

De evaluatie-aanpak in het GEON-project

K. M. STOKKING
Evaluator GEON

Samenvatting

Het artikel beschrijft de opzet en achterliggende uitgangspunten van de evaluatie van het GEON-project (1973-1980). GEON heeft als taak experimentele innovatie, richt zich op groepen kinderen in achterstandssituaties, en werkt vooral via inservicetraining van leerkrachten. Het aksent bij de evaluatie lag op summatieve en proces-evaluatie. De werkwijze was quasi-experimenteel, waarbij oordelen over interne en externe geldigheid gebaseerd worden op een aantal expliciete argumenten. Daartoe is een bepaald meet-design gehanteerd en vond uitgebreide proces-registratie plaats. Er is met name aandacht gegeven aan de problemen 'statistische regressie' en 'implementatie'. Ingewikkelde statistische analyse vond niet plaats, op grond van geldigheids- en bruikbaarheidsoverwegingen. Wel zijn replikaties ingebouwd. Het evaluatie-onderzoek was niet zozeer theoretietoetsend als wel effectmetend. In een bepaalde zin opgevat ging het om aktie-onderzoek.

1. Inleiding

Dit artikel is het eerste over de evaluatie van het GEON-project. In dit artikel stellen we ons voor alleen in te gaan op de opzet van de evaluatie. De ervaringen met de gehanteerde evaluatie-aanpak (evaluatie van de evaluatie) en de uitkomsten zullen in aparte artikelen omschreven worden.

Uiteraard beschrijven we de evaluatie-opzet vanuit de manier waarop we daar nu tegen aankijken. Weliswaar lag de opzet al in 1975 grotendeels vast, maar de motieven die aan de keuze ten grondslag lagen zijn destijds niet alle expliciet gemaakt. Ook ontstond pas gaandeweg een helder zicht op een aantal konsekventies. Dat proces komt in het kader van de evaluatie van de evaluatie-aanpak aan de orde. We geven nu een overzicht van de gevolgde werkwijzen en achterliggende opvattingen anno 1979. Om dat te kunnen doen geven we eerst

algemene informatie over het project.

1.1 Opbouw en taakstelling

GEON staat voor Programma GEdifferentieerd ONderwijs voor kinderen van 4 - 8 jaar. Er werd gewerkt op \pm 100 kleuter- en lagere scholen in een viertal plaatsen.

GEON is eind 1973 gestart. Bij de totstandkoming spraken verschillende betrokkenen verschillende wensen uit. Het Departement wilde een structureel antwoord op het toenemende aantal verwijzingen naar het buitengewoon onderwijs. De SVO wilde een model voor de toepassing van onderzoeksresultaten. De begeleidingsdiensten wilden een model voor kleuterschoolbegeleiding. Een algemener gevoel van diverse participanten was er een van ontevredenheid met de doorwerking van onderwijsvernieuwing en -begeleiding. De taakstelling luidde bij de start als volgt:

1. Beoordeling van het effect van de gebruikte middelen (kursussen of anderszins).
 2. Bijstelling hiervan op grond van de uitkomsten.
 3. Analyse van de voorwaarden waaronder de effecten zijn opgetreden.
 4. Opstelling en voorbereiding van een plan voor verdere toepassing van de bereikte inzichten.
 5. Beoordeling van de gekozen projektopzet als beleidsinstrument voor de overheid.
- Hier aan de orde zijn 1, 2, 3, en 4, met het aksent op 1.

GEON werkte in een aantal regio's: Amersfoort, Hilversum, Enschede en Arnhem. In Utrecht bevond zich de Centrale Staf (Projektleiding, Programma-ontwikkeling, Evaluatie, Administratieve ondersteuning). In elke regio werkte een team, bestaande uit een projektleider, een begeleidster en zes intermediaire krachten. De laatsten kwamen op de scholen. Dit team was te gast in de plaatselijke schooladviesdienst.

GEON werd uitgevoerd onder verantwoordelijkheid van de Commissie Experimentele Innovatie (CEI) waarin samenwerkten de Stichting Onderzoek

Leermoeilijkheden (RU-Utrecht), De Vereniging Samenwerkende Landelijke Pedagogische Centra en het Werkverband van Plaatselijke en Regionale Onderwijscentra, tesamen met de Schooladviesdiensten te Amersfoort, Arnhem, Enschede en Hilversum.

GEON is leerkracht gericht (in-service training), en kan gezien worden als een diepte-investering in het stimuleringsbeleid (zie het beleidsplan uit 1974).

GEON werkt op basis van subsidie van het Ministerie van Onderwijs en Wetenschappen en de Stichting voor Onderzoek van het Onderwijs, beiden te Den Haag.

Een meer uitgebreide verantwoording van de opzet van het GEON-programma is gegeven in het artikel van A. K. de Vries, GEON, opvoedings- en onderwijsperspektieven (Pedagogische Studiën, 1980 (57) 113-124).

1.2. Programma

GEON staat genoemd in de nota over het onderwijs aan groepen in achterstandssituaties uit 1974 als landelijk aktiveringsproject. Het programma is leerkracht gericht: het gaat om vaardigheden van leerkrachten. Het aksent ligt dus niet op materialen of onderwijsinhouden. Het programma bestaat uit een aantal kursussen voor leerkrachten, een viertal takenreeksen voor leerkrachten en kinderen en een werkmap met Buurt- Gezin-School suggesties. Elke kleuterschool die begeleidt wordt, krijgt ongeveer 2 dagen per week een intermediaire kracht (veelal een ex-kleuterleidster) op school. Deze heeft zich in een voorbereidingsperiode al met het programma vertrouwd gemaakt. Deze extra leidster is niet als deskundige buitenstaander bedoeld. Zij voert onderdelen van het programma in de klassen uit en neemt de klas af en toe over. Daardoor kan de klasseleidster werken aan verschillende kursussen die voor haar ontwikkeld zijn.

In de kleuterschool komt, uitzonderingen daargelaten, het hele programma in de loop van twee jaar aan de orde. Daarna trekt de leidster zich terug om in een andere kleuterschool te gaan werken. De begeleiding van de scholen die hebben meegedaan wordt voortgezet vanuit de schooladviesdienst.

Voor de uitvoering van het programma in de basisscholen is het uitgangspunt, dat de bij de kleuters begonnen begeleiding niet plotseling moet worden afgebroken. Daarom werd direkt bij het begin van de uitvoering van het projekt in de kleuterscholen kontakt gelegd met die basisscholen, waar de meeste kinderen uit deze kleuterscholen heen zullen gaan. De basisscholen zijn meestal bij

veel aktiviteiten van de schooladviesdiensten betrokken. Daarom is hier geen extra leerkracht ter beschikking. Wel zijn de kursussen beschikbaar.

Een meer uitvoerige beschrijving van de verschillende programma-onderdelen zal worden gegeven in het volgende artikel in de reeks.

1.3. Planning

In elke regio wordt gewerkt in twee experimentele rondes achter elkaar. Het projekt werkt daarin samen met zowel kleuterscholen als lagere scholen. In de eerste ronde participeren per regio een zestal kleuterscholen. Gedurende twee jaar worden de kleuterleidsters begeleid met behulp van het programma zoals hierboven aangeduid.

Na die twee jaar worden de aanvankelijk 4-jarigen nog twee jaar gevolgd, dus tot hun 8e jaar. Dit gebeurt via kontakten met de leerkrachten van de eerste en tweede klas. Een en ander gebeurt in nauwe samenwerking met medewerkers van de plaatselijke schoolbegeleidingsdienst. Na die eerste twee jaar begint op een zestal andere kleuterscholen in de regio een tweede experimentele ronde. De eerste ronde is in de eerste regio van start gegaan in augustus 1974; in de tweede en derde regio gebeurde dat in augustus 1975, in de vierde februari 1976. In augustus 1976 begon de tweede ronde in de eerste regio; in de 2e en 3e regio was dat in augustus 1977 het geval. In Figuur 1 is een en ander samengevat.

