Analfabetisme in Nederland?

Commentaar op de Voorstudie Periodieke Peiling van het Onderwijsniveau (PPON)

Inleiding

Een van de doelstellingen van de Voorstudie Periodieke Peiling van het Onderwijsniveau, die in de periode juli 1983 tot mei 1985 onder leiding van H. Wesdorp door de SCO (Stichting Centrum voor Onderwijsonderzoek van de Universiteit van Amsterdam) werd uitgevoerd, is het publiceren en doen becommentariëren van een eerste rapport waarin de resultaten van een deelpeiling van het niveau van het basisonderwijs worden gepresenteerd. Publikatie heeft inmiddels plaatsgevonden in vijf bij de SCO uitgegeven deelrapporten:

I: Lees- en schrijfprestaties van zesde klassers
II: Spreek- en luisterprestaties van zesde klassers
III: Het onderwijsaanbod voor moedertaal in de zesde klas
IV: Technische verantwoording
V: Lees- en schrijfprestaties van allochtone leerlingen in de zesde klas

en een door SVO/SCO/Ministerie van O. en W. uitgegeven samenvatting 'Goed onderwijs, wat is dat? Voorstudie Periodieke Peiling van het Onderwijsniveau'. De nu volgende kritische kanttekeningen bij vier van de vijf deelrapporten (deel III kon door de SCO nog niet ter beschikking worden gesteld) zijn bedoeld als een bijdrage aan de gewenste becommentariëring.

Het onderzoek wordt door Wesdorp en zijn medewerkers nadrukkelijk gepresenteerd als een proefonderzoek of 'haalbaarheidsstudie' voor het pas later te starten en periodiek te herhalen eigenlijke peilingsonderzoek. Het rapport besteedt overigens slechts enkele passages aan de haalbaarheid van de gevolgde procedure in toekomstige peilingen (IV, p. 159-163). Men zou bijvoorbeeld enige argumentatie verwachten voor de conclusie dat de bereidheid tot deelname redelijk groot was, terwijl toch niet meer dan 48% van de aangeschreven scholen heeft meegedaan en aanzienlijke percentages ontbrekende scores worden gerapporteerd. Een grondige analyse van de voor- en nadelen van alternatieve procedures en een kritische verwerking van eerdere ervaringen met peilingsonderzoek treft men in het rapport evenmin aan. Zoals de auteurs zelf aangeven, ontbrak daartoe de tijd: 'Daar de voor de Voorstudie PPON gebruikte toetsinstrumenten in betrekkelijk korte tijd moesten worden geconstrueerd en beproefd – in totaal zijn daaraan ongeveer zes maanden besteed – zijn lang niet alle verworvenheden uit de drie buitenlandse peilingsvoorbeelden in onze instrumenten doorgedrongen. Het verdient daarom aanbeveling in toekomstige peilingsonderzoeken meer tijd te reserveren voor de ontwikkeling van de toetsplannen en de concrete opgaven, inclusief de scorings- en analysevoorschriften. Er is nog veel te leren uit het Amerikaanse, Engelse en Canadese materiaal' (IV, p. 18).

Wat overblijft is de constructie en afname bij zesde-klassers van een aantal taalbeheersingsinstrumenten (3 leesinstrumenten bij telkens ±715 leerlingen, 2 schrijfinstrumenten bij telkens ±715 leerlingen, 1 spreekinstrument bij 199 leerlingen, 2 luisterinstrumenten bij resp. 188 en 651 leerlingen) en een taalattituden- en gebruikenvragenlijst (bij 2109 leerlingen). Daarnaast werden 2 vragenlijsten voor het taalonderwijsaanbod geconstrueerd en afgenomen bij telkens ±470 scholen. Bovendien werden bij ±75 scholen met allochtone leerlingen en bij 254 allochtone leerlingen een deel van de beheersingsinstrumenten en de vragenlijsten afgenomen. De uitvalpercentages zijn vaak aanzienlijk, soms zonder duidelijke redenen (bijv. bij de allochtonen). Wanneer overtuigend kan worden aangetoond dat bestaande instrumenten niet voldoen, is de constructie van goede peilingsinstrumenten uiteraard nuttig. Door tijdgebrek is echter weinig aandacht besteed aan de psychometrische kwali-

