk te rijmen met een 'Hawthorne'-interpretatie.
4. Het is niet zo dat méér begeleiding per se positiever blijkt uit te pakken.
5. De gegeven begeleiding zou weinig effect hebben gesorteerd als de leidsters niet hadden geleerd 'wat te doen'.

Ook Cook (1968) concludeert dat het met een 'Hawthorne-effect' in onderzoek in het onderwijs wel meevalt (zie Baker [1968, p. 343]).

16. *Toch kunt U Uw resultaten maar gedeeltelijk verklaren*

'We still know little about how individual children develop in these programs' (Takanishi, 1979, p. 158).

Het accent dat bij GEON ligt op testgegevens als evaluatiecriteria is al herhaaldelijk bekritiseerd. Eén vorm van kritiek zegt, dat de verklaring door middel van nauwkeurige procesgegevens ontbreekt. Nu hebben we wel vrij preciese registratie-gegevens, bijvoorbeeld aangaande deelname aan het programma en ook informatie rondom de afzonderlijke cursussen, zoals video-opnamen, maar dat neemt niet weg dat het verklaringsmodel deels een 'black box' blijft.

Elders bespreken wij daarbij verschillende aspec-

ten, hier willen we ingaan op de vraag of niet uitgebreidere observaties en interactie-analyses het gat tussen inservicetraining enerzijds en meting van effecten bij kinderen anderzijds hadden kunnen opvullen.

Allereerst moeten we meedelen dat dat praktisch gezien niet haalbaar was.De beschikbare tijd had uitgebreide observatietraining, observatie en analyse van de gegevens niet toegelaten. Bovendien zou dat bij een aantal kleuterleidsters een ontoelaatbare ingreep in de opbouw van de samenwerkingsrelatie hebben betekend (methodologisch gezien was een 'reactieve meting' het resultaat geweest, met alle representativiteits- en generaliseerbaarheidsproblemen van dien, nog afgezien van innovatie-overwegingen).

Vervolgens delen we het relatieve pessisme dat Groenendaal (1978, pp 11-12) verwoordt aangaande verklaringsmogelijkheden in zulk onderzoek: verbanden zijn vaak specifiek, respectievelijk: 'variabelen welke de directe ervaringsmogelijkheden van kinderen beschrijven, blijken een gering aandeel te leveren in de predicties.'

Colthof (1979, pp. 126 vlgg, 183 vlgg) blijkt ondanks minitieuze observatie van lesjes, minitieuze analyse van begeleidingsgesprekken, en grote aandacht voor niet-gehaalde implementatiedoelen niet in staat tot 'verklaring' van geconstateerde effecten bij kinderen op de manier die men lijkt te verlangen.

Nu zou men kunnen denken dat het wellicht zelfs bij Colthof nog niet grondig genoeg is geprobeerd, maar het streven naar predictiemogelijkheden via 'instruments that will produce usable indexes utilizing brief samples of classroom behavior' (Baker, 1968, p. 353) is ondertussen toch wel oud genoeg om zijn waarde bewezen te kunnen hebben.

Rosenshine en Furst (1973, p. 148) stellen dat de meeste gegevens zoals verzameld met observatie-instrumenten werden verkregen bij bepaling van een interbeoordelaarsbetrouwbaarheidscoëfficiënt. Daaruit zou je kunnen concluderen dat de bedoelde aanpak nog niet echt is geprobeerd, maar plausibel lijkt dat dat wèl het geval is, maar door teleurstellende resultaten vaak niet is gerapporteerd. Wellicht zijn velen al tot een conclusie gekomen als Rosenshine en Furst op p. 175: 'It is possible that the patterns of effective teaching for different ends are so idiosyncratic that they will never be isolated.' Dat zou dan *niet* betekenen dat we geen effectieve programma's zouden kunnen bouwen, maar dat er grenzen zijn aan de verklaringsmogelijkheden (zie ook McDill et al., 1969, p. 66: op verklaringen kan soms niet gewacht worden; Campbell, 1975, p. 182). Veelzeggend is in dit verband dat Rosenshine en Furst denken te kunnen stellen dat 'high inference' instrumenten én beter discrimineren tussen scholen/leerkrachten, én betere predicties kunnen leveren van effecten bij leerlingen, dan uitgebreide gedragstellingen (pp. 133, 136, 158).

Uit bovenstaande overwegingen zou je de conclusie kunnen trekken dat we meer 'high inference' metingen in de klas hadden moeten verrichten. Zie ook Lowyck (1979, p. 427), die eveneens stelt dat een beperking tot 'observeerbaar onderwijsgedrag' onvoldoende lijkt. Maar variabelen als bijvoorbeeld 'warmth' (McDill et al., 1969, p. 12) zijn juist weer moeilijk 'hard' te maken. En een oplossing van dit dilemma door over te stappen van observatietechnieken naar het vragen van het oordeel van de leerkrachten valt ook niet één twee, drie te verwachten, als bedacht wordt hoezeer meningen kunnen verschillen van andersoortige feiten, hoe onbetrouwbaar het geheugen van mensen kan zijn, hoezeer factoren als 'sociale wenselijkheid' een rol kunnen spelen. Zie voor het laatste b.v. Hofstee (1965). Dergelijke methodologische haken en ogen aan het gebruiken van meningen van betrokkenen als evaluatieve informatie zijn zeker ook op het GEON-project van toepassing. Ons vierde artikel (november 1980) was voor een niet onbelangrijk deel gebaseerd op oordelen van de intermediaire krachten van het project. Oordelen van intermediaire krachten kwamen op basis van bepaalde wijzen van 'in het project staan' tot stand. En het is niet alleen zo dat die oordelen de grondslag vormen van veel van in het kader van de evaluatie verzamelde gegevensmateriaal. De oordelen (bijvoorbeeld aangaande het 'lekker lopen' van een cursus, het al dan niet geslaagd zijn van een ouderavond) bepaalden mede het begeleidende werk zèlf! Reden genoeg om ze te evalueren. Dan blijkt het volgende.

Negatieve indicaties lagen voor de intermediaire krachten vooral in de organisatorische punten (niet halen van gemaakte planning), positieve in het gebruiken en blijven gebruiken van het geleerde en in een actieve houding van de leerkracht. Via de functie van intermediaire kracht hebben we ongetwijfeld meer zicht gekregen op een aantal processen dan anders het geval zou zijn geweest. Bovendien vormden de meningen die we dusdoende vrij systematisch konden verzamelen een checkmogelijkheid op conclusies die op basis van andere gegevens getrokken werden. Opvallend, en bemoedigend consistent met de analyse van de testresultaten respectievelijk de uitgevoerde contactenanalyses (zie De Vries et al. 1980) is dat de intermediaire krachten het vaakst de minicursus Zelfstandig Werken en het contact ou-

ders-school noemden als 'wel gelukt'.

Tot dusver bespraken we verklaringsmogelijkheden in relatie tot het beschikken over, respectievelijk gebrek aan gegevens. We kunnen 'verklaren' ook statistisch opvatten. Dan heet het verschijnsel dat het ene kind meer geprofiteerd heeft dan het andere: 'binnenvariantie' (cf. Light en Smith, 1970, pp. 6, 13, 15). Dat verschillen binnen experimentele condities vaak van dezelfde orde van grootte of zelfs groter zijn dan die ertussen is een droevige, maar helaas maar al te vaak geconstateerde zaak. Zie bijvoorbeeld ook Baker (1968, p. 345), Weikart (1975, p. 505), Berman en Pauly (1975, pp. VIII-IX), Groenendaal (1978, p. 135), House et al. (1978, p. 154). Als 'scholenfactor' kan dit aan personen worden toegeschreven, maar waarschijnlijk is veel afhankelijk van organisatorische condities, implementatievariatie e.d.