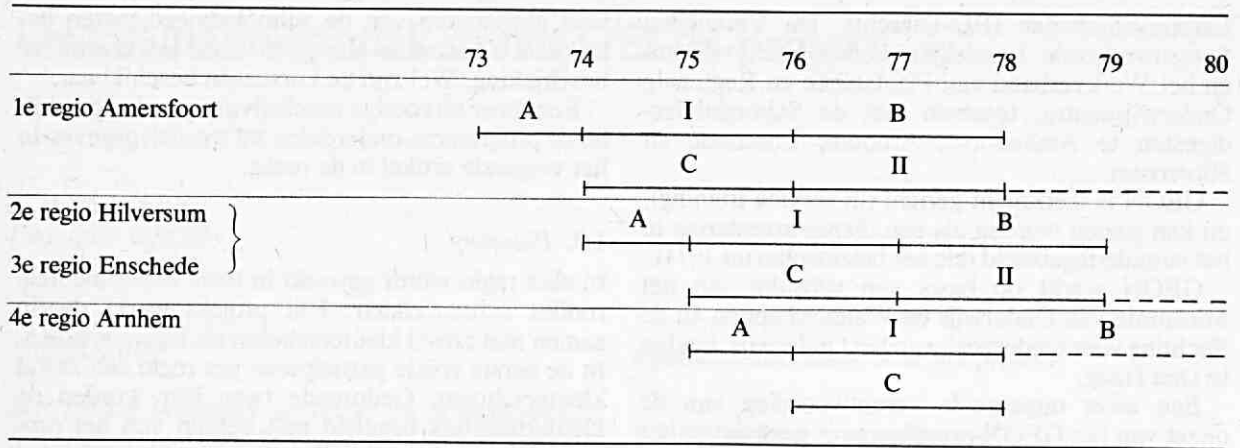
2. Soorten evaluatie

In de taakstelling zitten al diverse soorten evaluatie opgesloten:

- ad 1. Summatieve evaluatie (wat is er bereikt bij de betrokken kinderen, leerkrachten, scholen, regionale situaties);
- ad 2. Formatieve evaluatie (welke bijstellingen moet het programma ondergaan om hanteerbaarheid en effectiviteit te vergroten);
- ad 3. Procesevaluatie (wat was precies de beginsituatie, en hoe is vervolgens het programma uitgevoerd);
- ad 4. Beleidsevaluatie (implicaties van inzichten, besluitvorming over verder gebruik).

Het laatste vindt plaats via een aantal kontakten op regionaal en landelijk niveau en blijft hier verder buiten beschouwing.

Formatieve evaluatie is bij GEON niet strak opgezet. Er is een structuur voor interne informatie-uitwisseling. Over bijstelling van afzonderlijke kursussen zal apart worden gerapporteerd in een



Figuur 1 De planning

I = eerste ronde, II = tweede ronde

A = het voorbereidende jaar (aantrekken medewerkers en keuze van scholen)

B = volgen van de 1e ronde kinderen tot hun 8e jaar + sterk verminderde voortgezette begeleiding van de eerste groep projectscholen tijdens de 2e ronde (wel testafnames)

C = vergelijkingstestafnames op de scholen waarvan in het voorbereidende jaar in principe al besloten werd dat die tijdens de 2e ronde met het project zouden gaan meedoen

eindrapport.

Summatieve evaluatie en procesevaluatie hebben het meest aksent gekregen. Daarbij worden respectievelijk effectmeting en beschrijving van de uitvoering gemaximaliseerd. We komen daarop straks terug. Summatieve evaluatie is hier bedoeld inclusief follow-up evaluatie (dus aangaande beklivings-effecten).

Naast bovenstaande is in GEON nog onderscheiden:

- Begeleidingsevaluatie (ondersteuning van projectuitvoering);
- Produktevaluatie (bepaling van de waarde van elk programma-onderdeel);
- Vaktheoretische evaluatie (terugkoppeling naar onderwijskundige, psychologische en andere inzichten);
- Evaluatie van de evaluatie.

De begeleidingsevaluatie kreeg vooral gestalte via de interne structuur van het project: frekwente regionale werkbesprekingen, en inschakeling van centrale medewerkers op specifieke punten.

Produktevaluatie in de zin van voorspellingen doen over de werking van programma's is uiteraard een beetje veel gevraagd. We hopen per programma-onderdeel inventarisaties te maken van uitvoeringsaspecten. De ervaringen met elke cursus afzonderlijk zullen in een afzonderlijk artikel worden beschreven.

Vaktheoretische evaluatie lijkt wellicht een vanzelfsprekendheid, maar gezien de veelgehoorde klacht dat er zo weinig aan theorievorming gebeurt lijkt het geen overbodige luxe expliciet aandacht te besteden aan het trekken van konklusies op dit vlak.

Aan evaluatie van de evaluatie zal zoals gezegd een afzonderlijk artikel gewijd worden. Daarin zullen de genoemde soorten evaluatie weer de revue passeren, maar dan met het oog op hun functieervulling. Verder levert de evaluatie van de evaluatie achtergrondinformatie bij de effectiviteitskonklusies aangaande het programma.

Tenslotte zullen er implicaties zijn richting evaluatie-onderzoek zelf (methodologie, beleid).

We beperken ons in dit artikel tot de verzameling van gegevens ten behoeve van de summatieve, proces- en produktevaluatie, en de verwerking bij de summatieve evaluatie.

3. Uitgangspunten bij de summatieve, proces- en produktevaluatie

De vormgeving van de evaluatie van GEON is op een doorslaggevende manier beïnvloed door gedachten, zoals ontwikkeld sedert 1957 door D. T. Campbell. We gaan uit van een kausaal-analytisch ervaringsbegrip. Dat wil zeggen dat we in evaluatie proberen om zicht te krijgen op oorzaken en

gevolgen. We zijn van mening dat langs die weg ook het beste greep kan worden gekregen op de in onderwijs-onderzoek zo vaak aan de orde gestelde generaliseerbaarheid. Dit betekent dat we de interne geldigheid en de externe geldigheid van een onderzoek niet als kommuniserende vaten zien, zoals soms gesteld wordt. Anders gezegd: geen externe geldigheid zonder interne geldigheid.

We zijn er vanaf het begin van overtuigd geweest dat, hoe je effectmeting ook inricht, er altijd sprake zal zijn van argumentatie achteraf: hoe betrouwbaar zijn de gegevens? hoe waardevol als criterium? hoe duidelijk te koppelen aan het programma? In de evaluatiefase van onderzoek (als fase van de empirische cyclus) gaat het om een spel van argument en tegenargument. Daarin worden onderzoeksgegevens gekombineerd met algemene inzichten, waaronder ook oordelen over wat belangrijk is.

Het is voor de onderzoeker uiteraard zaak om van te voren al met een aantal tegenargumenten rekening te houden. In de volgende paragraaf laten we zien hoe we daadwerkelijk bij design-overwegingen gebruik hebben gemaakt van een inventarisatie van een aantal mogelijke valkuilen.

Een volgende overweging was dat elke wijze van gegevensverzameling 'zwakke punten' kent. Door een strategie van verschillende wijzen van verzameling te hanteren kunnen 'zwakke' punten van de ene gekompenseerd worden met 'sterke' punten van de andere. Ook dit heeft uiteraard te maken met mogelijke valkuilen of tegenargumenten. Daarom hebben we geprobeerd om diverse typen informatie te verzamelen. Vandaar zowel een quasi-experimenteel opgezet design met metingen als verschillende vormen van registratie.

