Het meten van de 'stelvaardigheid'

H. WESDORP
Research Instituut voor de Toegepaste Psychologie (R.I.T.P.) te Amsterdam

Samenvatting

Dit artikel vat een belangrijk deel van een onderzoek samen, dat tussen 1969 en 1973 aan het R.I.T.P. werd uitgevoerd. Er worden pogingen beschreven het meten van de 'produktief-schriftelijke taalvaardigheid' of 'stelvaardigheid' langs directe weg (opstel-beoordeling) te verbeteren. Tevens worden de mogelijkheden van toetsen ter (indirecte) meting van die vaardigheid verkend: de constructie en validering van twee typen toetsen worden beschreven. Geconcludeerd wordt, dat verbetering van de directe meting – door 'beoordelaars' – een arbeidsintensieve dus dure zaak is, wat voor veel Nederlandse examensituaties problemen zal opleveren. Toetsen zijn meettechnisch acceptabele instrumenten, die waarschijnlijk een betere indicatie van iemands 'schrijfvaardigheid' geven dan op één of enkele oordelen gebaseerde opstel-scores. Toetsen zijn echter slechts in noodsituaties toelaatbare instrumenten, omdat het 'backwash-effect' op het onderwijs- een negatieve terugslag op het curriculum – waarschijnlijk niet te vermijden is, als ze als enige eindtoetsing een rol zouden spelen.

1. *Achtergronden van het project 0123*

Het project 0123, dat van 1969–1973 aan het R.I.T.P. werd uitgevoerd, was gericht op het verkennen van beoordelingsmethoden voor de produktief-schriftelijke taalvaardigheid (ook wel stelvaardigheid of schrijfvaardigheid genoemd). De volgende overwegingen lagen ten grondslag aan het project:

a. De in Nederland gangbare methode ter beoordeling van de stelvaardigheid heeft ernstige gebreken. Meestal wordt een 'opstel' geschreven dat vervolgens wordt beoordeeld, meestal door één, soms door twee beoordelaars. Uit de buitenlandse literatuur betreffende dit onderwerp blijkt dat de in vele Nederlandse examens gevolgde procedure hoogst onbetrouwbaar moet zijn. Dit is zeker in situaties waarin t.a.v. leerlingen belangrijke beslissingen worden genomen, een onwenselijke zaak. Gezocht zou moeten worden naar mogelijkheden om de 'directe beoordeling' te verbeteren.

b. Sinds de introductie van objectieve studietoetsen in afsluitende programma's met een landelijk karakter, is het gevaar ontstaan dat slechts de middels objectieve toetsen meetbare doelstellingen hun status zouden behouden. Met name het project 'Schooltoetsen Basisonderwijs' van het C.I.T.O., dat jaarlijks ± 70.000 leerlingen der zesde klas toetst, kan door de beperkingen van de objectieve toetsvorm een negatief effect hebben op het basisschool-curriculum. Gezocht zou moeten worden naar mogelijkheden om de schrijfvaardigheid middels objectieve studietoetsen te meten, dus naar een 'indirecte beoordelingsmethode'.

De onder a. genoemde overweging is van algemeen belang voor de in Nederland te hanteren examenprocedures betreffende de stelvaardigheid. Daarnaast zou perfectionering van de 'directe methode' noodzakelijk zijn op grond van onderzoekstechnische overwegingen. De onder b. genoemde toetsen zouden immers slechts gevalideerd kunnen worden aan een 'direct criterium'. De onder b. genoemde overweging is slechts van

belang in de huidige situatie, waarin eindtoetsingen middels objectieve studietoetsen een belangrijke rol spelen. Zolang niet andere oplossingen van het overgangsprobleem B.O.–V.O. hun kans krijgen, is het zaak de toetsen die aan het eind van het basisonderwijs worden gehanteerd, zoveel mogelijk inhoudsvalide te doen zijn t.o.v. de doelstellingen.

In het volgende wordt eerst aandacht geschonken aan een aantal problemen betreffende de *directe beoordeling* (par. 2). De verschillende beoordelingsmethoden passeren de revue; een experiment met één dier methoden wordt beschreven. Vervolgens (par. 3) wordt de constructie van een tweetal instrumenten beschreven, n.l. een *objectieve schrijfvaardigheidstoets* – geheel in meerkeuzevorm – en een *interlineaire schrijfvaardigheidstoets*. Deze laatste toets zou men semi-objectief kunnen noemen. De toets bestaat uit een van bepaalde fouten voorziene tekst. De leerling corrigeert 'tussen de regels' eigenhandig. De scoring is uitsluitend gericht op de bewust in de tekst geïntroduceerde items, die overigens niet aangeduid worden voor de leerling. Er is een nauwkeurig scoringsvoorschrift. Vervolgens wordt het *validatie-onderzoek* beschreven, waarin de 'indirecte' instrumenten werden geconfronteerd met het 'directe' criterium: opstelbeoordelingen (par. 4). Tenslotte worden enkele resultaten summier besproken (par. 5).

2. *Directe beoordeling: op zoek naar een criterium*

2.1. *Inleiding*

In de beginfase van het onderzoek is vooral onderzoek gedaan naar de mogelijkheden een betrouwbaar en valide criterium te formeren, waaraan de te construeren toetsen zouden kunnen worden gevalideerd. Dit leidde tot een literatuuronderzoek (2.2) naar de mogelijkheden die de verschillende beoordelingstechnieken bieden en tot een eigen empirisch vooronderzoek (2.3) waarin getracht werd een aantal concrete vragen m.b.t. één der beoordelingstechnieken te beantwoorden.

Bij de beoordeling van de produktief-schriftelijke taalvaardigheid is sprake van tenminste twee 'proefpersonen': de schrijver die de vaardigheid moet tonen, en de beoordelaar van diens produkt, die een score moet vaststellen. Er zijn meerdere fouten-variantie-bronnen, die alle de betrouwbaarheid van de beoordeling van de schrijfvaardigheid beïnvloeden.

Gaat men ervan uit, dat een bepaalde proefpersoon ('schrijver' te noemen) in een bepaalde periode van zijn leven een bepaalde graad van produktief-schriftelijke taalvaardigheid 'heeft'[1] – zijn 'ware score' voor produktief-schriftelijke taalvaardigheid – dan kan men de volgende oorzaken van onbetrouwbaarheid aanwijzen bij pogingen om die graad van taalvaardigheid te meten:

1. *Instabiliteit van de schrijver*

Daar men meestal slechts een proeve van de produktief-schriftelijke taalvaardigheid beoordeelt, is het duidelijk dat men de 'performance' beoordeelt. Deze 'performance' van de schrijver fluctueert van dag tot dag, van situatie tot situatie. Dit betekent, dat bij het streven naar een betrouwbare scoring van de produktief-schriftelijke taalvaardigheid er gezocht moet worden naar een procedure, die ertoe bijdraagt de invloed van deze schrijverfluctuaties te minimaliseren.

2. *Instabiliteit van de beoordelaar*

Door de invloed van 'tijdelijke eigenschappen van het individu' fluctueren ook de beoordelaarsprestaties. De invloed van deze tijdelijke eigenschappen kan men door een weloverwogen onderzoeksdesign enigszins beperken. Het is ook het overwegen waard slechts die beoordelaars bij de beoordeling van de produktief-schriftelijke taalvaardigheid in te schakelen, die weinig of relatief weinig instabiel zijn bij de beoordeling in een bepaalde situatie. Dit zou kunnen leiden tot selectie van beoordelaars op hun stabiliteit, vooral

als men er niet zeker van is dat men de belangrijkste oorzaken van de instabiliteit door het design heeft weggenomen of sterk heeft verminderd.

3. *Gebreken in de intersubjectieve overeenstemming*

   Beoordelaars verschillen onderling in hun oordelen over bepaalde produkten. Zij hechten verschillende waarden aan de verschillende aspecten van de produktief-schriftelijke taalvaardigheid (een blijvende eigenschap, waarschijnlijk resulterend uit hun diverse opvattingen omtrent de doelstellingen van het onderwijs in de produktief-schriftelijke taalvaardigheid), maar kunnen b.v. ook verschillen doordat zij een verschillende mate van geoefendheid in het beoordelen van opstellen bezitten (een tijdelijke eigenschap, die na verloop van tijd verdwijnen kan).

4. *Non-equivalentie van taken,* hetzij t.o.v. schrijvers, hetzij t.o.v. beoordelaars, hetzij t.o.v. beiden. Het gaat hier om fluctuaties als gevolg van het onderwerp waarover geschreven wordt, het soort opstel (betoog, fantasieopstel), de tijd die ter beschikking staat, de examensituatie of de beoordelingssituatie. Verschillende taken leiden tot verschillende scores, hetzij door de verschillen zelf, hetzij doordat bepaalde leerlingen of bepaalde beoordelaars een zekere affiniteit tot bepaalde taken hebben. Het is vrij moeilijk om na te gaan welk van de onderscheidbare effecten het meest storend is. Toch is het in bepaalde situaties wel belangrijk inzicht te hebben in de verhoudingen tussen deze effecten, b.v. als bepaald moet worden of leerlingen de vrije keus van onderwerp moeten krijgen of elk dezelfde opdracht moet krijgen. Getracht moet worden de effecten van de non-equivalentie van taken te minimaliseren.

In ons literatuuronderzoek (2.2) hebben wij getracht informatie te verwerven omtrent de bovengenoemde 4 bronnen van onbetrouwbaarheid. De fluctuaties bij de schrijver zijn in een beperkt aantal studies onderzocht. Zij leiden tot schattingen van de 'schrijversstabiliteit'. De fluctuaties als gevolg van het onderwerp hebben iets meer aandacht in de research gekregen. Zij leiden tot schattingen omtrent het aantal onderwerpen, waarover men leerlingen moet laten schrijven om de storende invloed van de 'non-equivalentie van taken' tot acceptabele proporties terug te brengen. De research heeft de meeste aandacht besteed aan de fluctuaties bij de beoordelaar, leidend tot schattingen omtrent de 'beoordelingsstabiliteit', en aan de fluctuaties tussen beoordelaars, leidend tot schattingen betreffende de 'intersubjectieve overeenstemming'. Een aantal methoden is ontworpen om de 'instabiliteit van de beoordelaar' en de 'gebreken in de intersubjectieve overeenstemming' tot acceptabele proporties terug te brengen.

De betrouwbaarheid van beoordelingsmethoden wordt vaak uitgedrukt in een coëfficiënt die de mate van overeenstemming tussen individuele beoordelaars weergeeft, de intersubjectieve overeenstemming dus. Vaak wordt een gemiddelde intersubjectieve overeenstemming berekend tussen een groot aantal individuele beoordelaars. De aldus verkregen betrouwbaarheidscoëfficiënt houdt geen rekening met de andere drie genoemde foutenbronnen en is dus een onjuiste betrouwbaarheidsindex. Daarom hebben verschillende onderzoekers voorgesteld de leerling meer opstellen (over verschillende onderwerpen) te laten schrijven, en deze door meer beoordelaars te laten beoordelen; hiermee werden ook de 'instabiliteit van de schrijver' en de 'non-equivalentie van taken' aangevat. Met de introductie van meerdere beoordelaars, die in een 'team' werkten en een gesommeerd oordeel leverden, is bovendien een belangrijk probleem in de belangstelling gekomen, n.l. het probleem van de validiteit van een oordeel of een som van oordelen.