Coulson, (1978, p. 51) zegt n.a.v. een poging om succesvolle van niet-succesvolle 'sites' te onderscheiden, dat er sprake is van een continuum van geconstateerde effectiviteit. Dat ligt voor de hand; wat hij daarna meedeelt is veel ernstiger: 'There was a considerable shift in the composition of the "succesful" and "unsuccesful" groups; that is, some preciously succesful sites became unsuccesful, and vice versa'. Om dan maar af te zien van elke overall effectiviteitsbepaling, zoals wordt voorgesteld (Coulson, p. 17), gaat echter te ver. Weliswaar middel je dan over ongelijksoortige eenheden, maar de mogelijke beleidsimplicaties kunnen zodanig zijn dat een (uiteraard) niet geheel voorspelbaar zijn van effect voor lief wordt genomen. Vaak was de 'binnenvariantie' zo groot dat het eindoordeel er een van 'geen effect' werd, en dat is natuurlijk ook weer niet de bedoeling (cf. McDill et al., 1969, p. 2). Het is plausibel dat verschil in effectiviteit gerelateerd is aan verschil in leerkrachtgedrag, zie bijvoorbeeld McNeil (1968) en Coulson (1978, p. 52). Factoren die de laatste noemt, spreken ons zeer aan: gerichte instructie, en aandacht voor gelijke kansen. Wij spreken van: precisie en systematiek, en positieve discriminatie.

Uiteraard blijf je in statistische analyse met een zeker percentage 'onverklaarde variantie' zitten, wat in dit soort werk veelal betekent, dat er verschillen optreden van leerkracht tot leerkracht. Kwantitatief was zo'n 'leerkrachtfactor' echter niet te analyseren, deels doordat de kinderen bij meerdere leidsters zitten (en verschil in deelname aan het project er ook weer doorheen loopt), deels doordat bijvoorbeeld het voortestgemiddelde per leerkracht, dus eigenlijk per klas of althans groep testkinderen, ook op een en dezelfde school nogal uiteen kan lopen. Omstreden kwesties als het intervalschaalkarakter en het regressieverschijnsel verbieden dan als het ware verdere analyse.

Voor een meer filosofische verhandeling, tenslotte, omtrent 'verklaren', zie Beerling et al. (1970). Zie ook Van den Duyn (1979) en Stokking (1980[d]).

17. *Zijn Uw evaluatiecriteria niet een beetje eenzijdig?*

Naast de nadruk op 'overall impact assessment' is waarschijnlijk niets zo controversieel als de keuze van de meetinstrumenten c.q. evaluatiecriteria (cf. Coulson, 1978, p. 29). De bekende, grote evaluatiestudies in de VS (Head Start, Follow Through, e.a.) hanteerden steevast bestaande, gestandaardiseerde instrumenten. Argumenten die daarvoor pleiten waren: beschikbaarheid (relatief geringe kosten), gemakkelijke toepasbaarheid, interpreteerbaarheid via normen, en 'algemene aanvaardbaarheid' (House et al., 1978, p. 138; zie ook Coulson, p. 30 vlgg). Zulke instrumenten zijn niet curriculum-specifiek. Stevens vindt (1973, p. 29) dat daarmee dan ook geen evaluatie van een curriculum mogelijk is, maar daarbij heeft hij eerder een formatieve dan een summatieve evaluatie op het oog. De summatieve evaluatie bij GEON maakte ook gebruik van niet-curriculum-specifieke instrumenten. Hier willen we betogen dat dat niet toevallig is.

De Vries schrijft over de UCP's (1973, p. 1): 'In algemene termen kan gezegd worden dat de effecten waarnaar gestreefd werd niet voldoende operationeel gedefinieerd waren bij de start van de Utrechtse Compensatie Programma's (U.C.P.'s), hetgeen een te globale keuze van evaluatie instrumenten met zich mee bracht.' Dat kan niet gezegd worden in het geval van GEON. Effecten van de inservicetraining waren bedoeld algemeen van aard te zijn, dóórwerkend in het dagelijkse werk op allerlei zogenoemde 'kleine punten', onder handhaving van het 'gewone' kleuterschoolprogramma. Naar onze mening geldt dan de overweging, zoals Anderson et al. (1978, p. 163) die geven: 'Faced with a program having many goals of varying degrees of measurability and emphasized differently in its various organizational components, it seems appropriate to apply a common measure stick to each component...'.

Het GEON-programma kent meerdere doelen, die bovendien dynamisch van aard zijn (een ideaal leerkrachtgedrag wordt niet omschreven; verdere verbeteringen blijven steeds mogelijk). Je zou kunnen stellen dat we werken met een permanent verschuivend einddoel, een streefrichting. Gedragsgerichte doelformulering en daaraan gekoppelde evaluatie wordt dan lastig. (Vergelijk ook Light en Smith, 1970, p. 2 en Mellenbergh et al., 1968, p.

609).

De door ons gevolgde werkwijze lijkt consistent, als je mag verwachten dat algemene ontwikkelingsaspecten bij kinderen samen zullen hangen met de geïntegreerde, diverse werkmechanismen die we op 'leerkrachtniveau' op het oog hadden (vergelijk Madaus et al., 1979, pp. 221, 223). De algemene effecten bij kinderen zul je dan ook met algemene instrumenten dienen te meten. We gaan nu niet in op verschillen tussen 'achievement'- en 'aptitude' tests (zie House, 1978, p. 139; Madaus, 1979, pp. 207, 217, 220; Takanishi, 1979, p. 150).

Resumerend onderstrepen we de redenen die Stevens (1973, pp. 32-33) opsomt voor gebruik van intelligentietests bij evaluatie-onderzoek:

- het algemene karakter ('groot bereik');
- de voorspellende waarde t.a.v. schoolprestaties (zie bv. ook Hindley and Owen, 1978, p. 330);
- de overtuiging dat ontwikkelingsverschillen in belangrijke mate door de omgeving worden bepaald.

Uiteraard was het anno 1973 niet vanzelfsprekend om tests te gaan gebruiken. In sommiger ogen was dat een beetje een anachronisme. Maar het project wilde in ieder geval iets hebben aan kwantitatieve evaluatie-gegevens. Overigens was er ook binnen het project zelf geen consensus over. In twee van de vier regio's zag men het gebruik van tests niet zo zitten. In één daarvan was dat uiteindelijk ook wel te merken aan de gebrekkige wijze waarop met het vastgestelde design werd omgesprongen. In de andere verdween de weerstand overigens goeddeels toen met behulp van de testgegevens effect van eigen werk bleek te kunnen worden aangetoond.

Wat WPPSI en WISC-R zelf betreft: ook dáárvoor geldt de voorspellende waarde (zie bijvoorbeeld Bishop en Butterworth, 1979, p. 156, t.a.v. de WPPSI), de mogelijkheid om er experimentele effecten mee aan te tonen (zie bijvoorbeeld McDill et al., 1969, p. 64, t.a.v. de WISC) e.d.

Het bezwaar dat De Bruyn et al (1979) formuleren t.a.v. de WISC-R, namelijk mogelijk onverantwoord gebruik voor besluitvorming over individuen doordat een Nederlandse normering ontbreekt, geldt uiteraard niet bij vergelijkend evaluerend gebruik zoals in GEON. Gebruik van (delen uit) de WPPSI voor onderzoeksdoeleinden is in Nederland niet ongebruikelijk inmiddels, zie bijvoorbeeld: Van den Bos (1979), Schroots (1979), Mens (in Van Kemenade, 1973), Stevens (1973) (naast de WISC).

Gebruik van de WPPSI en WISC-R gezamenlijk is ook niet zo nieuw (zie onder meer het al genoemde artikel van Bishop en Butterworth, 1979, en Rasbury et al., 1977). Aan de keuze voor deze instrumenten lagen de volgende overwegingen ten grondslag (zie voor de WPPSI ook: Kohnstamm, 1968; en voor beide: Sattler, 1974):

- de in de UCP gebruikte AKIT had een voor GEON te beperkte range;
- de SON discrimineerde onvoldoende bij jongere kinderen;
- een deviatiescore is acceptabeler dan een MA-concept;
- door de beschikbaarheid van profielen is eventueel aan leerkrachten meer informatie te geven (dan één IQ);
- WPPSI en WISC sluiten op elkaar aan, waardoor follow-up evaluatie relatief minder problematisch is;
- de WPPSI is, ook voor kinderen, plezierig om te doen;
- de Wechsler-serie is internationaal bekend, wat de communicatie bevordert;
- de psychometrische eigenschappen zijn goed.

