Een internationaal vergelijkend onderzoek over wiskundige studieprestaties

H. FREUDENTHAL
Instituut Ontwikkeling Wiskunde Onderwijs, Utrecht

Inleiding

In de eerste helft van 1964 vond een internationaal vergelijkend onderzoek door middel van studietoetsen naar de wiskundige prestaties van schoolkinderen (13-jarigen en eindexamen-candidaten) in 12 landen plaats. Dit onderzoek was georganiseerd door de *Council of the International Project for the Evaluation of Educational Achievement* (IEA)*. De resultaten ervan zijn door T. Husén en zes medewerkers uitgegeven**. Over het Nederlandse deel bestaat een aparte publicatie***.

In kringen van wiskundigen en didactici der wiskunde is aan dat onderzoek geen of nauwelijks aandacht geschonken; het werd in de vakpers niet genoemd en in vakonderzoek niet of nauwelijks geciteerd; naar ik meen, ten onrechte, en daarom besteed ik er thans aandacht aan. De taak die ik me gesteld heb, wordt door de grote helderheid van het rapport in 't algemeen vergemakkelijkt; vaagheden, die er ook in voorkomen, zij het dan als *understatements* of in de vorm van parafrases, vergemakkelijken het werk evenzeer, omdat zij ongewild naar onvolmaaktheden verwijzen.

1. *De leiding van het projekt en de medewerkers*

De voorbereidingen voor het project startten in 1961. De algemene leiding en organisatie was – internationaal en nationaal – in handen van algemeen onderwijspsychologen; naar ik meen was er geen wiskundige of wiskundig didacticus bij betrokken.

* Later opgevolgd door de *International Association for the Evaluation of Educational Achievement* – ook IEA genaamd.

** International Study of Achievement in Mathematics – A comparison of twelve countries, I, II. Stockholm, Almqvist & Wiksell; New York, John Wiley. 1967.

*** S. Wiegersma, M. Groen, Resultaten van wiskundeonderwijs, Groningen 1968.

Voor secundaire taken werd er wel een beroep gedaan op wiskunde-onderwijskundigen. Er is sprake van een groep van experts in mathematisch onderwijs I.p. 40), die op een andere plaats als de *Working Committee of mathematicians and measurement specialists* (I. blz. 92) wordt aangeduid; maar slechts één ervan staat internationaal als deskundige voor wiskundeonderwijs bekend, geen van de andere namen komt – om een objectief criterium te gebruiken – voor op deelnemerslijsten van internationale congressen en symposia over wiskundeonderwijs. Op dezelfde plaats zijn twee panels voor de voorbereiding van tests met de namen van de leden genoemd, weer met als enige expert op internationale schaal voor wiskundeonderwijs de reeds eerder genoemde. Er waren verder twee niet-wiskundige 'test editors'; hun werk werd gereviseerd door twee 'mathematics educators' waarvan er één als zodanig bekend staat. Aan de internationale voorbereiding van de tests was voorafgegaan het verzamelen van nationale rapporten omtrent inhoud en doelstellingen van het wiskundeonderwijs in de betrokken landen – (I, blz. 92) 'requesting that a group of mathematicians and mathematics educators in that country formulate an authoritative statement of the objectives of mathematics instruction in secondary education in that country . . . a committee of mathematicians was set up in each country to formulate a national statement' waarbij Nederland met name genoemd wordt. Maar in Nederland werd deze taak in feite door het 'Nederlands Instituut voor Preventieve Geneeskunde' vervuld. Of er in andere landen wiskundigen of vakonderwijskundigen aan te pas zijn gekomen, kan ik in details niet vaststellen; onder de namen van de nationale correspondenten is er geen, die aan wiskunde of wiskundeonderwijs doet denken. Toch schijnt er in Engeland wel zo iemand te zijn geweest, althans iemand die van het Engelse wiskundeonderwijs op de hoogte was. Na de *try-out*, toen er alleen nog technische critiek op het voorgestelde toetsinstrument had mogen worden geoefend, kwam er van Engelse kant een fundamentele critiek op de inhoud van het instrument dat te

weinig aan het Engelse programma zou beantwoorden (I, blz. 98). De Engelse critici schijnen hun zin te hebben gekregen; hoe de andere landen hierop hebben gereageerd, wordt er niet bij vermeld.

Het nagenoeg uitsluiten van deskundigheid op het gebied van wiskunde en wiskunde-onderwijs doet te merkwaardiger aan, daar de auteurs met instemming stukken uit een resolutie van een internationaal colloquium over wiskunde-onderwijs (Budapest 1962) weergeven, onder meer:

> Om het peil van het wiskunde-onderwijs te verhogen is nauwe samenwerking geboden tussen wiskundigen, wiskunde-leraren, onderwijskundigen en psychologen met kennis van de moderne wiskundige begrippen . . .

De auteurs laten deze zinsnede voorafgaan door de verklaring (I, blz. 79):

> De IEA-studie is een stap en op internationale schaal wellicht de eerste naar de vervulling van de aanbeveling van het colloquium van Budapest . . .

Met 'een stap' en 'een eerste stap' is klaarblijkelijk bedoeld dat men het essentieel betrekken van wiskundigen en wiskunde-onderwijskundigen bij onderzoek omtrent wiskunde-onderwijs maar liever tot een volgende keer heeft uitgesteld.

2a. *De leerlingenpopulaties*

De populaties die men met het onderzoek op het oog had, waren

1a. De dertienjarigen,
1b. De leerlingen van het leerjaar waarin de meerderheid der dertienjarigen zich bevindt,
3a.* De leerlingen van het einde van het secundaire onderwijs – mathematische richting.
3b.* Idem – niet-mathematische richting.

De groepen 3 bevatten dus, wat Nederland betreft, leerlingen van het 11e schooljaar, terwijl dit bijvoorbeeld in de Bundesrepublik Deutschland leerlingen van het 13e schooljaar waren.

Met de populatie 1b schijnt iets mis te zijn geweest (zie ons nummer 9).

De toetsinstrumenten bestonden uit 58–70 items, die in 3–4 zittingen van telkens een uur moesten worden afgewerkt.

Het onderzoek werd aangevuld met vragenlijsten voor leraren en leerlingen, de onderwijsorganisatie, sociale factoren en attitudes betreffend. Op details hieromtrent komen we straks terug.

* Let wel: 3a betekent (ongeveer) onze B-leerlingen, 3b onze A-leerlingen.

3. *Scores*

Van de resultaten trekken allereerst aandacht de scores van de onderscheiden landen. Veel verrassingen vallen daar niet te beleven. Dat Israel en Japan aan de spits staan, is en was ook toen al geen verrassing, dat Australië, de Verenigde Staten en Zweden onderaan staan, evenmin; verklaringen hiervoor liggen voor de hand, maar ik zal ze niet geven, want ik ben er allesbehalve zeker van of ze even juist zijn als ze algemeen geaccepteerd zijn. Merkwaardigerwijs heeft nooit een onderzoek hiernaar plaats gehad. Het zou in elk geval veel dieper moeten graven dan statistisch mogelijk is.

Het enige werkelijk verrassende in de lijst van de scores is de plaats die België inneemt. In de populatie 1a direct na Japan, in 1b na Israel en Japan, maar nauwelijks significant hun mindere, in 3a achter Israel en Engeland en boven Japan, en alleen in 3b op een vijfde plaats, na Nederland, dat het trouwens in 't algemeen er ook niet slecht heeft afgebracht.

Zonder de Belgen een kwaad hart toe te dragen, mag men wel naar een verklaring zoeken. Daarstraks verzweeg ik de naam van de enige wiskundige van reputatie die – te midden van onderwijskundigen – het toetsinstrument heeft mogen bespelen. Het was mijn grote Belgische vriend Willy Servais – geen eminentere had men in wiskunde en wiskunde-onderwijs voor deze taak kunnen kiezen. Het kon niet anders of hij moest – als verreweg meest deskundige – op de toetsencollectie dezelfde stempel drukken als op het Belgische wiskunde-onderwijs en hiermee is – dacht ik – de Belgische positie aan de spits verklaard. Trouwens zal ook niemand die de incompatibiliteit van Servais' stijl en de Engelse kent, zich nog over de hevige reactie verbazen, die er van de overkant van 't kanaal op de oorspronkelijke toetscollectie kwam.