teiten van de geconstrueerde instrumenten. Voor de totaal- en combinatiescores waarop de belangrijkste conclusies met betrekking tot de taalbeheersing worden gebaseerd, zoekt men vaak tevergeefs naar interne consistentie-, betrouwbaarheids- en validiteitsmaten. Op de noodzaak tot perfectionering van de geconstrueerde functionele taaltoetsen en de ontbrekende validering wijzen de auteurs zelf (IV, p. 164 en p. 170-171).

Het rapport is grotendeels gewijd aan de resultaten van de taalbeheersingsmetingen, veelal in de vorm van gedetailleerde scoreverdelingen, en aan enkele verstrekkende conclusies die daaruit worden getrokken. In het bijzonder menen de auteurs te moeten concluderen dat 7% van de leerlingen wat hun leesvaardigheid betreft 'met recht "functionele analfabeten" genoemd kunnen worden' en dat van 9% der leerlingen 'de schrijfvaardigheid zo gering (is), dat ze in de meeste functionele schrijfsituaties hulpeloos zullen zijn' (Samenvatting, p. 23 en p. 50). Deze conclusies hebben uitgebreid aandacht gekregen in de landelijke dagbladpers. In NRC/Handelsblad van 24 oktober 1985 aarzelden de auteurs niet om de in hun ogen zorgelijke cijfers in verband te brengen met het zeer negatief beoordeelde moedertaalonderwijs in ons land. Op 28 november 1985 liet de minister van O. en W. in de Tweede Kamer het bijvoeglijk naamwoord 'functioneel' weg en was desondanks niet onder de indruk van het percentage analfabeten: 'In de VS is 16% van de 17-jarigen analfabeet. In Nederland 7% van de 13-jarigen. Wij moeten voorzichtig zijn met het trekken van conclusies'.

Voorafgaand aan een meer gedetailleerde evaluatie van de bijdrage die het onderzoek heeft geleverd aan toekomstig peilingsonderzoek, zal nu eerst de betekenis van de zo centraal gestelde percentages functionele analfabeten worden becommentarieerd.

x% functionele analfabeten

De nadruk op de gevonden percentages functionele analfabeten plaatst het onderzoek in de rij van sociaal-wetenschappelijke onderzoeken die in exacte percentages weten aan te geven welk gedeelte van de Nederlandse populatie aan een bepaald zorgwekkend verschijnsel lijdt: psychische stoornissen, zelfmoordneigingen, verslavingen, relatieproblemen, problemen op seksueel gebied enz. Om tot de dichotome uitspraak te komen dat respectievelijk x% wel en (100-x)% niet het zorgelijke verschijnsel vertoont, maken deze onderzoeken gebruik van sociaal-wetenschappelijke meetinstrumenten die bijna zonder uitzondering op beoordeling zijn gebaseerd, onbetrouwbaar zijn, en zelden een dwingende dichotomie opleveren. Wat empirisch resulteert is een scoreverdeling van hoog naar laag en niet de dichotomie met de daarbij gehanteerde terminologie. Deze wordt achteraf door de onderzoeker gekozen. Zoals dadelijk zal worden verduidelijkt, is ook de karakterisering van de laagste 7% in de leesscoreverdeling als functionele analfabeten een eigen toevoeging van de onderzoekers aan het empirisch gevonden materiaal.