Tot zover over de tests. Daarnaast waren er ook in GEON wel degelijk nog andere criteria op kindniveau.

We hebben steeds het begrijpend kunnen lezen aan het einde van de 2e klas GLO als een belangrijk evaluatiecriterium gezien. Daarvoor hebben we de toets Begrijpend Lezen II gehanteerd (zie Wit en Van Soest, 1975). We geven in Tabel 10 een overzicht.

Tabel 10 *Resultaten Begrijpend Lezen Toets II*

exp. groep	1e regio	2e regio	3e regio
pd	26 (30)	27 (19)	30 (16)
ps	32 (34)	32 (23)	31 (36)
vd	30 (13)	32 (9)	33 (7)
vs	–	35 (24)	35 (28)

Toelichting:
pd = projectscholen doelgroep
ps = projectscholen steekproef
vd = vergelijkingsscholen doelgroep
vs = vergelijkingsscholen steekproef
Tussen haakjes staan de aantallen.

Opvallend is uiteraard allereerst dat de aantallen kinderen waarvan we de resultaten op de BLT II konden verzamelen aanmerkelijk geringer zijn dan we eerder in het algemeen m.b.t. de follow-up evaluatie meld-

den. Dat komt grotendeels doordat besloten was om de toets alleen (klassikaal, door de leerkracht) te laten afnemen waar ten behoeve van de WISC-R afname steekproeven waren getrokken. Dat wil zeggen: op die lagere scholen waar de meeste kleuters terecht waren gekomen. Aangezien de verspreiding over de lagere scholen groter was dan we verwachtten (zie eerder in dit artikel) vallen de aantallen nu wat tegen. De BLT II laat weinig verschil zien tussen de onderscheiden experimentele groepen, zeker de maximumscore van 40 en de standaardafwijking van 7 in aanmerking genomen.

Vanuit het project is steeds gesteld dat verder een belangrijk evaluatiecriterium zou zijn of er van de begeleide scholen minder kinderen verwezen worden naar het buitengewoon onderwijs. Maar: minder dan wat? Je kunt kijken of er minder kinderen naar het BUO gaan dan van de vergelijkingsscholen. Maar het gaat om zulke kleine aantallen dat dat niet zoveel zegt. Je kunt kijken naar de percentages die in eerdere jaren golden. Maar die verschillen per regio, en regionale gegevens hieromtrent blijken niet steeds achterhaalbaar. Bovendien verschuiven die percentages in de loop van de tijd. Zou er in de tijd dat het project werkt algemener sprake zijn van afname in verwijzing naar BUO dan kunnen we een eventueel kleiner percentage t.a.v. begeleide kinderen niet op het conto van GEON schrijven. Een en ander komt er eigenlijk op neer dat het al dan niet verwezen worden, en daarmee ook de uiteindelijke percentages, van zoveel factoren afhankelijk zijn waar we onvoldoende zicht op hebben, dat we een dergelijk gemakkelijke vergelijking maar vergeten. Maar er zijn bij schoolloopbaan meer criteria dan alleen verwijzing naar BUO. Bijvoorbeeld ook het zittenblijven. De gegevens daaromtrent zijn wat exacter en completer te achterhalen. Maar ook dan blijven het kleine aantallen, te klein om er statistisch iets mee te doen.

We zijn blij dat we bovenstaande overwegingen expliciteerden (paper ORD'79), toen we nog slechts de beschikking hadden over de schoolloopbaan gegevens van de 1e regio. Die zagen er namelijk gunstig uit. Inmiddels zijn ook de gegevens van de 2e en 3e regio bekend, en nu blijkt dat we dit criterium als project niet gehaald hebben. In Tabel 11 staat een overzicht van de drie eerste rondes.

Behalve testgegevens, toetsgegevens en schoolloopbaangegevens hebben we nog andere criteria gehanteerd bij de evaluatie van ons werk. Die betroffen met name wat de volwassenen, die met de kinderen te maken hebben, *deden:* leerkrachten, ouders, e.a. Over observaties in de klas, contacten ouders-school, e.d. rapporteerden we echter al in De Vries et al, 1980. Sommige critici stellen dat gezien ons inservicetrainingsprogramma onze 'eigenlijke' evaluatiecriteria bij de leerkracht liggen. De reden om bij GEON te proberen een grondige evaluatie via effectmeting bij de kinderen van de grond te krijgen is de gedachte, dat een goede implementatie zulke effecten niet garandeert. Anders gezegd: ook al ben je overtuigd van de waarde van bepaald leerkrachtgedrag, het blijft een veronderstelling die getoetst moet worden (zie Rosenshine en Furst, 1973, pp 126 en 151). Overigens is een accent op 'pay-off evaluation' (zie bijvoorbeeld McDill et al, 1969, p. 47 – de term is van Scriven) naar onze smaak vaker terecht. Zie bijvoorbeeld de redenering van Smets (1979, p. 45): 'Over het geheel genomen ligt het zwaartepunt van de waargenomen resultaten in de attitudeverandering van de leerkrachten. De veranderingen in de individuele hulpverlening zijn de eerste vruchten van het stimuleringswerk die direct ten goede komen aan de kinderen'. Dit laatste is namelijk nog aan te tonen.

Tabel 11 *Schoolloopbaangegevens van drie eerste rondes*

	1e regio		2e regio		3e regio	
	proj.sch.	verg.sch.	proj.sch.	verg.sch.	proj.sch.	verg.sch.
aantal kinderen waarvan de gegevens, totaal	70	24	76	51	83	54
een jaar extra op de kleuterschool gezeten	3	4	2	0	2	0
in 1e klas blijven zitten	5	6	11	5	16	4
naar BLO verwezen	2	3	1	2	4	0

18. *Had U zich in Uw project wel zo moeten concentreren op summatieve evaluatie volgens de quasi-experimentele traditie? Er zijn toch nieuwere inzichten?*

Een recent artikel als dat van Glass en Ellett (1980) bespreekt een 7-tal evaluatieopvattingen. Als daaruit iets duidelijk wordt is het wel dat een simpele keuze voor 'de beste' niet mogelijk is. Er is geen 'beste' evaluatieaanpak. Een rangordening van de verschillende aanpakken naar kwaliteit of bruikbaarheid valt anders uit al naar gelang de criteria die je hanteert, waaronder een criterium als de specifieke kennisinteresse. (Zie ook Van den Bosch, 1975) Het GEON-project begon in 1972-1973 met een bij uitstek procesgerichte opvatting over evalueren:

- Buitenstaanders moeten kunnen beoordelen welke effecten zijn opgetreden en met welke oorzaak.
- De projectuitvoerenden moeten voortdurend feedback krijgen op hun handelen.

Daardoor waren een aantal gangbare onderzoekstechnische keuzen afgesneden. Zoals het gebruik van indirecte meetprocedures als attitudeschalen. En zoals het hanteren van meer geavanceerde mathematische analysetechnieken. Voor het laatste is immers een daarop toegesneden gegevensverzameling nodig. En dat zat er strikt genomen al daardoor niet in, omdat het gezien de relatie die de OBD's met het veld hebben het door hen onmogelijk werd geacht om bij de keuze van de scholen een aselecte procedure te volgen.

Wat ook wel genoemd mag worden is de geringe omvang van de evaluatiestaf. Alleen al daardoor moest er gekozen worden. Die keuzen werden in 1974-1975 gemaakt, dus zoals helaas maar al te gebruikelijk terwijl het project al draaide:

- het inbrengen van onderzoeksmatige overwegingen, naast de evaluatiedoelstellingen van feedback geven en beleid maken;
- het geven van prioriteit aan de uiteindelijke beleidsevaluatie (in de zin van effectbepaling), ten koste van de feedback-functie;
- het invoeren van een administratiesysteem om t.z.t. de waarschijnlijkheid van alternatieve verklaringen voor effecten te kunnen beoordelen;
- het achterwege laten van de ontwikkeling van een geformaliseerde formatieve evaluatie.