Het design kent voor- en nametingen, op projekt- en vergelijkingsscholen, bij doelgroepen en steekproeven. Op die manier denken we voldoende zicht te krijgen op factoren als 'autonome' ontwikkeling, statistische regressie, en dergelijke (zie ook Campbell and Erlebacher, 1970; Cicirelli, 1970, Thorndike, 1942). Daarop wordt behalve in paragraaf 4 nog verder ingegaan in paragraaf 7. De uitgebreide registratie is om zicht te krijgen op beginsituaties, besluitvorming, processen. Daarop gaan we in paragraaf 5 door.

We spreken van een instrumentele en een administratieve aanpak. In feite gaat het daarbij om een continuüm: vraagstellingen, tijdstippen, procedures en personen kunnen bij de gegevensverzameling meer of minder van te voren gepland worden. We proberen de gegevens aan het instrumentele uiteinde (tests) aan te vullen met gegevens omtrent

de kinderen, scholen, leerkrachten, alsmede met diverse procesgegevens. Vanaf het administratieve uiteinde proberen we case-studies aan te vullen met kategorisering en kwantificeringen.

Verder gebruiken we ook vragenlijsten en observatieschema's. De gegevens daaruit zullen tesamen met de registratiegegevens uitmonden in evaluaties per programma-onderdeel. Daarover zullen we in paragraaf 6 nog wat zeggen.

Vervolgens is er als notoir probleem de implementatie. Het ondersteunen van een schoolteam, het formatief evalueren van een programma, en het plegen van effectonderzoek sluiten niet vanzelfsprekend op elkaar aan. Het implementatieprobleem daarbij proberen we op te lossen door per kind en per leerkracht gegevens te verwerken als: bij welke leerkracht zit het kind in de klas, hoe groot is die klas, wat doet die leerkracht aan het programma; een en ander is per half jaar gekodeerd. Zie daarvoor verder paragraaf 8. Dit betekent dat de evaluatie zich aanpast bij de uitvoering zoals de leerkrachten die vormgeven. Grotere researchtechnische zuiverheid eisen zou juist verstorend werken.

Dit laatste geldt in onze ogen algemener. We hanteren bewust een quasi-experimentele strategie omdat je wat je bijvoorbeeld door randomisering aan statistische zuiverheid wint, weer verliest aan 'empirische' tegenargumenten, met name ten aanzien van externe geldigheid.

Hiermee zeggen we ook dat je niet in het algemeen kwaliteitsoordelen over methoden voor evaluatieonderzoek kunt vellen. Beslissingen daarover hebben te maken met vragen als *welk* programma (summatief) geëvalueerd moet worden en *wat* je ervan wilt weten. Een decisiegerichte aanpak meenden wij te moeten afwijzen als zijnde niet reëel.

Een en ander heeft er ook toe geleid dat we verhoudingsgewijs weinig ingewikkelde statistische analysetechnieken hebben gebruikt. Dat wordt in paragraaf 9 uitgewerkt. We meenden het meer te moeten zoeken in 'empirische' argumenten, en hebben daarom een aantal replikaties ingebouwd (zie paragraaf 1).

Tenslotte kozen we expliciet voor het afnemen van tests. We zoeken evaluatie-kriteria met name ook op kind-niveau om niet in de valkuil te stappen dat je denkt met het realiseren van andere onderwijsprocessen al je uiteindelijke doelstellingen te bereiken.

4. Problemen bij het trekken van konklusies

De aanpak van Campbell en Stanley lijkt een beetje

op de uitdrukking: een keten is net zo sterk als zijn zwakste schakel. Steines (zie Leinfeller en Köhler, 1974, pp. 417-430) noemt iets dergelijks het 'Strategic Combination Argument'. Met deze werkwijze wordt iets heel menselijks benadrukt: je gaat af op het al dan niet aanwezig zijn van tegenargumenten. Wetenschapsfilosofisch gezien betekent het een onderstreping van Poppers kriticisme. Hogben stelt terecht dat voor (een dergelijke opvatting van) experimenteren niet essentieel is of daadwerkelijk wordt ingegrepen en/of de gegevensverzameling al dan niet nog moet plaatsvinden (Hogben, 1957, pp. 377, 381). Het gaat inderdaad om het experiment als denkschema. Voorbeelden van dergelijke schema's zijn verder ook te vinden bij onder meer Solomon (1949) en Neale en Liebert (1973).

Campbell (Campbell, 1957; Campbell en Stanley, 1963) onderscheidt interne en externe geldigheid. Het gaat om geldigheid van de onderzoeksopzet, en daarmee van de uitkomsten. Interne geldigheid betreft de kausaliteit; externe geldigheid de generaliseerbaarheid. In genoemde publikaties wordt tevens aandacht geschonken aan een aantal bedreigingen voor geldigheid, vooral de interne. Enkele jaren later werden door Bracht en Glass (1968) de door Campbell al onderscheiden valkuilen ten aanzien van de externe geldigheid nog met een aantal uitgebreid (zie ook Rosenthal and Rosnow, 1969; Linschoten, 1964).

Hoewel de opsomming van dergelijke valkuilen regelmatig in de een of andere vorm werd herhaald kan niet worden gekonstateerd dat in de evaluatie-

fase van onderzoek het langslopen van een aantal van zulke factoren gewoonte is geworden. Nu gaat het bij die valkuilen deels om zaken die al veel langer bekend zijn, in de vorm van methodologische eisen altijd meespelen, en/of logisch gesproken voor de hand liggen. Voorbeelden zijn respectievelijk statistische regressie naar het gemiddelde (zie Thorndike, 1942), meet(on)betrouwbaarheid, interveniërende variabelen. Weliswaar zijn daarnaast nog factoren waarvan de ernst vooral pas algemener werd beseft door apart daarnaar verricht onderzoek, bijvoorbeeld aangaande de invloed die van het onderzoeken zelf kan uitgaan. Maar over het geheel genomen gaat het niet om bijzondere vondsten. Het gaat om konklusie-bedreigende factoren, met alternatieve verklaringen voor gevonden effecten.

Dat hier toch de aandacht daarvoor wordt gevraagd heeft drie redenen. Ten eerste lijkt het een goede zaak om de uitkomsten van elk (quasi) experiment te confronteren met een lijst van zulke factoren, om er redelijk zeker van te zijn dat je geen valkuil over het hoofd ziet. Het zijn er zo langzamerhand enkele tientallen, zodat de kans op 'vergeten' niet irreëel is. In de tweede plaats zou een dergelijke werkwijze expliciet maken (voor onderzoeker en lezer beiden) hoe de uitkomsten van een onderzoek in functie staan van gehanteerde methoden van onderzoeken en redeneren. En in de derde plaats zou, vooral indien in zo'n opsomming systematiek is aangebracht, met de verschillende factoren al bij de planning van een onderzoek rekening kunnen worden gehouden. Dan zou kunnen worden voor-

| | INTERNE GELDIGHEID | EXTERNE GELDIGHEID |
|-------------------|---|--|
| MEETINSTRUMENTEN | <ol style="list-style-type: none"> 1. Meetinstrumentenverandering 2. Meetinstrumentenongeldigheid 3. Meetinstrumentenonbetrouwbaarheid 4. Reactieve metingen | <ol style="list-style-type: none"> 14. Reactieve metingen 15. Onvergelijkbare metingen |
| ONDERZOEKSGROEPEN | <ol style="list-style-type: none"> 5. Onvergelijkbare groepen 6. Niet-representatief verlies 7. Autonome ontwikkeling 8. Statistische regressie 9. Steekproefspecifiteit | <ol style="list-style-type: none"> 16. Onvolledige bekende steekproefpopulatie verhouding 17. Niet-representatieve steekproef 18. Steekproefspecifiteit |
| BEINVLOEDING | <ol style="list-style-type: none"> 10. Beïnvloedingsverandering 11. Onduidelijke vergelijking 12. Diffusieverschijnselen 13. Unieke gebeurtenissen | <ol style="list-style-type: none"> 19. Meervoudige beïnvloeding 20. Onduidelijke beïnvloeding 21. Reactieve situatie 22. Unieke gebeurtenissen |

Figuur 2 *Overzicht van valkuilen*

komen dat achteraf blijkt dat onevenredig veel aandacht is besteed aan, bijvoorbeeld, de metingen, ten koste van de controle op de uitvoering van een programma of andere factoren. Uiteindelijk zou langs die weg wellicht zelfs een bijdrage aan onderzoeksbeleid geleverd kunnen worden. Er zou kunnen blijken uit een analyse van een aantal onderzoeken, dat bepaalde valkuilen wel bijzonder vaak plausibel zijn, hetzij door de eigenheid van het werkterrein (in dit geval onderwijs), hetzij doordat vaak de onderzoeksopzet dezelfde manko's vertoont (En als dat te vaak gebeurt werkt de communicatie in de evaluatiefase kennelijk onvoldoende).