4. *Internationale samenwerking t.a.v. wiskunde-onderwijs*

In geen vak zijn de internationale onderwijscontacten van ouds zo frekwent en nauw geweest als in de wiskunde, geen vak heeft zoveel internationaal vergelijkend onderzoek van invloedrijk karakter gekend als de wiskunde. De Internationale Commissie voor Wiskunde Onderwijs (in 't Duits IMUK, in 't Engels

ICMI, in 't Frans CIEM) werd al voor de eerste wereldoorlog opgericht en is dus ouder dan de Internationale Mathematische Unie, waarvan zij thans een Commissie is. De oudste zuster van de IMUK onder de onderwijscommissies bij de in de ICSU verenigde wetenschappelijke Unies is vijftig jaar jonger dan de IMUK zelf. Internationale congressen, conferenties, tijdschriften zoals voor wiskunde-onderwijs kent men voor geen ander vak. Nationale en internationale rapporten hebben over de grenzen heen kennis omtrent het wiskunde-onderwijs verschaft en er is een vrij grote kring van deskundigen met inzicht in het wiskunde-onderwijs op wereldwijde schaal. Hoe komt het dat de leiders van het onderhavige projekt geen beroep deden op deze ervaring en deskundigheid?

Het is een beetje als met de twee koningskinderen, die niet bij malkanderen konden komen. De brede rivier is er een van mentaliteitsverschil. Gemeenlijk stelt men zich een wiskundige voor als iemand die alles gaat kwantificeren, maar het schijnt me een juister beeld de wiskundige te zien als iemand die de grenzen van het kwantificeren – ook binnen de wiskunde – beter onderscheidt dan anderen. Het kan nauwelijks anders of een wiskundige zal een uitsluitend kwantificerend onderzoek als het onderhavige met gepaste scepsis bejegenen wat zijn *nuttige waarde* betreft. Kent hij de internationale verhoudingen, dan zal hij ook de *mogelijkheid* ervan in twijfel trekken. Het is een manier, zich als wiskundige impopulair te maken: anderen van hun blind geloof in de wiskunde trachten te beroven.

5. *Doelstellingen*

Wat een wiskundige bij dit onderzoek fundamenteel interesseert, is de vraag wat onderwijspsychologen bewogen heeft, zulk een onderzoek te entameren, wat ze ervan verwachtten en wat ze ermee dachten te bereiken. Wat kunnen de doelstellingen van de organisatoren zijn geweest?

De doelstellingen zijn goed geformuleerd (I, blz. 30–33), hoewel op een uiterst informele wijze, die men van onderwijspsychologen niet zou verwachten. Ik heb de indruk dat het a posteriori formuleringen zijn, waartegen ik trouwens geen bezwaar heb. Ik vertaal iets uit de uiteenzetting van de doelstellingen (I, blz. 30–31):

> . . . het algemeen doel was, met behulp van psychometrische technieken de uitkomsten in verschillende onderwijs*systemen* te vergelijken. Het feit dat deze vergelijkingen tussen de naties worden getrokken, mag hierbij niet als aanwijzing voor primaire belangstelling in bijv. nationale gemiddelden en dispersies in schoolprestaties op de gegeven leeftijd of schoolniveau gelden.
>
> . . . het hoofddoel is onderzoek naar de 'uitkomsten' van verschillende schoolsystemen door zoveel mogelijk relevante input-variabelen (voorzover bepaalbaar) te relateren tot de output, zoals bepaald met internationale toetsinstrumenten . . .
>
> Uit het volgende zal blijken, naar we hopen, hoe weinig het betekent alleen maar nationale niveaus van studieprestaties op verschillende leeftijden te vergelijken. Men mag verwachten, dat de verschillen tussen en binnen de landen niet alleen met betrekking tot de uitkomsten van het onderwijs (cognitief en niet-cognitief) variëren, maar ook met betrekking tot alle 'input' variabelen, zoals het technologische niveau en de urbanisatie van landen, de sociale achtergrond van de kinderen, de opleiding van hun ouders, de opleiding van hun leraren, het geld dat aan onderwijs besteed wordt, de jaren wiskunde-onderwijs, het aantal uren per week. Van speciaal belang is de structuur van het schoolsysteem, die bepaalt hoeveel kinderen hoe hoog de opleidingsladder mogen beklimmen . . .

Het is een aanvaardbaar idee. Om onderwijsvarianten te vergelijken, kan men *proeven* opzetten, maar men kan ook van de *gegeven* variabiliteiten gebruik maken. Een bezwaar van deze methode is dat men het beetje beheersing van de relevante factoren, dat de experimentele opzet toestaat, ook nog mist. Een voordeel is het gemakkelijker kunnen beschikken over grote populaties van leerlingen, hetgeen de statistische betrouwbaarheid van het onderzoek zou kunnen verhogen. Een groter aantal parameters komt hiermee in het bereik van een verantwoorde behandeling, laten we zeggen, zoveel parameters als de computer maar kan verwerken. Hier loert nu een gevaar, de verleiding, de parameters maar op goed geluk te kiezen. Als het er maar voldoende veel zijn, zullen enkele wel relevant blijken – zou een verleidelijke redenering kunnen zijn, die men zich in 't verleden, met minder geld en computercapaciteit, niet zou hebben gepermitteerd. Ik acht deze hoop geheel illusoir. Bij een onderzoek dat per definitie veel toevalsfactoren behelst, ook nog de opzet door het toeval te laten beheersen, is een aanlokkelijke, maar averechtse procedure.

6. *Het curriculum als parameter*

Aan een omvangrijk statistisch onderzoek als het

onderhavige dient in elk geval een degelijk kwalitatief onderzoek vooraf te gaan. In het verslag ontbreekt de neerslag van zulk onderzoek. Het geregeld als toevallige voorbeelden oproepen van te bestuderen parameters, zoals ook in het geciteerde stuk, is een veeg teken. Maar wat vooral in het geciteerde stuk bij het opnoemen van input-variabelen opvalt, is het ontbreken van één variabele, die men – geheel onbevooroordeeld – als de belangrijkste zou beschouwen: het genoten onderwijs, in 't bizonder het curriculum. Het zou op de geciteerde plaats een toevallige omissie kunnen zijn, maar hetzelfde verschijnsel doet zich ook in 't vervolg telkens weer voor. Telkens weer ligt de nadruk op wat ik zou willen noemen, bureaucratische parameters – ik bedoel dit 'bureaucratisch' niet als scheldwoord, maar denk aan parameters die zich door bureaucratische maatregelen laten beïnvloeden. Ik twijfel niet of schoolgrootte, klassengrootte, salarissen, toelatingseisen, enz. uiterst belangrijke variabelen zijn en ik ontzeg de organisatoren niet de bevoegdheid, die gerelateerdheid van de onderwijsuitkomst met deze variabelen te bestuderen. Maar indien mocht blijken dat verwaarloosde input-variabelen de gemeten verschillen in de output geheel of nagenoeg geheel voor hun rekening opeisen, dan zijn alle verklaringen van verschillen tussen de onderwijssystemen door middel van de wel aanvaarde variabelen van geen reële betekenis.

Onbevooroordeeld zou men menen dat – nationaal èn internationaal – het curriculum de input-variabele met de grootste invloed op de variabiliteit van de output is. Onder de oppervlakte van het rapport neemt men een controverse waar tussen hen die de tests moesten voorbereiden en die de verschillen van curriculum breed uitmeten, en de organisatoren van het project, die de verschillen wel erkennen, maar trachten te bagatelliseren en blijk geven van hun overtuiging dat de curriculum-input constant was, althans als een constante kon worden behandeld. De niet te loochenen – en ook niet geloochende – curriculum-verschillen moesten en konden door het toetsinstrument worden opgevangen – dit moet op de achtergrond van het niet erkennen van het curriculum als een variabele hebben gespeeld. Het geloof in toetsen kan bergen verzetten. Ik citeer weer (I, blz. 65):

> Er is één gebied, waar sinds 1920 een hoge mate van finesse (sophistication) is bereikt: studietoetsen. Men kan tegenwoordig de mate en aard van begrip van de leerling voor leerstof met een te voren ongekende subtiliteit meten. Moderne objectieve studietoetsen stellen, als ze behoorlijk ontwikkeld en geïnterpreteerd worden, een van de machtigste werktuigen voor, die er voor onderzoek van onderwijs zijn. Door hun gebruik zijn ontdekkingen gedaan die ver uit gaan boven wat gezond verstand kan bereiken.