Een andere beperkende factor was gelegen in de interne structuur van het project, met alleen al door de geografische verhoudingen een forse decentralisatie. De regionale autonomie was uiteraard betrekkelijk. Maar vanuit de evaluatie gezien was er weinig aan te doen dat men ter plekke liever het experimentele effect maximaliseerde dan de beoordeelbaarheid van dat effect. Een resultaat van een en ander was dat steeds onderhandeld moest worden met de regio's aangaande evaluatie. Hetgeen niet alleen tijd kostte, maar ook compleetheid, tijdigheid en vergelijkbaarheid van informatie bedreigde.

Toch is de vraag terecht of de gekozen opzet een goede keus was. Is het nodig om allerlei problemen zoals in het voorgaande op te werpen? Kunnen ze niet omzeild worden door de evaluatie anders in te richten?

Een belangrijke vorm van kritiek stelt dat men in zo'n project *te veel tegelijk* wil: actie, evaluatie, fundamenteel onderzoek, etc. (vergelijk Baker (1968; zie o.a. p. 354)). Sommigen zien in de combinatie juist voordelen (Anderson (1978, p. 227); Scriven (1972)). Een actuele controverse is verder in het onderscheid tussen *kwantitatieve en kwalitatieve gegevens* opgesloten (zie o.a. Stake (1974), Schutz (1968), Parlett en Hamilton (1972), Shipman (1973, p. 168), McCall en Simmons (1969), Meltzer et al. (1977), Duintjer (1977), Giorgi (1978)). Andere bezwaren bij de gangbare research-benadering komen voort uit de onduidelijkheid van de uitkomsten, waarbij vaak allerlei discussie plaats vindt over de beste *wijze van (statistisch) analyseren.*

Als men Cronbachs commentaar op de Head Start evaluatiediscussie leest (1976, p. 8.13): 'I may as well say at the outset that my preference among that analyses has changed more than once as I have studied the problem, and that I am not convinced that I now know what should have been done with these data', kan men zich voorstellen dat alleen al op grond dáárvan gepleit zou kunnen worden voor 'N=1 studies' (cf. Colthof, 1979, p. 214-215).

We zijn het er niet mee eens dat dan de generaliseerbaarheid zoek zou zijn. Ten eerste is, door moeizame selectie van onderzoekseenheden en experimenteel verlies bij vervolgmetingen de representativiteit van traditioneel uitgevoerde studies voor de nationale populatie ook twijfelachtig (vergelijk Smith/Bissell, 1970, p. 68). En bij dit soort experimenten gaat het ook eigenlijk niet om statistische generalisatie, maar om theoretische. Als men op grond van een case-studie inzicht heeft gekregen in 'oorzaak-gevolg relaties' is er een basis voor generalisatie. (Zie ook Groenendaal, 1978, p. 46; Weikart, 1975, p. 492).

Aanvankelijk zagen we een *beslissingsgerichte aanpak* (zie bijvoorbeeld Edwards en Guttentag, 1975; Stufflebeam et al., 1972) als een te kiezen alternatief. Gaandeweg bleek ons enerzijds dat een volledig doorvoeren daarvan een voorafexplicitering

van besluitvormingssituaties vereist (wie/wanneer/waarover/aan de hand waarvan) die verder gaat dan in zulk werk opgebracht kan worden. Aan de andere kant lijkt ons een scheiding tussen een conclusiegerichte en beslissingsgerichte aanpak kunstmatig aangezien er altijd sprake zal zijn van te nemen beslissingen, waarvoor informatie moet worden verzameld, respectievelijk waartoe onderzoek plaatsvindt.

Een onderscheid tussen *formatieve* en *summatieve* evaluatie is gangbaar (zie bijvoorbeeld Stevens, 1973, p. 31). Een bezwaar tegen een al te groot accent op het laatste is dat men er dan waarschijnlijk niet zoveel van kan leren. De grote evaluatie-onderzoekingen t.a.v. ondermeer Head Start waren sterk gericht op zgn. 'impact analyses' (cf. Coulson, 1978, p. 21), waarbij zelfs niet gericht gekeken werd welke programma's meer respectievelijk minder effectief waren (op. cit. p. 9; Smith en Bissell, 1970, p. 63), laat staan dat de aandacht gevestigd was op mogelijke verbeteringen in de programma's.

Al eerder werd aangegeven dat expliciet is afgezien van een geformaliseerde formatieve evaluatie-opzet, zoals bijvoorbeeld in Slavenburg (1978) beschreven. Dat werd enerzijds ingegeven door twijfels aan de bruikbaarheid van vooraf specificeren van (herschrijvings) beslissingen. Dat aspect – al dan niet 'decisionistisch' werken – laten we verder buiten beschouwing. Aan de andere kant was al snel duidelijk dat het onderscheid formatieve evaluatie–summatieve evaluatie voor ons lastig te handhaven was. De programma-variabelen waren moeilijk constant te houden; dat was zelfs niet de bedoeling. De cursussen werden gebruikt op een aan elke school en leerkracht aangepaste wijze. Anders gezegd: voorop stond de ondersteuning van het schoolteam, niet het zo gecontroleerd mogelijk onderzoeken van een (precies te omschrijven) programma op effectiviteit. Nu kan men met deze stand van zaken minimaal twee kanten op:

- het programma is geen constante, dus is summatieve evaluatie onmogelijk; de evaluatie kan dus alleen formatief zijn, sterker: de werkwijze is daar juist op gericht;
- door te middelen over allerlei uitvoeringsvariaties kan men wel degelijk van summatieve evaluatie spreken, temeer waar elk gebruik in de praktijk (bij verspreiding) toch ook op allerlei wijzen zal plaatsvinden. Door de variatie tijdens de experimentele periode kan men juist beter generaliseren; en voor formatieve evaluatie is juist méér nodig om precies te weten over welk (stukje) programma je praat; constantie of liever gezegd specificeerbaarheid, is daarbij – op een andere schaal! – juist essentiëler.

Ons standpunt was het tweede. Dat komt doordat we nogal onder de indruk zijn van allerlei beoordeelbaarheidsproblemen waar we juist bij onze summatieve evaluatie minder last van hebben vanwege het feit dat het daar om testafnames bij kinderen gaat. Zie voor deze discussie over onderzoekstypen en evaluatieopvattingen verder ook Scriven (1974), Stufflebeam (1974), Cronbach (1976), Cook and Campbell (1979), meer de relatie met het beleid betreffend: House (1976), Welters (1977), van de Vall (1979), en, over actie-onderzoek: Moser (1977) en Heinze (1975).

Het stemt tot nadenken dat er serieuze onderzoeksproblemen zijn die door dergelijke 'methodendiscussies' amper of niet geraakt lijken te worden. Eén van de belangrijkste is wellicht: het meten van veranderingen. We worstelden daarmee al bij 'statistische regressie' en 'testgerichte training'. We gaan verder niet in op het werken met verschilscores in een groeisituatie met ongelijke start (Campbell en Erlebacher, 1970; Campbell en Boruch, 1975), stijgende trends in IQ-gemiddelden (Hindley en Owen, 1978, pp 330, 335), en validiteitsproblemen (Bereiter, 1976, Roskam en van der Sanden, 1974).

19. *En tòch. Hebben we nu ineens iets positiefs? Is het wel generaliseerbaar?*

In Tabel 12 staan de testresultaten op de scholen, die in de eerste experimentele ronde projectscholen waren, tijdens de 2e ronde.

Ons eerste antwoord is: ons werk heeft in ieder geval een investeringskarakter, want waar amper of helemaal niet meer wordt begeleid (eerste en tweede regio), behalen de eerder begeleide kleuterleidsters, althans bij de 'doelgroep', dezelfde of zelfs nog betere testresultaten als tijdens de periode dat ze begeleid werden. En waar nog wel begeleiding plaatsvindt, met verminderde intensiteit (derde regio), profiteert weer ook de 'steekproef' fors (en dus de hele klas). Er is dus wel degelijk sprake van generaliseerbaarheid, in die zin dat niet blijvend (zeer intensieve) begeleiding nodig is. Anders zou onze werkwijze inderdaad niet geschikt zijn voor algemenere toepassing.

