Periodiek onderzoek naar de kwaliteit van het onderwijs: enige praktische ervaringen en methodologische problemen*

W. J. VAN DER LINDEN en
W. J. PELGRUM**
Technische Hogeschool Twente, Enschede

Samenvatting

Onlangs is van overheidswege om hernieuwde aandacht voor het kwaliteitsniveau van het onderwijs gevraagd. Een van de instrumenten voor de bewaking van de kwaliteit van het onderwijs is assessment-onderzoek. Dit zou dan opgenomen moeten zijn in een kwaliteitsbeheersingscyclus die van normstelling via vaststelling van de mate van normrealisering naar de keuze van beleidsmaatregelen dient te lopen. De bedoeling van dit artikel is om eerst een methodologische kenschets van assessment-onderzoek te geven en een aantal problemen te inventariseren. Daarna wordt een aantal bevindingen uit het Nederlandse aandeel van de IEA Second Mathematics Study gepresenteerd, een type onderzoek dat met uitzondering van de longitudinale aspecten volledig aan de opzet van assessment-onderzoek beantwoordt. Tenslotte wordt aangegeven dat de methodologische problemen van assessment-onderzoek opgelost lijken te kunnen worden door middel van nieuwe toepassingen van de itemresponstheorie. De algemene conclusie is dat kwalitatief goed assessment-onderzoek mogelijk is mits praktische ervaringen en nieuwere methodologische inzichten met elkaar verbonden worden.

1 *Inleiding*

In 1981 verscheen van de hand van de toenmalige bewindslieden van Onderwijs en Wetenschappen de nota 'Kwaliteit van het onderwijs'. Door middel van deze nota vroegen zij om een bezinning op de problematiek van de kwaliteit van het onderwijs.

Opmerkelijk aan de nota is dat deze niet in algemene kwaliteitsbeschouwingen blijft steken maar ruime aandacht schenkt aan mogelijke instrumenten voor de bepaling van de kwaliteit van het onderwijs. Bestaande regelingen worden gememoreerd en bij wijze van voorbeeld worden, zowel voor het school- als voor het landelijk niveau, enige aanvullingen op het bestaande instrumentarium genoemd. Een van de aanvullingen waarop gezinspeeld wordt, is die van periodiek onderzoek naar de kwaliteit van onderwijs, of – zoals dit in de Angelsaksische landen reeds lang bekend staat –: assessment-onderzoek. In Nederland is reeds door middel van enige, op specifieke probleemstellingen gerichte onderzoeksprojecten inzicht verkregen in het functioneren van sommige sectoren van ons onderwijssysteem. Met name kan hier gewezen worden op de volgende projecten van de *International Association for the Evaluation of Educational Achievement* (IEA) waaraan door Nederland is deelgenomen: First Mathematics Study, Six Subject Study, Second Mathematics Study, Written Composition Study. Grootschalig opgezet onderzoek naar de kwaliteit van onderwijs dat op periodieke basis opgezet en herhaald wordt kennen wij in Nederland evenwel nog niet. Een gevolg hiervan is bijvoorbeeld dat effecten van curriculuminnovaties en van wijzigingen in het onderwijsstelsel (Mammoetwet!) op landelijk niveau amper zichtbaar zijn geworden. De recente, van overheidswege ondernomen inventariserende activiteiten om de relevantie en haalbaarheid van assessment-onderzoek in ons land te beproeven, lijkt dan ook een welkome ontwikkeling.

Het is de bedoeling van dit artikel om eerst in te gaan op enkele methodologische problemen die aan assessment-onderzoek kleven. Vervolgens worden mogelijkheden voor assessment-onderzoek geïllustreerd aan ervaringen en gegevens uit het Nederlandse aandeel in de IEA Second Mathematics Study (Tweede Wis-

* Omwille van ruimtewinst is dit artikel teruggebracht tot een korte versie van het oorspronkelijke manuscript.
Het volledige manuscript is op aanvraag bij de auteurs verkrijgbaar.

** Beide auteurs hebben gelijk aan de totstandkoming van dit artikel bijgedragen.

kunde Project). Tenslotte worden enkele aanbevelingen gedaan die de kwaliteit van assessment-onderzoek ten goede kunnen komen. In de Nederlandse literatuur werd eerder over assessment-onderzoek geschreven in een beschouwing door Lagerweij (1982) en in de reisverslagen van Sandbergen (1979) en Wijnstra (1982). Het is de bedoeling van dit artikel om deze publikaties te completeren door praktische ervaringen te presenteren en in te gaan op enkele methodologische problemen van assessment-onderzoek.

2 *Assessment-onderzoek*

Assessment-onderzoek is een grootschalige verzameling van empirische gegevens over de resultaten van onderwijs die periodiek herhaald wordt en meestal een systematische afwisseling over leerstofgebieden en/of onderwijssectoren vertoont. Omdat telkens een volledige dwarsdoorsnede van een sector en/of een leerstofgebied beoogd wordt, moet volstaan worden met steekproeftrekking en (statistische) generalisatie van de uitkomsten. In dit opzicht – en mede ook door de praktische problemen die zich bij de realisatie van het onderzoek voor kunnen doen –, lijkt assessment-onderzoek veel op survey-onderzoek, zij het dat de laatste veelal minder op resultaatmetingen gericht is en overwegend voor andere doeleinden gebruikt wordt (verzamelen van gegevens over opinies, marktgedrag, e.d.). Een ander verschil met survey-onderzoek is het longitudinale karakter van assessment-onderzoek, dat essentiële randvoorwaarden oplegt aan iedere afzonderlijke dwarsdoorsnede.