Daarbij vergeleken zijn de parameters, die we gaarne zouden bestuderen ruw – gaan de auteurs door. Herhaaldelijk* wordt beklemtoond, dat dankzij het hoge peil van het toetsontwikkelen het vervaardigen van een internationaal toetsinstrument voor het wiskunde-onderwijs maar kinderspul is vergeleken bij het epineuze werk van het definiëren van de input-variabelen, waarbij dan weer bureaucratische parameters zijn bedoeld. Hier gaan de auteurs door:

> Andere hoofdstukken van dit rapport beschrijven op welke gronden de leerlingen der participerende landen vergelijkbaar genoemd kunnen worden, hoe de toetsen, die ze gemaakt hebben, vergelijkbaar zijn en bovendien (valide) en betrouwbaar (reliable), hoe de daten omtrent attitudes tegenover de wiskunde, de school en tegenover hun eigen toekomst vergelijkbaar zijn en hoe de onderwijspolitiek (der diverse landen) vergeleken kan worden.

De vergelijkbaarheid van de curricula wordt niet eens genoemd; van de meeste andere punten komt de vergelijkbaarheid trouwens ook nimmer aan de orde. De zogenaamde betrouwbaarheid – een vergoelijkende functie van de toets-steekproeffout – loopt van 0.958 tot 0.732 (I. blz. 107), hetgeen, tussen twee haakjes, respectievelijk beantwoordt aan steekproeffouten van liefst 20% tot 50%. De validiteit van het toetsinstrument is vastgesteld (I, blz. 108) door correlatie met Britse examenresultaten en dit bij een instrument dat er uitdrukkelijk niet zou zijn om gemiddelden en standaarddeviaties van een studietoets voor één land of afzonderlijke landen veilig te stellen, maar om de invloed van variabelen op de studieprestaties op internationale schaal te onderzoeken. Hoe was zo'n blunder mogelijk? Wel, de auteurs van Hoofdstuk 4 (Bloom & Foshay) waaruit ik blz. 95 citeerde, waren, zoals meermalen blijkt, van de details van het onderzoek zeer onvolmaakt op de hoogte, in 't bijzonder niet van de toetsconstructie die in Hoofdstuk 5 (door Thorndike) wordt beschreven. In Hoofdstuk 5 heeft men zich – weer bij wijze van understatement – voorzichtiger uitgedrukt t.a.v. de validiteit. Dat bij de gegeven omstandigheden ook de 'betrouwbaarheid' geen

* bijv. II, blz. 287.

aangewezen maatstaf is, werd echter niet opgemerkt.

Het probleem van de incompatibiliteit der curricula wordt herhaaldelijk aangeraakt en breed uitgemeten (in 't bijzonder I, blz. 83–85), waar de testvoorbereiders aan het woord zijn, maar altijd om meteen gebagatelliseerd te worden door degenen die het voor het zeggen hebben.

Uiteindelijk bleek de constructie van een toetsinstrument dat door verschillende landen gebruikt kon worden en toch een hoge curriculum-validiteit voor alle landen en programma's had, onmogelijk. De lezer van dit rapport moge begrijpen, dat bij een internationale pioniersstudie van dit soort de beslissing om de toetsen niet aan de afzonderlijke programma's aan te passen, de enige mogelijke en de meest geschikte was voor een experimentele bepaling van de diverse prestaties.

Wel, iedereen die het wiskunde-onderwijs op internationale schaal kende, had dit van meet af aan kunnen vertellen. Ik beweer niet, dat dit meteen tot verwerping van het idee zelf had moeten leiden. Met een tijdige kijk op de moeilijkheden had men veeleer de opzet van het toetsinstrument a priori zo kunnen bepalen, dat men a posteriori, waar nodig, de factor van diversiteit van de curricula had kunnen elimineren of ervoor corrigeren om aan de variabiliteit van de andere variabelen significantie te verlenen. Dit is niet geschied. Ik veronderstel, dat men zich pas zeer laat heeft gerealiseerd, dat bij de gekozen opzet het curriculum de beslissende en uiteindelijk enige significantie opleverende variabele was. Van dit inzicht is er in deel II de neerslag te vinden onder een hoogst vreemde titel, waarop ik nog terugkom.

7. *De invloed van het curriculum*

Vrijwel in 't begin van mijn analyse van dit onderzoek en eer ik aan de net genoemde plaats in deel II toe was, heb ik mezelf de taak gesteld, de invloed van de variabele 'curriculum' op de scores te schatten. Ik heb hierbij de volgende ruwe procedure gekozen. Ik heb de toetsen, item voor item, bekeken en nagegaan hoeveel tot het Nederlandse curriculum tot en met het 1e jaar voortgezet onderwijs, respectievelijk de eindexamenklas behoren; met het verschil van de schooltypen heb ik hierbij enigszins rekening gehouden. Het aantal toetsen, die tot het curriculum behoorden, was volgens mijn telling

voor 1b: 43 van de 70, dus 60%,
voor 3a: 39 van de 69, dus 57%,
voor 3b: 42 van de 58, dus 70%.

Veronderstellende dat deze getallen niet te ver zullen afwijken van de *gemiddelden* van 'bekende leerstof' voor de deelnemende landen, heb ik hierop de standaarddeviaties berekend en in de interlandelijke standaarddeviatie σ op de landelijke scoregemiddelden uitgedrukt:

voor 1b: $4{,}1 = 0{,}8\ \sigma$
voor 3a: $4{,}1 = \sigma$ resp. $1{,}5\ \sigma$
voor 3b: $3{,}4 = 0{,}6\ \sigma$ resp. 2σ;

hierbij hebben de gegevens achter 'resp.' betrekking op de landengroep na uitsluiting van de Verenigde-Staten voor 3a en van de Verenigde Staten en Zweden voor 3b. Houdt men met een moeilijkheidsgraad van 50% voor de onbekende leerstof rekening, dan blijkt nog steeds dat een substantieel deel van de interlandelijke variabiliteit (zo niet de gehele) door het curriculum verklaard wordt. Het lijkt derhalve uiterst dubieus of men uit de interlandelijke variabiliteit conclusies kan trekken omtrent de invloed van andere input-variabelen dan het curriculum op de output. Voor *dit* doel is, voorzover ik zie, de validiteit van het toetsinstrument nihil.

Ik heb dit zo uitvoerig uiteengezet, om aan te tonen dat er wel middelen zijn om met de invloed van het curriculum rekening te houden. Natuurlijk moet dit niet met zulke grove methoden geschieden. Bij een meer doelbewuste opzet, gepaard gaande met een experimenteel vooronderzoek, was er vermoedelijk wel significantie uit te halen geweest. De geciteerde zinsnede waarbij een beroep op de lezer wordt gedaan om de feitelijke opzet maar te accepteren, kan moeilijk anders worden geïnterpreteerd dan als een excuus achteraf.

8. *Gelegenheid om te leren*

Ik zei al dat de auteurs in deel II toch proberen, numeriek de invloed van het curriculum te evalueren. Dit geschiedt te midden van het onderzoek van allerlei variabelen, zoals interesse, attitudes, opleiding van de leraren, aantal lesuren, enz. De variabele die ik bedoel, heet *gelegenheid om te leren* (opportunity to learn). Vraag 8 van de lijst voor leraren luidde – ik laat een mij onbegrijpelijk deel ervan weg:

Onderzoek bij elke toetsvraag en geef aan of volgens u

A. alle of de meeste van uw leerlingen (ten minste 75%) in de gelegenheid waren, om dit type opgave te leren.
B. sommige (25% tot 75%) van deze groep leer-

lingen in de gelegenheid waren, om dit type opgave te leren.

C. weinig of geen (minder dan 25%) van deze groep leerlingen in de gelegenheid waren, om dit type opgave te leren.

De numerieke gegevens die hieruit voor de diverse landen zijn afgeleid, lijken liefst nog ongunstiger voor de validiteit van het totale onderzoek dan mijn schattingen. Nederland en België ontbreken, helaas, in de lijst. (Een interlandelijk regressie-onderzoek met de 'gelegenheid om te leren' als factor ontbreekt jammer genoeg; het zou echt de moeite waard geweest zijn.)