Hierbij enkele kanttekeningen. Zo bestaat de indruk dat kleuterleidsters soms actiever worden als de school niet langer begeleid wordt. Daarvoor zijn diverse oorzaken aan te geven, zoals: dingen worden niet meer voor ze gedaan, en: ze beseffen dan pas goed het nut van een aantal zaken. Verder waren ook de intermediaire krachten van oordeel dat het werk hen in de tweede ronde beter afging, en dat toen meer dingen 'lukten'. Tenslotte wijzen we erop dat wat

Tabel 12 *Testresultaten 1e ronde-scholen tijdens 2e ronde*

regio	exp. groep	aantal kinderen	voortest $\bar{x}$	voortest s_x	natest $\bar{x}$	natest s_x	verschil	S_{ed}
1	d	16	78	11	92	14	+ 14	4,9
	s	49	98	16	104	15	+ 6	5,7
2	d	10	90	15	108	13	+ 18	4,3
	s	44	110	17	114	15	+ 4	4,9
3	d	48	80	16	96	16	+ 16	3,9
	s	57	95	18	107	16	+ 12	4,2

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

Smith en Bissell (1970, p. 98) constateren: 'The history of pre-school education for disadvantaged children has shown consistently that programs become more effective over their first few years of operation as initial problems are worked out', ook bij ons in de testresultaten tot uitdrukking kwam.

Het is jammer dat follow-up evaluatie, zowel naar de leidsters als naar de kinderen toe, slechts mogelijk was voor de eerste rondes. Van de tweede rondes zou je immers nog 'meer' kunnen verwachten.

Toch blijft de generaliseerbaarheidsvraag staan. Nu kan men langs meerdere wegen generaliseren. Men kan veralgemenen naar *situaties en personen* toe. Wat dat betreft kunnen we inmiddels wel vertrouwen hebben dat we gewerkt hebben in situaties die representatief zijn voor achterstandssituaties, en houden we alleen de persoonsfactor over (leerkracht, schoolteam) als moeilijk in te schatten voorwaarde.

Maar je kunt ook veralgemenen naar een *criterium* toe. Dat betekent dan dat je uit toegenomen testscores tot andere resultaten probeert te concluderen, via voorspellende waarde voor schoolsucces, via psychologische theorie omtrent transfer, via maatschappelijke betekenis etc. Ook dat is niet eenvoudig.

Tenslotte kun je veralgemenen naar *werkwijze:* als zo'n begeleiding met zulke programma's effect had, wat mag je dan verwachten van deels andere begeleiding (b.v. minder intensief), deels andere programma's, e.d.? Je vergelijkt natuurlijk altijd met àndere activiteiten. De vraag of een programma effect boekt moet dan ook aangevuld worden met de vraag of het meer, of belangrijker, of goedkoper, effect boekt dan alternatieven. Nu is vergelijking tussen verschillende programma's niet gemakkelijk. De 'program characteristics' verschillen vaak; voorzover dezelfde uitdrukkingen worden gebezigd hebben die soms ook een heel andere betekenis, en/of lopen de resultaten uiteen.

Generaliseren is vooral zinvol als er zicht is op daadwerkelijke verspreiding, èn op verspreidingscondities. Hoe liggen elders de beginvoorwaarden, zijn de verwachtingen niet te hoog gespannen (of: wel hoog genoeg...)? Het is onjuist om van evaluatieonderzoek in het onderwijs generaliseerbaarheidsbewijzen te verlangen. Toch zit voor velen de research in die positie. Vandaar ook het stof dat een toch niet zo opzienbarend artikel als dat van Lewin (1977) weer doet opwaaien, waarin tot positieve lange-termijn-effecten van Head Start wordt geconcludeerd. Ook Lazar et al. (1977) komen tot de conclusie dat er programma's aan te wijzen zijn die tot significante lange termijn effecten op schoolprestaties leiden. Er vindt minder verwijzing naar vormen van buitengewoon onderwijs plaats; de kinderen blijven minder zitten (p. 12). (Opvallend is dat zij menen te kunnen constateren dat van projecten met de meest doortimmerde evaluatieopzet de meest indrukwekkende resultaten verwacht mogen worden (p. 13). Bij dit wel vaker geconstateerde verschijnsel kan men zich afvragen of een goede evaluatie-opzet indicatief is voor de kwaliteit van het programma, als zijnde ook een onderdeel van hetzelfde project. Of dat grondiger evaluatie de door Campbell en Erlebacher (1970) en Campbell en Boruch (1975) geanalyseerde onderschattingsmogelijkheden voor het experimentele effect minder kans biedt.)

Vanwaar dit (voorwaardelijke) *optimisme?* We 'weten' toch dat 'compensatie-programma's' aan

strenge eisen afgemeten slechts zeer geringe effecten sorteren? Dat is althans een vaak geveld oordeel, zie bijvoorbeeld Van der Kley, in Van Kemenade (1973, pp 64). Van der Kley lijkt echter niet scherp genoeg rekening te houden met het feit dat bij 'Head Start', 'Follow-Through', e.d., federale bestedingsprogramma's geëvalueerd worden, geen eenduidige onderwijsprogramma's (zie p. 68). En ook komt de rol van de onderwijzer als één van 'strategisch belang'er maar bekaaid af.

Van de vele factoren die meebepalen of het oordeel over programma's in deze leeftijdsperiode positief of negatief uitvalt is de factor *'keuze van variabelen'* wellicht onderwijskundig/onderzoeksmatig de belangrijkste. Zo stellen Madaus et al. (1979, p. 225) dat Coleman (die in 1966 tot de conclusie kwam dat schoolse factoren weinig of geen invloed hebben op het prestatieniveau van veel minderheidsgroepen) tot een andere uitkomst was gekomen, als hij in plaats van karakteristieken als grootte, materiële faciliteiten, kwalificaties van leerkrachten, meer aandacht zou hebben gegeven aan het klimaat en de activiteiten op school. We delen Madaus' mening dat aandacht daarvoor interessanter is dan het verder uitvlooien van de gehanteerde statistische analyse.

Factoren zoals Coleman meenam zijn overigens uiteraard heel begrijpelijk waar overheidsbeleid geevalueerd wordt: de 'hardware' kant is het meest tastbaar en direct gerelateerd aan jaarlijkse begrotingen! Groenendaal noemt met Coleman als brengers van 'het slechte nieuws' in één adem Jensen (1969) en Jencks (1972) (1978, p. 23), naast de evaluatierapporten van nationale projecten zoals Head Start. McNeil (1968) noemt verder nog het Plowden Rapport (1967).

Naar onze mening kan uit deze wetenschappelijke produktie worden geconcludeerd dat het hem meer zit in de *processen* (zowel op school als thuis) dan in de gemakkelijker grijpbare variabelen van meer materiële of 'labelende' aard. Daarbij zijn twee belangrijke nevenconclusies mogelijk:

- uitkomsten zijn afhankelijk van de wijze van statistische analyse (als je bijvoorbeeld 'family-status'-variabelen als eerste in een regressiemodel opneemt zullen die onevenredig veel variantie aan zich trekken en mede dáárdoor belangrijk worden) (Zie bijvoorbeeld Madaus et al. 1979, pp 221 en 214);
- uitkomsten zijn afhankelijk van de wijze waarop omgevingsfactoren worden gemeten èn gevarieerd (waar de variatie in omgevingsfactoren niet wordt gecontroleerd en/of veel geringer is dan denkbaar, zijn quasi-definitieve conclusies over onderwijsmogelijkheden weinig zinvol).

Voor een deel is dit soort werk afhankelijk van *overtuigingen*. Smith en Bissell (1970, p. 5) gaan uit van de veronderstelling dat achter Head Start de notie steekt 'that intelligence is malleable'. (Zie ook De Groot, 1980; Vroon, 1980; Terwee, 1980; Van den Borne en Backus, 1979; Fatke, 1970; Sauer, 1970; Schusser, 1970).