In het GEON-project werd het overzicht van valkuilen gehanteerd zoals in Figuur 2 is aangegeven.

Indien elke verandering in de gegevens zou duiden op verandering van de afhankelijke variabele(n) ten gevolge van de onafhankelijke, en elk ontbreken van verschil in de meetresultaten zou betekenen dat ook in de afhankelijke variabele(n) geen wijziging is opgetreden, dan zou deze paragraaf overbodig zijn. Het kan echter gebeuren dat de meetresultaten niet overeenkomen terwijl er toch geen verandering in de afhankelijke variabele(n) heeft plaatsgevonden (zie de nummers 1, 2, 3, 5, 6, 8). Ook het omgekeerde is mogelijk: het ten onrechte niet optreden van verschil in de gegevens, hetgeen slechts op toevallige en meer klinische wijze konstaterbaar zou zijn.

Bovendien kunnen de meetresultaten terecht uiteenlopen, zonder dat dit geweten mag worden aan de daarvoor bedoelde activiteiten (zie de nummers 4, 7, 10, 12, 13, 14, 19, 20, 21, 22).

De meeste standaard-alternatieve verklaringsmogelijkheden hebben een systematisch karakter. Niet systematisch zijn: 1, 6, 10, 12, 13, 22.

Zeer belangrijk in de onderwijsresearch is het verschijnsel van de 'self-fulfilling prophesy', die kan inwerken op (onderdeel kan uitmaken van) 4, 5, 11, 15, 18, 19, 20, 21.

Bovenstaande bedreigingen zijn niet alle even reëel en kunnen ook niet alle op dezelfde wijze worden onschadelijk gemaakt.

- Het idee 'Steekproefspecifiteit' (9) lijkt meer het produkt van iemand die slim wil zijn, dan een werkelijke bedreiging.
- Sommige bedreigingen zijn reeds min of meer ondervangen door de kenmerken van het project, zoals reactieve metingen (14) en niet-representatieve steekproef (17) (respectievelijk: zelfbeoordelingslijsten e.d. zijn een integraal onderdeel van het programma, en: de vier regio's garanderen een

redelijke representativiteit op landelijke schaal).

- Andere zijn betrekkelijk eenvoudig te ondervangen, zoals 6, 15 en 18 (niet-representatief verlies behoeft geen bedreiging te vormen als de gegevens op individueel niveau beschikbaar zijn en blijven; onvergelykbare metingen zullen zich niet voordoen als bij generalisatie wordt aangegeven op welke wijze resultaten mogen worden beoordeeld/gemeten; steekproefspecifiteit is – althans methodologisch! – geen probleem doordat kan worden aangegeven naar welke groepen gegeneraliseerd mag worden).
- Weer andere zijn per definitie niet geheel te vermijden: 12, 18, 21 en 22 (respectievelijk diffusieverschijnselen; onbekende steekproefspecifiteit; reactieve situatie; unieke gebeurtenissen). Dergelijke factoren kan men ten hoogste trachten te beoordelen op de invloed die ze wellicht kunnen hebben.
- Van de resterende bedreigingen kunnen worden gekeerd:
 - . door middel van een adekwate onderzoekopzet (design): 1 (meetinstrumentverandering), 10 (beïnvloedingsverandering) en 11 (onduidelijke vergelijking);
 - . door middel van statistische analyse: 5 (onvergelykbare groepen) en 8 (extreme steekproef);
 - . door zo goed mogelijke meetinstrumenten te gebruiken: 2 en 3 (meetinstrumentenongeldigheid en -onbetrouwbaarheid).
- Problematisch kunnen dan nog worden:
 - . 16 (onvolledig bekende steekproef-populatieverhouding): dat wat het project hieraan kan doen beperkt zich voornamelijk tot aangeven van de vermoedelijke relevante aspecten in de hier bedoelde vergelijking;
 - . 19 en 20 (meervoudige en onduidelijke beïnvloeding): het eerste betekent dat de oorzaken van effecten niet geheel isoleerbaar zullen zijn naar afzonderlijke programma-onderdelen (hoewel door middel van frekwente voor- en nametingen, gesprekken en observaties, en verschil in projectuitvoering tussen scholen en regio's, waarschijnlijk de merites van de afzonderlijke onderdelen vrij goed beoordeelbaar zullen zijn); het tweede houdt de noodzaak in de programma-onderdelen en innovatie-activiteiten zo goed mogelijk te expliceren (hetgeen moeilijk vooraf kan worden gedaan, aangezien veel activiteiten zullen afhangen van de plaatselijke situatie en de ontwikkeling daarin, en aangezien het van de programma-onderdelen juist de bedoeling is dat ze creatief gebruikt worden en in de praktijk geïntegreerd).

Uitgaande van de ervaringen van het SVO-project 0 83 levert de komponent 'onderzoeksgroepen' de grootste problemen op voor de interne geldigheid (met name valkuil 5-8), terwijl ten aanzien van de externe geldigheid het grootste probleem is gelegen in de komponent 'beïnvloeding' (met name valkuil 19-21). Gezien de gemaakte analyse heeft voor het GEON-project de werkelijk serieuze methodologische problematiek vooral betrekking op het laatste konglomeraat: de bedreigingen 19-21. Dit vormt ook de oorspronkelijke en belangrijkste aanleiding voor systematisering van de zogenaamde administratieve benadering. (De verschillende vormen van registratie).

Het laatste stel problemen wordt wel aangeduid met 'implementatie', notoir in onderwijsresearch. Daarop gaan we in paragraaf 8 verder in (zie ook Anderson, et al., 1978; Cicirelli, et al., 1970; Coulson, 1978; House, et al., 1978; Light and Smith, 1970; Smith and Bissell, 1970). In het licht van bovenstaande krijgt het toetsen op statistische significantie iets minder zwaar aksent. Dat komt in paragraaf 9 nog aan de orde.

5. De administratieve aanpak

Hierboven (paragraaf 3) spraken we van een continuüm: je kunt het verzamelen van gegevens op verschillende aspecten meer of minder van te voren plannen. Als je van te voren vastlegt wie/wanneer/hoe/waarop let, spreek je van meetinstrumenten. In die sektor hanteren we in GEON vragenlijsten, observaties, tests en toetsen. Zie daarvoor de paragrafen 6 en 7. Daarnaast hebben we een aantal registratiemiddelen gehanteerd. De hulpmiddelen voor registratie bestonden uit kaarten, lijsten en formulieren, aan de hand waarvan:

- gegevens van kinderen, leerkrachten, e.d. konden worden vastgelegd;

- de beginsituatie per buurt en school kon worden geïnventariseerd;
- het verloop per programma-onderdeel kon worden aangegeven;
- de contacten met en van het schoolteam konden worden vastgelegd (begeleidingsgesprekken, ouderavonden, besprekingen k.s. - l.s., e.a.);
- observaties konden worden verricht;
- meningen konden worden verzameld;
- bepaalde gegevens voor verdere verwerking worden klaargemaakt.