Nu is de variabele 'gelegenheid om te leren' geenszins identiek met de door mij bedoelde variabele 'curriculum'; kenmerkend is dat ze in de catalogus (ouders, leraren, school, leerling) onder de lerarenvariabelen wordt gerangschikt, omdat het hier volgens de auteur om een subjectieve opvatting (rating) van de leraar zou gaan.

Waarom heeft men nu deze vage variabele in het oog gevat in plaats van een keihard – harder dan de meeste andere – te definiëren variabele 'curriculum'? Als ik een vermoeden mag uiten, zou ik het als een compromis willen interpreteren tussen hen die meenden dat het toetsinstrument de stoot moest kunnen opvangen, en hen, die de bezwaren van een uniform toetsinstrument beter hadden onderkend. De eerste groep kreeg de overhand want het curriculum als variabele te erkennen zou in strijd met de filosofie van het totale onderzoek zijn geweest, en had men het curriculum als variabele erkend, dan was men er wellicht ook niet aan ontkomen, de invloed van deze variabele interlandelijk te evalueren. Aan de andere kant was een critiek op het feit, dat men met het curriculum als variabele totaal geen rekening had gehouden, gegrond genoeg om er zich niet lichtvaardig aan bloot te stellen. Men heeft dus voor een compromis gekozen: een variabele, die aan 'curriculum' doet denken, maar die van meet af aan zo vaag gedefinieerd is, dat ze gemakkelijk kan worden gebagatelliseerd.

Toch vind ik, wat hier is geschied, geen bagatelle. Vooral de naam, die men aan de variabele heeft gegeven, stoort me intens, en ik zal zeker niet de enige zijn die het zo voelt. Krijgen de leerlingen werkelijk in de diverse landen zo uiteenlopende gelegenheden om wiskunde te leren? Neen, natuurlijk niet. Laat het waar zijn dat bijv. Nederlandse leerlingen (in die tijd) in de gelegenheid gesteld werden, om maar 60% of 70% van de in het toetsinstrument vertegenwoordigde stof te leren, dan is het evenzeer waar, dat ze ook nog andere dingen in de wiskunde leerden – ik schat à drie keer zoveel als hetgeen in het toetsinstrument vertegenwoordigd is. Met leerlingen uit andere landen zal het natuurlijk analoog gesteld zijn geweest. Ja, het spijt me – suggereert de variabele 'gelegenheid om te leren' – dit telt niet mee. Waarom niet? Is het waardeloos? Ten dele zeker, maar voor een niet onbelangrijk deel (bijv. de bij ons traditionele nadruk op ongelijkheden en eliminatie) is het wel degelijk van waarde. Maar dit doet er eigenlijk niet toe. Wie bepaalt dat het toetsinstrument samenvalt met de 100% gelegenheid te leren? Commissies waarvan één wiskundige van naam lid is? Een ondeskundige groep van testbewerkers? Of het land, dat het luidste protesteert?

Natuurlijk was de naam van deze variabele niet zo bedoeld, maar hij suggereert het wel degelijk. En wie garandeert dat iedereen die zich op dit onderzoek beroept – gelukkig zijn die er tot nu toe niet – precies naleest hoe die variabele bedoeld was; zo nauwkeurig plegen toch zelfs de auteurs van een verzamelwerk wederzijds de bijdragen niet te lezen.*

8. *Andere variabelen*

Laat ik nu nog de overige onderzochte variabelen in ogenschouw nemen. Ik stap echter heen over variabelen die alleen maar tot vragen naar de bekende weg aanleiding geven. We weten langzamerhand wel, dat sociale positie en opleiding van de ouders van invloed zijn op de schoolprestaties van de kinderen, dat jongens beter leren dan meisjes, dat belangstelling voor een vak het leren bevordert – ga zo maar door – en ik voel geen behoefte, dit in de tweede decimaal gepreciseerd te zien. Ik wil veeleer van een aantal gekozen variabelen nagaan of ze enige nuttige informatie zouden kunnen behelzen.

Allereerst: de grootte van de school blijkt een positieve factor. Ik zou zeggen – uit het Nederlandse perspectief – dat haal je de koekoek: De grote scholen (in 1965), dat waren de Lycea en HBSen, de kleine, dat waren Mulo's en LBO-scholen, en de betere leerlingen zitten uiteraard in de eerste groep.

De grootte van de klas blijkt een negatieve factor waar men gevorderde leerlingen in kleine klassen verenigt, en een positieve waar men juist achtergebleven leerlingen dit intensere onderwijs verstrekt.

* In een dithyrambisch artikel in School Review (May 1974) beweert B. S. Bloom, een der medewerkers, dat *opportunity to learn* de feitelijk onderwezen fractie van het *officiële* leerplan (niet van de IEA-tests) was; vrijwel alles in dit artikel berust op vage herinneringen i.p.v. citaten.

Niet verbazingwekkend.

De duur van de leraarsopleiding is een positieve factor en 'in-service-training' is een negatieve factor, maar dit heeft uiteraard noch met de *duur* van de opleiding noch met het *feit* van in-service-training te maken.

De financiële uitgaven per leerling zijn (bij 13-jarigen) een positieve factor, vooral in Nederland. Dit haal je weer de koekoek. Leraren VHMO worden belangrijk beter betaald dan leraren (en onderwijzers) in Mulo en LBO.

Wel, dit soort statistiek – ik zou er nog een poos mee kunnen doorgaan – doet denken aan de befaamde redeneringen uit wijlen Chr. Rümke's Vehicologie: Treinen lopen op de mankracht van de passagiers, want als men de treinen indeelt naar 'rijdend en stilstaand' en naar 'bezet en leeg', blijkt dat 'rijdend' overwegend gepaard gaat met 'bezet' en 'stilstaand' met 'leeg'.

Het is duidelijk dat de *aard* van de school hier, een veel belangrijker rol speelt dan die variabelen waarop men zijn oog heeft laten vallen. Maar wat de *aard* van de school aangaat heeft men alleen het verschil van *comprehensive* en niet-*comprehensive* bekeken en er veel aandacht aan besteed, hoewel het in 1965 *internationaal* nog van geen betekenis was.

Wat de opleiding van de leraren betreft, heeft men ook pardoes een verkeerde variabele gekozen, en dientengevolge zijn de tabellen hieromtrent (I, blz. 266–269) een samenstel van cijfers, waaraan geen moeite is besteed ze internationaal vergelijkbaar te maken – en toch betrekt men zoiets in correlatie- en regressieberekeningen (II, blz. 180–181, 270–271). In plaats van duur van de opleiding (hoe bepaal je zo iets in het Nederlandse actensysteem en analoge buitenlandse systemen?) zou de enige juiste variabele – en een keiharde – zijn geweest: het wiskundig niveau van de genoten leraarsopleiding. In sommige landen, waar men met het wiskunde-onderwijs minder tevreden is, wordt in wiskundige kringen het ontoereikende wiskundige niveau van de leraren namelijk als essentiële factor in de slechte resultaten aangewezen. Het zou zelfs niet te gek zijn geweest te vragen of de *leraar* de *leerlingen*-toetsen kon maken; in sommige landen werd namelijk beweerd dat vele leraren, die wiskunde moesten onderwijzen, zelf niet meer wiskunde dan tot plm. het 10e schooljaar gehad hebben.

Een tweede vraag, minder keihard, die men had kunnen stellen, zou er een geweest zijn naar het aandeel algemene en vak-onderwijskunde in de opleiding van de leraar (om de invloed ervan op de studieresultaten van de leerlingen vast te stellen). In sommige landen, waar dit aandeel groot is, werd namelijk door wiskundigen – terecht of ten onrechte – beweerd, dat de invloed ervan op leerlingenprestaties negatief is.

Een redelijke vraag, het onderwijs betreffende, zou geweest zijn die naar de invloed van psychometrische technieken zoals objectieve studie-toetsen; in de toepassing hiervan bestaan immers grote verschillen van land tot land. Men heeft echter liever gegevens omtrent het karakter van het door de leerlingen ontvangen onderwijs door een lijst van vragen trachten te achterhalen, waarop de leerling zijn leraar moest beoordelen – op zich zelf een gezond idee als de vragen gezond gekozen zijn, maar in feite is het het type van vragen, waarop de leerling het hem gunstig dunkende antwoord geeft als de leraar hem aanstaat, en het ongunstige als hij de leraar niet kan uitstaan. Ook als dit niet het geval is, loopt men de kans, dat bijv. een op begrip gestelde leerling de neiging heeft te veel *rekenwerk* te constateren in het onderwijs dat hij ontvangt terwijl de zwakkere leerling allicht vindt, dat zijn leraar te veel *begrip* vraagt. Men krijgt dan een negatieve correlatie tussen begripsmatig onderwijs en studieprestatie.