Voor de in assessment-onderzoek te gebruiken meetinstrumenten geldt dat ze een goede dekking van het curriculum van het betreffende leerstofgebied of vak moeten garanderen. Hoe dit bereikt kan worden is een van de moeilijke opgaven waar men voor komt te staan. Assessment-onderzoek is transversaal onderzoek, d.w.z. gericht op het maken van dwarsdoorsneden van (delen) van het onderwijssysteem. De methodologische vraag die dit met zich meebrengt is hoe met een minimum aan kosten en 'verstoring' een zo nauwkeurig mogelijke doorsnede gemaakt kan worden. Assessment-onderzoek is evenwel ook longitudinaal onderzoek, wat van het begin af aan eisen aan de opzet stelt en onder meer de noodzaak van equivalente metingen met zich meebrengt. Adequate rapportage van de resultaten is van eminent belang en geeft, hoe wonderlijk dit misschien op het eerste gezicht moge lijken, lastige technische problemen. De verzamelde gegevens moeten bijvoorbeeld van dien aard zijn dat aggregatie op het gewenste niveau (regio, sector, nationaal, e.d.) mogelijk is, wat eisen aan de meet- en analysemethoden stelt. Tenslotte is assessment-onderzoek een gigantische organisatorische taak met vaak onvermoede problemen van plannings- of logistieke aard waarvoor management door onderzoekers met ervaring in grootschalig onderzoek vereist is. In het vervolg gaan we op ieder van bovengenoemde aspecten afzonderlijk in. Aan de orde komen dus achtereenvolgens curriculum-, transversale, longitudinale, rapportage- en praktische aspecten van assessment-onderzoek. De behandeling zal overwegend probleemstellend zijn. Nadat enige ervaringen uit het IEA Tweede Wiskunde Project gepresenteerd zijn komen we in een slotparagraaf op deze aspecten terug aan de hand van enkele nieuwe methodologische ontwikkelingen en doen we aanbevelingen. Hoewel we dus eerst afzonderlijk op de genoemde aspecten van assessment-onderzoek ingaan zullen we vrijelijk aan grensoverschrijding doen. De lastigste problemen liggen namelijk juist in het feit dat de gezamenlijke aanwezigheid van deze aspecten soms onverenigbare eisen met zich mee lijkt te brengen. Zo vragen bijvoorbeeld curriculumwijzigingen om veranderingen in het instrumentarium terwijl dit juist constant zou moeten blijven vanwege het longitudinale karakter van het onderzoek. Verder brengt goede rapportage van assessment-onderzoek publikatie van toetsitems met zich mee, maar lijkt (opnieuw) het longitudinale karakter van het onderzoek te eisen dat deze voor de volgende afnamen achtergehouden worden. De kunst om goed assessment-onderzoek op te zetten ligt in het met methodologische creativiteit verenigen van dergelijke op het eerste gezicht tegenstrijdige eisen.

2.1 *Curriculumaspecten*

De keuze en samenstelling van de meetinstrumenten wordt in assessment-onderzoek meestal bepaald in samenspraak met panels van curriculumdeskundigen die wenselijke (na te streven) doelstellingen identificeren. Hier worden vervolgens toetsopgaven, vragen, e.d.

bij geconstrueerd. Door middel van eventuele pilot testing en opnieuw beoordeling door een panel (nu curriculumdeskundigen èn psychometrici) kunnen hier dan nog wijzigingen in aangebracht worden. Grofweg wordt deze methode in Nederland ook met betrekking tot eindexamenconstructie gevolgd. Een in het oog springend nadeel van dit deskundigheidsmodel (waarin de doelstellingeninventarisatie doorgaans door een beperkt aantal participanten geschiedt) is dat zo'n ondervraging van sleutelfiguren alleen maar goed kan werken in een gecentraliseerd onderwijssysteem. Immers, in een dergelijk systeem zijn uniforme, voorgeschreven curricula aanwezig die daadwerkelijk uitgevoerd worden en waarop de controle scherp is. In dergelijke systemen kan door beleidsinstanties ook veel waarde gehecht worden aan discrepanties tussen verwachte en feitelijke opbrengsten. Hoewel in het Nederlandse onderwijssysteem een zekere uniformiteit in curricula bewerkstelligd wordt door centrale examens, ministeriële richtlijnen en beschikbare methoden van educatieve uitgevers, is er toch sprake van een relatief grote autonomie bij de curriculumplanning van de individuele school of docent. Om deze reden zal in ons land de registratie van de feitelijk aangeboden curriculuminhouden mede geboden zijn. Bij assessment-onderzoek in Nederland is het onderscheid tussen de volgende niveaus dus van eminent belang:

- het bedoelde curriculum;
- het feitelijk uitgevoerde curriculum;
- de opbrengst van het curriculum.

Wie wel eens serieus een studietoets geconstrueerd heeft weet dat het probleem van de dekking van de leerstof (of zo men wil: de inhoudsvaliditeit van de toets) altijd één van de moeilijke punten is. In assessment-onderzoek verhevigt dit probleem zich. Allereerst komt dit door de omvang van het leerstofgebied. In assessment-onderzoek gaat het niet om de dekking van de leerstof van een beperkt aantal lessen of een korte cursus maar van een geheel schoolvak. Dit betekent enerzijds een kwantitatieve vergroting van het dekkingsprobleem; men staat nu voor de taak een geheel vak te dekken. Anderzijds wordt er kwalitatief meer gevraagd. Ieder vak heeft zijn rijkdom aan gedragsniveaus waarop onderwezen wordt, en er kan niet volstaan worden met een enkele toetsingsvorm die aan alle leerlingen voorgelegd wordt. Verder is het onder schrijvers van leerdoelen en testitems een bekende ervaring dat men door kan gaan met verdere detaillering en verfijning van de curriculuminhoud. Een fraai verslag van de problemen die zich hier aan de testconstructeur voor kunnen doen is te vinden in Popham (1980).

Toch stelt ook hier de praktische haalbaarheid van het onderzoek zijn randvoorwaarden. Men kan niet alle leerlingen alles voorleggen. Bovendien zal de afname met een minimum aan inbreuk op het klassegebeuren plaats moeten kunnen vinden, dus veelal met korte schriftelijke instrumenten en met opgaven die de leerlingen zelfstandig kunnen maken. De statistiek maakt echter steekproeftrekking van curriculumelementen mogelijk waardoor niet alles aan iedere leerling voorgelegd hoeft te worden en nauwelijks informatie verloren gaat.

2.2 *Transversale aspecten*

Nadat een goede dekking van het curriculum bereikt is, staat men voor de taak om een goede dwarsdoorsnede te maken van de opbrengsten van het curriculum voor een sector van het Nederlandse onderwijs. Dit betekent dat men de prestaties moet meten van populaties van soms tienduizenden leerlingen, met hun interne structuur aan klassen, schooltypen, pedagogisch-didactische opvattingen, regio's, enz. Praktisch gezien zijn er oneindig veel onderverdelingen van de totale populatie mogelijk, en op welke onderverdelingen men het houdt hangt met name af van de beleidsvragen waarop een antwoord gegeven moet worden en de niveaus waaraan gerapporteerd moet worden.