Deze gegevens zijn en worden gebruikt bij:

- het maken van periodeverslagen per regio;
- het maken van case-studies;
- het maken van deelevaluatie per programma-onderdeel;
- het kwantitatief analyseren, mede in relatie tot testgegevens (zie paragraaf 8).

Bij de evaluatie van de evaluatie-aanpak hopen we meer gedetailleerd op de verschillende middelen en procedures in te gaan. We volstaan nu met geven van een overzicht van de middelen in Figuur 3.

De kaarten zijn vooral bedoeld voor het neerslaan van allerlei 'zakelijke' gegevens, zoals leeftijden en aantallen.

De checklijsten zijn een ondersteuning bij het inventariseren van beginsituaties.

De vragenlijsten vóór en na elke cursus bevatten een aantal open vragen, waarmee planning (vooraf) en evaluatie (direkt na afloop) worden ondersteund.

De registratieformulieren maken het mogelijk om in één oogopslag het verloop van een cursus op een school te overzien (data, problemen, e.a.). De klassikale overzichtsformulieren zijn vooral bedoeld voor verdere verwerking in kwantitatieve zin van diverse gegevens.

De contactenkaart is misschien wel het belangrijkste registratiehulpmiddel. Daarop wordt van elk gesprek en activiteit aantekening gehouden. Het gaat een heel eind in de richting van een dagboek.

Klaskaart
Kindkaart
Leerkrachtkaart
Schoolkaart

Checklist Buurt
Checklist Buurtinstanties
Checklist School
Checklist Gezin

Open vragenlijst vóór en na de cursussen, voor elke cursus apart
Registratieformulieren bij de cursussen
Klassikale overzichtsformulieren voor test- en toetsgegevens
Kontaktenkaart (notulering per gesprek, contact, activiteit)

Figuur 3 Overzicht van hulpmiddelen voor procesregistratie

Alle hulpmiddelen functioneren in handen van de extra leidster (zie paragraaf 1). Logistiek gezien betekent dat een keus voor een gedecentraliseerde opslag. De extra leidster benut diverse formulieren ook zelf om haar werk goed te kunnen doen. Methodologisch is het een keus voor een vorm van participerende observatie.

6. De evaluatie per programma-onderdeel

De belangrijkste evaluaties van een cursus vinden plaats als een leerkracht die cursus doorwerkt, begeleid door de extra leidster. De neerslag van allerlei opmerkingen, problemen en ervaringen is terug te vinden op contactkaarten. En verder als een medewerker van de plaatselijke onderwijsbegeleidingsdienst in samenwerking met een schoolteam en een regionaal project-medewerkster, die cursus leert begeleiden.

Deze ervaringen monden uit in verbetering, eventueel aanvulling, van de cursus. Die verbetering heeft prioriteit gekregen boven het op schrift zetten van evaluaties per programma-onderdeel voor rapportagedoeleinden. De formatieve evaluatie zal afzonderlijk verantwoord worden. Produktievevaluatie per programma-onderdeel zal vooral bestaan uit inventarisatie van uitvoeringsaspecten (zie ook paragraaf 2): wat is er daadwerkelijk aan/mee gedaan, wat waren de problemen, wat waren (neven)effecten. Deels als ondersteuning voor de summatieve evaluatie, deels als bron voor begeleidingssuggesties.

'Gemeten' wordt per cursus alleen via vragenlijsten en observaties. De vragenlijsten zijn in multiple choice vorm en worden na afloop door de leerkracht ingevuld. Onderwerpen zijn: bedoeling van de cursus, tekst, hulpmateriaal, nut, toepasbaarheid, begeleidingsaspecten, contacten, problemen. Observatieformulieren worden gebruikt bij het bekijken

van videobanden. In het kader van diverse cursussen worden namelijk video-opnamen gemaakt. Een ander heeft zowel begeleidings- als evaluatiefuncties. Het observeren gebeurt meestal door leerkracht en projectmedewerkster samen. Het evalueren is daarmee onderdeel van de cursus. Dat schept geen generalisatieproblemen omdat een belangrijk deel van de vaardigheden waar het bij de inservice-training in GEON om gaat juist het gericht verzamelen en benutten van informatie betreft (over individuele kinderen, de klasseorganisatie, de gezinsachtergrond, e.a.). In elke cursus zijn dan ook formulieren opgenomen, met behulp waarvan de leerkracht zichzelf een feedback omtrent haar eigen handelen kan voorzien. We vatten een en ander samen in Figuur 4.

De cursus Sociale Waarneming en Vaardigheid is nog in ontwikkeling. Bij de cursus Voor het Lezen zit eenzelfde type werkljst als bij Voorspellen en Reageren, waar deze cursus een mogelijk vervolg op is. Tenslotte vermelden we dat bij elke takenreeks (zie paragraaf 1, maar ook paragraaf 7) een toets hoort, voor plaatsings- zowel als (summatieve) evaluatiedoeleinden.

7. Afname van tests

Summatieve evaluatie vindt plaats ten aanzien van de cursussen (door te letten op gedrag, meningen en vaardigheden bij de kleuterleidsters), de programma's voor de kleuters (door middel van observaties en van takenreeks-toetsen), en de activiteiten-suggesties (door te letten op de verandering van contactstructuren). Het geheel wordt, tesamen met de gegeven begeleiding (ook voor zover deze zich uitstrekt tot de lagere schoolperiode) summatief geëvalueerd door middel van de Wechsler Preschool and Primary Scale of Intelligence (afname op 4-jarige en op 6-jarige leeftijd) en de Wechsler Intelligence

| <i>Verwoorden van Probleemgedrag</i> | <i>Zelfstandig Werken</i> | <i>Voorspellen en Reageren</i> |
|--|---|--|
| Vragen per hoofdstuk. | Observatieformulier bij video-opnames. | Werklijst met vragen en opdrachten. |
| Opdrachten per hoofdstuk. Observatieschema's. | Vragenlijst bij klasselessen. Observatielijsten bij klasselessen. Zelfbeoordelijsten. | Schema ontwikkelingsgebieden. Opdrachten per hoofdstuk |
| Vragenlijst, multiple choice, na afloop. | Vragenlijst multiple choice na afloop. | Observatieformulier bij video-opnames. Vragenlijst multiple choice na afloop. |

Figuur 4 Evaluatiemiddelen per cursus, te gebruiken/in te vullen door de leerkracht en/of de extra leidster

Scale for Children (afname op 8-jarige leeftijd), alsmede een toets voor begrijpend lezen (afname eveneens aan het eind van de 2e klas van de lagere school). (Begrijpend Lezen II).

Andere toetsgegevens worden, indien beschikbaar, mee verwerkt (Leesvoorwaardentoets; Begrijpend Lezen I).

Het afnameschema is weergegeven in Figuur 5.

| Instrument | 4 jaar | 6 jaar | 8 jaar |
|--------------|-------------|--------|--------|
| WPPSI | x | x | |
| TRT I t/m IV | xx xx xx xx | | |
| WISC | | | x |
| BLT | | | x |

Figuur 5 Afname-schema tests en toetsen

De tijd tussen voorafname en na-afname is bij de WPPSI gemiddeld 19 maanden, bij de takenreeksen gemiddeld 3 maanden. De WISC wordt afgenomen aan alle kinderen (projektscholen én vergelijkingsscholen) waar we ook WPPSI-gegevens van hebben. Bovendien worden op die lagere scholen waar de meeste kleuters terecht zijn gekomen ook steekproeven getrokken en met WISC getest; dit om de vergelijkbaarheid over de 6-8 periode te bevorderen.

Kinderen die zijn blijven zitten of naar BUO verwezen, wordt óók de WISC afgenomen. De BLT (toets Begrijpend Lezen II) wordt uiteraard alleen in de 2e klas gehanteerd.