De reeds vroeg geconstateerde fluctuaties tussen beoordelaars hebben geleid tot pogingen deze fluctuaties tot acceptabele proporties terug te brengen. De in 2.2 behandelde technieken zijn dan ook voornamelijk op dit punt gericht. Wiseman (1949) heeft erop gewezen, dat de

overeenstemming tussen een beoordelaar en een gesommeerd oordeel van een aantal anderen evenzeer een validiteits-coëfficiënt is als een betrouwbaarheidscoëfficiënt. Zeker als een 'ware score' wordt gepostuleerd in de vorm van het gesommeerde oordeel van een oneindig aantal beoordelaars, is deze benadering duidelijk. Als echter deze redenering wordt aanvaard, is een hoge intersubjectieve overeenstemming tussen de leden van een team niet meer noodzakelijkerwijze wenselijk: een zekere mate van gebrek aan overeenstemming kan dan juist nastrevenswaardig zijn voorzover daarmee de diversiteit in standpunt ten opzichte van de waardering van de complexe produktief-schriftelijke taalvaardigheid wordt gerespecteerd. De bijdrage, geleverd door de individualiteit van de beoordelaar, dus door datgene wat hem juist onderscheidt van de andere beoordelaars, wordt daarmee positief gewaardeerd. Dit heeft gevolgen voor onze waardering van de validiteit van oordelen, die hoog intercorreleren door invoering van in 2.2 behandelde technieken. Wij zullen daar in de evaluatie van elk der behandelde technieken op terug komen. Daar verschillen in opvatting van belang zijn met het oog op het verkrijgen van een valide som-oordeel van een team van beoordelaars, is een apart gedeelte van ons empirisch vooronderzoek aan dit punt gewijd.

2.2. *Literatuuronderzoek beoordelingsprocedures*

2.2.1. *Beoordelingsschalen*

Reeds in de eerste decennia van deze eeuw heeft men gepoogd de oorzaken van de gebreken in de 'intersubjectieve overeenstemming' onder controle te krijgen, opdat daardoor de intersubjectieve overeenstemming tussen beoordelaars zou worden verhoogd. Men construeerde 'beoordelingsschalen' (composition scales). In een dergelijke schaal komt een aantal in kwaliteit oplopende voorbeeldopstellen voor, die alle voorzien zijn van een beoordelingscijfer. De docent kan bij het beoordelen van het werk van zijn leerlingen hun opstellen vergelijken met de in de schaal aanwezige opstellen.

Beoordeling met behulp van schalen heeft reeds vroeg kritiek ondervonden. Die kritiek was gericht op de validiteit van de resulterende oordelen, maar ook op de betrouwbaarheid, d.w.z. het gebrek aan intersubjectieve overeenstemming. De uit de literatuur blijkende winst op dit gebied is niet van dien aard, dat andere, nog te bespreken methoden inferieur lijken. Wel kan gezegd worden dat in massale examenprogramma's schalen een enigszins stabiliserende functie hebben. De schaal doet een gemeenschappelijk referentiekader ontstaan.

De beste resultaten met het gebruik van schalen zijn echter bereikt na soms intensieve training. Daartegen kan nu juist weer bezwaar worden gemaakt: het is niet onwaarschijnlijk dat de inhoudsvaliditeit van het individuele oordeel door training vermindert, althans vanuit het standpunt van de groep docenten, die dezelfde visie heeft als de docent vóór de training. Docenten met een niet algemeen aanvaarde visie op de waardering van verschillende aspecten van schriftelijke taalvaardigheid zullen door training en discussie hun standpunt herzien. Zeker als met behulp van schalen een team-oordeel wordt geveld, is het van belang de afzonderlijke visie der individuele beoordelaars te respecteren. Het is de vraag of niet juist gestreefd moet worden naar team-beoordeling, waarin verschillende visies hun kans krijgen (vgl. 2.1).

Onderwijskundige voordelen biedt het gebruik van schalen zeker: de STEP-essay-test[2] is o.m. geaccepteerd omdat door deze evaluatiemethode de nadelen van objectieve toetsen ter meting van de produktief-schriftelijke taalvaardigheid konden worden voorkomen. Immers, deze laatste zouden kunnen leiden tot een vermindering van de aandacht voor het onderwijs in het schrijven van essays. Dit gevaar ontstaat niet bij het gebruik van schalen. In de kritiek wordt weinig aandacht aan de betrouwbaarheid van de STEP-essay-test geschonken, maar des te meer aan de inhoudsvaliditeit van de ermee tot stand gekomen oordelen. Hieruit zou kunnen worden afgeleid dat de critici de onderwijskundige voordelen van het gebruik van schalen veel belangrijker achten

dan de meettechnische nadelen.

Voor het in het project 0123 noodzakelijke criterium is niet gekozen voor gebruikmaking van een schaal. Ten eerste zou een schaal moeten worden geconstrueerd: een arbeidsintensieve en kostbare zaak. Ten tweede bleken de individuele oordelen bij gebruikmaking van schalen niet tot een acceptabele overall-betrouwbaarheid te kunnen leiden. Ten derde zou, bij gebruikmaking van schalen door een team eenzelfde of beter resultaat bereikt kunnen worden met andere beoordelingsmethoden (zie 2.2.3), tenzij tot training of selectie van beoordelaars zou worden overgegaan. De bezwaren hiertegen zijn al uiteengezet.

2.2.2. *Beoordeling m.b.v. analytische schema's; globale beoordeling*

Voortbouwend op de ervaringen van sommige schaalconstructeurs zijn in de jaren twintig en dertig analytische beoordelingsschema's ontwikkeld. Een analytisch beoordelingsschema tracht de complexe produktief-schriftelijke taalvaardigheid in een aantal deelvaardigheden af te breken. De beoordelaar moet de (soms grote aantallen) subvaardigheden apart beoordelen. Sommige deelvaardigheden zijn belangrijk en krijgen een groot gewicht bij de bepaling van de totaalscore, andere wegen minder mee.

Na ± 1925 werd echter ook de weg ingeslagen van de globale beoordeling. Bedoeld wordt hiermee een onmiddellijke beoordeling van de geleverde prestatie als geheel, zonder het referentiekader van een schaal en zonder analyse in aspecten. Deze methode wordt ook wel de 'short impression method', de 'impressionistic method' of 'wholistic marking' genoemd.

Het is niet zinvol de kwaliteiten van beide methoden – 'analytische' en 'globale' – afzonderlijk te behandelen. Veel onderzoek is n.l. gericht geweest op een vergelijking van beide methoden.

Hoewel niet alle vergelijkende onderzoekingen geheel gelijkluidende resultaten opleveren, is de tendens zeker waarneembaar dat analytische schema's gemiddeld tot iets hogere intersubjectieve overeenstemming en beoordelaarsstabiliteit leiden. Vooral als slechts één beoordelaar de beoordelingstaak uitvoert en de beoordelingstijd geen rol speelt, kan dit belangrijk zijn. Tegen de validiteit van met behulp van analytische schema's tot stand gekomen oordelen is vaak bezwaar gemaakt. Tegenover het voordeel dat elk analytisch schema biedt – nauwkeurige informatie omtrent de te beoordelen subvaardigheden en daarmee een zekere specificatie van het begrip-zoals-bedoeld – staat het nadeel dat elk analytisch schema slechts een opvatting omtrent de wenselijke doelstellingen weerspiegelt, en niet meer dan dat.

De analytische schema's verschillen alle; sommige leggen de nadruk op formele taalvaardigheden, andere zijn meer op de inhoudelijke aspecten gericht. Soms wordt een der aspecten totaal genegeerd. Ook de wegingen die sommige analytische schema's voor de verschillende subvaardigheden voorschrijven, zijn een uiting van meningen omtrent de wenselijke doelstellingen van het onderwijs in de produktief-schriftelijke taalvaardigheid. De training, die noodzakelijk is voor het goed leren hanteren van analytische schema's, vergroot onze bezwaren tegen de validiteit van de met deze schema's tot stand gekomen oordelen. De essentie van deze bezwaren is dezelfde als die tegen de training in het gebruik van beoordelingsschalen. Overigens blijkt zelfs dat ervaring in het gebruik van een analytisch schema niet leidt tot werkelijk onafhankelijke beoordeling der subvaardigheden. Sommige onderzoekers (b.v. Bonnardel, 1946) signaleerden een sterk halo-effect.

Analytische schema's blijken slechts resultaten te hebben voor de 'lagere' vaardigheidsaspecten (spelling, interpunctie, zinsbouw). De 'hogere' vaardigheden (stijl, creativiteit, organisatie) blijken, ondanks de soms vrij nauwkeurige voorschriften dienaangaande zeer subjectief en intersubjectief verschillend te worden beoordeeld. Men kan hoogstens zeggen dat er enige aanwijzing is dat analytische schema's enig perspectief bieden op dit gebied. De analytische methode is tenslotte een dure methode: analytische beoordelaars moeten zeer nauwkeurige voorschriften hanteren, soms uitvoerige notatiesystemen aan-

leren en toepassen, wat weer vaak weerstanden oproept bij de beoordelaars, die zich deze weelde in de schoolpraktijk niet kunnen veroorloven (Lamb, 1953). Analytische beoordelaars werken van 2 tot 15 maal zo langzaam als globale beoordelaars.

Voor het in het project 0123 noodzakelijke criterium is op grond van bovenstaande overwegingen niet gekozen voor het opstellen en gebruiken van een analytisch beoordelingsschema.

De globale beoordelingsmethode is door sommige critici een 'oppervlakkige' beoordelingsmethode genoemd, die het beoordelen van de minder technische aspecten (stijl, originaliteit, ordening) zou doen verwaarlozen. Tegenover deze kritiek staan echter de researchresultaten. Daaruit blijkt juist, dat globale beoordelaars deze kwaliteiten wel degelijk trachten te beoordelen. Zij doen dit alleen niet in gelijke mate (Diederich, French en Carlton, 1961), wat behalve nadelen ook voordelen heeft. De voordelen hangen samen met de in 2.2.1. al besproken bezwaren tegen training, die bij de beoordeling met een schaal en met behulp van een analytisch schema onontbeerlijk zijn. Bij de globale methode is training, met zijn mogelijke nadelige invloed op de validiteit van een gesommeerd jury-oordeel, overbodig. De globale methode is daarom – èn omdat in vele gevallen minder beoordelingstijd per opstel nodig is – goedkoper dan de analytische.

Voor het in het project 0123 noodzakelijk criterium is globale beoordeling door één beoordelaar afgewezen: een dergelijke methode zou zonder meer onbetrouwbaar en beperkt valide zijn. Wel leek de globale methode, toegepast door een jury van onafhankelijke beoordelaars, een acceptabele methode ter verkrijging van een aan hoge eisen voldoend criterium.