De vragenlijsten voor leerlingen, die op de attitudes van leraren en leerlingen doelen, bevatten nauwelijks items, die voldoende specifiek voor wiskunde zijn. Er worden zeer algemene vragen gesteld, zoals iedereen die tegenwoordig – volgens sjablones – kan verzinnen. Men was blijkbaar niet toe aan het idee, attitudes op te sporen door middel van scherpe vragen omtrent concrete mathematische problemen – iets waarvoor uiteraard deskundige medewerking vereist was geweest.

Het is merkwaardig, dat aan een zeer invloedrijke, een schoolbestel karakteriserende variabele – trouwens een bureaucratische variabele bij uitstek – geheel geen aandacht is geschonken: aan het examen (eindexamen of toelatingsexamen tot de universiteit). Deze variabele verschilt nogal van karakter in de diverse landen, maar dit karakter laat zich dan ook keihard formuleren. Het ligt voor de hand te vermoeden dat de internationale verschillen in prestaties van de populaties 3a en 3b in hoge mate door het examen en de eisen daarvan verklaard worden. Hoe kon men zo'n variabele over het hoofd zien!

In 't algemeen toont de keuze van leerlingen- en leraarvariabelen aan, dat geen wiskundige aan de vaststelling ervan te pas is gekomen. Ze lijken ontstaan doordat in een algemeen schema op de aangewezen plaatsen het woord wiskunde is ingevuld. Maar ook in onderwijskundig opzicht laten ze niet meer zien dan een bureaucratische benadering. Men kan zich voorstellen dat over de exacte formulering

van deze of gene item uren is gediscussieerd, maar niets wijst op het aansnijden en doordenken van enige problematiek.

Dit geldt ook voor de schoolvariabelen, echter met het verschil dat de organisatoren hier geobsedeerd waren door de vraag van al-of-niet-selectiviteit van een schoolsysteem. De gebruikelijke Amerikaanse misvattingen op dit gebied zijn door de organisatoren voetstoots aanvaard. Selectiviteit wordt steeds als een al-of-niet, i.p.v. een gamma van mogelijkheden geïnterpreteerd. De auteurs hebben een begrip 'rententivity' geschapen, dat ook weer alleen parafraserend gedefinieerd wordt (II, p. 116), als 'de fractie leerlingen die tot het pre-universitaire stadium schoolgaan'. Dat werd dan geïnterpreteerd als naar een eindexamen toegaan dat recht geeft tot het afleggen van de academische examens – een interpretatie die alleen uit onbekendheid met de internationale situatie verklaard kan worden. Zodoende komen absurde lijsten van 'indices of retentivity' (Table 3.35, II, p. 117) tot stand, waarop het hele vervolg is gebaseerd.

Op de vraag 'betekent meer – slechter?', die positief wordt beantwoord, volgt de vraag 'betekent meer – slechter over de hele lijn', die – zoals te verwachten was – negatief wordt beantwoord. Het probleem van al of niet selectief onderwijs tot 18 jaar wordt wel erg formalistisch behandeld; over de essentialia stapt men gewoon heen. Ik bedoel het volgende:

In de Verenigde Staten behaalde in groep 3b (18-jarigen)
43% der leerlingen een score van $\leqslant 5$ punten uit 58,
67% der leerlingen een score van $\leqslant 10$ punten uit 58,
80% der leerlingen een score van $\leqslant 15$ punten uit 58.

'$\leqslant 5$ punten' betekent vermoedelijk (indien het niet totaal op toeval berust), dat deze leerlingen zich in wiskunde (rekenen) op het peil van ten hoogste 't vierde leerjaar basisschool bevinden – dit zijn er dus 43% van het totaal der 18-jarigen in 3b; 'van 5 tot 10 punten' zal een dragelijke vertrouwdheid met het elementaire rekenen en met ietwat wiskundige routines betekenen – dit zijn er dus 24%: bij in 't geheel 80% blijft het nog bij een vrij primitieve vertrouwdheid met wiskunde. Met de discussie over selectief of niet-selectief onderwijs stapt men over de essentiële kwestie heen, en het is jammer dat zulk een rapport er ook geen bijdrage toe poogt te leveren: wat voor zin heeft het deze 67% of 80% leerlingen met wiskunde te plagen – preciezer met *het soort wiskunde* dat dit toetsinstrument vertegenwoordigt – of over tests te laten broeden die volmaakt Chinees voor ze zijn? Zeer zeker – men kan dezelfde vraag ook bij 25% van de 3a-leerlingen in de Verenigde Staten stellen, bij 35% resp. 67% van de Zweedse 3b-leerlingen en bij 10% van de 3b-leerlingen in sommige andere landen, maar het meest urgent is de vraag wel in landen met niet-selectief onderwijs. Dit zijn ook zowat de enige significante cijfers van dit rapport.

9. *Andere facetten*

Zoals eerder medegedeeld zouden de populaties 1a en 1b als volgt zijn gedefiniëerd:

1a. de 13-jarigen.

1b. de leerlingen van het leerjaar waarin de meerderheid der 13-jarigen zich bevindt.

Men zou verwachten dat 1b gemiddeld ouder en langer op school was dan 1a (bij wijze van schatting een half jaar), maar in elk geval niet jonger en korter op school. Men zou dus ook gemiddeld scores van 1b verwachten, die hoger zijn (in elk geval niet lager) dan die van 1a. Dat gaat ook voor alle landen op, behalve voor Nederland, waar 1b liefst $2\frac{1}{2}$ punten ($= \frac{1}{2}$ van het interlandelijke σ) minder scoort dan 1a. Wat zit hier achter?

Volgens de tabel van de gemiddelde leeftijden van de diverse populaties (I, blz. 271) is de gemiddelde leeftijd van 1b Nederland 13 jaar 1 maand, hetgeen een minimum record betekent; het maximum wordt door 1b Engeland met 14 jaar 4 maanden behaald. 1b Nederland is dus liefst een half jaar jonger i.p.v. ouder dan 1a Nederland. Merkwaardig is ook de hoge standaarddeviatie van 1a, te weten 11,6 m. Hoe kan dit? Aangezien er in de normaal 13-jarige klas ook 14-jarigen zullen zitten (zie de hoge standaarddeviatie), doet het gemiddelde van 13 jaar 1 maand vermoeden, dat in de klasse waarin het merendeel der 13-jarigen zit meer 12- dan 13-jarigen zitten. Onmogelijk is zoiets niet, maar waarschijnlijk lijkt het ook niet. Even vreemd, of nog vreemder is het geval van Engeland met het gemiddelde 14 jaar 4 maanden, dat wijst op een grote meerderheid van 14-jarigen in de klasse waar de meeste 13-jarigen zitten.

Nu is er ook een verantwoording omtrent de samenstelling van populatie 1b, die men niet zo gauw ontdekt, omdat zij onder de averechtse titel 'De plaats van de 13-jarigen in het systeem' is verstopt (I, blz. 234–235) – eerder merkten we al een geval van vreemd goochelen met opschriften op. Volgens deze lijst bestaat de populatie 1b Nederland uit de leerlingen van het 6e leerjaar (basisschool), de populatie 1b Engeland uit die van het 8e leerjaar; de andere landen zitten er meestal tussen, (d.w.z. 7e

leerjaar) maar er is er geen bij behalve Nederland waar 1b het 6e schooljaar beslaat. Met de fraaie formuleringen, die we eerder (I,blz. 65) citeerden

> Andere hoofdstukken van dit rapport beschrijven op welke gronden de leerlingen der participerende landen vergelijkbaar genoemd kunnen worden . . .

was vermoedelijk niet het hoofdstuk bedoeld, waar de samenstelling van populatie 1b vermeld wordt. Ik bereid de lezer echter alvast voor op de onthulling dat de twee nu geciteerde gegevens in strijd met de waarheid zijn, die trouwens in zeker opzicht nog erger is.

„De plaats van de 13-jarigen in het systeem' is een onverteerbare hutspot. Ik ben er niet uitgekomen omdat telkens twee dingen door elkaar lopen. Men kan twee vragen stellen:

Welke fractie leerlingen in het met 1b aangeduide leerjaar zijn 13 jaar oud?