Zien we af van de empirische pretenties, dan is zo'n dichotome uitspraak nauwelijks nog interessant. Ze degradeert tot het kaliber van: 'Vijftig procent van de bevolking is klein van stuk en tien procent is extreem klein'. De percentages krijgen pas betekenis in de zin van veel of weinig bij vergelijking met voorgaande tijdstippen in een tijdreeks (periodieke peiling van de prestaties aan de hand van hetzelfde meetinstrument) of bij internationale vergelijking. De periodieke peiling moet echter nog starten, zodat het gevonden percentage extreem laag zou kunnen zijn en met behulp van hetzelfde meetinstrument in de toekomst niet meer gehaald wordt. Ook internationale vergelijking levert op dit moment nog geen houvast omdat in landen waar peilingsonderzoek wordt gedaan (de VS, Canada en Engeland) andere instrumenten worden gebruikt en niet wordt gerapporteerd in termen van dezelfde analfabetisme-dichotomie.

In feite zijn de onderzoekers zich van het vrijblijvende karakter van hun dichotomieën terdege bewust. Zij twijfelen in die mate aan de mogelijkheid om in peilingsrapportages de grens tussen een voldoende en een onvoldoende prestatie te beschrijven, dat zij nader onderzoek hiernaar noodzakelijk achten (IV, p. 183). Jammer is, dat hun zorg over het lage taalvaardigheidsniveau van de leerlingen sterker in de publiciteit is doorgedrongen dan hun twijfel aan de mogelijkheid hierover in het uitgevoerde peilingsonderzoek uitspraken te doen.

Hun voorstel om in de toekomst een meer op de mening van bredere groepen berustende grens tussen onvoldoende en voldoende te bepalen lijkt op het eerste gezicht uitvoerbaar door deze groepen (bijv. ouders, leerkrachten en specialisten) te vragen naar de vaardigheden die in diverse taalsituaties beheerst zouden moeten worden. Dit is in feite gebeurd bij een groep van 349 respondenten (overigens een respons van 35% op de in totaal 1008 verzonden vragenlijsten) en leverde bijvoorbeeld op dat 90% van de respondenten het lezen van een telefoongids wenselijk en 94% haalbaar achtte voor de leerlingen (IV, p. 44). Hoewel een dergelijk onderzoek een belangrijke bijdrage kan leveren aan de inhoudsvaliditeit van de te construeren instrumenten, is de betekenis voor de normstellingsproblematiek beperkt. Enige ervaring met de constructie van schoolprestatietoetsen leert dat eenzelfde te toetsen kennisitem in heel verschillende vragen kan worden omgezet met even zoveel verschillende moeilijkheidsgraden. De auteurs schijnen zich dat onvoldoende te realiseren. Zij concluderen bijvoorbeeld niet dat hun vragen betreffende het raadplegen van een telefoonboek wellicht wat te moeilijk zijn uitgevallen voor de onderzochte zesde-klassers maar menen met zekerheid te kunnen stellen dat het hanteren van het telefoonboek boven het vermogen van zesde-klassers ligt (IV, p. 73-74). De moeilijkheidsgraad van een vraag of toets wordt in feite door talloze factoren beïnvloed, is uiterst lastig a priori in te schatten en is zeker niet alleen een functie van het aan de orde gestelde kennisitem. Als gevolg hiervan is ook de grens tussen voldoende en onvoldoende niet a priori afleidbaar uit de wenselijk of noodzakelijk geachte vaardigheden.