2.2.3. *Jury-beoordeling*

In de jaren veertig introduceerde Wiseman (1949) in Engeland een systeem van jury-beoordeling. De grondgedachte daarbij – ontleend aan de testtheorie – was dat een aantal onafhankelijk jurerende beoordelaars gezamenlijk een score tot stand brachten, die een benadering zou zijn van de 'ware score': de gemiddelde score toegekend door een oneindig aantal beoordelaars. Volgens hem had het geen zin verder te experimenteren met analytische methoden, daar een verhoging van de beoordelaarsstabiliteit als gevolg van die methode niet aanzienlijk was. Wel achtte hij het raadzaam globale beoordelaars op hun stabiliteit te selecteren. Hij achtte de beoordelaarsstabiliteit een betere index voor de waarde van een beoordelaar dan zijn intercorrelatie met andere beoordelaars (intersubjectieve overeenstemming), omdat – als de beoordelaars competent zijn – een zekere mate van onenigheid juist wenselijk is. Deze onenigheid wijst op het feit, dat de beoordelaars verschillende standpunten innemen ten aanzien van de beoordeling van de complexe produktief-schriftelijke taalvaardigheid. Een sommering van een aantal van die diverse oordelen zou een beter beeld geven van de waarde van de te beoordelen opstellen.

Met deze gedachtengang kunnen wij geheel meegaan. Hij sluit aan bij wat hierboven al gesteld werd. Sommering of middelen van niet specifiek getrainde maar wel stabiele globale beoordelaars is met het oog op de validiteit van het door ons gezochte criterium van belang: niet slechts één visie op de kwaliteiten van het te beoordelen produkt wordt geaccepteerd of nagestreefd (d.m.v. training, een nauwkeurig analytisch schema, of eliminatie van dissidente beoordelaars), maar meerdere visies worden gewaardeerd en in de vorming van het eind-oordeel betrokken.

Voor het in het project 0123 gezochte criterium leek beoordeling door een team van globale beoordelaars een acceptabele methode: zeker indien meerdere opstellen per leerling zouden worden gebruikt – om de gevolgen van de instabiliteit van de schrijver en de non-equivalentie van taken tot acceptabele proporties terug te brengen – zou een voldoende grote jury, die globaal beoordeelt, een somoordeel kunnen leveren, dat zowel een hoge betrouwbaarheid als een hoge validiteit bezit.

2.2.4. *Computer-beoordeling*

Volledigheidshalve vermelden wij hier ook de laatste ontwikkeling op het gebied van de essaybeoordeling: de scoring van essays per computer. Enkele recente onderzoeken hebben duidelijk gemaakt dat het mogelijk is per computer oordelen te vellen over opstellen, die niet of nauwelijks te onderscheiden zijn van die van competente beoordelaars (Page, 1969).

2.3. *Oriënterend criteriumonderzoek*

In 2.2.3. concludeerden wij, dat globale jurybeoordeling een acceptabele methode zou zijn ter verkrijging van een voor het project 0123 noodzakelijk criterium. Een aantal vragen moest echter eerst beantwoord worden, voor een besluit genomen zou kunnen worden ten aanzien van 1. het aantal noodzakelijke beoordelaars per opstel, 2. het aantal noodzakelijke opstellen per leerling en eventuele selectie van beoordelaars 3. op grond van hun stabiliteit en 4. op grond van het 'type', waartoe zij behoren. De eerste drie problemen zijn problemen met betrekking tot de betrouwbaarheid van het criterium, het laatste probleem betreft de validiteit van het criterium.

In een proefonderzoek, waarin 16 onderwijzers elk 105 opstellen van zesdeklassen (globaal) beoordeelden en na enkele maanden herbeoordeelden werd informatie ingewonnen om de volgende vragen te beantwoorden:

1. *Hoeveel beoordelaars moeten per opstel worden ingeschakeld* om de invloed van de gebreken in de intersubjectieve overeenstemming tussen beoordelaars tot acceptabele proporties terug te brengen?

3. *Is selectie van beoordelaars op grond van hun beoordelaarsstabiliteit wenselijk?* Is het organiseren van een 'stabiliteitsproef' met het oog op selectie van juryleden zinvol, m.a.w. zijn de oordelen van geselecteerde jury's substantieel betrouwbaarder dan die van ongeselecteerde jury's?

4. *Zijn er verschillende 'typen' van beoordelaars te herkennen*, zodat bij de formering van een jury rekening gehouden kan worden met de standpunten der leden, opdat een zo veelzijdig mogelijke jury wordt geformeerd, met het oog op de validiteit van het criterium?

Op basis van een literatuuronderzoek werd een antwoord gezocht op de vraag

2. *Hoeveel opstellen moeten per leerling in het criterium worden opgenomen* om de invloed van de 'non-equivalentie van taken' en de 'instabiliteit van de schrijver' tot acceptabele proporties terug te brengen?

Bij de beantwoording van bovenstaande vragen zijn wij uitgegaan van de wenselijkheid voor het valideringscriterium een *beoordelingsbetrouwbaarheid* van ± .95 en een *score-betrouwbaarheid* van ± .85 te bereiken[3].

2.3.1. *Het aantal beoordelaars per opstel*

Indien een beoordelingsbetrouwbaarheid (juryequivalentie) van minstens .95 wordt nagestreefd, zal een groot aantal individuele beoordelingen moeten worden gesommeerd. Op basis van de resultaten van ons oriënterend onderzoek kunnen wij een schatting maken van de gemiddeld te verwachten intersubjectieve overeenstemming. Tabel 1 geeft de intercorrelatiematrix tussen de 16 beoordelaars: de resultaten van de eerste beoordeling (november 1969) staan *boven* de diagonaal; de resultaten van de tweede beoordeling (februari 1970) staan *onder* de diagonaal.

De intersubjectieve overeenstemming varieert in november 1969 tussen .17 en .67, in februari tussen .27 en .67 .De gemiddelde intersubjectieve overeenstemming is in november .45, in februari .48 (berekend via z-omzetting).

Wij mogen dus een gemiddelde intersubjectieve overeenstemming van . 45 verwachten tussen de juryleden die aan de totstandkoming van het criterium zullen medewerken. De som van de

Tabel 1. Intercorrelaties tussen de 16 beoordelaars van 105 opstellen van zesde klassers bij eerste beoordeling (november 1969; boven de diagonaal) en bij herbeoordeling (februari 1970; onder de diagonaal).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	—	.33	.26	.49	.41	.34	.46	.51	.33	.42	.35	.45	.47	.24	.43	.32
2	.45	—	.30	.45	.42	.36	.49	.47	.41	.53	.46	.55	.58	.39	.56	.56
3	.51	.38	—	.36	.17	.34	.39	.33	.38	.35	.26	.38	.39	.23	.28	.33
4	.60	.37	.54	—	.50	.55	.47	.48	.62	.55	.58	.54	.62	.56	.63	.61
5	.48	.45	.48	.52	—	.42	.47	.43	.45	.32	.38	.44	.65	.22	.54	.45
6	.46	.39	.28	.49	.48	—	.41	.40	.40	.35	.45	.30	.46	.52	.42	.32
7	.62	.59	.51	.46	.56	.55	—	.65	.49	.45	.44	.53	.59	.27	.53	.60
8	.53	.31	.37	.46	.43	.29	.49	—	.50	.53	.36	.46	.56	.34	.51	.47
9	.47	.31	.30	.56	.47	.42	.43	.46	—	.34	.47	.43	.51	.36	.52	.44
10	.56	.42	.41	.46	.45	.41	.61	.42	.37	—	.36	.46	.50	.37	.43	.34
11	.49	.37	.39	.51	.51	.50	.54	.39	.51	.51	—	.38	.55	.44	.67	.49
12	.55	.67	.51	.45	.51	.42	.59	.45	.32	.46	.48	—	.58	.30	.54	.49
13	.49	.60	.41	.54	.53	.58	.59	.39	.27	.44	.46	.65	—	.34	.66	.59
14	.61	.35	.50	.46	.49	.38	.48	.38	.36	.56	.45	.48	.49	—	.53	.19
15	.59	.50	.47	.61	.58	.50	.63	.49	.51	.53	.54	.50	.54	.63	—	.62
16	.46	.49	.37	.39	.43	.38	.53	.36	.43	.41	.37	.49	.45	.34	.56	—

oordelen van 25 juryleden zou een beoordelingsbetrouwbaarheid (jury-equivalentie)) van .95 garanderen[4].

2.3.2. *Selectie van beoordelaars o.g.v. hun stabiliteit*

De correlatie-berekeningen tussen de novembern en de februari-resultaten van het oriënterend onderzoek leverden de stabiliteitscoëfficiënten. Zij staan, met enkele andere karakteristieken van de beoordelaars vermeld in tabel 2.

Op grond van berekening van de beoordelingsbetrouwbaarheid van meer en minder stabiele jury's en de vergelijking daarvan, kon worden geconstateerd, dat weinig verschillen waren te verwachten tussen de beoordelingsbetrouwbaarheid van op stabiliteit geselecteerde en niet-geselecteerde jury's.

2.3.3. *Selectie van beoordelaars op gehanteerde maatstaven*

In de literatuur wordt melding gemaakt van het bestaan van min of meer duidelijk te onderscheiden opvattingen ten aanzien van de weging der verschillende aspecten bij de beoordeling van de complexe produktief-schriftelijke taalvaardigheid. In het licht van de eerder aangehaalde opvatting, dat diversiteit onder de beoordelaars gewenst is, om bij team-beoordeling verzekerd te zijn van een 'veelzijdig' totaal oordeel, dat meer aanspraak kan maken op validiteit dan een nauwkeurig gestuurd oordeel, is het van belang meer informatie te verwerven omtrent het bestaan van verschillende 'typen' van beoordelaars. Daar zou dan eventueel rekening mee gehouden kunnen worden bij de samenstelling van de aan de totstandkoming van het criterium meewerkende jury.

Rekening houden met de eventueel bestaande diversiteit in beoordelingsgedrag en dus met de diversiteit die kennelijk bestaat in de opvattingen betreffende de onderwijsdoelstellingen betekent een keuze. Een keuze vóór het accepteren van die diversiteit, die leidt tot het opnemen van zoveel mogelijk – of indien beperking noodzakelijk is: de belangrijkste – opvattingen in de criteriumjury. Een keuze tégen het vastleggen van de doelstellingen vooraf – als wij daartoe al bij

Tabel 2. Gemiddelden en standaardafwijkingen per beoordelaar bij eerste beoordeling (november 1969) en tweede beoordeling (februari 1970) van 105 opstellen van zesdeklassers. Stabiliteits-coëfficiënten; aantal jaren leservaring.

Beoord.	Eerste beoord.		Tweede beoord.		Stab.	Aantal
	Gem.	S.A.	Gem.	S.A.	coëff.	dienstjaren
1	6,4	1,5	5,9	1,4	.17	0
2	6,0	1,2	5,9	1,4	.53	25
3	6,0	1,2	5,9	1,0	.32	3
4	6,3	1,5	6,0	1,6	.48	16
5	6,0	1,7	6,2	1,1	.50	39
6	6,4	1,7	6,8	1,3	.47	10
7	5,4	1,6	6,0	1,5	.40	5
8	6,4	1,5	6,1	1,5	.60	0
9	5,6	1,4	5,5	1,2	.48	8
10	6,0	2,3	5,9	2,2	.32	4
11	5,5	1,6	5,6	1,6	.42	0
12	6,4	1,5	6,5	1,4	.34	13
13	6,0	1,4	6,3	1,1	.66	31
14	6,6	1,4	6,0	1,2	.45	40
15	5,7	1,6	5,9	1,3	.59	28
16	5,8	1,6	6,0	1,5	.57	16

machte waren – en dientengevolge tegen besturing van het beoordelingsgedrag d.m.v. training en discussie vooraf, of d.m.v. selectie op intersubjectieve overeenstemming achteraf. Wij volgen met deze keuze de gedachtengang van Wiseman (1949).