Het zal duidelijk zijn dat niet aan alle leerlingen alles voorgelegd kan worden. De taak waar men bij de opzet van assessment-onderzoek voor staat is om op efficiënte wijze en met zo laag mogelijke kosten nauwkeurige informatie te verzamelen. Praktisch betekent dit dat men de hulp van de statistiek in zal moeten roepen en met steekproeftrekking en schatting van populatiegrootheden moet volstaan. Bij geschikt gekozen steekproefgrootten hoeft dit slechts tot een verwaarloosbaar verlies aan nauwkeurigheid te leiden. Er is reeds opgemerkt dat assessment-onderzoek door zijn transversale aspecten verwantschap vertoont met survey-onderzoek. In dergelijk onderzoek wordt veelal met steekproeftrekking van respondenten gewerkt (zie bijv. Albinski, 1974, hfd. 5). Zoals we eerder reeds aangaven moet ook steek-

proeftrekking uit de inhoud van het curriculum plaats vinden om dekking van het curriculum te garanderen (binnen randvoorwaarden als bijvoorbeeld de beschikbare toetstijd in de klas). Beide vormen van steekproeftrekking leiden logisch tot een techniek die bekend staat als 'multiple-matrix sampling', waarbij aselecte steekproeven van items aan aselecte steekproeven personen toegewezen worden. Is men geïnteresseerd in (sub)populatiegemiddelden, dan biedt deze techniek vele voordelen. Lord (1962) was één van de eersten die het effect van matrix sampling op de vereiste steekproefgrootte onderzocht. Een eenvoudig overzicht van de beschikbare theorie is te vinden in Sirotnik (1974). Een voorbeeld van de praktische winst die men kan boeken is te vinden in het California Assessment Program, waar uit berekeningen bleek dat men bij overgang van gewone survey sampling op matrix sampling per onderzochte leerling slechts een kwart van het aantal items nodig had om dezelfde nauwkeurigheid te bereiken (Pandey & Carlson, 1976). Wanneer beleidsrelevante onderverdelingen van de populatie gegeven zijn en rapportage op deze niveaus gewenst is, kan gedacht worden aan het gestratificeerd trekken van matrixsteekproeven. Voor verdere reductie van steekproefomvang zou het gebruik van Bayesiaanse statistiek overwogen kunnen worden. Mogelijke verfijningen van de techniek van matrix sampling zijn nog niet alle onderzocht. Hier ligt nog een gebied braak dat voor assessment-onderzoek van belang is.

2.3 *Longitudinale aspecten*

Wil men assessment-onderzoek gebruiken voor de evaluatie van innovaties of voor de vaststelling van de effecten van bijvoorbeeld een teruglopend leerlingenaantal, dan zijn het juist de longitudinale aspecten die dit onderzoek zo geschikt maken voor zijn doel. Effecten van innovaties en demografische ontwikkelingen manifesteren zich als een discontinuïteit of een geleidelijke trend in de resultaten van opeenvolgende metingen. Om in te zien dat empirische gegevens van deze aard nuttig zijn, hoeft men slechts te denken aan het 'test-score decline' debat in de Verenigde Staten. Met een verwijzing naar dit debat zitten we ook in het hart van de problemen die aan longitudinaal sociaal-wetenschappelijk onderzoek kleven. Vermoedens over teruglopende testprestaties kunnen ontstaan wanneer er geen vergelijkingsbasis aanwezig is, en deze ontbreekt wanneer metingen niet equivalent zijn. Om praktische redenen kan men dezelfde toetsitems slechts één keer gebruiken. In assessment-onderzoek is dit bijvoorbeeld zo omdat items moeilijk geheim gehouden kunnen worden en dit om rapportageredenen zelfs ongewenst is (zie par. 2.4). Wil men bij volgende gelegenheden gerichte training op het in de items gevraagde voorkomen, dan zullen nieuwe items gebruikt moeten worden. Het probleem dat men dan echter ontmoet is dat van de testafhankelijke scoring (zie voor de betekenis van dit probleem voor evaluatie-onderzoek Van der Linden, 1978). Zou bijvoorbeeld de rapportage plaats vinden in de vorm van geschatte item p-waarden, dan is het niet duidelijk of een verschil – of een overeenkomst! – tussen p-waarden toegeschreven moet worden aan een verschil in opbrengsten van het onderwijs en/of aan een verschil in de moeilijkheid van de items. Zonder verdere voorzieningen ter voorkoming van deze onvergelijkbaarheid, zou de zin van assessment-onderzoek vervallen.

In ander longitudinaal sociaal-wetenschappelijk onderzoek is dit probleem meestal oplosbaar door de scores van instrumenten die geacht worden hetzelfde te meten te 'equivaleren' met een van de beschikbare conventionele methoden. Deze bestaan hieruit dat in een apart (voor)onderzoek een tweetal instrumenten bij dezelfde steekproef-personen afgenomen wordt. Vervolgens wordt de bivariate scoreverdeling benut om een transformatie te vinden die de scores van het ene instrument omzet in die op het andere. Ook kan men proberen enkele gemeenschappelijke items (*ankeritems*) aan beide instrumenten mee te geven die wel geheim blijven en gebruikt worden om de juiste transformatie te vinden. In het algemeen is de eerste methode de minst praktische en de tweede de minst betrouwbare. Een extra complicatie bij de tweede methode is dat niet bekend is hoe deze zich met multiple-matrix sampling verdraagt; voor zover wij weten is conventionele equivalering van tests met behulp van scores verkregen door multiple-matrix sampling een nog niet benaderd probleem. In het meeste sociaal-wetenschappelijke longitudinale onderzoek kunnen wel oplossingen gevonden worden voor de hierboven beschreven praktische problemen en is een min of meer bevredigende equivalering mogelijk. Bij

assessment-onderzoek is er evenwel een apart probleem. Omdat daar als eindprodukt onder andere een rapportage op itemniveau wordt beoogd, ligt de interesse niet in de equivalering van *toetsscores* maar in de mogelijkheid om de prestaties tijdens opeenvolgende metingen op *itemniveau* in elkaar te vertalen. We zullen straks zien dat nieuwere methoden beschikbaar zijn waarmee deze en andere problemen opgelost lijken te kunnen worden.

