

sequentiële redundantie (SR) en het gebruik van de context bij woordherkenning (WD) sterk gerelateerd aan het verschil in leesvaardigheid tussen de groepen L en H, terwijl het vormen en toetsen van zins hypothesen (ZH) in weerwil van de top down-voorspelling niet tussen de groepen differentieerde. Goede lezers zijn vooral goed in woordherkenning, en de toevoeging van context vergroot die voorsprong nog. Wat het eerste punt betreft: algemene deelvaardigheden uit het SI-model, zoals Cognitie en Evaluatie, werden in de toetsscores teruggevonden. Verband tussen de deelvaardigheden en de processen is er zeker maar het is onvoldoende specifiek om een bepaalde deelvaardigheid te kunnen relateren aan een bepaald cognitief proces. Om zo'n gedifferentieerde samenhang vast te stellen zal verder onderzoek nodig zijn, waarin ook andere operationalisaties van de SI-vaardigheden en andere cognitieve processen opgenomen zijn. Lukt het ook dan niet, dan verliest het SI-model aan waarde als taxonomie van taalvaardigheden. Lukt het wel, dan is in het procesonderzoek steun te vinden voor de validiteit van de SI-vaardigheden.

Noten

1. Dit artikel is een sterk bekorte en ingedikte weergave van ons doctoraal-onderzoek. Het proces-onderzoek daarbinnen was een 'altered replication' van Frederiksen (1980). Het onderzoek werd uitgevoerd onder supervisie van Dr. H.L.M. Meuffels.
2. De tekst van deze *Inleiding* berust in belangrijke mate op het werk van Meuffels (1982).
3. In de analyses is de invloed van eventuele verschillen in werkt tempo uitgebannen door te werken met gewogen scores. Per toets is voor elke leerling het aantal goed gemaakte items uitgedrukt als percentage van het totale aantal door die leerling gemaakte items (goed of fout).
4. De oorspronkelijke opzet om de correlatiematrix van de 50 items te toetsen aan een model met de zes SI-toetsen als factoren kon niet uitgevoerd worden, omdat de 50*50 matrix singulier bleek te zijn.
5. Van de Cognitie-toetsen wordt verwacht dat ze niet laden op Evaluatie en Snelheid en omgekeerd wordt van de Evaluatie-toetsen verwacht dat ze niet laden op Cognitie en Snelheid. Deze negatieve verwachtingen worden in het model gefixeerd door de betreffende ladingen op nul stellen. De factor Snelheid wordt maar door één variabele gemeten en daardoor is de lading niet te schatten. Deze lading is arbitrair op .50 gesteld.

Literatuur

- Frederiksen, J.R.
 1980 'Component skills in reading: measurement of individual differences through chronometric analysis'. In: R.E. Snow, P. Federico en W.E. Montague (eds), *Aptitude, learning and instruction*. (Vol. 1: 'Cognitive process analysis of aptitude'.) Hillsdale (N.J.): Erlbaum, 105-138.
- 1980 'Component skills in reading: measurement of individual differences through chronometric analysis'. In: R.E. Snow, P. Federico en W.E. Montague (eds), *Aptitude, learning and instruction*. (Vol. 1: 'Cognitive process analysis of aptitude'.) Hillsdale (N.J.): Erlbaum, 105-138.
- Meuffels, B.
 1982 *Studies over taalvaardigheid*. Amsterdam (Diss. Universiteit van Amsterdam).
- Stanovich, K.E.
 1980 'Toward an interactive-compensatory model of differences in the development of reading fluency'. *Reading Research Quarterly* XVI, 32-71.

De ontwikkeling van een functionele taaltoets voor eindleerplichtige leerlingen van het LBO en MAVO

Henk Blok

1. Inleiding

Wat is het doel van het taalonderwijs? Ontwikkelen jonge mensen hun taalvaardigheid voor de school of voor het leven? Is de kwaliteit van het taalonderwijs aan verandering onderhevig? Voor dergelijke vragen is het project *Functionele Taalvaardigheid* van belang. Het project kan gezien worden als een doelstellingenonderzoek voor het moedertaalonderwijs. Het project wordt verricht op het niveau 'einde leerplicht' en concentreert zich op jongeren die na het behalen van een MAVO- of LBO-diploma het volledige dagonderwijs verlaten. Het doelstellingenvraagstuk wordt aan de orde gesteld vanuit het evaluatie-perspectief: wat willen we met het onderwijs bereiken en lukt dat? Doel van het onderzoek is het bepalen van een aantal wenselijke onderwijsdoelen en het nagaan of bij schoolverlaters deze doelen bereikt worden. De aandacht is vooral gericht op die vaardigheden die van groot praktisch belang zijn in het naschoolse leven. Dergelijke vaardigheden worden wel aangeduid als *functionele taalvaardigheden*.