Om meer informatie te verwerven met betrekking tot de theoretische maatstaven der beoordelaars werd een instrument ontwikkeld, bestaande uit een lijst met opstel-kenmerken, die met behulp van de Q-sort-methode (Block, 1961) door de beoordelaars gerangschikt moesten worden naar hun belangrijkheid bij het beoordelen (Van Peet, 1970). Op de intercorrelatiematrix van de Q-sorteringen der 16 beoordelaars werd – evenals op die van de eerste en tweede opstelbeoordeling– hiërarchische clusteranalyse toegepast. Vergelijking van de drie resulterende analyses leidde tot de conclusie dat er weliswaar o.g.v. theoretische maatstaven clustervorming optrad en ook o.g.v. de praktische cijfergeving, doch 'theorie' en 'praktijk' bleken bij lange na niet parallel te lopen. Op grond van de resultaten van dit vooronderzoek werd besloten geen pogingen te doen beoordelaars in de jury op te nemen o.g.v. het 'type' dat zij zouden vertegenwoordigen.

2.4.4. *Het aantal opstellen per leerling*

Uit de literatuur bleek, dat de te verwachten fluctuaties van de schrijver als gevolg van de 'non-equivalentie van taken', aanzienlijk zou zijn. Wanneer twee 'titels' worden beoordeeld door twee verschillende beoordelaars mag tussen de twee reeksen oordelen geen sterk verband worden verwacht. Waarschijnlijk moeten wij rekening houden met te verwachten coëfficiënten van .25 tot .35. Indien jury's de verschillende titels beoordelen, zijn de coëfficiënten aanmerke-

lijk hoger. Waarschijnlijk mogen wij, als een jury van ± 5 beoordelaars wordt geformeerd, op correlatie-coëfficiënten van .50 – .60 rekenen tussen de jury-oordelen over verschillende titels.

Indien wij een score-betrouwbaarheid (jury x taak – equivalentie) van ± .85 willen bereiken, moeten wij ± 5 verschillende opstellen door jury's van ± 5 laten beoordelen[5]. Een dergelijke procedure zou, gezien de overwegingen aan het slot van 2.3.2. eveneens een beoordelingsbetrouwbaarheid (jury-equivalentie) van ± .95 garanderen.

Besloten werd daarom het in het valideringsonderzoek noodzakelijke criterium te doen bestaan uit 5 opstellen over verschillende onderwerpen, elk opstel beoordeeld door een jury van 5 beoordelaars.

3. *Constructie van de 'indirecte instrumenten'*

Voor de toetsconstructie is inzicht in de inhoud van de doelstellingen van het stelonderwijs noodzakelijk. Betreffende deze doelstellingen voor het Nederlandse basisonderwijs is informatie in de literatuur vrij summier. Ook andere literatuur is derhalve geraadpleegd om tot concrete toetsspecificaties te komen. Ten eerste hebben wij een poging gedaan meer informatie te verwerven omtrent de categorisering van de *doelstellingen* van het onderwijs in de produktief-schriftelijke taalvaardigheid in andere landen. Ten tweede bleek de literatuur betreffende analytische *beoordelingsschema's* een belangrijke bron van informatie: elk beoordelingsschema is een poging tot explicitering der doelstellingen. Ten derde is de literatuur over de *analyse van beoordelaarsgedrag* een mogelijke informatiebron: resultaten van zulk onderzoek kunnen aanwijzingen geven omtrent de in de toets op te nemen deelvaardigheden. Ten vierde bieden de bestaande *toetsen ter meting van de produktief-schriftelijke taalvaardigheid* uiteraard belangrijke informatie. Zulke toetsen zijn reeds expliciteringen van doelstellingen op het vaardigheidsgebied-in-kwestie en bieden, mits goede handleidingen aanwezig zijn, enig inzicht in de structuur en inhoud van het begrip-zoals-bedoeld. Ten vijfde is de literatuur, die inzicht biedt in de *geconstateerde tekorten op het gebied van de produktief-schriftelijke taalvaardigheid* ook informatief: doelstellingen kunnen concreet gemaakt worden door de te elimineren tekorten op het gebied van de vaardigheid-in-kwestie te inventariseren. In verband hiermee hebben wij enig onderzoek verricht op dit gebied. Door middel van een foutenanalyse van een aantal opstellen van 12-jarigen is enig inzicht verworven omtrent de aanwezigheid van tekorten. De hieruit voortkomende informatie is gebruikt bij de besluitvorming met betrekking tot de inhoud der te construeren toetsen.

Het (overigens noodzakelijkerwijs beperkte) literatuuronderzoek heeft geleid tot het opstellen van 'toetsspecificaties': uitgewerkte opsommingen van de in de te construeren toetsen op te nemen aantallen items per subvaardigheid. In deze toetsspecificaties zijn niet alle doelstellingen die in de literatuur vermeld worden of daaruit afgeleid kunnen worden vertegenwoordigd. Er moesten enkele beperkingen worden aangebracht. Die noodzakelijke beperkingen waren aanleiding om de geconstrueerde instrumenten 'toetsen ter meting van *aspecten* van de produktief-schriftelijke taalvaardigheid' te noemen, en niet meer, zoals in de oorspronkelijke doelstelling van het project 0123, te spreken van 'toetsen ter meting van *de* produktief-schriftelijke taalvaardigheid'. Over de beperkingen kunnen wij het volgende opmerken:

Ten eerste werden zelden genoemde doelstellingen of zelden in beoordelingsschema's of 'test-specifications' voorkomende vaardigheden niet opgenomen. Alleen vrij algemeen voorkomende vaardigheden werden opgenomen.

Ten tweede werden sommige meermalen genoemde doelstellingen of vaardigheden niet opgenomen, omdat zij slechts doelstellingen of vaardigheden waren voor zeer bepaalde vormen van produktief-schriftelijke taalvaardigheid. Daarom is b.v. geen poging gedaan de vaag geformuleerde eigenschap 'suspense' in het schema op te nemen onder de categorie 'Inhoud', omdat deze eigenschap slechts wenselijk is voor zeer

bepaalde opstellen. Hetzelfde geldt m.n. voor eigenschappen als 'directness', 'forcefullness', 'richness'.

Ten derde zijn enkele doelstellingen weggelaten omdat zij naar onze mening in Nederland niet algemeen beschouwd worden als subdoelstellingen van het onderwijs in de produktiefschriftelijke taalvaardigheid. Het vertonen van een 'goed handschrift' en 'netheid' vallen hier bijvoorbeeld onder.

Ten vierde zijn een aantal in de literatuur genoemde doelstellingen niet opgenomen omdat de operationalisering ervan een technisch te zware opgaaf is bij de huidige stand van toetsconstructietechniek. Het betreft een aantal vaardigheden die merendeels in de categorie 'Inhoud' van ons schema zouden vallen. (Originaliteit, creativiteit, e.d.).

Behalve de bovenstaande gronden, die tot beperking van het aantal in het schema op te nemen (sub)vaardigheden moesten leiden, is er nog een algemene psychometrische reden om dit aantal te beperken. Het is namelijk weliswaar mogelijk nominaal een groot aantal verschillende subvaardigheden te onderscheiden, maar uit een oogpunt van meting blijken deze elkaar gewoonlijk zo sterk te overlappen, dat op meettechnische gronden bepaalde subvaardigheden best gemist kunnen worden. Ook uit deze overweging zijn beperkingen van de in de hoofdcategorieën van het schema op te nemen (sub)vaardigheden voortgevloeid.

De in de toetsen te operationaliseren (sub) vaardigheden zijn in het schema gedefinieerd in termen van 'tekorten'. Dit heeft de volgende praktische reden: bij de constructie van meerkeuzetoetsen doet de toetsconstructeur er goed aan zijn 'afleiders' weloverwogen te kiezen. Een gangbare procedure is hierbij vaak gemaakte fouten als afleiders te hanteren. De ervaren docent is bij de toetsconstructie daarom in het voordeel: zijn ruime praktijkervaring biedt hem de mogelijkheid te putten uit het tekortengebied van de leerling, dat hij goed kent.

Voor de toetsconstructie werd het volgende schema gebruikt:

I. *Tekorten op het gebied der 'conventies'*

1. Fouten in de spelling van 'gewone' Nederlandse woorden.
2. Fouten in de spelling van zgn. bastaardwoorden.
3. Fouten in de spelling van werkwoordsvormen.
4. Fouten in het gebruik van hoofdletters.
5. Fouten tegen de regels der interpunctie.
6. Diverse fouten (b.v. foute afkortingen; te klein getal in cijfers; verkeerd afbreken, enz.).

II. *Tekorten op het gebied van de grammaticale bouw van de zin*

1. Woord ontbreekt, waardoor zinsstructuur gebrekkig is.
2. Verkeerd woord (veelal: verwijswoord, geslachtsaanduidend woord of voorzetsel).
3. Verkeerde werkwoordstijd of -vorm.
4. Incongruentie (veelal: onderwerp-persoonsvorm, echter soms ook onderwerp-naamwoordelijk deel, enz.).
5. Verkeerde volgorde van zinsdelen.
6. Verkeerde beknopte bijzin (ook fouten met 'om' en 'zodoende').
7. Contaminatie van zinnen.
8. Onjuiste samentrekking van zinnen; verbinding van zinnen gebrekkig.
9. Foute lijdende vorm.
10. Algemeen: syntactisch gebrekkige zinsbouw.

III. *Tekorten op het gebied van het 'passend taalgebruik', het 'effectief taalgebruik'*

1. Woord te veel, overbodig.
2. Storende herhaling van een woord.
3. Niet passend, niet effectief woord.
4. Contaminatie van woorden, uitdrukkingen.
5. Storende herhaling van een (stuk van een) zin.
6. Omslachtige uitdrukking.
7. Verkeerde uitdrukking.
8. Passe partout: nietszeggend, onduidelijk van bedoeling.
9. Cliché, foute of afgesleten beeldspraak.
10. Niet acceptabele praattaal, telegramstijl.
11. Te kort en onduidelijk geformuleerd.

IV. *Tekorten op het gebied van de organisatie van het betoog*

1. Er ontbreekt noodzakelijke informatie. (Welke?)
2. Er ontbreekt bepaalde informatie (Waar plaatsen?).
3. Er is bepaalde informatie overbodig.
4. De informatie staat in de verkeerde volgorde, op de verkeerde plaats.

V. *Tekorten op het gebied van de inhoud*

1. Er wordt een onlogische gedachtensprong gemaakt.
2. Er wordt een onjuiste conclusie getrokken.
3. Er is tegenspraak tussen mededelingen in twee zinnen.
4. De informatie is in strijd met (algemeen bekende) feiten.
5. Beweringen worden niet of nauwelijks geargumenteerd.
6. Het betoog is verward.