Welke fractie van alle 13-jarigen zitten in het met 1b aangeduide leerjaar?

Dit zijn uiteraard twee a priori (en a posteriori) verschillende fracties. De titel suggereert iets in de richting van de tweede vraag, maar de gegevens slaan veelal op de eerste. Dit is immers ook hetgeen men uit de bij het onderzoek verzamelde data kan vaststellen, terwijl voor het beantwoorden van de tweede vraag *algemeen* statistische data nodig zijn.

Voor Engeland wordt hier vermeld, dat de fractie la (dus echt 13-jarigen) in de 1b-populatie 84% is. Aan dit gegeven zou met vrij grote zekerheid beantwoorden een gemiddelde leeftijd van 13 jaar 8 maanden voor 1b Engeland, in strijd met de elders vermelde 14 jaar 4 maanden. Omgekeerd beantwoordt aan de opgegeven gemiddelde leeftijd van 14 jaar 4 maanden veeleer een fractie $\frac{5}{6}$ (= 84%) *veertien*-jarigen (ipv. dertienjarigen) in 1b. Eigenlijk is de overeenstemming ($\frac{5}{6}$ = 84%) veel te goed – ik vermoed dat die 84% niet een statistisch gegeven, maar berekend is en dat bij vergissing de 16% ~ $\frac{1}{6}$) door zijn supplement 84% is vervangen. Het lijkt gek, maar het kan nog gekker.

Er is in het internationale rapport een speciaal hoofdstuk over de Engelse component van het onderzoek. Daar staat een verhaal dat ik wel twintig keer vergeefs heb trachten te begrijpen en dat mij nu bij het onderzoek naar de samenstelling van populatie 1b duidelijk is geworden. Uit dit verhaal blijkt, dat er langdurig en verward gediscussieerd is over definities van populatie 1b – er zijn er blijkbaar een heleboel geweest –, en dat tenslotte de Engelsen maar de knoop doorgehakt en er het 8e schooljaar voor genomen hebben. Was het hierbij gebleven, dan had de gemiddelde leeftijd van de klasse waarin de meerderheid van de 13-jarigen zat, voor Engeland dichter bij de 15 dan de 14 gelegen en had het aantal 13-jarigen in deze populatie inderdaad zowat bij de daarstraks door mij vermoede 16% gelegen. Jammer genoeg waren ze in Engeland door al dit geharrewar de kluts kwijt geraakt en is tenslotte feitelijk de groep 1b gedefinieerd als de vereniging van de

> *leerlingen in het 8e leerjaar van 14 jaar of ouder,*
> *leerlingen in het 7e leerjaar van 13 tot 14 jaar.*

Mijn voorspelling van daarstraks dat het nog gekker kan, is dus goed uitgekomen.

Wat Engeland aangaat, kan men met veel speurwerk er achter komen, dat bepaalde gegevens in het rapport in strijd met de waarheid zijn en dat zeer ernstige fouten zijn begaan. Hoe is het nu met Nederland gesteld? Hier blijkt het analoge (in omgekeerde richting) nergens uit het internationale rapport, maar wel uit het Nederlandse (dat zich trouwens van het internationale in elk opzicht voordelig onderscheidt). Ik citeer eruit (blz. 9):

> De ene uitzondering betrof Nederland. Doordat de frequentie van doubleren in ons land groter is dan in enig ander deelnemend land, is er geen enkel leerjaar aan te wijzen, waarin zich meer dan de helft van de 13-jarigen bevindt. Halverwege het schooljaar treft men ruim 20% van de 13-jarigen in de 6e klas van het lager onderwijs en iets minder dan 50% in het eerste leerjaar van het voortgezet onderwijs aan. Van de anderen bevinden zich ruim 10% nog op het niveau van de 5e klas lagere school; de overigen (20% – De Schrijver) zijn meest leerlingen van het tweede jaar van het voortgezet onderwijs, die in de tweede helft van het schooljaar 14 jaar worden . . .
>
> Om aan deze moeilijkheid het hoofd te bieden is in Nederland als tweede populatie gekozen alle leerlingen die zich hetzij in de 6e klas van de lagere school, hetzij in het eerste schooljaar van het voortgezet onderwijs bevinden.

Niets hiervan is in het internationale rapport terug te vinden; daar staat als Nederlandse definitie van 1b het zesde leerjaar aangegeven. Evenals in 't geval van Engeland is men ervoor teruggedeinsd openlijk te laten uitkomen dat twee (of meer nog) landen in strijd hebben gehandeld met de definitie van 1b als één leerjaar. Maar ook van die algemene definitie van 1b, die de Nederlandse uitweg zou hebben geprovoceerd, is geen aanduiding in het internationale rapport te vinden. Nergens staat er dat meer dan 50% der 13-jarigen in de populatie 1b vertegenwoor-

digd moet zijn. En dit is ook – nu in strijd met de bewering in de Nederlandse tekst – bij andere landen niet het geval geweest. Het geval Engeland met slechts een kleine fractie der 13-jarigen heeft ons al beziggehouden. In Schotland zat 40% van de 13-jarigen in het 7e leerjaar en 35% in het 6e en heeft men het 7e als 1b gekozen; België kwam met 56% niet ver boven de 50% uit; van Frankrijk zijn geen gegevens beschikbaar maar volgens de aanduidingen moet het percentage beneden de 50% hebben gelegen, en voor de Verenigde Staten komt men tot een dergelijke schatting.

Bij overmaat van ramp wordt bij geen enkele van deze cijfers in 'De plaats van de 13-jarigen in het systeem' vermeld voor welk tijdstip van het schooljaar ze berekend zijn; het is zelfs niet zeker of 'gemiddelde leeftijd in 1b' en 'fractie van alle 13-jarigen die in 1b zitten', 'fractie van 1b die 13-jarig is' op 't zelfde tijdstip zijn genomen. 'Halverwege het jaar' – zoals in het Nederlandse rapport staat – is een peildatum, die nergens in het internationale rapport voorkomt; het is òf 'op het ogenblik van het toetsen' of 'drie maanden voor 't eind van het schooljaar' of 'op 1 juli van het schooljaar' – ik kom op deze verwarring nog terug.

Het is dus niet juist dat Nederland in een speciale positie was; mogelijk was het enigszins uniek in die zin dat men hier tijdig de onderwijsstatistiek had geraadpleegd en van de gegevens die men daar vond een probleem heeft gemaakt. Het is ook niet juist dat de Nederlandse afwijking door het grote aantal zittenblijvers is veroorzaakt; volgens de gegevens in het internationale rapport is dat in Frankrijk en België groter en in de Verenigde Staten niet veel kleiner. Wat wel een speciale positie schept is het (niet gemakkelijk verklaarbaar) grote percentage van leerlingen die één jaar op hun leeftijdsgroep voor zijn – liefst 20%; deze groep heeft in feite tot het – kunstmatig geschapen – probleem aanleiding gegeven. Het gevolg was, dat men gemiddeld een half jaar naar beneden corrigeerde en de populatie Nederland 1b een half jaar jonger dan 1a (in plaats van een half jaar ouder) werd en het er 2½ punten slechter afbracht. Het verbaast me achteraf, dat het zo weinig scheelde en ik moet eerlijk zeggen, dat ik de zaak nog niet helemaal vertrouw en het vermoeden niet van me af kan zetten, dat er nog ergens een steekje los is.

Er bestond geen enkele reden voor de Nederlandse correctie naar beneden, zeker niet waar sommige landen liefst één jaar naar boven corrigeerden. Wel had er alle aanleiding bestaan om de Nederlandse afwijking in het internationale rapport te vermelden. Of waren er zoveel afwijkingen, dat er geen beginnen aan was?

Zijn de principes waarop de Nederlandse definitie voor 1b berust door de Nederlandse medewerkers zo maar uit de lucht gegrepen? Ik geloof van niet. Ik vermoed dat 'meer dan 50%' en 'halverwege het schooljaar' de oorspronkelijke formule (of een der oorspronkelijke formules) is geweest:* Wat men omtrent de internationale formule uit het rapport kan opmaken is namelijk weinig verheffend.