Bij dit onderzoeksdoel hoort een opzet waarin drie delen onderscheiden kunnen worden.

A. Om de wenselijkheid van moedertaalonderwijsdoelen te bepalen zijn bij jongeren die hun schoolloopbaan inmiddels al enige jaren hebben afgerond, gegevens verzameld over hun taalgebruik. Over verschillende typen van taalgebruikssituaties zijn de meningen van ruim 500 oud-leerlingen van LBO en MAVO verkregen. De meningen hebben betrekking op drie centrale vragen die bij elk van de in totaal 69 typen taalgebruikssituaties gesteld zijn: *hoe vaak heb je dit gemiddeld gedaan?*; *hoe goed denk je dat je dit nu kunt?*; en *hoe belangrijk vind je het om dit te kunnen?* Op grond van de verzamelde gegevens zijn lijsten samengesteld van typen taalgebruikssituaties die frequent voorkomen en van situaties die moeilijk gevonden worden terwijl het belang van een goede beheersing groot wordt geacht (Blok & De Gloppe 1983).

B. In het vervolg zijn enkele van de meest wenselijke doelen geoperationaliseerd in de vorm van een takenproef, waarmee bij individuele leerlingen gepeild kan worden in welke mate deze doelen bij hen gerealiseerd zijn. Daarvoor zijn acht typen taalgebruikssituaties uitgekozen. Sommige omdat ze zeer frequent blijken voor te komen, andere omdat ze moeilijk maar toch belangrijk gevonden worden. Het betreft de volgende situatietypen:

- voor jezelf korte notities of aantekeningen maken;
- een brief schrijven aan een officiële instantie waarin je gevraagde informatie verschaft of zelf om inlichtingen vraagt;
- een lange/dikke handleiding lezen van een formulier dat je moet invullen;
- het algemene nieuws in de krant lezen;
- informatie opzoeken in gidsen of naslagwerken;
- mondeling informatie vragen;
- uitleg, instructies of informatie geven of beluisteren;
- luisteren (en kijken) naar informatieve programma's.

Vanaf begin 1983 tot het najaar is in samenwerking met twee docententeams gewerkt aan de constructie van de takenproef (Blok 1984). Besloten is twee (parallel)programma's te maken, in elk waarvan de acht taalgebruikssituaties thematisch aaneengesmeed zijn. Het ene programma heeft als thema 'een tienertoevakantie'. Het andere programma bevat opdrachten rond het thema 'werken bij een uitzendbureau'. De situaties waarmee de leerlingen in de opdrachten geconfronteerd worden, zijn in vergaande mate realistisch. De leerlingen lezen een echte krant, schrijven een brief, zoeken informatie in het spoorboekje, vullen een aanrijdingsformulier in, enz. Momenteel zijn we bezig met proefnamen van beide programma's. Met elke programma-afname is een volledige dag gemoeid. Gedurende zo'n dag worden 12 leerlingen door een testleider plus twee testassistenten met alle onderdelen geconfronteerd. Op het moment dat dit geschreven wordt, is de afname van één programma voltooid: 208 leerlingen hebben hieraan deelgenomen. De afname van het andere programma is gepland in oktober en november 1984. Daarna worden de antwoorden van de leerlingen gescoord en statistisch verwerkt. Deze fase wordt afgesloten met een rapport over de meettechnische kwaliteiten van de beide programma's.

C. Als straks de programma's klaar zijn en de meettechnische kwaliteit voldoende is bevonden, dan komen we toe aan het laatste onderdeel van het project *Functionele Taalvaardigheid*. Dat is de afname van de programma's bij een steekproef van leerlingen aan het einde van een LBO- of MAVO-opleiding. Uit de prestaties van de leerlingen kunnen we leren hoe het er met de taalvaardigheid van deze leerlingen in een aantal functionele situaties voorstaat. Daarnaast zullen ook andere vragen aan bod kunnen komen als: hoe groot is de variatie in leerling-prestaties en zijn er associaties tussen de prestaties van de leerlingen en de soort opleiding of bepaalde andere kenmerken van de leerlingen? De peiling zal veel informatie opleveren die van belang is voor de meningsvorming over doelen van het moedertaalonderwijs aan het einde van de leerplicht.