Op grond van de hiervoor genoemde foutenanalyse van opstellen van zesde klassers werd de mate waarin de verschillende hoofdcategorieën in de toetsen zouden moeten zijn vertegenwoordigd, als volgt bepaald. Aan de categorie I (conventies) zouden aparte toetsen (n.l. voor spelling en voor interpunctie) worden gewijd. Aan categorie II (grammaticale bouw) zou ± 35% der items worden gewijd, en aan categorie III (passend en effectief taalgebruik) ± 40%. Aan de categorieën IV en V tesamen zou ± 25% der items worden gewijd.

Nadat de instrumenten waren geconstrueerd volgde een pretest-fase. Na itemselectie kwam een aantal toetsen gereed, dat aan hoge eisen t.a.v. de toetsbetrouwbaarheid zou kunnen voldoen. Er werden twee objectieve toetsen ter meting van (aspecten van) de schrijfvaardigheid gereed gemaakt (OA en OB), twee interlineaire (IA en IB), een spellingtoets (SP) en een interpunctietoets (IP).

4. *Validatie- en controle-validatie-onderzoek*

4.1. *Het validatie-onderzoek*

Bij 211 leerlingen der zesde klas van 7 over het gehele land verspreide scholen – waarvan de taalprestaties niet afweken van die van de CITO-schooltoets-populatie (± 60.000 ll.) werd het validatie-onderzoek uitgevoerd.

Als *criterium* werden 5 opstellen per leerling afgenomen, geschreven over 5 vastgestelde onderwerpen. Elk opstel werd door 5 beoordelaars globaal beoordeeld. Zo ontstonden 5 × 5 scores per leerling, die tesamen het opstelcriterium (OC) vormden (6).

Twee groepen predictoren werden onderscheiden. Ten eerste de in het project 0123 ontwikkelde *experimentele predictoren*: objectieve toetsen (OA, OB), interlineaire toetsen (IA, IB) en een interpunctie-toets (IP). Ten tweede de reeds lang in het project Schooltoetsen Basisonderwijs gehanteerde instrumenten: *traditionele predictoren.* Het betrof een spellingtoets (SP), een stilleestoets (SL) en een toets 'gemengde taalopgaven', voornamelijk bestaand uit semantische opgaven (GT).

4.2. *Analyse van het opstel-criterium*

De voor het onderzoek noodzakelijke totstandkoming van een betrouwbaar en valide criterium op bovenomschreven wijze nodigde uit tot een variantie-analytische bewerking van de resulterende 211 × 5 × 5 cijfers. Een aantal hypothesen betreffende te verwachten effecten werd geformuleerd. In het volgende worden de resultaten van de nadere analyse van het opstelcriterium in het kort besproken. Uitgangspunt is de in tabel 3 weergegeven variantie-analyse van de opstelcijfers.

Om enig inzicht te verkrijgen in de orde van grootte van de verschillende effecten, is een schatting van de grootte der verschillende variantiecomponenten verricht[7].

De grootste componenten zijn: Leerlingen (69,65) en residu (67,17). Deze zijn praktisch

Tabel 3. Variantie-analyse van de opstelcijfers*

Variantie-bron	Kwadratensom	Vrijheidsgr.	Gemiddelde kwadratensom	F-ratio	Quasi F-ratio***
leerlingen	422896,23	210	2014,79		7,37**
onderwerpen	33262,69	4	8315,67		7,08**
beoordelaars	63314,47	4	15828,62		14,53**
leerl. × ond.	182523,47	840	217,29	3,23**	
leerl. × beoord.	103683,29	840	123,43	1,84**	
onderw. × beoord.	16383,35	16	1023,96	15,24**	
residu	225701,69	3360	67,17		
Totaal	1047765,19	5274			

* De opstelcijfers zijn bij de uitvoering der variantie-analyse om verwerkings-technische redenen alle met 10 vermenigvuldigd.
** Significant op 1%-niveau.
*** Voor de hoofdeffecten zijn Quasi-F-ratio's berekend (zie WINER 1972, p. 200 e.v.).

even groot. De interactie leerlingen × onderwerpen (30,02) is ook nog aanzienlijk; deze interactie behoeft echter geenszins als ongewenst te worden opgevat: wij komen hierop terug bij de behandeling van hypothese 5. Twee componenten kunnen nog als substantieel worden aangemerkt: beoordelaars (13,98) en leerlingen × beoordelaars (11,25); dit laatste is opmerkelijk, omdat juist zoveel mogelijk getracht is deze interactie te minimaliseren door b.v. de identificatiegegevens van de opstellen te verwijderen en beoordelaars te kiezen die de leerlingen niet kenden. Tenslotte zijn de componenten onderwerpen (6,77) en onderwerpen × beoordelaars (4,53), hoewel significant, verwaarloosbaar in vergelijking met de eerder genoemde componenten.

De belangrijke leerlingenvariantie en de aanzienlijke leerlingen × onderwerpen-variantie kunnen wij als bedoelde of gewenste variantie beschouwen. Doch de andere componenten – vooral de nog substantiële beoordelaars- en leerlingen × beoordelaars-variantie – zijn uit het oogpunt van objectiviteit in het algemeen niet gewenst. Hetzelfde geldt uiteraard voor de belangrijke residu-component. In ons geval behoeft een deel van de 'niet gewenste' componenten echter niet als betrouwbaarheid-verlagend te worden gezien, als men de wijze waarop het criterium tot stand kwam in aanmerking neemt.

Betrouwbaarheid van het opstelcriterium: de beoordelingsbetrouwbaarheid (jury-equivalentie) kan geschat worden met de formule[10] van Maxwell en Pilliner (1968).

Het resultaat is een coëfficiënt van .939. Dit betekent dat als een tweede – net zo competente – groep van 5 beoordelaars de opstellen zou beoordelen, de correlatie tussen de twee series totaalscores ongeveer .94 zou zijn. De scorebetrouwbaarheid (jury × taak-equivalentie) kan geschat worden door gebruik te maken van de formule voor de coëfficiënt α. Dit levert een correlatiecoëfficiënt van .885 op. Deze scorebetrouwbaarheid is een schatting van de correlatie die verwacht mag worden tussen de somoordelen van een nieuwe jury van vijf over vijf nieuwe opstellen (over nieuwe onderwerpen) en de originele somoordelen van de huidige jury van vijf over de huidige vijf opstellen. De beoordelingsbetrouwbaarheid van een éénmaal beoordeeld opstel is de (gemiddelde) correlatie tussen de oordelen van verschillende beoordelaars over hetzelfde opstel. De 50 berekende correlaties lopen van .36 tot .78. Gemiddeld is de samenhang .58. Dit is een aanzienlijk hogere co-

Tabel 4. Gemiddelden en standaardafwijkingen van de vijf beoordelaars bij elk van de vijf opstellen.

Opstel	Beoord. A	Beoord. B	Beoord. C	Beoord. D	Beoord. E	Totaal
1. Gem.	5,99	6,40	6,39	5,91	5,61	30,29
S.A.	1,24	1,11	1,20	1,49	1,42	5,46
2. Gem.	6,09	6,18	6,50	6,11	5,66	30,54
S.A.	1,24	1,17	1,22	1,52	1,41	5,41
3. Gem.	5,89	6,60	6,38	5,97	5,35	30,17
S.A.	1,40	1,07	1,27	1,60	1,60	5,85
4. Gem.	5,11	5,77	6,19	4,83	5,24	27,15
S.A.	1,20	1,31	1,06	1,66	1,48	5,34
5. Gem.	5,77	6,17	6,44	5,29	5,55	29,18
S.A.	1,17	0,99	1,08	1,61	1,42	4,95
Totaal	28,85	31,12	31,90	28,11	27,41	
Gem.	5,77	6,22	6,38	5,62	5,48	147,28
S.A.	1,30	1,19	1,17	1,66	1,48	22,37

efficiënt dan die welke resulteerde uit het oriënterend criteriumonderzoek. Toen werden bij eerste en bij tweede beoordeling respectievelijk coëfficiënten van .45 en .48 geconstateerd (zie 2.3.1.). De score-betrouwbaarheid voor een éénmaal beoordeeld opstel is de (gemiddelde) correlatie tussen de oordelen van verschillende individuele beoordelaars over verschillende opstellen, dus: verschillende titels. De 200 berekende correlaties lopen van .10 tot .62. Gemiddeld is de samenhang .40. Deze coefficiënt kunnen wij niet vergelijken met resultaten uit de criteriumstudie, waarin slechts één opstel per leerling aanwezig was. Wel kan de coëfficiënt van .40 worden vergeleken met die uit het onderzoek van Godshalk, Swineford en Coffman (1966). Zij constateerden een score-betrouwbaarheid van (gemiddeld) .26; daarbij zij echter opgemerkt, dat hun onderzoek gericht was op de schrijfvaardigheid van leerlingen op het niveau eind High School.

De relevantie van een aantal effecten voor het criterium

In het volgende worden de resultaten van een aantal hypothese-toetsingen gegeven en het belang van de geconstateerde effecten voor zowel het project 0123 als de 'examensituatie in Nederland'.

1. Er is een over-all verschil tussen de scores voor de opstellen over verschillende onderwerpen. De cijfers verschillen van onderwerp tot onderwerp. Als de leerlingen allemaal over elk van een aantal onderwerpen een opstel schrijven en als hun score de somscore van deze opstellen is, dan zijn de verschillen, toe te schrijven aan het onderwerp, niet relevant. De betekenis van ons criterium wordt dan ook niet aangetast door dit effect. De verschillen tussen de scores als gevolg van de factor 'onderwerp' zijn in bepaalde situaties echter wèl belangrijk, b.v. in examensituaties, waarin men gewend is de leerlingen te laten kiezen uit een aantal onderwerpen. In zo'n situatie zal de score die een leerling krijgt mede afhankelijk zijn van het onderwerp, dat hij kiest.

Tabel 4 geeft inzicht in de grootte van een aantal effecten.

2. De meeste verschillen tussen de totaalscores van de opstellen zijn significant. Vooral het verschil tussen de 'gemakkelijkste' opgave (op-

stel 2, fantasie-opstel) en de 'moeilijkste' opgave (opstel 4, een 'verhandeling') is vrij groot (0,68 punt gemiddeld). De geconstateerde verschillen tasten de betekenis van ons criterium niet aan, daar alle leerlingen over alle onderwerpen schreven; voor andere situaties – b.v. examen-situaties – zijn de geconstateerde effecten van belang.

3. Er is verschil tussen het beoordelingsgedrag der verschillende beoordelaars: hun gemiddelden verschillen en de verdeling der cijfers verschilt. Het geconstateerde verschil in gemiddelden tussen beoordelaars tast de betekenis van ons criterium niet aan: alle beoordelaars beoordeelden alle opstellen. Wèl wijzen wij op het grote belang van dit effect voor examensituaties, waarin verschillende groepen van leerlingen door verschillende examinatoren worden beoordeeld. In die situatie zal de score die een leerling krijgt in belangrijke mate afhankelijk zijn van de ene docent, die zijn opstel beoordeelt.

De verdeling der cijfers verschilt, d.w.z.: de verschillen tussen de standaardafwijkingen zijn significant voor alle gevallen, behalve het geval B–C.