2.4 *Rapportage*

Men kan stellen dat het succes van assessment-onderzoek afhangt van de mate waarin men er in slaagt een op de gebruiker afgestemde rapportage plaats te doen vinden. Zo bleek uit een evaluatie-onderzoek naar de effectiviteit van de beginperiode van het National Assessment of Educational Progress Project (NAEP) dat men er kennelijk niet in geslaagd was adequaat te rapporteren. De door het project aangedragen informatie vond men te weinig afgestemd op de gebruiker, te globaal, te weinig voorzien van beleidsvoorstellen en werd te weinig doorgespeeld naar lokale niveaus (Lapointe & Koffler, 1982). De reisverslagen van Sandbergen (1979) en Wijnstra (1982) vermelden vergelijkbare bevindingen. Dit kan aan het project gelegen hebben maar ook aan het feit dat de in de nota 'Kwaliteit van het onderwijs' zo bepleite fase van normstelling ontbroken heeft. Met andere woorden, men zou minder de deductieve werkwijze gevolgd hebben die in de reeds besproken kwaliteitsbeheersingscyclus besloten ligt en eerder inductief te werk zijn gegaan door de beleidsadviezen vanuit het verzamelde materiaal op te laten borrelen. Het komt ons voor dat de in de nota voorgestelde deductieve werkwijze betere garanties biedt voor adequate rapportage aan het beleid doordat beleid en onderzoek vanaf het begin op elkaar gebonden zijn. Eveneens onmisbaar zijn rapportages afgestemd op andere mogelijke gebruikers (scholen, verzorgingsstructuren, vakverenigingen, leermiddelenindustrie, bijscholingsinstanties, en niet te vergeten individuele docenten) die hier ieder maatregelen aan kunnen ontlenen voor hun eigen taken. Tenslotte is er voor onderwijsonderzoekers de mogelijkheid om het verzamelde en geanalyseerde materiaal voor onderzoek te gebruiken.

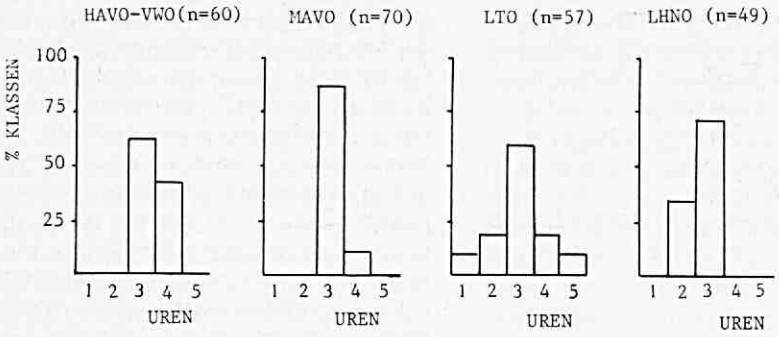
Rapportage van curriculumopbrengsten kan in het algemeen normgeoriënteerd- of criteriumgeoriënteerd van aard zijn. In het eerste geval rapporteert men de resultaten met een relatieve interpretatie en krijgen de prestaties van bijvoorbeeld een groep scholen of een regio reliëf tegen de achtergrond van de verdeling van de prestaties van de hele populatie. In het tweede geval worden de (groeps)scores absoluut geïnterpreteerd door ze te beschouwen tegen de achtergrond van een gedragscontinuum. Voor een korte beschrijving van de verschillen en een opsomming van beschikbare criteriumgeoriënteerde methoden verwijzen we naar Van der Linden (1982a, 99-101).

Het zal duidelijk zijn dat assessment-onderzoek om een criteriumgeoriënteerde rapportage vraagt. In de eerste plaats kunnen totaalresultaten van een sector van het onderwijs niet van een relatieve interpretatie voorzien worden (tegen welke populatie zouden deze vergeleken moeten worden?). Verder vraagt het onderwijsbeleid en eventuele te nemen beleidsmaatregelen om de vergelijking van feitelijke en gewenste resultaten op een interpreteerbare, 'absolute' schaal. Natuurlijk is het wel zinvol om lokale resultaten mede van relatieve interpretaties te voorzien.

Recent is er in assessment-onderzoek een vorm van rapportage in gebruik genomen die bekend staat als 'scale-score reporting' en die een criteriumgeoriënteerde interpretatie van (groeps)scores op het niveau van de individuele items mogelijk maakt. We doelen hier op het werk van Bock en Mislevy in het California Assessment Project (CAP). Voorlopers van deze rapportagevorm, die een bijzondere diagnostische waarde heeft en in principe voor groepsscores op ieder gewenst aggregatieniveau toepasbaar is (in het CAP wordt deze voor rapportage aan individuele scholen gebruikt), zijn elders te vinden. 'Scale score reporting' berust op psychometrische modellen die ook bruikbaar zijn voor de oplossingen van eerder genoemde problemen in assessment-onderzoek. We komen hier in onze slotparagraaf op terug.

2.5 *Praktische aspecten*

Assessment-onderzoek brengt een aantal praktische problemen met zich mee waarvan we er enkele noemen: logistieke problemen (verspreiding en retournering van grote hoeveelheden toetsmateriaal), planning en bewaking van de werkzaamheden, hantering van omvangrijke databestanden, tijdige rapportage. We noemen deze omdat ze het succes van



Figuur 1 *Verdeling van het aantal uren wiskunde per week per schooltype*

assessment-projecten kunnen maken of breken. Het management van assessment-onderzoek vraagt om ervaring die minstens op het niveau van een groot survey-onderzoek ligt.

3 *Het IEA Tweede Wiskunde Project*

In de volgende paragrafen wordt een kort overzicht gegeven van het Nederlandse aandeel in het TWP dat aan de T.H. Twente uitgevoerd is[1]. De bedoeling is om aan de hand van echt onderzoek naar de opbrengsten van wiskunde-onderwijs in Nederland enige problemen te illustreren. De nadruk ligt daarbij op het aspect van de dekking van het curriculum terwijl ook ingegaan wordt op transversale aspecten. Een longitudinale opzet was in het Nederlandse aandeel aan dit project niet aan de orde. Het Tweede Wiskunde Project richt zich op de populatie van alle leerlingen uit de tweede klassen van HAVO-VWO, MAVO, LTO en LHNO. De zeer grote bereidheid om aan het onderzoek mee te werken resulteerde in een volledig gerealiseerde steekproef van 236 scholen met 5500 leerlingen. Wat betreft de allocatie van instrumenten was sprake van multiple-matrix sampling. Voor meer gedetailleerde informatie verwijzen we naar de verschenen voortgangsrapporten en overige geplande rapportages (Eggen, 1979, 1980a, 1980b; Eggen & Pelgrum, 1981; Pelgrum, 1981, 1982).