In ieder geval is de wijze waarop de auteurs de normstellingsproblematiek op a priori basis proberen op te lossen weinig overtuigend. Met betrekking tot de leesvaardigheid gaan zij bijvoorbeeld als volgt te werk. Voor ieder van de drie toetsboekjes (De Natuurclub 1, De Natuurclub 2, Schoolreis naar Waddenoog) die aan verschillende groepen van leerlingen zijn voorgelegd, vermelden zij de situatiebepaalde deelvaardigheden die daarin worden getoetst en de aantallen items per deelvaardigheid (I, p. 30 en p. 95). Daarbij valt op dat de toetsboekjes aanmerkelijk verschillen wat de opgenomen deelvaardigheden en de aantallen items per deelvaardigheid betreft. Deelvaardigheid 7 (telefoongids lezen en raadplegen) is bijvoorbeeld uitsluitend in toetsboekje A opgenomen, terwijl de enige overal opgenomen deelvaardigheid 3 (leerboeken lezen) respectievelijk 9, 8 en 6 items bevat. De zorgvuldig voorbereide inhoudsvaliditeit wordt op die manier bij de samenstelling van de toetsboekjes weer teniet gedaan, zodat het begrip 'functionele analfabeet' bij ieder van de toetsboekjes een verschillende inhoud krijgt (met name bij B dat bijna geen overlap vertoont met A en C). Omdat geen correlaties tussen de deelvaardigheden worden vermeld, is moeilijk in te schatten wat het effect is van het weglaten van deelvaardigheden.

Vervolgens geven de auteurs in tabelvorm per toetsboekje en deelvaardigheid de scores die naar hun mening indicatief zijn voor 'minimaal lezen'. De keuzen in deze belangrijkste stap bij de bepaling van functioneel analfabetisme worden niet beargumenteerd. In de tekst wordt slechts opgemerkt: 'Daarbij hebben wij uiterst lage scores nog als bewijs van 'kunnen lezen' geïnterpreteerd. Anders gezegd: onze indicaties voor 'minimale lezers' zijn... uiterst voorzichtig' (I, p. 98). Niet duidelijk is onder meer waarom deelvaardigheden A7 en A8 met respectievelijk 6 en 9 items dezelfde 'minimale' scores 0 en 1 krijgen toegewezen en waarom bij B9 met 6 items ook 3 nog 'minimaal' is.

In de laatste stap worden ten slotte die leerlingen functionele analfabeten genoemd die minstens op 4 van de 6 deelvaardigheden (toetsboekje A) of 3 van de 5 (toetsboekje B en C) minimale *of* ontbrekende scores hebben. In plaats van de testtheoretisch eenvoudig te hanteren somscore over de deelvaardigheden hanteert men bij het toewijzen van een leerling aan de functionele analfabeten of functionele alfabeten impliciet een tamelijk ingewikkelde combinatie-score. De psychometrische karakteristieken hiervan, in het bijzonder de betrouwbaarheid, zijn moeilijk te bepalen en in ieder geval niet gelijk aan die van de somscore. Opnieuw geeft de tekst geen inhoudelijke argumentatie voor de gemaakte keuzen die niettemin een belangrijke invloed hebben op de resulterende percentages. De ingewikkelde combinatieprocedure blijkt in feite bedoeld om bij de berekening

van de percentages functionele analfabeten een oplossing te bieden voor de aanzienlijke percentages ontbrekende scores (12%, 4% en 4% voor de somscores van resp. toetsboekje A, B en C). Mede omdat alleen op somscore-niveau informatie wordt gegeven over de ontbrekende scores en niet voor deelscores en combinaties van deelscores, is het succes hiervan moeilijk vast te stellen. Gevreesd moet worden dat met name bij toetsboekje A, waar het percentage ontbrekende scores dat van de functionele analfabeten ver overtreft, nogal wat leerlingen vanwege 4 of meer ontbrekende deelscores bij de functionele analfabeten terecht zijn gekomen terwijl het in feite goede lezers kunnen zijn. Het onderzoeksrapport geeft hierover geen informatie. Wel wordt aangegeven dat uit niets blijkt dat de leerlingen met ontbrekende scores de betrokken opgaven hebben proberen te maken.