De WPPSI is een betrekkelijk nieuw instrument.

Bij de konstruktie ervan is aangesloten bij de al langer bestaande WISC. De WPPSI is geschikt voor gebruik bij kinderen van 4 tot 6½ jaar. In het land van herkomst is het instrument geijkt op een steekproef van 1200 kinderen. Daardoor zijn naast de ruwe scores ook gewogen scores beschikbaar. De WPPSI bestaat in standaardvorm uit 10 subtests, 5 'verbale' en 5 'performale' (=handelingsbetrokken). De Nederlandse vertaling is van A. Voswinkel. Een Nederlandse normering is nog niet beschikbaar.

De WISC die we in aansluiting bij 8-jarigen gebruiken is de Nederlandse vertaling van de WISC-R (revised) door P.P. van Haasen.

De BLT is de CITO-toets Begrijpend Lezen II van O. de Wit en W. van Soest. De toetsprocedure gedurende de kleuterschoolperiode was als weergegeven in Figuur 6.

Op de geschetste manier kunnen we enigszins zicht krijgen op het belang van de volgende vier factoren:

1. de ontwikkeling van de kinderen als ze op een kleuterschool gaan: de vooruitgang in de scores van de steekproef op de vergelijkingsscholen;
2. de mate waarin statistische regressie optreedt: wat de doelgroep op de vergelijkingsscholen méér vooruitgaat dan de steekproef op de vergelijkingsscholen;
3. het effect van het zitten in een klas waar de juf een GEON-begeleiding krijgt: wat de steekproef op de projektscholen méér vooruitgaat dan de steekproef op de vergelijkingsscholen;
4. het effect van de eventuele extra activiteiten: wat de doelgroep op de projektscholen dan nog extra vooruitgaat.

| Exp. konditie | Instrument | VOOR | NA |
|----------------------|--------------------|--|---|
| Projektscholen | WPPSI | doelgroep ¹ (pd) + steekproef (ps) | dezelfde kinderen als bij de voortest |
| | takenreeksstoetsen | alle kleuters ² | doelgroep ³ + steekproef |
| Vergelijkingsscholen | WPPSI | zoals op de projektscholen (vd resp. vs) | |
| | takenreeksstoetsen | doelgroep ¹ + steekproef | dezelfde kinderen als bij de voortoets |

Figuur 6 Toetsprocedure kleuterschoolperiode

- ad 1. bepalen van de doelgroep vóóraf geschiedt door middel van gezamenlijke beoordeling van het functioneren van de kinderen;
- ad 2. de te toetsen kleuters vormen hier een leeftijdsgroep, afhankelijk van de takenreeks in kwestie;
- ad 3. doelgroep is hier de groep die (mede op grond van de voortoetsgegevens) werd betrokken in de extra activiteiten.

Een en ander per takenreeks, en voor de WPPSI.

Bij deze testgegevens werden tot dusverre steeds de volgende aanvullende gegevens betrokken (zie ook paragraaf 8):

- van het kind (geboortedatum, geslacht);
- van de afname (testdata, proefleiding, experimentele konditie);
- van de ouders (beroep, opleiding);
- van de leerkrachten (leeftijd, ervaring, deelname aan het project);
- van de klassen (klassegrootte; wanneer bij wie in de klas);
- van de school (schoolgrootte, teamstructuur, ouderkontakten).

De keuze voor ook niet-curriculumspecifieke instrumenten zullen we verantwoorden in het artikel over de summatieve evaluatie-uitkomsten.

8. Verwerking van gegevens en het implementatieprobleem

De summatieve evaluatie beslaat dus met name testgegevens van kinderen. De interpretatie wordt ondersteund door systematische procesgegevens (dit om te voorkomen dat je resultaten meet zonder te weten waarvan). Het design is het quasi-experimentele 'nonequivalent pretest-posttest control group design'.

Het implementatieprobleem geldt in GEON des te meer omdat een pakket programma-onderdelen wordt gehanteerd, waarbij de begeleide leerkrachten zelf mee bepalen welk onderdeel ze wanneer/hoe lang/hoe intensief doorwerken. Daarbovenop komt dan nog het probleem dat als je evaluatiecriteria zoekt bij de kinderen, die kinderen a) niet allemaal evenveel extra aandacht krijgen (mede door het genoemde, maar ook los daarvan, door positieve diskriminatie namelijk, hetgeen juist beoogd werd), en b) niet allemaal twee jaar lang bij dezelfde leerkracht in de klas zitten.

Zoals in paragraaf 4 staat uitgewerkt vormt dit methodologisch bezien een van de lastigste zaken bij de evaluatie van GEON: dat het programma bestaat uit een pakket (i.p.v. één cursus) dat dan ook nog eens wordt uitgevoerd aangepast aan de school, dus op flexibele wijze. De oplossing voor dit probleem proberen we te vinden in een tweevoudige aanpak:

- I. Gedetailleerde procesbeschrijvingen (Case-studies);
- II. Een kobinatie van onze administratieve en instrumentele aanpak.

Hier is II aan de orde. Van de kinderen komen testgegevens beschikbaar. Per leidster wordt bijgehouden wanneer ze welk onderdeel doet, hoe

intensief, etc. Verwerkt is nu langs kwantitatieve geautomatiseerde weg (we hebben het hier alleen over de deelnamegegevens uit het lijstje aan het eind van de vorige paragraaf):

- per leidster welke cursussen ze helemaal doorwerkte, welke gedeeltelijk, en welke takenreeksen ze deed, een en ander per halfjaar gespecificeerd;
- per kind: welke cursussen etc. gedaan werden door de leidsters waarbij het kind in de klas zat, 1) op het moment dat het kind in die klas zat, 2) voordat het kind bij die leidster kwam; een en ander ook weer per halfjaar gespecificeerd.

Dit onderscheid is gemaakt, omdat er rekening mee is gehouden dat het verschil kan uitmaken of een kind 'aan den lijve' een leidster meemaakt die op dat moment een cursus doet, of bij een leidster in de klas komt die - eerder - een cursus deed. Je zou kunnen zeggen dat we met 2) aan retentiemeting doen t.a.v. de leidster. Een en ander is afgezet tegen de testresultaten. Daarbij bleek het geen overbodige luxe dat niet alleen per school (team en leidsters) is bezien hoe de deelname lag, maar ook per kind waarvan we de testgegevens verwerken rekening wordt gehouden met de vraag wanneer elk kind bij welke leidster zat. Het deelnamepatroon ten aanzien van de kinderen ziet er, op deze manier gekoppeld, namelijk toch wel weer anders uit als dat ten aanzien van de leidsters. Op die manier konden we per kind een aantal deelnamegegevens specificeren, die in feite dus slaan op de leerkracht(en) waar het bij in de klas zat. De volgorde waarin leerkrachten onderdelen deden, alsmede de situatie dat een leerkracht vóórdat een kind in de klas komt een cursus gedaan kan hebben, worden zo tevens meegenomen in de analyse. Daarbij werden dan niet alleen de testgegevens betrokken, maar ook aanvullende gegevens als beroep en opleiding van de ouders, klassegrootte, leeftijd en ervaring van de leerkracht, en zo meer.

Strategie was om:

1. via korrelatieberekeningen en kruistabelleringen zicht te krijgen op relaties tussen de diverse randfactoren (deelnamegegevens en aanvullende gegevens) onderling;
2. via significantieberekeningen en kruistabelleringen zicht te krijgen op relaties tussen randfactoren en testgegevens.

9. Statistische analyse

Aan het gebruik van (ko-)variantie-analyse en verwante technieken zijn in dit verband twee zaken belangrijk. In de eerste plaats ligt daarbij het aksent nogal op toetsen. En vervolgens wordt gewerkt met

zogenoemde parametrische methoden, waarbij nogal wat assumpties behoren. (Zie ook Bloxham, 1976; Brewer, 1972; Carver, 1978; Cox, 1977; Hogben, 1957; Lord, 1967 & 1969; Morrison and Henkel, 1970; Neyman, 1977; Porter and Chibucos, 1974; Verberk, 1970).