3.1 *Dekking van het curriculum*

In het project werd allereerst internationaal geïnventariseerd welke curriculumelementen van belang zouden kunnen zijn. Vervolgens werden door middel van beoordelingen in alle deelnemende landen elementen geïdentificeerd die van voldoende communaal belang waren. Voor Nederland werd voor deze beoordelingen gebruik gemaakt van een tweetal panels van deskundigen: een nationale begeleidingscommissie (met deskundigen in wiskunde-onderwijs, inspectie, Vereniging van wiskundeleraren, CITO, wiskunde-vakdidactici) en een commissie van wiskundedocenten uit de diverse schooltypen. Voor de uiteindelijk gekozen elementen werden cognitieve toetsen geconstrueerd. Daarnaast werden instrumenten gemaakt voor attitudemeting, achtergrondsmeting en de bepaling van de relatieve nadruk die curriculumelementen in de scholen krijgen.

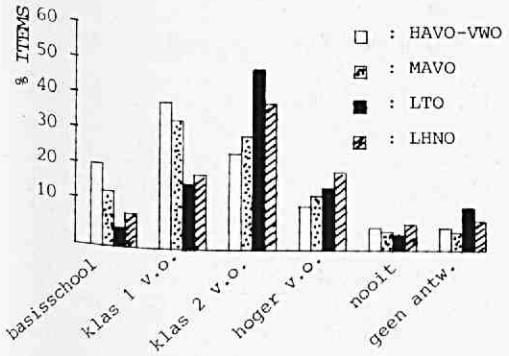
3.2 *Feitelijke wiskunde-curriculum*

Hoewel ministeriële richtlijnen tot op zekere hoogte voorschrijven hoeveel uren wiskunde per week gegeven dient te worden, valt op dat in de onderwijspraktijk vrij veel variatie bestaat (zie Figuur 1). Vooral op het LTO is de spreiding opvallend groot.

Uit de relatieve besteding van uren over wiskunde-onderdelen blijkt dat er vrij grote verschillen bestaan tussen het algemeen vormend onderwijs (AVO) en het lager beroepsonderwijs (LBO): bijvoorbeeld formules en vergelijkingen komen op het AVO relatief veel voor. Op het LBO worden relatief meer dan op het AVO onderwerpen als procenten en statistiek behandeld.

3.3 *Leerstofaanbod*

Iedere docent in het onderzoek heeft per toetsitem aangegeven in welke periode de voor de beantwoording van het toetsitem noodzakelijke leerstof onderwezen wordt. De beoordelingen werden kenbaar gemaakt op een schaal met de categorieën 1 basisschool, 2 klas 1 V.O.,



Figuur 2 *Verdeling van de in de items gevraagde stof over de schoolfasen volgens beoordeling van de 236 docenten*

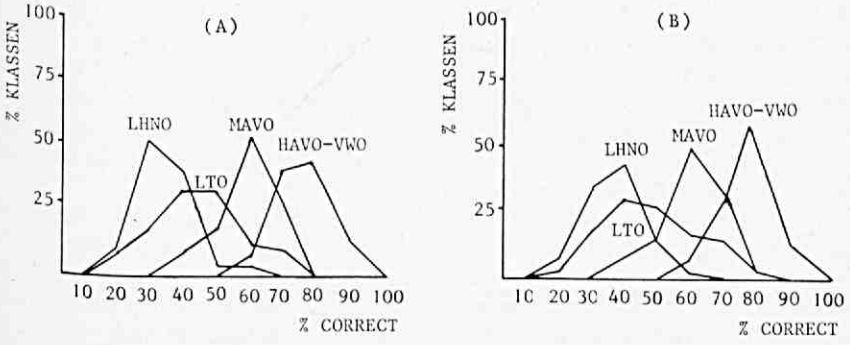
3 klas 2 V.O., 4 hoger V.O., en 5 nooit. Het percentage beoordelingen per categorie, gemiddeld over alle items, werd berekend. In Figuur 2 is de verdeling van deze gemiddelde percentages weergegeven. Uitgesplitst naar schooltype geven deze gemiddelden aan dat de leerstof uit de toetsitems volgens de oordelen van docenten eerder wordt aangeboden op HAVO-VWO en MAVO dan op het LTO en LHNO.

Uit de tot nu toe gepresenteerde gegevens blijkt dat zich in het uitgevoerde wiskundecurriculum aanzienlijke verschillen tussen schooltypen voordoen. Hiermee is een mogelijke verklaring gegeven voor verschillen in prestaties tussen scholen die gevonden werden.

3.4 *Kennis van wiskunde*

In Figuur 3 is de wiskunde-kennis van leerlingen op klasse-niveau per schooltype weergegeven (uitgaande van de 40 items die aan alle leerlingen zijn voorgelegd). Uit de figuur blijkt dat grote verschillen in centrale tendentie tussen de schooltypen bestaan en dat vooral in het LTO van een vrij grote spreiding sprake is. Verwacht zou kunnen worden dat deze resultaten verwijzen naar verschillen in het aangeboden curriculum of anders gezegd: verschillen in de mate waarin leerlingen de gelegenheid hebben gehad om de stof behorende bij de items te leren. Figuur 3(A) nodigt uit tot een wellicht als oneerlijk te kwalificeren vergelijking tussen scholen (of klassen) en/of schooltypen. Om deze reden zou een schatting van de wiskunde-kennis van leerlingen beter gebaseerd kunnen worden op die items waarvan naar het oordeel van de docent de betreffende leerstof in klas 1 of klas 2 V.O. onderwezen is. Figuur 3(B) geeft een overzicht van deze berekening (het minimum aantal items waarop deze berekening gebaseerd is bedraagt 10). Opvallend is dat het weinig uitmaakt of de oordelen van docenten over het al of niet onderwezen zijn van leerstof verdisconteerd worden bij het berekenen van de totale toetsscore. De gegevens uit Figuur 3(B) zouden kunnen wijzen op het bestaan van soms deplorabele opbrengsten van het wiskundeonderwijs. Deze interpretatie is vooralsnog evenwel aanvechtbaar, daar onder andere niet uit te sluiten is dat de oordelen van docenten over het al dan niet onderwezen zijn van leerstof invalide en/of onbetrouwbaar kunnen zijn.