De betekenis van de gevonden percentages wordt ook nog sterk gerelativeerd door twee statistische overwegingen. In de eerste plaats hebben de auteurs veel moeite gedaan om de steekproefgrootte zodanig te kiezen, dat redelijk betrouwbare uitspraken over de percentages in de populatie mogelijk worden. Niettemin houden zij geen rekening met het berekende 95% betrouwbaarheidsinterval van x% ± 5.2% (IV, p. 103). Bij een steekproefwaarde van 7% zou het werkelijke percentage tussen 1.8% en 12.2% liggen. Dit impliceert een onzekerheid waarbij de afwijking zowel naar boven als naar beneden kan uitvallen. Volgens een tweede statistische overweging, samenhangend met de meetbetrouwbaarheid van de scores, is het ware percentage bij aanhouding van dezelfde alfabetisme-analfabetisme dichotomie met zekerheid kleiner dan 7%. Onbetrouwbaarheid leidt tot een grotere variantie var(x) van de geobserveerde scores x, bij benadering gelijk aan de som van de ware-score variantie var(w) en de meetfout-variantie var(e): var(x) = var(w) + var(e). De betrouwbaarheid r = var(w)/var(x) van de combinatiescores is niet precies bekend maar uitgaande van .85 en normaalverdelingen voor de geobserveerde en ware scores zou het percentage van 7% slinken tot 5.5% in de ware-score verdeling en het interval 1.8% - 12.2% tot 1.1% - 10.5%. Dus rekening houdend met de betrouwbaarheid zou het ware percentage functionele leesanalfabeten dicht bij de 5% liggen, terwijl tevens rekening houdend met de steekproeffluctuaties een waarde van bijna 1% allerminst uitgesloten is.

Al bij al moet worden geconcludeerd dat de wetenschappelijke betekenis van de in de publiciteit zo sterk benadrukte percentages functionele analfabeten uitermate gering is. Met deze conclusie wordt natuurlijk geen uitspraak gedaan over het taalbeheersings- en taalonderwijsniveau in Nederland. Allerlei reacties in de samenleving wijzen erop dat het onderwijs in Nederland verslechtert. Een wetenschappelijke evaluatie van het beheersings- en onderwijsniveau in Nederland kan echter beter worden uitgesteld tot meerdere peilingen voorhanden zijn. Periodiek peilingsonderzoek is daar juist voor bedoeld. De resultaten van één enkele peiling vormen een te zwakke basis.

Bijdrage aan toekomstig peilingsonderzoek

In de volgende bespreking van de bijdrage die de proefpeiling heeft geleverd aan toekomstig periodiek te herhalen peilingsonderzoek zal speciaal op het aspect van de vergelijkbaarheid van opeenvolgende peilingen worden gelet. Het rapport vermeldt terecht als een der belangrijkste taken van peilingsonderzoek de identificatie van prestatieverschuivingen in de loop der tijd (IV, p. 181). Deze taak stelt hoge eisen aan de planning. Vereist is allereerst dat een reeks van peilingen, bijvoorbeeld over een periode van tien jaar, als een geheel wordt opgezet. De haalbaarheid is dan ook allerminst bewezen doordat bij de proefpeiling in korte tijd vele instrumenten geconstrueerd konden worden voor 'moeilijk toetsbare' taalvaardigheidsaspecten (IV, p. 159). Waar het om gaat is de keuze van een beperkt aantal helder definieerbare vaardigheden waarvan is te verwachten dat ze ook in de toekomst door de meest betrokken groepen als belangrijk ervaren worden. Vervolgens moet gekozen worden voor simpele, weinig tijd vergende instrumenten waarbij geen enkele twijfel bestaat dat de scoringsprocedure over de tijd constant gehouden kan worden. Toekomstige peilingsonderzoeken zullen zich immers niet kunnen beperken tot uitsluitend de taalvaardigheden. Daarnaast maken ingewikkel-

de, moeilijk scoorbare instrumenten grote kans uit het meetinstrumentarium verwijderd te worden waardoor de vergelijkbaarheid in gevaar komt. Een belangrijk facet van de vergelijkbaarheid is ook dat het testmateriaal en de gekozen itemvoorbeelden in principe actueel moeten blijven over de hele peilingsperiode. Gevreesd moet worden dat bij de constructie van onderdelen zoals met betrekking tot Greenpeace de bruikbaarheid in toekomstige peilingen niet is gegarandeerd.