Ik moet eventjes weer op de stijl van het internationale rapport ingaan. Het is van 'de geest van de IEA studie was coöperatief' (I, blz. 64) (Lees: wat zijn wij toch aardige jongens onder elkaar), 'beïnvloed door de bedoelingen van het project', 'onze uiteindelijke beslissingen werden gedragen door de volgende criteria' (I, blz. 74), 'het proces van opstellen en reviseren van hypothesen was coöperatief' (I, blz. 75), en zo blijft het bij algemeenheden. Fundamentele beslissingen worden te vaak niet woordelijk, maar geparafraseerd weergegeven, mogelijk omdat die beslissingen inmiddels door de praktijk gelogenstraft zijn.

Nu de definities van de populatie 1b: Na een inleiding (I, blz. 45), waarvan ook de meest zorgvuldige lezer het fijne moet ontgaan, volgt eindelijk (I, blz. 46) iets wat – gecursiveerd – nu toch een keiharde definitie lijkt (en geen parafrase):

> In september 1963 werden de volgende definities van deelpopulaties geformuleerd:
> *Populatie 1a – Alle leerlingen, die 13.0–13.11 zijn op de dag van de toetsing.*
> *Populatie 1b – Alle leerlingen in dat leerjaar, waar de meerderheid van de 13.0–13.11 oude leerlingen zitten.*

Merkwaardigerwijs wordt bij 1b geen peildatum vermeld. Men zou aannemen dat het ook de 'dag van de toetsing' zou zijn, maar bladert men terug naar blz. 45, waar hetzelfde verhaal geparafraseerd staat, dan leest men

> afgesproken werd dat het het leerjaar zou zijn, waarin binnen drie maanden voor het eind van het schooljaar de meerderheid van de 13-jarigen zat.

Ook dit is geen ondubbelzinnige definitie, maar het vreemde is dat ze alleen als geparafraseerde afspraak wordt vermeld. Het is best mogelijk dat die

* Door de heer S. Wiegersma wordt mijn vermoeden bevestigd, dat 'meer dan 50%' en 'halverwege het schooljaar' de definitie-formule was.

'afspraak' pas ná september 1963 is tot stand gekomen.

Nu is september 1963 als datum voor een beslissing over een fundamentele definitie, op de drempel van de toetsingscampagne ook al een merkwaardige zaak. Het is nauwelijks denkbaar dat nationaal hiermee nog rekening kon worden gehouden – dat blijkt er bij voorbeeld uit, dat de Nederlandse groep de definitie niet kende en van een totaal andere uitging. Het is echter moeilijk, om de internationale organisatie voor dit gat te vangen. *Nergens* in het rapport *wordt beweerd* dat het in september 1963 beslotene de *definitie* is die aan het onderzoek *ten grondslag ligt*. Er wordt uitsluitend gezegd dat dit definities waren, maar niet dat ze ook toegepast zouden worden of toegepast werden. Aan de andere kant wordt ook nergens vermeld dat ze *niet* toegepast zijn.

Op zichzelf is het geen kwaad idee, naast leeftijdspopulaties ook schoolleeftijdspopulaties te beschouwen. Ik zou zelfs zeggen, dat het bij dergelijk onderzoek het enige juiste is. Wat hier mis is, is de wijze van definiëren – achter het bureau i.p.v. van de praktijk uitgaande en er op gericht. Wat zou eenvoudiger zijn geweest dan voor te schrijven 'het zevende leerjaar volgens die en die nationale tellingen' (want dit was de bedoeling); voor andere tellingen (vroeger of later schoolbegin) was het beantwoordende leerjaar hieruit met groot gemak af te leiden. Uniformiteit zou hierdoor ook niet gegarandeerd zijn geweest; landen die volgens de leeftijd toelaten zoals Nederland, kregen in elk leerjaar een enkele maanden oudere schoolbevolking dan die volgens het kalenderjaar toelaten. Maar er zou op deze wijze toch wel een hoge mate van uniformiteit zijn bereikt en voor de verschillen had men – aangezien ze bekend zijn – gemakkelijk kunnen corrigeren.

Als men daartegenover een bureau-definitie prefereert, dan is men verplicht, om er een te kiezen die deugt. Wat men hier gedaan heeft, is het ergste dat men kon doen: een *discontinue* definitie geven. Geheel toevallig kon de meerderheid der 13-jarigen van een deelnemend land bijv. op 29 februari 1964 in het achtste en op 1 maart 1964 in het zevende leerjaar zitten. Ook als men precies een peildag voorschrijft, blijft de procedure onaanvaardbaar. Het is onaanvaardbaar dat in een onderzoek van de aard van het onderhavige op zulk een punt een zwaarwegend kanselement wordt geïntroduceerd. Is dit nooit aan de orde gesteld? Er schijnt over modaliteiten gepraat te zijn, maar niet over het principe. Tenslotte in tijdnood moest iedere nationale groep de knoop maar doorhakken – twaalf knopen op twaalf manieren doorgehakt. Het internationale rapport zwijgt er als het graf over. Waar de definitie van 1b ter sprake komt, is het altijd bij wijze van parafrase. Stilletjes is 'meerderheid der 13-jarigen' veranderd in 'de meeste 13-jarigen', waarmee de verleiding 'absolute meerderheid' wordt uitgesloten, of 'de grootste fractie 13-jarigen', dat nog een tikkeltje duidelijker is. Nergens wordt er de peildatum aan toegevoegd; nergens wordt vermeld welke peildatum en welke definitie de diverse nationale groepen hebben toegepast. Behalve natuurlijk de Engelsen, die in hun rapport duidelijk 1 juli 1964 als peildatum aangeven, maar desniettemin als 1b het 8e leerjaar hebben gekozen dat dan nog maar een zesde van de 13-jarigen omvat.

10. *Zijn fouten onvermijdelijk?*

Op critiek zoals hier geoefend volgt steevast het antwoord: 'Natuurlijk worden bij zulk een mammoet-onderzoek ook fouten gemaakt; natuurlijk zijn er onder duizenden cijfers ook een aantal die niet kloppen; dit is onvermijdelijk".

Inderdaad, ik heb het wel eens de bedrijfsongevallen van de onderwijsstatisticus genoemd. Maar die duizenden cijfers staan niet los van elkaar; door ze te relateren, wil men conclusies trekken, en dan kan al één fout alles bederven. Toen Snellius met een tot dan toe ongekende nauwkeurigheid geodetische metingen deed om de omtrek van de aarde te bepalen, heeft hij maar één fout gemaakt op vele duizenden metingen, maar dan met het gevolg dat alle moeite, aan de nauwkeurigheid van die metingen besteed, vergeefs was.

Door hoeveel fouten is het onderhavige onderzoek verontsierd? Ik heb alleen maar een kleine steekproef genomen. Wel, het was geen toevalssteekproef. Ik kwam cijfers tegen die ik niet vertrouwde. 'Een significant verschil', zou de onderwijsstatisticus triomfantelijk uitroepen. De reactie van de wiskundige is dan veeleer 'hier is iets loos'. En dan begint de speurtocht. Ik kan meer van die vreemde cijfers aanwijzen, maar heb geen zin, me op nog meer speurtochten te begeven.

Onderzoek als het onderhavige is tot falen gedoemd en is niet te redden door mathematische finesses als in I, Hoofdstuk 9 uiteengezet die op zulk materiaal toegepast alleen maar tot hoofdschudden uitnodigen. De centrale organisatie van zo'n onderzoek bestaat uit mensen, die af en toe een vergadering kunnen bijwonen, die er maar een fractie van hun tijd aan kunnen besteden en die op zijn minst al daarom niet in staat zijn, het behoorlijk op te zetten en er een kijk op te houden. Bovendien

hangen zij af van een dozijn perifere centra, waarop ze niet de minste controle hebben. Deze perifere centra hangen in hun activiteit af van honderden scholen en duizenden leraren en de kleinste vergissing in de opzet kan tot onherstelbare fouten leiden. Veel te groots en te ambitieus opgezet moet zo'n projekt bij alle toewijding na korte tijd uit de hand lopen, en wat er uit komt kan alleen een rijstebrijberg van getallen zijn, waarop men ver van de bron geen kijk meer heeft en die daarom ook niet te verwerken is. Alleen uit dit gebrek aan overzicht valt een rapport te verklaren, vol van tegenstrijdigheden en met hele hoofdstukken die zonder kennis van het onderzoek zijn geschreven.