De mate waarin de leerstof uit de toetsen naar het oordeel van de docenten onderwezen is, hangt nauwelijks samen met de prestatie van leerlingen (r=.22). Interessant is echter dat docenten wel degelijk inzicht hebben in de vraag of leerlingen de betreffende leerstof beheersen (zie Figuur 4). Uit de figuur blijkt immers dat er



Figuur 3 *Relatieve verdelingen van het percentage correcte items (40) per klas voor 4 schooltypen, respectievelijk niet (A) en wel (B) gecorrigeerd voor de leerstof die volgens de docenten niet onderwezen is*

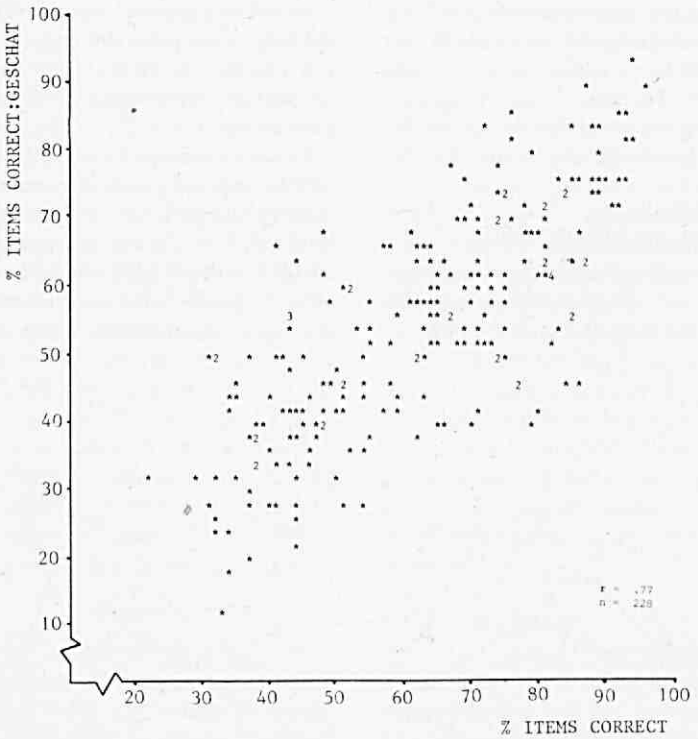
globaal gezien sprake is van samenhang. De spreiding per kolom in de figuur is echter nog steeds vrij aanzienlijk. Dit komt overeen met het feit dat ons elders gebleken is dat leraren amper in staat zijn moeilijkheden van items te schatten (Van der Linden, 1982b). Dat er niettemin een globale samenhang is zou mogelijk kunnen liggen aan de grote verschillen tussen schooltypen.

Verder bleek het verband tussen oordelen over de gelegenheid om te leren en de moeilijkheidsschatting van items door docenten zwak te zijn. Uit deze gegevens kan geconcludeerd worden, dat onder aanname van voldoende validiteit van het gebruikte instrument om de gelegenheid om te leren te meten, docenten redelijk goed in staat zijn om te beoordelen welke leerstof aangeboden wordt en in staat zijn leerstofdomeinen te signaleren waarop zich leerachterstanden kunnen manifesteren. Door middel van deze informatie kunnen maatregelen ter optimalisering van het onderwijs onderbouwd worden.

4 *Nieuwe methodologische ontwikkelingen*

Uit het Tweede Wiskunde Project is gebleken dat het mogelijk is om de opbrengsten van een schoolvak voor een zekere populatie te peilen. Ondanks aanzienlijke verschillen tussen schooltypen en curricula bleek het mogelijk om een instrument te ontwikkelen dat een redelijke dekking van de betrokken curricula gaf en ook nog binnen een beperkte tijd bij schoolklassen afgenomen kon worden. Dit lag niet in de laatste plaats aan de toepassing van multiple-matrix sampling.

Elders zullen meer gedetailleerde kwantitatieve overzichten van de opbrengsten gepubliceerd worden. Hoewel daarmee de mogelijkheid van dergelijke overzichten aangetoond is, moet bedacht worden dat deze opgesteld zijn met het oog op internationale vergelijking. Dit betekent dat het gaat om opbrengstmaten met een hoog aggregatieniveau. Zou het project van het begin af aan niet als aandeel in een internationaal vergelijkend onderzoek maar als



Figuur 4 *Strooidiagram van het gemiddelde door docenten geschatte percentage correct versus het percentage correct beantwoorde items per klas*

'national assessment study' in een beleidsonderzoek ingebed zijn geweest, dan zou dit niveau veel te hoog zijn, want aggregatie van gegevens over totale itemverzamelingen middelt verschillen uit. Assessment-onderzoek is pas beleidsrelevant als deze verschillen zichtbaar worden doordat de resultaten naar curriculumelementen en onderwijsniveaus (scholen, regio's, sectoren) uitgesplitst kunnen worden. Dit maakt het noodzakelijk dat resultaten op itemniveau voor individuele scholen aanwezig zijn en dat deze vervolgens tot ieder gewenst niveau geaggregeerd kunnen worden.

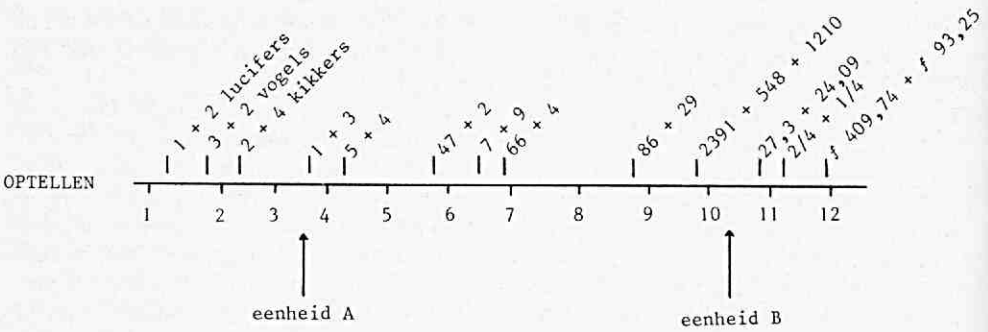
De voorafgaande alinea geeft een verschil aan tussen survey- en assessment-onderzoek. Eerder werd reeds een aantal andere kenmerken van goed assessment-onderzoek genoemd. We vatten de belangrijkste hier nog eens samen:

1. De resultaten moeten op ieder door het beleid gewenst niveau geaggregeerd en uitgesplitst kunnen worden. Deze niveaus lopen in principe van individuele scholen tot het nationale niveau.
2. Er moet een volledige dekking van het curriculum bereikt worden, waarbij de afnametijd van de instrumenten evenwel kort dient te zijn.
3. De resultaten moeten absoluut en niet alleen relatief geïnterpreteerd kunnen worden, wat het vrijgeven van items tot gevolg heeft.
4. Curricula kunnen zich over de jaren heen wijzigen en de opzet van assessment-onderzoek en de samenstelling van de instrumenten dienen zich hierbij aan te passen.
5. Om longitudinale vergelijking mogelijk te maken, moeten de opeenvolgende afnamen equivalente of equivaleerbare metingen opleveren. Omdat de items vrijgegeven moeten worden, betekent dit dat er apart geëquivaleerd moet worden. De noodzaak van absolute interpretatie van de scores brengt de noodzaak van equivalering op itemniveau met zich mee.