Vanuit de boven omschreven eisen kunnen vraagtekens worden geplaatst bij de beslissing om functionele taalvaardigheden te meten. Het begrip functionele taalvaardigheid krijgt slechts in enkele passages een summiere omschrijving. Het gaat om taalvaardigheden 'die als geheel in de alledaagse praktijk kunnen voorkomen' (IV, p. 33). In de overige formuleringen wordt het begrip afgezet tegen de traditionele instrumenten die de aandacht zouden concentreren op details, op subvaardigheden zoals het kunnen gebruiken van voorzetsels, het kunnen spellen van werkwoordsvormen. Aan het eind van deel IV vindt men de volgende verduidelijking: 'Zoveel mogelijk was voorkomen dat slechts schoolse "subvaardigheden" in de instrumenten zouden worden geoperationaliseerd' (IV, p. 163). Als enig argument voor de keuze van 'functionele taaltaken' wordt aangevoerd: 'Het leggen van de nadruk op deelvaardigheden is in zekere zin "schools", en kan bij een publiek van geïnteresseerde leken niet op veel belangstelling rekenen' (IV, p. 25).

Op zichzelf valt op het streven naar aansluiting bij functionele taalsituaties niets aan te merken. Twijfel ontstaat pas als dit zou moeten impliceren, dat schoolse (sub)vaardigheden bewust buiten de te construeren instrumenten worden gehouden. Hoe kan peilingsonderzoek van het onderwijs en het effect van onderwijs op de prestaties overtuigingskracht blijven houden bij geïnteresseerden, in het bijzonder bij leerkrachten, als niet mede aandacht wordt geschonken aan schoolse (sub)vaardigheden, al dan niet in functionele taalsituaties? Niettemin zijn er aanwijzingen dat de geconstrueerde instrumenten juist op dit punt tekorten vertonen. Zo bleek bij een aantal schrijftaken de telling van taalgebruiksfouten (grammaticale en idiomatische fouten) opvallend lage interbeoordelaars-betrouwbaarheden te bezitten (IV, p. 140). De nadruk op de functionaliteit van de taken, of liever van de beoordelingscriteria, kan ertoe leiden dat uiteindelijk meer intelligentie dan taal wordt gemeten. In de pers is herhaaldelijk bekritiseerd dat nogal wat taalkundig correcte schrijfprodukten toch als uitingen van functioneel analfabetisme worden aangemerkt in het onderzoek. Een voorbeeld is het volgende briefje van een kind aan haar ouders: 'Mam, ik ben even naar een lid van de club. Ze had iets in haar hoofd en daar zou ze over praten. Doei!' Dit kon niet door de beugel van de functionele criteria. Hierbij wreekt zich het gebrek aan valideringsonderzoek van de functionele taalvaardigheidsinstrumenten, waardoor niet duidelijk is of en in welke mate zij onderscheiden zijn zowel van intelligentie-maten als van de traditionele taalvaardigheidsinstrumenten.

Volgens de onderzoeksopdracht van de proefpeiling moesten de instrumenten voor de gegevensverzameling in verband met de beschikbare tijd zoveel mogelijk ontleend worden aan reeds eerder ontwikkelde instrumenten. Hieraan hebben de onderzoekers zich niet gehouden, omdat er naar hun zeggen voor productieve taalvaardigheid (schrijven en spreken) geen instrumentarium voorhanden was. In feite heeft de instrumentconstructie een overheersende plaats gekregen in het onderzoek. Onduidelijk is, waarom men heeft gemeend dat ook voor de receptieve taalvaardigheid (lezen en luisteren) nieuwe instrumenten ontwikkeld moesten worden. De thematische aanpak alleen kan hiervoor toch geen voldoende argument zijn. Met name voor lezen zijn er goede instrumenten op de markt die eventueel na geringe aanpassingen bruikbaar moeten zijn in peilingsonderzoek. Onduidelijk is ook, waarom voor ieder van de aspecten schrijven en lezen drie verschillende instrumenten (toetsboekjes A, B en C) werden ontwikkeld. Zoals eerder vermeld zijn de toetsboekjes niet als parallelvormen ontwikkeld zodat de resultaten niet met zekerheid vergelijkbaar zijn. De auteurs hadden er beter aan gedaan om, bijvoorbeeld via een uitgewogen procedure van domeinsampling, één instrument te construeren waarvan de inhoud representatief zou zijn geweest voor het geheel van deelvaardigheden in het domein. Door dit ene instrument