Hoe nu *de waarde van de GEON-uitkomsten* te beoordelen? Moeten we scherp letten op statistische significantie, of is dat niet nodig gezien de replicaties, en de vraagtekens die zoals gezegd bij significantieberekeningen gezet kunnen worden? Moeten we op zoek naar een maat voor 'educational significance'? (House et al. stellen die op ¼ standaardafwijking (1978, p. 145).). Maar wat betekent een gemeten winst in testscores van zeg 10 punten tussen het 4e en het 6e levensjaar, individueel en maatschappelijk? Betekent het voor aanvankelijk lager scorenden meer dan voor de rest, zodat je een bepaalde kritische grensscore als evaluatie-criterium zou moeten hanteren (denk aan de toelatingsnormen voor BUO)? Of mag je afgaan op de voorspellende waarde die testscores voor schoolprestaties blijken te hebben, en er verder maar het beste van hopen? Geldt die voorspellende waarde ook voor verschilscores (vooruitgang)? (Zie Zimiles, 1970, p. 244; Stevens, 1973, p. 36).

Gelukkig hebben wij over *kritiek* niet te klagen gehad. Zo vroeg J. A. van Kemenade zich in 1978 al af of er bij GEON geen sprake was van een 'Hawthorne-effect', veronderstelde S. J. Sandbergen begin 1979 testgerichte oefeningseffecten, verwees H. J. Groenendaal tijdens de ORD '79 naar statistische regressie, vroeg minister Pais najaar 1979 naar de generaliseerbaarheid van de conclusies, en verlangde M. Zwarts (1979) naar een meer gedetailleerde verklaring van de opgetreden effecten.

Naar onze mening ligt de achilleshiel, bij alle duidelijke en consistente (zich replicerende) uitkomsten, bij de laatstgenoemde kanttekening. Het verklaringsmodel blijft, voor een deel, noodgedwongen (zie paragraaf 16), een 'black box'. Dit betekent vooral dat de voorspelbaarheid van resultaat op het niveau van het individuele kind waarschijnlijk niet zo groot is. Dat hangt mede samen met het (bekende) verschijnsel van gebrek aan harde theorie.

Toch trokken we meerdere, duidelijke conclusies. De naar onze mening belangrijkste is dat de invloed van routinematige gedragingen erg groot kan zijn; gedragingen die slechts via gerichte training gewijzigd kunnen worden. Deze slotconclusie is in essentie reeds te vinden bij Mellenbergh et al. (1968, pp 615-616): 'Aangezien men sneller onderwijsresultaten bereikt bij goede dan bij slechte leerlingen, fungeren de goede leerlingen als bevestigers van het on-

derwijsgedrag van de lesgevers, waardoor ook op deze manier een voorkeursbehandeling waarschijnlijker wordt. (...). Hoe dan ook, in plaats van een negatieve terugkoppeling in het onderwijs vindt men hierdoor een positieve terugkoppeling: aanvankelijk kleine verschillen worden op ongerechtvaardigde wijze vergroot'. (Zie ook Colthof, 1979, p. 197; Lowyck, 1979, p. 441; Hermanns, 1979; Groenendaal, 1978, pp 14-15).

Het bestek van dit artikel staat niet toe verder in te gaan op effectiviteitsvoorwaarden bij het GEON-programma. We herhalen alleen dat we die duidelijk gekoppeld zien aan disseminatiemogelijkheden. Varianten voor en condities bij verspreiding zijn gedurende de laatste jaren van het project verregaand uitgewerkt in een nota over verspreidingsaspecten (1977), een verspreidingsadvies (1978, waarvan een uittreksel is aangeboden door het projectbestuur aan de minister), een beleidsnotitie voor overleg met de onderwijsinspectie (1979), en een opzet voor een verspreidingsexperiment. Sedert december 1978 is het project in de publiciteit getreden, en zijn er vragen gesteld en een motie en een amendement aangenomen in de Tweede Kamer. Op moment van schrijven van dit artikel (mei 1980) is bekend dat de LPC's de minister hebben toegezegd een plan voor verdere implementatie te willen opzetten.

Epiloog

De in Stokking, 1980[b], aangekondigde evaluatie van de evaluatie-aanpak is gedeeltelijk in bovenstaand artikel verwerkt. Er bestaat een uitgebreider rapport over (Stokking, 1980[a]). We hopen er in artikel- of boekvorm nog op terug te komen. Verder zijn er de nodige deelevaluatierapporten per programmaonderdeel, en enkele case-studies.

Als contactadres fungeert het IPAW te Utrecht.

Literatuur

Anderson, R.B. et al., Pardon us, but what was the question again? A response to the critique of the Follow Through evaluation, *Harvard Educational Review,* 1978 (48) 161-170.

Anderson, S.B., From textbooks to reality. Social researchers face the facts of life in the world of the disadvantaged, in: I. Hellmuth (Ed.), *Disadvantaged Child,* vol. 3 New York, 1970, pp 226-237.

Baker, R.L., Curriculum Evaluation, *Review of Educational Research,* 1968 (39,3) 339-358.

Beerling, R.F., S.L. Kwee, J.J. A. Mooij en C.A. van Peursen, *Inleiding tot de wetenschapsleer.* Utrecht, 1970.

Bereiter, C., Some persisting dilemma's in the measurement of change, in: Ch.W. Harris (Ed.), *Problems in measuring change.* London, 1976.

Berk, K.N. and J.S. Francis, A review of the manuals for BMPD and SPSS, *Journal of the American Statistical Association,* 1978 (73,361) 65-71.

Berman, P. and E.W. Pauly, Federal programs supporting educational change, Vol. II *Factors affecting change agents projects.* Rand, Santa Monica, 1975.

Bishop, D. and G.E. Butterworth, A longitudinal study using the WPPSI and WISC-R with an English sample, *British Journal of Educational Psychology,* 1979 (49) 156-168.

Borne, C. van den, en M. Backus. Is intelligentie erfelijk? *Revoluon* 1979 (5,1) 14-53.

Bos, K.P. van den, Leerstoornissen en WISC of WISC-R profielen, *Pedagogische Studiën,* 1979 (56,10) 397-408.

Bosch, L.J. van den, Evalueren van onderwijsinnovaties, *Pedagogische Studiën,* 1975 (52) 128-140.

Bracht, G.H. and G.W. Glass, The external validity of experiments, *American Educational Research Journal,* 1968 (5) 437-474.

Bronfenbrenner, U., Een experimentele ecologie van de menselijke ontwikkeling, in: W. Koops en J. J. van der Werff, *Overzicht van de ontwikkelingspsychologie.* Groningen, 1979, pp 407-423.

Bruyn, E.E.J. de, W. Heinrichs, H. Oosterbaan, Niveauverschillen op nietgenormeerde Nederlandse versies van de intelligentietests HAWIK en WISC-R, *Tijdschrift voor Orthopedagogiek,* 1979.

Campbell, D.T., 'Degrees of freedom' and the case study, *Comparative Political Studies,* 1975 (8,2) 178-193.

Campbell, D.T. and R.F. Boruch, Making the case for randomized assignment to treatments by considering the alternatives: Six ways in which quasiexperimental evaluations in compensatory education tend to under-estimate effects, in: C.A. Bennet and A.A: Lumsdaine (Eds.), *Evaluation and Experiment.* New York, 1975, 195-296.

Campbell, D.T. and A. Erlebacher, How regression artifacts in quasiexperimental evaluation can mistakenly make compensatory education look harmfull, in: I. Hellmuth (Ed.), *Disadvantaged Child,* volume 3. New York, 1970, 185-210.

Cicirelli. V.G., J.W. Evans, and J.S. Schiller, The impact of Head Start: A Reply to the Report Analysis, *Harvard Educational Review,* 1970 (40,1) 105-129.

Cohen, J., Multiple regression analysis as a general data-analytic system, *Psychological Bulletin,* 1968 (70,6) 426-443.

Cohen, J., *Statistical power analysis for the behavioral sciences.* New York, 1969.