Eerder werd op de mogelijkheid van multiple-matrix sampling gewezen. Multiple-matrix sampling is echter geen antwoord op ieder van de hierboven samengevatte wensen. Zo is het bijvoorbeeld niet goed mogelijk om bij multiple-matrix sampling rekening te houden met curriculumwijzigingen. Iedere wijziging van het curriculum is in feite een wijziging van het itemdomein, en scores op instrumenten die uit verschillende domeinen getrokken zijn, zijn onvergelijkbaar.

Er zijn nieuwe ontwikkelingen binnen de psychometrie die aan het laatste probleem tegemoet komen doordat ze geen steekproeftrekking van items veronderstellen en toch een zeer inzichtelijke, absolute interpretatie van toetsscores mogelijk maken. Bij deze modellen kunnen items in principe aan domeinen toegevoegd of eruit verwijderd worden, zonder dat dit de interpretatie van scores aantast. Bovendien zijn deze modellen zeer geschikt voor gebruik in longitudinale studies doordat prestaties op instrumenten die uit verschillende items uit het domein samengesteld zijn in scores op dezelfde schaal vertaald kunnen worden. We doelen hier op modellen uit de zogenaamde itemresponstheorie. Centraal in deze modellen staat de kans dat een persoon op een gegeven item een goede respons produceert. Het voert hier te ver om aan te geven hoe deze modellen precies in elkaar zitten en wat door het gebruik van deze modellen allemaal aan toepassingen mogelijk is. De geïnteresseerde lezer raadplege bijvoorbeeld Lord (1980). Voor ons doel is het hier van belang om slechts op te merken dat itemresponsmodellen de kans op een succesvol antwoord op een item als een functie van de te meten variabele beschouwen. Dit opent de mogelijkheid om de betekenis van een score op deze variabele te interpreteren met behulp van de inhouden van de items. Figuur 5 geeft een voorbeeld van een gedragscontinuum dat dan ontstaat. Dit fictieve voorbeeld is geïnspireerd door de wijze van rapporteren die behoort bij een diagnostische rekentoets die bekend staat onder de naam Keymath (1976). Dit instrument, dat bestemd is voor de eerste klassen van het lager onderwijs, heeft niet alleen een schaal voor optelling zoals in Figuur 5 maar ook nog 13 andere rekenvaardigheidsschalen. Met behulp van de rekentoets worden leerlingen op het continuum geplaatst en door hun positie op het continuum is duidelijk welke van de kort omschreven vaardigheden zij beheersen. In Figuur 5 is dit voor 2 leerlingen (eenheden) aangegeven.

Itemresponsmodellen laten niet alleen dergelijke, voor het onderwijsveld zeer inzichtelijke interpretaties van toetsscores toe maar zijn eveneens geschikt voor equivalering van toetsscores in longitudinale studies. Het is mo-



Figuur 5 *Voorbeeld van een gedragscontinuum voor rekenen (optellen)*

gelijk om – zowel op hetzelfde als op opeenvolgende tijdstippen – verschillende leerlingen verschillende items behorende bij zo'n gedragscontinuum als in Figuur 5 te geven en alle leerlingen toch op dit continuum te plaatsen. Hiervoor is geen steekproeftrekking van items nodig en kunnen items zelfs bewust zo gekozen worden dat dit voor groepen van wisselend niveau zo nauwkeurig mogelijk gebeurt.

Met de ontwikkeling van bovenstaande modellen zijn we er nog niet. Wat deze namelijk nog niet toelaten is aggregatie van individuele gegevens tot op voor het beleid relevante niveaus. Ook hier zijn echter belangwekkende recente ontwikkelingen die voortgekomen zijn uit het California Assessment Project (zie bijvoorbeeld Bock, Mislevy & Woodson, 1982; Mislevy, Reiser & Zimowski, 1983; Reiser, 1980, 1982). Deze auteurs hebben speciaal voor assessment-onderzoek versies van itemresponsmodellen ontwikkeld die op groepsniveau toegepast kunnen worden en in feite een combinatie van itemresponsmodellen en multiple-matrix sampling technieken zijn. Deze modellen plaatsen groepen van ieder gewenst niveau rechtstreeks op gedragscontinua als in Figuur 5. Scores kunnen daarbij naar believen geaggregeerd worden. De eenheden A en B in Figuur 5 zouden dus net zo goed scholen of steden als leerlingen kunnen zijn. Bovendien impliceert de afname van items volgens een multiple-matrix opzet dat uiterst efficiënt met de tijd van leerlingen omgesprongen wordt. Technische details kunnen in de bovengenoemde referenties gevonden worden. Wij geven hier alleen een voorbeeld uit het Californische project dat afkomstig is uit Mislevy, Reiser en Zimowski (1983). In een deel van het project werden schoolprestaties in 61 curriculumelementen gemeten. Dit betekent dat scholen maar liefst op 61 verschillende gedragscontinua van het type in Figuur 5 geplaatst konden worden, waardoor voor iedere school een uiterst waardevol profiel ontstond, dat inzicht in de eigen zwakke en sterke punten gaf. Niettemin vroeg dit per leerling minder dan een uur toetstijd. Deze efficiëntie werd bereikt doordat de curriculumelementen over 30 toetsversies van ieder 34 items werden verdeeld. Iedere versie behoefde niet meer dan 1 item per element te bevatten. Iedere aselect getrokken leerling werd 1 aselect getrokken toetsversie voorgelegd. Hoewel het model op schoolniveau werd gedefinieerd konden de resultaten vervolgens op hogere niveaus geaggregeerd worden (schooldistricten, steden, counties en de gehele staat California). Prestaties op ieder van deze niveaus konden dus met behulp van 61 gedragscontinua als in Figuur 5 beschreven en vergeleken worden. Het komt ons voor dat dit een uiterst efficiënte wijze van verzamelen van precies die informatie is waar men bij beleidsgericht assessment-onderzoek belang in stelt.

5 *Conclusie*

Praktische ervaringen met onderzoek als in het Tweede Wiskunde Project wijzen erop dat assessment-onderzoek in Nederland organisatorisch haalbaar is. Als dergelijke ervaringen verbonden worden met recente methodologische vorderingen ontstaat een type assessment-onderzoek dat zowel op efficiënte als informatieve wijze aan de beheersing van de kwaliteit van het onderwijs bij kan dragen.