voor te leggen aan alle ±2100 leerlingen had men tevens nauwkeurigere schattingen van de populatiewaarden verkregen.

De constructie van de beide luistertoetsen is aan het CITO uitbesteed. Door de aanmerkelijke uitval, mede als gevolg van de gebrekkige organisatie bij de afname (IV, p. 158), en de geringe betrouwbaarheid van .61 voor beide luistertoetsen heeft het luistergedeelte van de proefpeiling niet aan de verwachtingen voldaan. Het is overigens onduidelijk waarom alleen over de resultaten van luistertoets I en niet over die van luistertoets II wordt gerapporteerd (II, p. 97).

Wat de overige instrumenten betreft valt op, dat van de in totaal 27 onderdelen voor lezen en schrijven er slechts 5 objectief scoorbaar zijn en daarbij betrouwbaarheden opleveren van .35 tot .66 met een uitschieter van .90 (IV, p. 158). Alle overige onderdelen doen in meer of mindere mate een beroep op relatief arbeidsintensieve beoordelings- en scoringsprocedures. Aangezien voor deze procedures 'kenmerkend is dat de beoordelaar verstand van de te beoordelen vaardigheid moet hebben, en vaak moet interpreteren en een keuze moet maken' (IV, p. 124), kan getwijfeld worden aan de bruikbaarheid in toekomstige peilingsonderzoeken. Deze twijfel vloeit niet alleen voort uit de te verwachten kosten wanneer vergelijkbare procedures ook gebruikt gaan worden voor andere gebieden dan taal, maar heeft vooral ook betrekking op de constanthouding van de beoordelingscriteria in toekomstig onderzoek. Wie garandeert dat in komende jaren andere beoordelaars met andere ervaringen en andere opvattingen over onderwijs en eisen te stellen aan leerlingen tot dezelfde interpretaties en keuzen komen? De onderzoekers moesten al in de proefpeiling, op een en hetzelfde tijdstip dus, grote moeite doen om ongewenste effecten zoals normverschuiving, invloed van persoonlijke tendenties enz. te vermijden. Veelbetekenend is in dit verband dat bij het beoordelen van opstellen, de beoordelaars kritiek hadden op de jury's die eerder de schaalopstellen hadden beoordeeld en daarvan de schaalwaarde hadden vastgesteld: 'De schaalwaarde van bepaalde schaalopstellen is, ook na nauwkeurige analyse en onderling overleg, niet goed te begrijpen voor de beoordelaars. Zij suggereren dat de oorspronkelijke jury van 9 die de schalen construeerde, zich heeft laten beïnvloeden door een aantal niet ter zake doende aspecten van de opstellen (bv. spelling)' (IV, p. 135).

A fortiori geldt de genoemde twijfel voor het spreekinstrument. De gevolgde procedure is hier zo complex en arbeidsintensief, zowel wat de afname als de scoring betreft, dat het instrument slechts bij 200 leerlingen kon worden afgenomen. Weliswaar vermelden de auteurs voor een van de twee beoordeelde aspecten (Inhoud) interbeoordelaarsbetrouwbaarheden van .73 tot .88 (voor Taalgebruik en opbouw golden lagere waarden van .41 tot .77) maar deze waarden werden pas bereikt nadat was besloten 3 van de 9 oorspronkelijke beoordelaars op grond van hun afwijkend beoordelingsgedrag buiten beschouwing te laten en in de beoordelingsprocedure van een der taken aanmerkelijke vereenvoudigingen aan te brengen.