Colthof, J., *Opleiding en begeleiding van taakleidsters.* Diss., Utrecht, 1979.

Cook, T.D., D.T. Campbell, *Quasi-experimentation.* Chicago, 1979.

Coulson, J.E., National Evaluation of the emergency school aid act (ESAA): A review of methodological issues, *Journal of Educational Statistics,* 1978 (3,1) 1-60.

Cronbach, L.J., *Research on classroom and schools: formulation of questions, design and analysis.* Stanford, 1976.

Cronbach, L.J., Beyond the two disciplines of scientific

psychology, in: G.V. Glass (Ed.), *Evaluation Studies Review Annual.* Vol. 1, 1967, Beverly Hills, London, 15-27.
Dronkers, J. (Red.), *Onderwijs en economie.* Bijdragen tot de onderwijsresearchdagen 1978. SVO-reeks nr. 10 Den Haag, 1978.
Duintjer, O.D., *Rondom regels.* Meppel/Amsterdam, 1977.
Duyn, P. van den, Wie geneest heeft gelijk, *Intermediair,* 4 mei 1979, 45-47.
Edwards, K.I., Summative Evaluation: Some basic consideration, in: G.D. Borich (Ed.), *Evaluating educational programs and products.* Englewood Cliffs, 1974, 373-399.
Edwards, W. and M. Guttentag, Experiments and evaluations: a reexamination, in: C.A. Bennet, and A.A. Lumsdaine (Eds.), *Evaluation and Experiment.* New York, 1975, 409-463.
Eisner, E.W., Instructional and expressive educational objectives, in: M.D. Merrill (Ed.), *Instructional Design, Readings.* London, etc., 1971, 97-101.
Evans, S.H. and E.J. Anastasio, Misuse of analysis of covariance when treatment effect and covariate are confounded, *Psychological Bulletin,* 1968 (69,4) 225-234.
Fatke, R., Zur Kontroverse um die Thesen A. Jensens, *Zeitschrift für Pädagogik,* 1970 (16,2) 219-226.
Feldt, L.S., A comparison of the precision of three experimental designs employing a concomitant variable, *Psychometrika,* 1958 (23,4) 335-353.
Freudenthal, H. *Weeding and Sowing, Preface to a science of mathematical education.* Dordrecht/Boston, 1978.
Giorgi, A., *Fenomenologie en de grondslagen van de psychologie.* Meppel/Amsterdam, 1978.
Glass, G.V. and F.S. Ellett, Evaluation Research, *Annual Review of Psychology,* 1980 (31) 211-228.
Groen, H.K., Leerlingen uit verschillende sociale milieus hebben geen gelijke kansen, *Tijdschrift voor Onderwijsresearch,* 1975 (1,1) 40-42.
Groenendaal, H.J., *Vroegtijdige hulpverlening aan zwakfunktionerende kleuters.* Verslag van een onderzoek. Diss. Amsterdam, 1978.
Groot, A.D. de, en A.A.J. van Peet, Enkele kanttekeningen bij het proefschrift van J.L. Peschar: Milieu, school en beroep, *Tijdschrift voor Onderwijsresearch,* 1975 (1,1) 36-39.
Groot, A.D. de, en A. van Peet, Nogmaals de invloed van regressie-effecten, *Tijdschrift voor Onderwijsresearch,* 1976 (1,3) 133-137.
Groot, A. D. de, De betekenis van 'intelligentie' en 'aanleg'. (afscheidscollege), *Intermediair,* 8 februari 1980 (16,6) 11-17 + 29.
Haassen, P.P. van, *WISC-R, Nederlandse handleiding.* Amsterdam, 1976.
Heinze, T. u.a., *Handlungsforschung im pädagogische Feld.* München, 1975.
Hermanns, J., Het ontstaan van schoolproblemen. Een longitudinaal onderzoek in kleuter- en lagere school, *Pedagogische Studiën,* 1979 (56) 348-357.
Hindley, C.B. and C.F. Owen, The extend of Individual Changes in I.Q. for Ages between 6 months and 17 years, in a British Longitudinal Sample, *Journal of Child Psychology and Psychiatry,* 1978 (19) 329–350.
Hofstee, W.K.B. De vragenlijstsituatie, *Nederlands Tijdschrift voor de Psychologie,* 1965 (10) 592-602.
House, E.R., Justice in Evaluation, in: G.V. Glass, *Evaluation Studies Review Annual,* Vol. 1, 1976, London, 75-100.
House, E.R. et al., No Simple Answer: Critique of the Follow Through Evaluation, *Harvard Educational Review,* 1978 (48,2) 128-160.
Huck, S.W. and R.A. McLean, Using a repeated measures ANOVA to analyze the data from Pretest-Posttest design: a potentially confusing task, *Psychological Bulletin,* 1975 (82) 511-518.
Irvine, J., J. Miles and J. Evans, *Demystifying social statistics.* London, 1979.
Kemenade, J.A. van, *Bijdragen uit de onderwijswetenschappen.* Alphen a/d Rijn, 1973.
Kemenade, J.A. van, *Als de smalle weegbree bloeit.* Amsterdam, 1979.
Kenny, D.A., A Quasi-experimental Approach to Assessing Treatment Effects in the Nonequivalent Control Group Design, *Psychological Bulletin,* 1975 (82) 345-362.
Kohnstamm, G.A., Over de AKIT, de WPPSI en de UTANT, Drie nieuwe tests voor 4-7 jarigen, *Nederlands Tijdschrift voor de Psychologie,* 1968 (23) 269-290.
Kuhn, T.S., *De noodzakelijke spanning. Traditie en vernieuwing in de wetenschap.* Meppel/Amsterdam, 1979.
Lazar, J. et al., *The persistence of preschool effects.* A long-term Follow-Up of fourteen infant and preschool experiments. Summary Report, US Dep. of HEW, october 1977.
Lewin, R., 'Head Start' pays off. *New Scientist,* 3 March 1977, 508-509.
Light, R.J. and P.V. Smith, Choosing a Future. Strategies for Designing and Evaluating New Programs, *Harvard Educational Review,* 1970 (40,1) 1-28.
Lord, F.M., Elementary models for measuring change, in: Ch. W. Harris (Ed.), *Problems in measuring change.* London, 1967.
Lord, F.M., Statistical adjustments when comparing preexisting groups, *Psychological Bulletin,* 1969 (72,5) 336-337.
Lowyck, J., Procesanalyse van het onderwijsgedrag, *Pedagogische Studiën,* 1979 (56) 427-440.
Madaus, G.F. et al., The sensitivity of measures of school effectiveness, *Harvard Educational Review,* 1979 (49,2) 207-230.
Marjoribanks, K., *Families and their learning environments. An empirical analysis.* London, 1979.
McCall, G.J. and J.L. Simmons, *Issues in participant observation.* Reading, 1969.
McDill, E.L. et al., *Strategies for success in compensatory education: an appraisal of evaluation research.* Baltimore/London, 1969.
McNeil, J.D., Forces influencing curriculum, *American Educational Research Journal,* 1968 (5) 293-318.
Mellenbergh, G.J. et al., Vaardigheden en tekorten, een strategie voor het analyseren van onderwijsdoelen, *Nederlands Tijdschrift voor de Psychologie,* 1968, 609-631.

Meltzer, B.N., J.W. Petras and L.T. Reynolds, *Symbolic interactionism, Genetics, varieties and criticism.* London, 1977.

Molenaar, I.W. en A. Thomas, Psychometrics in subgroups, or Regression to the mean revisited, *Tijdschrift voor Onderwijsresearch,* 1968 (3,4) 152-160.

Moser, H., *Methoden der Aktionsforschung.* München, 1977.

Nie, N.H. et al., *Statistical Package for the Social Sciences.* New York, etc., 1975².

Overall, J.E. and J.C. Klett, *Applied multivariate analysis.* New York, 1972. (Chapter 8: Complex Least-squares Analysis of Variance).

Overall, J.E. and D.K. Spiegel, Concerning Least Squares Analysis of Experimental Data, *Psychological Bulletin,* 1969 (72,5) 311-322.