4. Er is een significante interactie tussen leerlingen en onderwerpen. Sommige leerlingen schrijven over sommige onderwerpen betere opstellen, terwijl andere leerlingen over andere onderwerpen betere opstellen schrijven. Men zou dit feit kunnen beschouwen als een argument vóór het aanbieden van een lijst met onderwerpen, waaruit de leerling zèlf zijn keuze maakt, zoals op de meeste examens pleegt te geschieden. Het is echter de vraag of de leerling in staat is te beoordelen wèlk onderwerp hèm voordeel zal opleveren. Bovendien zou – àls de leerlingen dit zouden kunnen – tezelfdertijd een kwantitatief niet groter, maar niet te verwaarlozen foutenbron worden geïntroduceerd, n.l. de reeds hiervoor ter sprake gebrachte factor 'onderwerpen'.

Men zou uiteraard kunnen stellen dat het proportioneel verschil tussen de besproken interactie (30,02) en het hoofdeffect 'onderwerpen' (6,77) in die orde van grootte ligt, dat gekozen dient te worden voor het aanbieden van een zo groot mogelijke verscheidenheid van onderwerpen, opdat de leerling kieze – ten eigen voordeel. Dàn echter dient het kunnen kiezen van het meest bevoordelende (= het meest bij het eigen schrijftalent passende) onderwerp ook expliciet tot onderwijsdoel te worden gemaakt. Hierop is in principe niets tegen: tenslotte schrijft ook de professionele schrijver over onderwerpen die hem 'liggen', waarover hij de nodige voorkennis heeft enz. Men zou echter de doelstellingen voor het onderwijs in de produktiefschriftelijke taalvaardigheid ook kunnen afleiden uit de taak, die de gemiddelde abituriënt van het basisonderwijs of het voortgezet onderwijs moet kunnen verrichten. Hoe dan ook: het is zeer de vraag of op dit moment reeds voldoende het 'kunnen kiezen van het beste onderwerp' als onderwijsdoelstelling wordt gezien. Pas als hieromtrent duidelijkheid bestaat, kan keuze gemaakt worden tussen het aanbieden van verscheidene onderwerpen en het aanbieden van één verplicht onderwerp.

De betekenis van ons criterium wordt door de gesignaleerde interactie tussen leerlingen en onderwerpen niet ernstig aangetast: wij hebben ernaar gestreefd een gevarieerd aanbod aan onderwerpen te doen, opdat iedere leerling de mogelijkheid zou krijgen over hem passende onderwerpen te schrijven. De onderwerpen zijn in overeenstemming met wat in verschillende moderne taalmethoden geacht wordt doelstelling te zijn van het onderwijs in het produktief-schriftelijk taalgebruik.

5. Er is een significante interactie tussen beoordelaars en onderwerpen. Hoewel dit effect bijna verwaarloosbaar is in vergelijking met de andere effecten, is het verschijnsel relevant voor vele praktische (examen-) situaties in Nederland.

Uit de variantie-analyse blijkt – geheel in overeenstemming met wat de literatuur te dien aanzien meldt – dat vele effecten invloed hebben op de totale opstel-scores. Wel gelden de hierboven besproken bevindingen voor de beoordeling van opstellen van leerlingen der zesde klassen basisonderwijs, maar het is waarschijnlijk, dat dezelfde effecten in sterkere of minder sterke mate een rol spelen bij de totstandkoming van bijv. eind-

Tabel 5. Correlaties tussen de 5 experimentele predictoren, de 3 traditionele predictoren en het onderwijskundig relevant criterium OC.

	OA	OB	IA	IB	IP	SP	GT	SL	OC
OA	—	.697	.732	.665	.593	.535	.681	.670	.682
OB	.697	—	.616	.672	.654	.447	.737	.721	.667
IA	.732	.616	—	.703	.553	.486	.656	.605	.639
IB	.665	.672	.703	—	.509	.409	.715	.721	.665
IP	.593	.654	.553	.509	—	.473	.545	.454	.590
SP	.535	.447	.486	.409	.473	—	.457	.323	.488
GT	.681	.737	.656	.715	.545	.457	—	.815	.625
SL	.670	.721	.605	.721	.454	.323	.815	—	.574
OC	.682	.667	.639	.665	.590	.488	.625	.574	—

examen-opstelcijfers (Jansen en Wesdorp, 1973). Is het mogelijk bij examens-op-grote-schaal een enigszins betrouwbare opstelscore te verkrijgen? Het antwoord moet zijn: nauwelijks, want de maatregelen, die noodzakelijk zijn om de storende effecten te elimineren of tot acceptabele proporties terug te brengen, zijn organisatorisch en financieel slechts haalbaar, als men zich veel moeite en hoge kosten wil getroosten (zie hierover 5.1.).

Het in het project 0123 tot stand gekomen criterium kon door de beperkte grootte van de steekproef, door de aanwezigheid van honoreringsmogelijkheden voor de beoordelaars en door de terbeschikkingstelling van voldoende afnametijd per leerling wèl volgens een procedure ontstaan, die garandeerde dat de betrouwbaarheid van de opstel-totaal-score door de genoemde effecten zo weinig mogelijk werd aangetast. Het criterium voldoet aan hoge eisen van betrouwbaarheid en is ook – gezien de wijze waarop het tot stand kwam – qua validiteit weinig aanvechtbaar.

4.3. *De validiteit van afzonderlijke toetsen en van toetscombinaties*

Uitgangspunt voor de toetsing van een aantal hypothesen betreffende de validiteit van afzonderlijke toetsen is de in tabel 5 weergegeven correlatiematrix.

De voornaamste resultaten laten zich als volgt samenvatten:

De validiteiten der toetsen variëren, zij het minder sterk dan verwacht werd. De geheel op meting van de stelvaardigheid gerichte experimentele toetsen OA, OB, IA en IB slagen daarin i.h.a. beter dan de niet met dat doel geconstrueerde traditionele stilleestoets (SL). Doch de verschillen met de andere traditionele toets GT zijn niet significant. Enigszins onverwacht zijn ook de vrij hoge validiteiten van op zeer specifieke subvaardigheden als spellen en interpungeren gerichte toetsen (SP en IP).

T.a.v. combinaties van toetsen laten zich de resultaten als volgt samenvatten:

Combinaties van twee experimentele toetsen zijn i.h.a. meer valide dan combinaties van twee traditionele toetsen. De validiteiten voor tweetoetscombinaties variëren van .688 (OB+IP) tot .732 (OA+IB).

Combinaties van drie experimentele toetsen variëren van .731 (OA+IA+IP) tot .757 (OA+IB+IP). Combinaties van drie experimentele toetsen prediceren het opstelcriterium beter dan de combinatie van de drie traditionele predictoren (SL+GT+SP): .673.

Alle combinaties van vier predictoren voorspellen het criterium betrekkelijk goed. De validiteiten van vier-toets-predictoren variëren van .706 tot .765.

4.4. *Verschillen in validiteit tussen toetscombinaties en opstellen als predictoren van een vier-opstellen-criterium.*

Daar wij de beschikking hadden over een 5-opstellen-criterium kon een vergelijking gemaakt worden tussen de predictieve validiteit van toetscombinaties en die van een opstel ten opzichte van een 4-opstellen-criterium. Daartoe is beurtelings elk opstel als predictor beschouwd van het (resterende) 4-opstellen-criterium. De resultaten konden worden vergeleken met de validiteitscoëfficiënten van toetscombinaties t.o.v. deze 4-opstellen-criteria. (N.B.: de predicerende opstellen zijn 5 × gescoord, het criterium eveneens).

De validiteit van opstel 1 t.o.v. het resterende 4-opstellen-criterium blijkt .698 te zijn, die van opstel 2 .720, die van opstel 3 .787, die van opstel 4 .672 en die van opstel 5 .755. De mediane validiteit van één (5 × gescoord) opstel voor de voorspelling van een 4-opstellen-criterium (5 × gescoord) is .720. Een dergelijke validiteit is door toetscombinaties moeilijk substantieel te overtreffen. Toetscombinaties bereiken een validiteit van ± .71 ten opzichte van een 4-opstellen-criterium. Toetscombinaties prediceren zulk een criterium wèl beter dan een éénmaal gescoord opstel: dat bereikt een validiteit van ± .60.

4.5. *Controle-validatie-onderzoek*

Eén jaar na het validatie-onderzoek werd bij een steekproef van 120 leerlingen op 20 scholen, verspreid over het gehele land, en als totale groep qua taalprestaties niet afwijkend van de CITO-schooltoets-populatie (± 70.000 ll.), een controle-validatie-onderzoek verricht.

Als *criterium* fungeerde weer de somscore van 5 opstellen, elk 5 maal beoordeeld, echter nu niet door steeds dezelfde beoordelaars, doch door aselect gekozen teams van 5 uit een totaal van 25, waarbij ervoor werd gezorgd dat elk team éénmaal elke leerling beoordeelde.

Als *predictoren* werden slechts de 3 experimentele toetsen OA, IB en IP gebruikt, benevens de traditionele toets SP.

Alle validiteitscoëfficiënten van enkele toetsen en van combinaties van toetsen lagen hoger dan de bij het validatie-onderzoek geconstateerde. De validiteit der enkele toetsen bedroeg ± .70, die van combinaties van twee toetsen varieerden van .63 (IP+SP) tot .765 (OA+IP). Combinaties van drie varieerden van. 742 (OA+IP+SP) tot .781 (OA+IB+IP).

Ook in het controle-validatie-onderzoek werd een vergelijking gemaakt tussen de predictieve validiteit t.o.v. een 4-opstellen-criterium van een enkel opstel (5 × gescoord) en toetscombinaties. Nu bleek dat toetscombinaties wèl superieur waren aan enkele opstellen bij de predictie van 4-opstellen-criteria. De validiteit van opstellen varieerde van .59–.68, die van 2-toets-combinaties van .70–.74, die van 3-toets-combinaties van .73–.77. Deze van het validatie-onderzoek verschillende resultaten kunnen verklaard worden uit de andere wijze waarop het criterium tot stand kwam. Tijdens het validatie-onderzoek was het criterium door de medewerking van steeds dezelfde 5 docenten homogener; het onderzoeksmateriaal was m.b.t. het onderzochte validiteits-verschil 'gecontamineerd'; de resultaten van het controle-validatie-onderzoek zijn minder betwijfelbaar.

5. *Discussie*

Zonder in te gaan op de vele details van het zeer summier beschreven onderzoek, willen wij toch twee belangrijke punten in deze paragraaf aansnijden, n.l. de resultaten van het project betreffende het opstel-beoordelingsprobleem en die op het gebied van de toetsconstructie.