Het laatste heeft te maken met geldigheidsvragen. Als aan assumpties niet is voldaan toets je een andere hypothese dan je wilt en/of toets je tegen een ander significantieniveau dan je wilt. Weliswaar is er veel, geruststellend, zogenaamd Robuustheidsonderzoek gedaan (zie o.a. Glass, et al., 1972; Havlicek and Peterson, 1974), maar zulk onderzoek is zelf aan beperkingen onderhevig, en de mogelijkheden om in een concreet onderzoek na te gaan in hoeverre aan diverse voorwaarden is voldaan, zijn eveneens beperkt.

Ten aanzien van het aksent op toetsen spelen bruikbaarheidsvragen een rol. Het belang van significantieberekening kan in het algemeen worden gerelativeerd, deels gezien het ontbreken van enige vorm van aselektering (geen aselekte steekproeven, geen randomisering), deels door de in het geval van GEON ingebouwde replikaties. Effekten die zich in andere regio's/rondes herhalen zullen betrouwbaar zijn. We hopen op een en ander in artikelvorm terug te komen, en volstaan hier met een verwijzing naar de dissertatie van de auteur (Stokking, 1979).

10. Slot

Ter afsluiting willen we graag nog even aandacht besteden aan de vraag wat evaluatie-onderzoek zoals hierboven beschreven nu in feite voor soort onderzoek is: toetsend onderzoek? explorerend onderzoek? beschrijvend onderzoek? In onze opinie komt het erop neer dat je vraagt wat voor functie het evaluatie-onderzoek heeft (zie ook Scriven, 1972; Blankertz, 1977; Stufflebeam, et al., 1972; Suchmann, 1967).

In de nota evaluatie-aanpak (1975) wordt opgemerkt dat tot op zekere hoogte uit dit project inzichten kunnen worden verwacht naar de mate waarmee ze er worden ingestopt. Immers: als je niet weet waar je naar moet kijken zie je niets of minstens ook een aantal irrelevante dingen.

GEON heeft zoveel mogelijk een en ander open willen houden; er is bijvoorbeeld pas gaandeweg geëxpliciteerd welke zaken de moeite waard waren om beschreven te worden. Dat geldt met name voor de administratieve poot onder de evaluatie. De instrumentele was al in een vroeger stadium gepland; maar ook ten aanzien daarvan heeft GEON tot dusver weinig theoretisch geëxpliciteerd.

Resumerend: er is geen sprake van theorie-toetsend onderzoek in de zin van Popper cs. Weliswaar wordt er statistisch getoetst, maar dat is meer bij wijze van effectmeting, Je zou kunnen spreken van het toetsen van veronderstellingen over de realiseerbaarheid van veranderingen. Aangezien het projekt ingrijpt in de (sociale) werkelijkheid is er geen sprake van puur beschrijvend onderzoek, maar we proberen wel beginsituaties te beschrijven. Waar het ons nu om gaat is dat naar onze mening sprake is van een vorm van onderzoek die je met aktie-onderzoek zou kunnen aanduiden, en dat zulk onderzoek zowel toetsend, beschrijvend als explorerend van karakter is. Dat betekent niet dat allerlei methodologische standaarden worden losgelaten (zie ook Nauta, 1977).

Wat meestal genoemd wordt als kenmerken van aktie-onderzoek zijn in onze ogen niet de meest essentiële kenmerken. Genoemd worden wel:

- je wilt iets bereiken;
- je werkt 'procesmatig';
- je werkt planmatig;
- je maakt gebruik van mechanismen die in de researchmethodologie als ongewenst worden beschouwd, zoals het op zich beïnvloedend zijn van metingen;
- je bent ook gericht op theorievorming;
- de 'doelgroep' is betrokken bij het onderzoek.

Voor een deel gaat het hier weliswaar om belangrijke aspecten, maar die belangrijkheid wordt niet altijd geëxpliciteerd. Als je dat wel doet kunnen de positieve kanten meer naar boven komen, maar óók de problematische. Zo betekent het procesmatige ongetwijfeld dat de onderzoeksvraagstelling kan verschuiven. Daarmee ontstaat een al dan niet bewust gebruikte, al dan niet als positief ervaren speelruimte. Maar tevens wordt het lastig om vooraf einkriteria voor de evaluatie op te stellen. En het planmatige geeft een zekere herhaalbaarheidsgarantie, maar het benutten van bijvoorbeeld de reactiviteit van metingen betekent dat de onderzoekende en de beïnvloedende functies ook wel eens onontwaaarbaar verstrengeld kunnen raken.

Het meest essentiële echter is naar onze mening gelegen in het werken vanuit een streefrichting, het gaandeweg, en steeds opnieuw, aantonen van mogelijkheden. Je kreëert 'nieuwe feiten' (althans dat probeer je). Het is misschien ook wel merkwaardig om in de sociale werkelijkheid die in zo grote mate door mensen (samen) gemaakt wordt alleen dat te willen zien (onderzoeken) wat er is. Voorzover je uiteraard moeilijk anders kunt bedoelen we wat bijvoorbeeld Kohnstamm (1976, p. 222; beschrijft als aandacht voor wat ooit maximaal mogelijk bleek.

Daarbij is wat Van Calcar (1976) aksentueert: het zichtbaar maken van objectieve mogelijkheden tegen vooroordelen in, niet essentieel, in de zin dat niet per se sprake behoeft te zijn van vooroordelen. Essentieel is : nagaan in hoeverre die sociale werkelijkheid inderdaad gemaakt kan worden. We gaan verder niet in op de rol van de specifieke vraagstelling of streefrichting; het is duidelijk dat er niet steeds evenveel valt te onderzoeken (Zie onder meer Hofstee, 1975; Kuypers, 1968; Lepenies and Nolte, 1971). Wat tenslotte belangrijk is, is de vraag naar de betekenis voor onderwijsbeleid. Anders gezegd: wat wordt er met de, meestal in projecten verkregen, resultaten gedaan?

Slotopmerkingen

Naar onze mening wordt onderzoeksinformatie niet op zo'n manier bij beleidsbeslissingen betrokken dat gebruik van een decisiemodel daarvoor zinvol is. De waardering van de resultaten hangt van veel factoren af: beschikbare alternatieven, actuele problemen, de wijze waarop de resultaten verkregen zijn, overkoepelende beleidsdoelstellingen. Generalisatie is uiteindelijk niet alleen een kwestie van onderzoeksmatig generaliseerbaarheid aantonen, maar van het scheppen van kondities. Beslissend is dan of de wil bestaat om ruimer gebruik te bevorderen.