Bij experimenteel onderzoek zijn fouten onvermijdelijk. Het is moeilijk alle foutenbronnen van te voren te verkennen. Bij natuurwetenschappelijke experimenten is het normaal, dat na negen aanlopen pas de tiende slaagt. En dan mag je nog van geluk spreken, want het kan ook totaal mislopen.

De principiële fout van mammoet-onderzoek als het onderhavige is dat het niet herhaalbaar is; loopt er bij alle toewijding die er zeker was, iets mis, dan kun je niet een tweede keer over de millioenen beschikken, die het heeft gekost. Die moeten bovendien verantwoord worden en dat verklaart rapporten als het onderhavige. Een rapport, waar eens alle fouten uit de doeken worden gedaan, zou ten minste methodologische waarde hebben, men zou er uit kunnen leren, maar wie overziet al die fouten nog?

11. *Conclusies*

Om tot de conclusie te komen:

1. Door de fundamentele tekortkoming van het als een constante behandelen van de variabele curriculum zijn de op statistische vergelijking van internationale data gebaseerde conclusies voor 't grootste deel twijfelachtig.
2. De overige gekozen variabelen zijn nauwelijks specifiek voor het wiskunde-onderwijs.
3. Ze zijn bovendien veel te zeer gericht op het onderwijs als een abstracte entiteit en te weinig op het onderwijs zoals het reilt en zeilt.
4. Werkelijk essentiële variabelen zijn over 't hoofd gezien.
5. Er zijn zeer ernstige fouten begaan, maar niet vermeld.
6. De tekortkomingen zijn het gevolg van te weinig kwalitatieve kennis omtrent het wiskunde-onderwijs en het onderwijs in 't algemeen op internationale schaal.

12. *Slotopmerking*

Op sommige plaatsen spreekt uit dit rapport een zelfvoldaanheid met de bereikte resultaten, die vooral dan vreemd aandoet, als het om resultaten gaat die nergens in het rapport te vinden zijn. Het is me niet bekend of het rapport ergens critischer is bekeken dan door zijn leiders. De IEA heeft op dit onderzoek er één omtrent moedertaal, natuurwetenschappen en maatschappijleer laten volgen, waarvan mij thans twee delen bekend zijn. Hiernaar te oordelen zijn de fouten van het eerste onderzoek bij het tweede versterkt herhaald.

13. *APPENDIX: Het Nederlandse aandeel*

Zoals eerder vermeld, doet het Nederlandse rapport weldadig aan, vergeleken bij het internationale met zijn zelfvoldaanheid en neiging om tekortkomingen in understatements en parafrases te verbergen. Ik zal het hier echter niet op zijn eigen merites waarderen; ik beperk me tot zijn relatie tot het internationale rapport.

Ook aan het Nederlandse aandeel werkten geen wiskunde-onderwijskundigen mede. Aan de geciteerde vereiste, dat een groep van wiskundigen en wiskundeleraren de doelstellingen van wiskunde-onderwijs in ons land autoritatief zou formuleren, werd *niet* voldaan. Van enige medewerking van wiskundigen, zelfs op de meest bescheiden schaal, wordt niet gerept.

Over het internationaal onderzoek schrijft men (blz. 5):

> Een werkgroep van drie wiskundigen kreeg . . . tot taak de nationale analyses te vergelijken en vervolgens series opgaven samen te stellen, die internationaal bruikbaar zouden zijn voor onderzoek naar de resultaten van het wiskunde-onderwijs. Het grote voordeel van het gekozen studieobject bleek nu: over de totaliteit van het onderwijs gezien zijn de verschillen in de inhoud van het wiskunde-onderwijs tussen de deelnemende landen gering. Er zijn uiteraard verschillen naar de plaats en tijd waar bepaalde gedeelten van de wiskunde worden behandeld, maar het is een uitzondering, dat bepaalde onderdelen in een land geheel niet aan de orde komen . . .

Over de 'werkgroep van drie wiskundigen' heb ik eerder het nodige gezegd. In hun rapport (I, Hoofdstuk 4) komt niets voor dat ook maar in de verte lijkt op het hierboven beweerde. Men kan daar veeleer het tegendeel lezen (gedeeltelijk door mij

uit I, blz. 83–85 geciteerd). Het in het Nederlandse rapport gestelde is veeleer afkomstig van plaatsen in het internationale rapport, geschreven door medewerkers die van de inhoud van I, Hoofdstuk 4 niet op de hoogte waren of deze trachtten te bagatelliseren.

De verschillen van inhoud van het wiskunde-onderwijs in de diverse landen waren bepaald niet gering. In de inventaris waarop de Nederlandse medewerkers doelen (I, App. 1) werd gewerkt met de categorieën

U universal (op alle scholen onderwezen)
R restricted (op sommige scholen onderwezen)
E experimental
N nil (nergens in het schoolsysteem onderwezen)

Het 'geheel niet aan de orde komen' is in de uiterst beperkte zin van N bedoeld. Dit N vindt men bij minstens één land (maar niet alle) voor

24 van de 41 onderwerpen van de onderbouw (1a-b)
22 van de 39 onderwerpen van de bovenbouw (3a-b)

Zoiets zou ik geen 'uitzondering' noemen.

Ik ga nog nader op de Nederlandse vertaling van de toetsen in. De Noordnederlandse vertaling – moet ik eigenlijk zeggen. Want terwijl Engelsen, Schotten, Amerikanen en Australiërs met één (Engelse) versie konden volstaan, waren er voor het Nederlandse taalgebied twee nodig. Immers voor de buitenwereld zijn 'Dutch' en 'Flemish' twee verschillende talen. Mijn oordeel slaat op de preventief-geneeskundige versie: voor de groepen 1 matig, voor de groepen 3 slecht.

Allereerst bestond er een neiging, de toetsen in de vertaling te vergemakkelijken. De taal van het origineel is soms ingewikkeld, blijkbaar met opzet, namelijk om ook de bekwaamheid van de leerlingen in het lezen van ingewikkelde taalkundige constructies te onderzoeken; deze complicaties zijn vrij systematisch weggewerkt.

Vereenvoudigd werden
bij 1a-b: 11 van de 70 vraagstukken
bij 3a : 11 van de 69 vraagstukken.

Hierbij was in 1a-b twee keer de vereenvoudiging sterk (A 20, A 22 door weglaten van de misleiding); bij 3a was dit 4 keer het geval (VI 5, VII 16, VIII 3 door wegwerken van een misleiding, VIII 13 door verbetering van een slechte Engelse tekst).

Een keer kwam het voor, dat de Nederlandse tekst moeilijker was (G 6).

In 1a-b komen voor:

1 taalkundig foutieve vertaling (B 5) zonder consequentie,
2 logisch foutieve vertalingen (C 18–19),
1 logisch slechte vertaling (A 23),
1 onduidelijke vertaling (B 18),
1 mathematisch foutieve vertaling (B 22)
1 mathematische nonsense vertaling (C 17).

In 3a-b komen (bovendien) voor

1 taalkundig foutieve vertaling (VI 7) zonder consequenties,
1 taalkundig foutieve vertaling (VII 7) mèt consequenties,
5 logisch foutieve vertalingen (VII 7, VII 11, VIII 11, IX 12, IX 7_3),
soms grove denkfouten behelzend, waardoor de zin geheel gewijzigd wordt,
2 onbegrijpelijke vertalingen (V 17, V 19),
1 misleidende vertaling (VII 9),
1 onbegrijpelijke vertaling van een slecht te begrijpen Engelse tekst (VIII 15),
1 (toen) ongebruikelijke terminologie, die volgens de internationale bepalingen in de in Nederland gebruikelijke had moeten worden omgezet (VII 14).

Het toppunt is zonder twijfel de preventief geneeskundige vertaling van het Engelse 'if and only if' door

'alleen en slechts alleen dan als'.

Ik heb dit niet opgesomd om op elke slak zout te leggen. Ik had een verderstrekkende bedoeling. De IEA heeft inmiddels onder meer een internationaal onderzoek naar moedertaalbeheersing gedaan. Hiervoor werd een instrument geschapen, om tekst- en woordbegrip te toetsen. Ik vraag me af hoe men zoiets adekwaat in een vreemde taal kan overbrengen, wanneer al het adekwaat vertalen van wiskundige toetsten op zulke enorme moeilijkheden stuit.*

* Het is me inmiddels gebleken, dat het probleem van adekwaat vertalen internationaal niet eens gesteld is.