Samenvattend kan men zich in de eerste plaats afvragen of het begrip 'functionele vaardigheid' al voldoende uitgekristalliseerd is om ook in toekomstig peilingsonderzoek als uitgangspunt voor de constructie van instrumenten te dienen. In de tweede plaats vertonen de reeds geconstrueerde instrumenten enkele eigenschappen, met name de bewerkelijkheid en het moeilijk in de tijd constant te houden beoordelingskarakter, waardoor ze niet erg geschikt lijken voor periodiek te herhalen peilingsonderzoek.

Slotopmerkingen

In het bovenstaande commentaar op de Voorstudie Periodieke Peiling van het Onderwijsniveau is voornamelijk ingegaan op de percentages functionele analfabeten die de auteurs met behulp van de geconstrueerde taalbeheersingsinstrumenten hebben menen te kunnen vaststellen en op de bruikbaarheid van de taalbeheersingsinstrumenten in toekomstig peilingsonderzoek. Geconcludeerd is dat de wetenschappelijke betekenis van de gevonden percentages gering is en de bruikbaarheid van de taalbeheersingsinstrumenten in toekomstig peilingsonderzoek twijfelachtig.

Tot besluit volgen nu enkele opmerkingen en suggesties met betrekking tot de vergelijkbaarheid van de gegevens in periodiek peilingsonderzoek.

Bij lezing van het rapport krijgt men sterk de indruk dat de auteurs zich onvoldoende hebben gerealiseerd dat hun peiling een eerste stap zou moeten vormen in een reeks van peilingen en dat het succes van zo'n eerste peiling mede afhankelijk is van een goede planning van de hele reeks. Zij achten in feite peilingsonderzoek haalbaar omdat de proefpeiling in hun ogen haalbaar is gebleken. De vraag of en op welke wijze in toekomstige peilingen gegevens verzameld kunnen worden die vergelijkbaar zijn met die van de proefpeiling krijgt in het rapport geen aandacht. Deze cross-sectionele gerichtheid kan overigens niet alleen de auteurs worden aangerekend. Ook in de onderzoeksopdracht ging het meer om het opdoen van 'ervaring in een landelijk niveaubepalend onderzoek' en om de 'resultaten van een deelpeiling van het niveau van het basisonderwijs' dan om een strakke planning van opeenvolgende peilingen zodat vergelijkbare gegevens resulteren.

Het laatste is niettemin een complex probleem dat een aparte voorstudie waard zou zijn en waarvan in het bijzonder de psychometrische kanten belicht zouden moeten worden. Vragen die zich daarbij aandienen zijn de volgende. Is het in het Nederlands onderwijssysteem haalbaar om in opeenvolgende peilingen exact dezelfde instrumenten aan leerlingen van eenzelfde leerjaar voor te leggen? Moet verwacht worden, dat testingeffecten optreden doordat leerkrachten de leerlingen in bekende peilingsinstrumenten oefenen, en zijn schattingen te maken van deze testingeffecten? Kunnen parallelle instrumenten ontwikkeld worden om de vergelijkbaarheid in opeenvolgende peilingen te garanderen? Kan een grotere flexibiliteit in de samenstelling van de instrumenten voor opeenvolgende peilingen worden verkregen door uit te gaan van Rasch-homogene itempools? Dit soort vragen verdient serieuze aandacht, wil een van de hoofdtaken van periodiek peilingsonderzoek, het vaststellen van prestatieverschuivingen in de tijd, uitvoerbaar worden.

J. H. L. Oud

(Instituut voor Orthopedagogiek K.U Nijmegen)