5.1. *Opstelbeoordeling*

Mogen wij erin geslaagd zijn een binnen ons onderzoek realiseerbaar beoordelingssysteem te vinden, dat een zekere beoordelingsbetrouwbaarheid en -validiteit zou garanderen, voor de praktijk buiten het onderzoek staan wij eigenlijk met lege handen. Immers, geen enkele der besproken beoordelingsprocedures is in de Nederlandse praktijk – b.v. de examen-praktijk bij het MAVO, HAVO en VWO – zonder meer realiseerbaar. 'Analytische' beoordeling en 'globale' jury-beoordeling zijn slechts realiseerbaar, als meer beoordelingstijd ter beschikking komt, wat in de huidige situatie nauwelijks haalbaar lijkt. Hoogstens kan op dit moment met de voorhanden gelden een globale jury-beoordeling op kleine schaal worden georganiseerd. Daarbij zouden de docent-examinator en één externe beoordelaar – afkomstig uit een door een centraal bureau beheerde 'pool' van competente beoordelaars – elk onafhankelijk een globaal oordeel kunnen uitspreken. In bepaalde gevallen – b.v. bij een scoreverschil van 2 of meer punten of bij een kleiner verschil in de buurt van de grens voldoende-onvoldoende – zouden een derde en vierde beoordelaar uit de pool kunnen worden ingeschakeld. Een dergelijk systeem leidt niet tot voldoende betrouwbare jury-oordelen, daar jury's van twee veelal te klein zijn om enige garanties dienaangaande te bieden. Toch zou het een stap in de goede richting zijn: het is te verwachten dat het percentage onjuiste beslissingen in elk geval wordt verminderd. Bovendien heeft een dergelijk centraal georganiseerd scoringssysteem voordelen van andere aard. Door publicatie van opstellen die tot nogal uiteenlopende oordelen aanleiding geven, kan de discussie over de doelstellingen van het stelonderwijs gestimuleerd worden. Op dit moment ontbreekt elke informatie omtrent de dimensies waarlangs op eindexamens opstellen worden beoordeeld.

Andere beoordelingsmethoden hebben slechts zeer beperkte voordelen (beoordelingsschalen) of zijn volstrekt onrealiseerbaar (computerbeoordeling). Dat betekent, dat wij slechts het volgende kunnen constateren: Er is geen praktische, eenvoudige en 'totale' oplossing voor het opstelbeoordelingsprobleem. In het Nederlandse toelatings- maar vooral eindexamensysteem is geen der bestaande opstelbeoordelingstechnieken op eenvoudige wijze op ideale wijze te integreren: de in het systeem passende methoden leveren onbetrouwbare leerlingscores op; de meer betrouwbaarheid garanderende methoden zijn vooralsnog om financiële en organisatorische redenen binnen het systeem niet volledig realiseerbaar. Dat deze conclusie niet alleen geldt voor 'eindexamens' van het basisonderwijs, maar zeker ook voor eindexamens van het voortgezet onderwijs, hopen wij door onze literatuurstudie (zie 2.2.) en door enig eigen onderzoek (Jansen en Wesdorp, 1973) aannemelijk gemaakt te hebben.

Als het mogelijk zou zijn meer docenten-arbeid te investeren in de beoordeling van examen-opstellen, dan zouden twee technieken in aanmerking komen, n.l. 1. de beoordeling met behulp van een analytisch schema (behandeld in 2.2.2.) en 2. beoordeling door een jury van 'globaal' beoordelende docenten (zie 2.2.3.). Beide methoden hebben, afgezien van de financiële en organisatorische problemen die zij oproepen, hun voor- en nadelen.

Ten aanzien van de analytische beoordeling kan gesteld worden, dat het zeer moeilijk zal zijn een voor alle betrokkenen aanvaardbaar waarderingsschema op te stellen. In wezen is de opstelling van een dergelijk schema een flinke stap in de goede richting: het vergt doelstellingen-explicitering. Maar juist omdat de doelstellingen van het onderwijs in de produktief-schriftelijke taalvaardigheid, zeker op het hogere niveau (eind MAVO, eind HAVO, eind VWO)

Tabel 6. De intercorrelaties tussen de officiële cijfers voor 103 examenopstellen HAVO (1972) en de cijfers van 8 juryleden voor die opstellen. Gemiddelde en standaardafwijking per beoordelaar. Percentages onvoldoenden. (Onderzoek JANSEN en WESDORP, 1973).

	Officieel examen-opstelcijfer	Juryleden								Jury-gemid-delde
		A	B	C	D	E	F	G	H	
Officieel ex. cijfer	—	.43	.49	.46	.37	.47	.42	.44	.35	.60
A	.43	—	.52	.33	.27	.42	.57	.56	.35	.71
B	.49	.52	—	.43	.52	.61	.49	.61	.46	.79
C	.46	.33	.43	—	.36	.37	.40	.39	.35	.66
D	.37	.27	.52	.36	—	.43	.44	.54	.25	.66
E	.47	.42	.61	.37	.43	—	.50	.53	.28	.71
F	.42	.57	.49	.40	.44	.50	—	.51	.38	.75
G	.44	.56	.61	.39	.54	.53	.51	—	.39	.79
H	.35	.35	.46	.35	.25	.28	.38	.39	—	.62
Jurygemiddelde	.60	.71	.79	.66	.66	.71	.75	.79	.62	—
Gem.	6.08	6.03	5.85	5.45	5.27	5.85	5.50	6.20	6.21	5.79
Standaardafw.	0.89	1.17	0.75	1.24	1.06	0.96	1.02	0.98	1.14	0.74
% onvoldoenden (=5½ of minder)	29	33	29	35	57	34	49	27	30	45

op dit moment zeer vaag omschreven zijn, heerst er bij alle betrokkenen een zeer duidelijke heterogeniteit van opvattingen. Derhalve lijkt de opstelling van een analytisch beoordelingsschema een moeilijke taak. Overhaaste, door het 'bevoegd gezag' bewerkstelligde doelstellingen-explicitering zal weerstanden opwekken. De validiteit van de met behulp van een analytisch schema tot stand gekomen oordelen zal altijd worden betwijfeld. Voorts is controle op de hantering van een dergelijk schema een vereiste.

Ten aanzien van globale jury-beoordeling gelden minder bezwaren. De vraag is echter: hoe groot dient een jury te zijn om de beschikking te krijgen over een opstel-score, waarop men staat kan maken? Bij de schatting van het aantal benodigde juryleden kan men uitgaan van de gegevens in de volgende tabel, die enige resultaten van een onderzoek van Jansen en Wesdorp (1973) samenvat, waarin de officiële cijfers van een steekproef van 103 eindexamenopstellen HAVO werden geconfronteerd met de door een competente jury van 8 gegeven cijfers.

De gemiddelde intercorrelatie tussen de 8 juryleden in dit onderzoek bleek .44 te zijn. Indien men een beoordelingsbetrouwbaarheid van .80 wil bereiken, is een jury van 5 noodzakelijk. Stelt men zijn eisen hoger en wil men een beoordelingsbetrouwbaarheid van .85 bereiken, dan zijn minstens 7 onafhankelijke juryleden nodig[8]. De validiteit van een dergelijk jury-oordeel behoeft minder weerstanden op te wekken dan die van een volgens een (opgelegd) analytisch beoordelingsschema tot stand gekomen score. Overigens zal uiteraard ook bij jury-beoordeling controle moeten worden uitgeoefend op de uitvoering. De totstandkoming der oordelen dient onafhankelijk plaats te vinden. Ook het voorkomen van zeer extreme 'dissidente' beoordelaars dient te worden gesignaleerd, opdat niet bepaalde groepen van leerlingen, die door een jury worden beoordeeld, waarin één of meer dissidente beoordelaars zitting hebben, gedupeerd worden.

5.2. *De toetsconstructie*

De in het project geconstrueerde experimentele schrijfvaardigheidstoetsen hebben blijkens de

resultaten van het validatie- en controle-validatie-onderzoek een *predictieve validiteit* van ± .70 ten opzichte van een opstelcriterium. Vergeleken met de blijkens buitenlands onderzoek 'haalbaar' te achten validiteitscoëfficiënten (voor beide typen toetsen eveneens op ± .70 geschat) zijn wij dus geslaagd. Wel moet verhoging van deze validiteitscoëfficiënten mogelijk zijn, althans voor toetsen voor het niveau eind basisonderwijs, waarop de leerlingen nog een zeer heterogene groep vormen. Ook als de validiteitscoëfficiënten van de toetsen worden vergeleken met die van éénmaal of zelfs vijfmaal gescoorde opstellen, blijken de toetsen een redelijk figuur te slaan. In onze opzet zijn wij dus redelijk geslaagd: de experimentele toetsen behalen een validiteitsniveau, dat haalbaar geacht werd, en dat superieur is aan dat van enkele opstellen.

Toch zijn, zeker met betrekking tot de *inhoudelijke validiteit* van de toetsen nog wel enkele kritische kanttekeningen te maken. De toetsen zijn slechts 'indirecte' meetinstrumenten. Ze meten een aantal subvaardigheden, die blijkens onze (summier beschreven) analyse aan de complexe produktief-schriftelijke taalvaardigheid ten grondslag liggen. Dat deze analyse op zichzelf niet geheel invalide is, blijkt uit de vrij hoge predictieve validiteit, die de geconstrueerde instrumenten bleken te bereiken. Toch is deze analyse niet compleet en voor discussie vatbaar. Wat is immers het geval? Gedeeltelijk door de relatieve oppervlakkigheid van de analyse – in betrekkelijk korte tijd moest een volledig helder 'zicht' op de doelvaardigheid worden verkregen – gedeeltelijk door de waarschijnlijk zeer heterogene opvattingen omtrent de doelstellingen van ons schrijfvaardigheidsonderwijs, zijn de eruit voortkomende toetsen voor velerlei kritiek vatbaar. Zo is het mogelijk, dat in het onderwijs weinig waarde wordt toegekend aan b.v. het 'formele, correcte schrijven', waaraan de toetsen vooral aandacht schenken. Zo is b.v. ook weinig aandacht geschonken aan het 'communicatieve aspect', hoewel sommige items daaraan wel gewijd zijn. Dit gebrek wordt veroorzaakt door het feit dat de specificaties voor de toetsen zijn ontstaan in een periode, waarin althans voor het Nederlandse basisonderwijs de waarde van het 'communicatieve schrijven' zeker nog niet in discussie was. Hoe dan ook: de tekorten in de toetsen zijn verklaarbaar, en waarschijnlijk ook aanvulbaar. Wel zijn wij van mening dat de toetsen de doelstellingenopvattingen in het onderwijs dienen te volgen, en niet dienen vooruit te lopen op de verschuivingen in het doelstellingendenken. Een doelstellingenonderzoek is daarom een eerste vereiste.

Maar geheel los van het bovenstaande, en meer betrekking hebbend op de constructie-procedure, die in het project 0123 is gevolgd, staat de hierboven in de eerste plaats geformuleerde kritiek: de analyse en de concretisering in toetsen zijn vrij oppervlakkig en snel tot stand gekomen; de toetsen zijn in ongeveer een half jaar tijds geconstrueerd èn gepretest. Een team met meer ervaring en meer tijd ter beschikking zou waarschijnlijk op een doordachtere wijze vorm gegeven hebben aan de items en zou waarschijnlijk bij de operationalisering van sommige subvaardigheden zijn geslaagd, waar nu gefaald werd. Met andere woorden: meer inzicht in de te meten vaardigheid (dus: een betere, op de huidige doelstellingenopvattingen geënte analyse) en meer toetsconstructie-ervaring zouden hebben kunnen leiden tot verhoging van de inhoudelijke validiteit der toetsen. Helaas zijn in het project 0123 deze voorwaarden niet vervuld. De toetsen hebben hun gebreken, en kunnen, als eerste in hun soort, niet aan de genoemde hoge eisen voldoen. Wij zien twee wegen om tot een verbeterde visie op de te meten vaardigheid te komen. Nauwkeurige doelstellingenanalyse, waarbij vrij recente publicaties op dit terrein behulpzaam kunnen zijn en de meningen van belangrijke groepen betrokkenen t.a.v. de doelstellingen worden geïnventariseerd, zou een eerste mogelijkheid zijn. Linguistische analyse van de (relatieve) belangrijkheid van bepaalde subvaardigheden, die een rol spelen bij het zich schriftelijk uitdrukken,zou een tweede mogelijkheid zijn.