Noten

1. SVO-project 0.452. Projectleiding Drs. W. J. Pelgrum.

Literatuur

Albinski, M., *Survey research*. Utrecht/Antwerpen: Het Spectrum, 1967.

Bock, R. D., R. J. Mislevy, An item response curve model for matrix-sampling data: The California grade-three assessment. *New Directions for Testing and Measurement*, 1981, *10*, 65-90.

Bock, R. D., R. J. Mislevy & C. Woodson, The next stage in educational assessment. *Educational Researcher*, 1982, *11*, 4-11.

Eggen, Th., *IEA Tweede Wiskunde Project Voortgangsrapport no. 1*, Enschede: T.H.-Twente, Onderafdeling Toegepaste Onderwijskunde, 1979.

Eggen, Th., *IEA Tweede Wiskunde Project Voortgangsrapport no. 2*, Enschede: T.H.-Twente, Onderafdeling Toegepaste Onderwijskunde, 1980a.

Eggen, Th., *IEA Tweede Wiskunde Project Voortgangsrapport no. 3*, Enschede: T.H.-Twente, Onderafdeling Toegepaste Onderwijskunde, 1980b.

Eggen, Th. & W. J. Pelgrum, *IEA Tweede Wiskunde Project Voortgangsrapport no. 4*, Enschede: T.H.-Twente, Onderafdeling Toegepaste Onderwijskunde, 1981.

Keymath diagnostic arithmetic test. Circle Pines, Minn.: American Guidance Service, 1976.

Kwaliteit van het onderwijs. 's-Gravenhage: Staatsuitgeverij, 1981.

Lagerweij, N. A. J., Kwaliteitsbepaling van onderwijs: een noodzakelijke nota. *Pedagogisch Tijdschrift / Forum voor Opvoedkunde*, 1982, *7*, 77-80.

Lapointe, A. E. & S. L. Koffler, Your standards or mine? The case for the National Assessment of Educational Progress. *Educational Researcher*, 1982, *11*, 4-9.

Linden, W. J. van der, Het klassieke testmodel, latente trekmodellen en evaluatie-onderzoek. *(VOR publikatie nr. 7)* Amsterdam: Vereniging voor Onderwijsresearch, 1978.

Linden, W. J. van der, Criterion-referenced measurement: Its main applications, problems, and findings. In: W. J. van der Linden (ed.), Aspects of criterion-references measurement, *Evaluation in Education: An international Review series*, 1982a, *5*, 97-118.

Linden, W. J. van der, A latent trait method for determining intrajudge inconsistency in the Angoff and Nedelsky techniques of standard setting. *Journal of Educational Measurement*, 1982, *19*, 295-308 (b).

Lord, F. M., Estimating norms by item sampling. *Educational and Psychological Measurement*, 1962, *22*, 259-267.

Lord, F. M., *Application of item response theory to practical testing problems*. Hillsdale, New Jersey: Erlbaum, 1980.

Mislevy, R. J., M. R. Reisner & M. Zimowsky, Scale-score reporting of national assessment data. *Science Education*, 1983, (in druk).

Pandey, T. J. & D. Carlson, Assessing pay-offs in the estimation of the mean using multiple matrix sampling designs. In: D. N. M. de Gruijter & L. J. T. van der Kamp (Eds), *Advances in Educational and Psychological Measurement*. London: John Wiley & Sons, 1976.

Pelgrum, W. J., *IEA Tweede Wiskunde Project Voortgangsrapport no. 5*, Enschede: T. H.-Twente, Onderafdeling Toegepaste Onderwijskunde, 1981.

Pelgrum, W. J., *IEA Tweede Wiskunde Project Voortgangsrapport no. 6*, Enschede: T.H.-Twente, Onderafdeling Toegepaste Onderwijskunde, 1982.

Popham, W. J., Domain specification strategies. In: R. A. Berk (Ed.), *Criterion-referenced measurement: The state of the art*. Baltimore, MD: The Johns Hopkins University Press, 1980.

Reiser, M., *A latent trait model for group effects*. Unpublished doctoral dissertation, University of Chicago, 1980.

Reiser, M., *An item response model for the estimation of group fixed effects*. Paper presented at the 1982 Annual Meeting of the American Educational Research Association, New York, March 19-23, 1982.

Sandbergen, S., Wat hebben ze nou geleerd op school? Verslag van een oriëntatiereis over assessment. 's-Gravenhage: Ministerie van Onderwijs en Wetenschappen, BO/VB, 1979. (2 delen).

Sirotnik, K. A. Introduction to matrix sampling for the practitioner. In: W. J. Popham (Ed.), *Evaluation in Education*. Berkeley, Calfornia: McCutchan, 1974,

Wijnstra, J. M., Wat leren kinderen op school? *Pedagogische Studiën*, 1982, *59*, 223-232, 277-287.

Curricula vitae

W.J. van der Linden (1948) legde doctoraalexamens af in de psychologie en de sociologie (cum laude). Promoveerde aan de Universiteit van Amsterdam op het proefschrift *Psychometric contributions to the analysis of criterion-referenced measurements* (cum laude). Was van 1973-1977 als wetenschappelijk medewerker verbonden aan de vakgroep Psychometrie, Statistiek en Modelvorming van de Subfaculteit Psychologie, Rijksuniversiteit Utrecht. Sinds 1977 in dienst van de Onderafdeling Toegepaste Onderwijskunde, T.H. Twente, waar hij thans als hoogleraar verbonden is aan de vakgroep Onderwijskundige Meetmethoden en Data-analyse. Publiceerde in diverse binnen- en buitenlandse tijdschriften en is Ad-

visory Editor van *Applied Psychological Measurement* en *Journal of Educational Measurement*.

W. J. Pelgrum (1950) studeerde tussen 1970 en 1977 psychologie aan de Rijksuniversiteit te Groningen (specialisatie psycholinguïstiek). Sinds 1978 is hij als onderzoeker verbonden aan de Onderafdeling Toegepaste Onderwijskunde van de T.H. Twente. Sinds 1981 houdt hij zich als projectleider bezig met het uitvoeren van het Nederlandse aandeel aan de *Second International Mathematics Study* en de *Second International Science Study*. Beide projecten worden uitgevoerd in het kader van de *International Association for the Evaluation of Educational Achievement.*

Adres: Technische Hogeschool Twente, Onderafdeling der Toegepaste Onderwijskunde, Postbus 217, 7500 AE Enschede

Manuscript aanvaard 25-10-'83