Overall, J.E., D.K. Spiegel, and J. Cohen, Equivalene of Orthogonal and Nonorthogonal Analysis of variance, *Psychological Bulletin,* 1975 (82) 182-186.

Parlett, M. and D. Hamilton, *Evaluation as illumination: a new approach to the study of innovatory programs.* Occ. paper no 9, CRES, Edinburgh, 1972.

Peschar, J.L., De invloed van de regressie-effecten in het Milieu-School-Beroep onderzoek: Een antwoord aan A.D. de Groot en A.A.J. van Peet, *Tijdschrift voor Onderwijsresearch,* 1976a (1,2) 49-58.

Peschar, J.L., Andermaal de invloed van regressie-effecten, *Tijdschrift voor Onderwijsresearch,* 1976b (1,3) 137-138.

Peschar, J.L., Educational opportunity within and between Holland and Sweden: The semi-experimental approach, *Sociologische Gids,* 1978 (25,4) 273-296.

Porter, A.C. and T.R. Chibucos, Selecting analysis strategies, in: G.D. Borich, (Ed.), *Evaluating educational programs and products.* Englewood Cliffs, 1974, 415-464.

Rasbury, W. et al., Relations of scores on WPPSI and WISC-R at a one-year interval, *Perceptual and Motor Skills,* 1977 (44) 695-698.

Rosenshine, B. and N. Furst, The use of direct observation to study teaching, in: R.M.W. Travers (Ed.), *Second Handbook of research on teaching.* Chicago, 1973, 122-183.

Roskam, E.E. en A.L.M. van der Sanden, Factor analytische modellen in longitudinaal onderzoek, *Nederlands Tijdschrift voor de Psychologie,* 1974 (29) 67-94.

Sandbergen, S., 'Testslimheid', De invloed van oefening en coaching op testscores, *Nederlands Tijdschrift voor de Psychologie,* 1968 (23) 502-529.

Sattler, J.M., *Assessment of children's intelligence.* Philadelphia, 1974, section 4.

Sauer, W., Stand der Zwillungsforschung in pädagogischer Sicht, *Zeitschrift für Pädagogik,* 1970 (16,2) 173-202.

Scriven, M., Die Methodologie der Evaluation, in: Chr. Wolf (Hrsg.), *Evaluation.* München, 1972.

Scriven, M., Evaluation perspectives and procedures, in: W.J. Popham (Ed.), *Evaluation in education. Current applications.* Los Angelos, 1974, 1-93.

Schroots, J.J.F., *De Leidse Diagnostische Test, experimentele versie.* Diss., Amsterdam, 1979.

Schusser, G., Vererbung, Intelligenz und Schulleistung, *Zeitschrift für Pädagogik,* 1970 (16,2) 203-218.

Schutz, R.E., Methodological issues in curriculum research, *American Educational Research Journal,* 1968 (5) 359-366.

Sherrets, S. and M. Quatrocchi, WISC-R differences - Fact or Artifact? *Journal of Pediatric Psychology,* 1979 (4,2) 119-127.

Shipman, V.C., Disadvantaged children and their first school experiences. ETS-HEAD START longitudinal study, in: J.C. Stanley (Ed.), *Compensatory education for children ages 2 to 8.* Baltimore, 1973, 145-195.

Slavenburg, J.H. (Red.), *Het Project Onderwijs en Sociaal Milieu.* Tilburg, 1978.

Smets, P., *Beginfase begeleidingswerk stimuleringsscholen.* Deelverslag 1, Evaluatie Onderwijsstimulering. ITS, Nijmegen, 1979.

Smith, G. and J. James, The effect of preschool education: Some American and British evidence, *Oxford Review of Education,* 1975 (1,3) 223-240.

Smith, M.S. and J.S. Bissell, Report Analysis: The impact of Head Start, *Harvard Educational Review,* 1970 (49,1) 51-104.

Stake, R.E., Program Evaluation, particularly responsive evaluation, in: *New trends in evaluation.* Report from the institute of education. Göteborg, 1974.

Stevens, L.M., *Curriculum Schoolrijpheid, Deel II. Een evaluatieonderzoek.* Den Bosch, 1973.

Stokking, K.M., *Toetsend Onderzoek.* (Dissertatie). Groningen, 1979.

Stokking, K.M., *Evaluatie-onderzoek Inservicetraining GEON-project.* Paper voor de onderwijsresearchdagen 1980 (1980a).

Stokking, K.M., De evaluatie-aanpak in het GEON-project, *Pedagogische Studiën,* 1980 (57,4) 182-194 (1980b).

Stokking, K. M., Statistische regressie: enkele methodologische notities, *Tijdschrift voor Onderwijsresearch,* 1980 (5) 271-279 (1980c).

Stokking, K. M., *Evaluatie van de evaluatie,* Utrecht, 1980 (1980d).

Stufflebeam, D.L., Alternative approaches to educational evaluation: A self-study guide for educators, in: W.J. Popham (Ed.), *Evaluation in education. Current Applications.* Los Angelos, 1974, 95-143.

Stufflebeam, D.L. et al., *Educational evaluation and decision making.* Itasca, 1972.

Takanishi, R., Evaluation of early childhood programs: toward a developmental perspective, in: L.G. Katz (Ed.), *Current topics in early childhood education,* vol. II. Norwood N.J., 1979.

Terwee, S., Heeft A.D. de Groot gelijk? *Intermediair,* 14 maart 1980 (16,11) 45-47 + 61.

Thorndike, R.L., Regression fallacies in the matched groups experiments, *Psychometrika,* 1942 (7,2) 85-102.

Vall, M. van de, en B. van Dijkum-de Jong, *Het rendement van sociaal beleidsonderzoek.* Alphen a/d Rijn, 1979.

Verberk, A.J.A., *Variantie-analyse in de gedragswetenschappen, in het perspektief van andere multivariate analysetechnieken.* Groningen, 1970.

Vries, A.K. de, *Evaluatie SVO-projekt 083. Tussentijdse*

globale rapportage. Utrecht, 1973.
Vries, A.K. de, GEON, opvoedings- en onderwijsperspektieven, *Pedagogische Studiën,* 1980 (57,3) 113-124.
Vries, A.K. de, Inhouden en achtergronden van de inservice-trainingskursussen in het GEON-project. *Pedagogische Studiën,* 1980 (57,10) 449-459.
Vries, A.K. de, M.H. Kramer-van Walderveen, K.M. Stokking en L.C. Thierens, GEON: Ervaringen met en evaluatiegegevens over de programma-onderdelen, *Pedagogische Studiën,* 1980 (57) 504-531.
Vroon, P.A., Enkele kanttekeningen bij het onderzoek naar de herkomst van intelligentieverschillen, *Tijdschrift voor Onderwijsresearch,* 1978 (3,6) 284-291.
Vroon, P., Erfelijkheid en I.Q.: Chaos in onderzoek, *De Volkskrant,* 10 januari 1980.
Wechsler, D., *WPPSI-manual.* New York, 1967.

Manuscript aanvaard 22-8-'80

Wechsler, D., *WISC-R manual.* New York, 1974.
Weikart, D.P., Über die Wirksamkeit vorschulischer Erziehung, *Zeitschrift für Pädagogik,* 1975 (21,4) 489-509.
Welters, L., Het beleid en de zachte sektor van het sociaal onderzoek, in: L. Brunt (Red.) *Anders bekeken.* Meppel/Amsterdam, 1977 pp 109-126.
Wit, O. en W. van Soest, *Handleiding Begrijpend lezen II.* CITO, Arnhem, 1975.
Woodward, J.A. en J.E. Overall, Multivariate analysis of variance by Multiple Regression Methods, *Psychological Bulletin,* 1975 (82) 21-32.
Zimiles, H. Has evaluation failed compensatory education? in: J. Helmuth (Ed.), *Disadvantaged Child,* Vol. 3 New York, 1970, 238-245.
Zwarts, M., *Verslag van de meta-evaluatie van de evaluatie van het GEON-projekt,* Utrecht, 1979.

Curriculum vitae: zie Pedagogische Studiën 1980, *57,* 194.