De toetsen zijn ontstaan vanuit de problematiek rond de Schooltoetsen Basisonderwijs. Ze zijn bruikbaar binnen de gegeven situatie, die een eindtoetsing van het basisonderwijs wenselijk

maakt en die een vrije doorstroom van leerlingen naar diverse vormen van voortgezet onderwijs praktisch onmogelijk maakt. Ze zijn geschikt voor 'selectief' en 'evaluatief' gebruik. Het is naar onze mening ook zinvol de objectieve schrijfvaardigheidstoets onderdeel te laten zijn van de Schooltoets, ondanks de bezwaren die er zijn tegen de toetsvorm in het algemeen of de wellicht gebrekkige inhoudelijke validiteit van de schrijfvaardigheidstoets in het bijzonder. Zolang de Schooltoets qua status erg veel lijkt op een afsluitend examen – zonder dat een bevredigende regeling getroffen is voor de 'examinering' van al die doelvaardigheden, die in het geheel niet of liever niet voor objectieve toetsing in aanmerking komen – is het zaak de Schooltoets zoveel mogelijk diverse onderwijsdoelstellingen te doen representeren. Uit het bovenstaande volgt: a) Wij zijn niet gelukkig met de bestaande situatie, waarin van vrije doorstroming van leerlingen van het basisonderwijs naar het voortgezet onderwijs geen sprake is. Wij staan daarin niet alleen: praktisch de gehele onderwijswereld is ongelukkig met de bestaande situatie; vooralsnog zien wij geen snelle oplossing van het probleem, al bieden experimenten met vormen van interne differentiatie en plannen t.a.v. de middenschool misschien enige hoop. b) Wij zien de opname van een schrijfvaardigheidstoets in het programma Schooltoetsen Basisonderwijs als noodmaatregel. Zolang geen andere examenfilosofie zijn kans krijgt, waarin de expliciete toetsing van alle doelstellingen als prestatie-onderdeel niet meer nodig is (vgl. De Groot, 1968) of zolang het doorstromingsprobleem niet structureel is opgelost, zullen schoolvorderingentoetsen – waarvan de Schooltoetsen Basisonderwijs slechts één voorbeeld zijn – aan het eind van het basisonderwijs nodig zijn. Slechts in die noodsituatie, waarbij het praktisch onmogelijk is op andere wijze een eindtoetsing voor de schrijfvaardigheid te garanderen, achten wij het gebruik van objectieve schrijfvaardigheidstoetsen acceptabel. In 1973 is derhalve een objectieve schrijfvaardigheidstoets in de CITO-schooltoets opgenomen. Niet de predictieve validiteit t.o.v. een opstelcriterium, die blijkens ons onderzoek vrij acceptabel geacht mag worden, is daarbij het hoofdargument geweest. De Schooltoets zou, zo was de redenering, door opname van een dergelijke toets een iets breder gebied van doelstellingen representeren dan hij deed. Tot nog toe werden de stilleesvaardigheden, de spellingvaardigheid (facultatief), de vaardigheid op het gebied van de traditionele grammatica (facultatief) en een aantal vooral semantische vaardigheden (in de subtoets Gemengde Taalopgaven) gemeten. De schrijfvaardigheidstoets verruimt – ondanks zijn gebreken – naar onze mening het totale door de Schooltoets bestreken taalvaardigheidsgebied. Verwacht wordt dat de expliciete aandacht die het stelonderwijs in de eindtoets krijgt een positieve invloed op het curriculum zal hebben, m.a.w. dat in vergelijking met voorheen – toen géén directe of indirecte toetsing van de schrijfvaardigheid plaats vond – meer aandacht aan het stelonderwijs zal worden geschonken. Op zijn minst wordt verwacht dat de stelvaardigheid niet geheel zal worden veronachtzaamd, wat zonder expliciete aandacht wellicht wel het geval zou zijn.

De interlineaire schrijfvaardigheidstoetsen zijn door hun scoringssysteem niet geschikt voor onderzoekingen op grote schaal, zoals de Schooltoets. Zij lijken echter bijzonder geschikt voor onderzoekingen op kleinere schaal, waarbij de beschikbaarheid van een betrouwbaar instrument vereist wordt, omdat belangrijke beslissingen t.a.v. de leerlingen moeten worden genomen. Temeer daar interlineaire toetsen veel sneller – ook door niet professionele toetsconstructeurs – zijn te vervaardigen, lijken ze het aangewezen instrument om daar, waar voor een gering aantal leerlingen behoefte bestaat aan een schrijfvaardigheids-meetinstrument, dienst te doen. Zo zouden interlineaire toetsen kunnen functioneren in toelatingsprocedures, in determinatieprocedures in de brugklas. Interlineaire toetsen zouden ook een rol kunnen spelen bij de evaluatie van de produktief-schriftelijke vaardigheid in een vreemde taal, b.v. tijdens het schoolonderzoek.

Noten

1. Er is geen noodzaak om hiervan uit te gaan. Met name is de alternatieve opvatting dat 'produktief-schriftelijke taalvaardigheid' niet een eenheid is, maar eventueel beter gesplitst kan worden naar verschillende soorten taken (en dus soorten schrijfvaardigheden) zeker houdbaar. In dat geval is de onder 4 hieronder genoemde 'non-equivalentie van taken' geen foutenbron maar een systematische factor.

2. Een door de Educational Testing Service (1957) uitgegeven test die gebruik maakt van een composition-scale, in tegenstelling tot de ook door E.T.S. uitgegeven STEP-writing-test en vele elders in Amerika verschenen objectieve schrijfvaardigheidstoetsen.

3. Deze termen vergen enige toelichting. Als we uitgaan van het verschil tussen stabiliteit en equivalentie, zoals gedefinieerd door Cronbach (1960), dan betreft het hier niet alleen de equivalentie van items of toetsen (zoals in de testleer), dus in ons geval: equivalentie van taken, maar ook equivalentie van jury's. De betrouwbaarheden, die we hierboven introduceerden zijn dan te beschrijven als *jury-equivalentie* en *jury* × *taak-equivalentie*.
   Onder de '*beoordelingsbetrouwbaarheid*' of '*jury-equivalentie*' moet dan worden verstaan: de correlatie tussen het somoordeel van een jury over een aantal opstellen per leerling en het somoordeel van een andere – even competente – jury over diezelfde opstellen per leerling. In de Amerikaanse literatuur wordt deze correlatie de 'reading-reliability' genoemd. Onder de '*scorebetrouwbaarheid*' of '*jury* × *taak-equivalentie*' moet dan worden verstaan: de correlatie tussen het somoordeel van een jury over een aantal opstellen per leerling en het somoordeel van een andere jury over een zelfde aantal opstellen per leerling over àndere onderwerpen. In de Amerikaanse literatuur wordt deze correlatie 'score-reliability' genoemd.

4. Bij deze schatting (m.b.v. de Spearman-Brown-formule) wordt het eventueel effect van de variatie in onderwerpen buiten beschouwing gelaten.

5. Deze schatting (m.b.v. de Spearman-Brown-formule) gaat uit van een scorebetrouwbaarheid van ± .55 voor het oordeel van een jury van 5.

6. Behalve dit voor het onderwijs relevante opstelcriterium werd een linguistisch relevant criterium geconstrueerd en in het validatie-onderzoek betrokken. Het bestond uit een viertal toetsen. Ondanks de belangwekkende resultaten is in dit korte overzicht dit linguistische criterium geheel buiten beschouwing gelaten.

7. Voor de schatting der variantie-componenten wordt verwezen naar bijlage 5 van Wesdorp, 1974.

8. Het is de vraag in hoeverre een dergelijke jurering praktisch realiseerbaar is. Dat overigens beslist *iets* moet worden ondernomen op dit gebied blijkt uit de data in tabel 6 en uit de volgende 2 resultaten:
   a. De standaardmeetfout van een individuele beoordelaar kan geschat worden op 1½ scorepunt.
   b. In ± 30% van de gevallen bleken jury (gemiddeld) en examinator van mening te verschillen over het 'voldoende' of 'onvoldoende' zijn van het opstel.

Literatuur

Block, J., *The Q-sort method in personality assessment and psychiatric research*. Springfield Ill.: Thomas, 1961.

Bonnardel, R., Application de la méthode d'analyse factorielle de Thurstone à l'étude de la notation des copies d'examens. *Le Travail Humain*, 1946, 9, 150–167.

Cronbach, L. J., *Essentials of psychological testing*. New York etc.: Harper and Row, 1960.

Diederich, P. B., J. W. French en S. T. Carlton, *Factors in the judgement of writing ability*. Research Bulletin R.B. 61–65. Princeton, N.J.: Educational Testing Service, 1961.

Godshalk, F. I., F. Swineford en W. E. Coffman, *The measurement of writing ability*. New York: College Entrance Examination Board, 1966.

Groot, A. D. de, Studietoetsen en Examens. *Levende Talen*, 1968, 244, 71–77.

Jansen, G. G. H. en H. Wesdorp, De waarde van eindexamen-opstel-cijfers. Een onderzoek betreffende de opstelcijfers van het H.A.V.O.-eindexamen 1972. *Levende Talen*, 1973, 297, 191–204.

Lamb, H., The English essay in secondary selection of examinations: a comparison of two methods of marking. *British Journal of Educational Psychology*, 1953, 23, 131–133.

Maxwell, A. E. en A. E. G. Pilliner, Deriving coefficients of reliability and agreement for ratings. *The British Journal of Mathematical and Statistical Psychology*, 1968, 21, 105–116.

Page, E. B., Grading essays by computer: Progress report. In: Educational Testing Service: *Proceedings of the Invitational Conference on Testing Problems*. Princeton, N.J.: E.T.S., 1967, 87–100.

Peet, A. A. J. van, *Theorie en praktijk van opstelbeoordeling*. Amsterdam: R.I.T.P., 1970.

Wesdorp, H., *Het meten van de produktief-schriftelijke taalvaardigheid. Directe en indirecte methoden: 'opstelbeoordeling' versus 'schrijfvaardigheids-toetsen'*. Purmerend: Muusses, 1974.

Winer, B. J., *Statistical principles in experimental design*. New York: Mc. Graw-Hill, 1972.

Wiseman, S., The marking of English composition in grammar school selection. *British Journal of Educational Psychology*, 1949, 19, 200–209.

Curriculum vitae

Hildo Wesdorp, geb. 1935, studeerde Nederlands aan de G.U. te Amsterdam. Deed doctoraalexamen in 1965. Sindsdien werkzaam bij het voortgezet onderwijs (1965–1969), het C.I.T.O. (1969–1973) en het R.I.T.P. (vanaf 1966). Promoveerde begin 1974 op een proefschrift: Het meten van de produktief-schriftelijke taalvaardigheid. Directe en indirecte methoden: 'opstelbeoordeling' versus 'schrijfvaardigheidstoetsen'. Promotor was prof. dr. A. D. de Groot.

Adres: R.I.T.P., Prinsengracht 303, Amsterdam.