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 (48,2) 161-170.
- Beekman, A. J., *Dienstbaar inzicht. Opvoedingswetenschap als sociale planwetenschap*. Groningen, 1972.
- Berk, K. N. and J. S. Francis, A Review of the manuals for BMDP and SPSS, *Journal of the American Statistical Association*, 1978 (73,361) 65-71.
- Birnbaum, A., The Neyman-Pearson Theory as Decision Theory, and as Inference Theory, *Synthese*, 1977 (36,1) 19-49.
- Blalock, H. M., *Causal inference in nonexperimental research*. New York, 1972.
- Blankertz, H., Zur Bewährungskontrolle erzieherischen Handelns, in: D. Benner (Hrsg.), *Aspekte und Probleme einer pädagogischen Handlungswissenschaft*. Kastellaun, 1977, 43-57.
- Bloxham, M. A., Note on publication and the value of significance tests, *Theory and Decision*, 1976 (7,1/2) 135-139.
- Bracht, G. H. and G. W. Glass, The external validity of experiments, *American Educational Research Journal*, 1968 (5) 437-474.
- Brewer, J. K., On the power of statistical tests in the American Educational Research Journal, *American Educational Research Journal*, 1972 (9,3) 391-401.
- Brink, W. P. van den, Psychologie, de wetenschap van het toeval? *Tijdschrift voor Onderwijsresearch*, 1978 (3,3) 133-136.
- Calcar, Co van, *Innovatieproject Amsterdam, I + II*. Amsterdam, 1976.
- Campbell, D. T., Validity of experiments in social settings, *Psychological Bulletin*, 1957 (54) 297-312.
- Campbell, D. T. and J. C. Stanley, *Experimental and quasi-experimental designs for research on teaching*. Chicago, 1963.
- Campbell, D. T. and A. Erlebacher, How regression artifacts in quasi-experimental evaluation can mistakenly make compensatory education look harmful, in: I. Hellmuth, (Ed.), *Disadvantaged Child*, volume 3. New York, 1970, 185-210.
- Carver, R. P., The case against statistical significance testing, *Harvard Educational Review*, 1978 (48,3) 378-399.
- 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.
- Cicirelli, V. G., The Relevance of the Regression Artifact Problem to the Westinghouse-Ohio Evaluation of Head Start: A Reply to Campbell and Erlebacher, in: I. Hellmuth, (Ed.), *Disadvantaged Child*, volume 3. New York, 1970, 211-215.
- Cohen, J., Some statistical issues in psychological research, in: B. Wolman, (Ed.), *Handbook of clinical psychology*. New York, 1965, 95-121.
- Cohen, J., *Statistical power analysis for the behavioral sciences*. New York, 1969.
- 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.
- Cox, D. R., The Role of Significance Tests, *Scandinavian Journal of Statistics*, 1977 (4) 49-70.
- Edwards, K. I., Summative Evaluation: Some basic considerations, in: G. D. Borich, (Ed.), *Evaluating educational programs and products*. Englewood Cliffs, 1974, 373-399.
- Elashoff, J. D., Analysis of covariance: a delicate instrument, *American Educational Research Journal*, 1969 (6,3) 383-399.
- Ellis, B., *Basic concepts of measurement*. Cambridge, 1968.
- Glass, G. V., P. D. Peckham and J. R. Sanders, Consequences of failure to meet assumptions underlying the fixed analysis of variance and covariance, *Review of Educational Research*, 1972 (42,2) 337-384.
- Guttman, L., *What is not what in statistics*. Paper presented to the Annual Meeting of the Israel Statistical Association in Jerusalem, June 28, 1976. (The Statistician, 1977.).
- Havlicek, L. L. and N. L. Peterson, Robustness of the t-Test: A guide for researchers on effect of violations of assumptions, *Psychological Report*, 1974 (34) 1095-1114.

- Hindley, C. B. and C. F. Owen, The extent 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 betrekkelijkheid van sociaal-wetenschappelijke uitspraken, *Nederlands tijdschrift voor de Psychology*, 1975 (30,7) 573-600.
- Hogben, L., *Statistical theory. The relationship of probability, credibility and error*. New York, 1957.
- House, E. R. et al., No Simple Answer: Critique of the Follow Through Evaluation, *Harvard Educational Review*, 1978 (48,2) 128-160.
- Kemphorne, O., Logical, Epistemological and statistical Aspects of Nature-Nurture Data Interpretation, *Biometrics*, 1978 (34,1) 1-23.
- Kenny, D. A., A Quasi-Experimental Approach to Assessing Treatment Effects in the Nonequivalent Control Group Design, *Psychological Bulletin*, 1975 (82) 345-362.
- Kohnstamm, D., *Had de proefkreche effect? Verslag van een wetenschappelijk onderzoek*. Nijmegen 1976.
- Kuypers, K., De relatie tussen kennen en maken als wetenschapstheoretisch beginsel, *Algemeen Nederlands tijdschrift voor Wijsbegeerte en Psychologie*, 1968 (60) 179-199.
- Leinfeller, W. and E. Köhler, *Developments in the methodology of social science*. Dordrecht, 1974.
- Lepenes, W. und H. Nolte, *Kritik der Anthropologie*. München, 1971.
- 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.
- Linschoten, J., *Idolen van de Psycholoog*. Utrecht, 1964.
- Lord, F. M., A paradox in the interpretation of group comparisons, *Psychological Bulletin*, 1967 (68,5) 304-305.
- Lord, F. M., Statistical adjustments when comparing preexisting groups, *Psychological Bulletin*, 1969 (72,5) 336-337.
- Morrison, D. E. and R. E. Henkel (Eds.), *The significance test controversy*. London, 1970.
- Nauta, L. W. (Red.), *Het neopositivisme in de sociale wetenschappen*. Amsterdam, 1973.
- Nauta, L. W., Context of discovery en context of justification. Notities in de marge van een onderscheid, *Kennis en Methode*. 1977 (1,1) 5-20.
- Neale, J. M. and R. M. Liebert, *Science and behavior. An introduction to Methods of research*. Englewood Cliffs, 1973.
- Neyman, J., Frequentist Probability and Frequentist Statistics, *Synthese* 1977 (36,1) 97-131.
- 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.
- Poynor, H., Selecting units of analysis, in: G. D. Borich (Ed.), *Evaluating educational programs and products*. Englewood Cliffs, 1974, 399-414.
- Rosenthal, R. and R. L. Rosnow (Eds.), *Artifact in behavioral research*. New York, 1969.
- Scriven, M., Die Methodologie der Evaluation, in: Chr. Wulf, (Hrsg.), *Evaluation*. München, 1972.
- Smith, M. S. and J. S. Bissell, Report Analysis: The impact of Head Start, *Harvard Educational Review*, 1970 (49,1) 51-104.
- Solomon, R. L., An extension of controlgroup design, *Psychological Bulletin*, 1949 (46) 137-149.
- Stokking, K. M., *Toetsend Onderzoek* (Dissertatie). Groningen, 1979.
- Stokking, K. M., *Evaluatie-onderzoek inservicetraining GEON-project*. Paper voor de onderwijsresearchdagen 1980.
- Strasser, S., Het probleem van een praktische wetenschap en de opvoedkunde, *Pedagogische Studiën*, 1971 (48) 499-512.
- Stufflebeam, D. L., Evaluation als Entscheidungshilfe, in: Chr. Wulf, (Hrsg.), *Evaluation*, München, 1972.
- Stufflebeam, D. L. et al., *Educational evaluation and decision making*. Itasca, 1972.
- Suchmann, E. A., *Evaluative Research*. New York, 1967.
- Thorndike, R. L., Regression fallacies in the matched groups experiments, *Psychometrika*, 1942 (7,2) 85-102.
- Verberk, A. J. A., *Variante-analyse in de gedragswetenschappen, in het perspectief van andere multivariate analysetechnieken*. Groningen, 1970.
- Zonneveld, L. (Red.), *De school in het kapitalisme*. Amsterdam, 1974.
- GEON: Nota Evaluatieaanpak (december 1975).
Notitie voor de SVO omtrent de afronding van de evaluatie (nov. 1978).
Paper voor de onderwijsresearchdagen (april 1979).
Nadere Kwantitatieve Analyse, I, II, III, IV, V, VI (maart 1977, november 1977, juli 1978, februari 1979, september 1979, november 1979).

Curriculum vitae

Stokking, K. M. (1948) studeerde na het eindexamen HBS-b en na de onderwijzersopleiding pedagogiek in Utrecht, alwaar hij in 1975 het doctoraalexamen aflegde. Van 1974 af evaluatie-medewerker GEON-project; promoveerde op een proefschrift over Toetsend Onderzoek; werkt momenteel aan de afronding van de evaluatie van het GEON-project, welke hij tevens coördineert.

Adres: Hernhutterslaan 20, 3402 GZ IJsselstein