In deze bijdrage wil ik niet verder ingaan op de grote lijn van het project. In plaats daarvan stel ik voor ons te concentreren op een tweetal interessante onderdelen van het project. In de eerste plaats wil ik dieper ingaan op de vraag: 'Wat zijn nu eigenlijk de kenmerkende eigenschappen van functionele

taaltoetsen?'. Na de behandeling van deze vraag zal ik één van de programma-onderdelen demonstren. Deze demonstratie maakt hopelijk duidelijk wat de moeilijkheden en de mogelijkheden zijn van de ontwikkeling van functionele taaltoetsen.

2. Wat is een functionele taaltoets?

Men kan een functionele taaltoets beschouwen als een soort onderwijskundig meetinstrument, een instrument dat bedoeld is voor het verrichten van metingen ten behoeve van onderwijs of onderwijskunde. Andere soorten onderwijskundige meetinstrumenten zijn bijvoorbeeld: toetsen, overhoringen, proefwerken, examens, tentamens. Omdat functionele taaltoetsen bedoeld zijn voor het verrichten van metingen, dienen ze aan eisen te voldoen waaraan ook andere meetinstrumenten dienen te voldoen. Dat zijn eisen op het terrein van de nauwkeurigheid of betrouwbaarheid, gevoeligheid, objectiviteit en efficiëntie. In elk handboek over onderwijskundig meten kan men deze eisen omschreven vinden (bijv. De Groot & Van Naerssen 1975).

Waarin een functionele taaltoets zich wel van veel onderwijskundige meetinstrumenten onderscheidt, is de *beoogde score-interpretatie*. In dat opzicht is de functionele taaltoets een criteriumtoets. Om dit te kunnen uitleggen grijp ik even terug op het onderscheid dat toetsdeskundigen sedert het einde van de zestiger jaren maken tussen absoluut meten en relatief meten. Van *relatief meten* spreekt men bij studietoetsen die resulteren in scores waarvan de interpretatie afhankelijk is van het niveau van de leerlingen die de toets gemaakt hebben. Van *absoluut meten* spreekt men als de interpretatie van de scores onafhankelijk is (of beter: kan zijn) van het niveau van de leerlingen, maar daarentegen afhankelijk is van het gedrag dat de toets beoogt te meten. Om scores te verkrijgen die rechtstreeks geïnterpreteerd kunnen worden in termen van gedrag, construeert men zogenaamde *criteriumtoetsen*. Hoe gaat dat in zijn werk?

De eerste stap bij de constructie van een criteriumtoets is de specificatie van een gedragsdomein, een reeks van gedragingen van leerlingen waarover men een uitspraak wil kunnen doen. Zo'n gedragsdomein kan op verschillende manieren omschreven worden, bijvoorbeeld in termen van taken of opdrachten die de leerling moet kunnen verrichten, of meer globaal in termen van doelstellingen. Een criteriumtoets is een set taken die representatief is voor het gedragsdomein. Daardoor is de score op een criteriumtoets rechtstreeks te relateren aan het bijbehorende domein. In plaats van een norm in termen van beheersing van anderen zoals bij relatief meten, maken we bij absoluut meten dus gebruik van een norm in termen van de opdracht zelf, het gevraagde gedrag of de bijbehorende onderwijsdoelstelling.

Voor sommige doeleinden is relatief meten een goede zaak, onder meer voor selectiedoeleinden. Bij examens bijvoorbeeld bewijzen relatieve metingen nuttige diensten. Door de caesuur, de voldoende-onvoldoende grens,

jaarlijks aan het niveau van de kandidaten aan te passen, blijft het percentage geslaagden onder controle. Maar examens waar relatief gemeten wordt, zijn ongeschikt voor een discussie over onderwijsdoelstellingen. Ze zeggen namelijk niets over het prestatieniveau. Scores die rechtstreeks in doelstelling-terminen vertaald kunnen worden, zijn voor zo'n discussie beter bruikbaar. Wil men de scores op een functionele taaltoets inbrengen in een doelstellingen-discussie, dan doet men er dus goed aan de toets als een criteriumtoets te construeren.

Nu is een functionele taaltoets niet zo maar een criteriumtoets. Het is een criteriumtoets waarvan het domein bestaat uit werkelijke taalgebruikssituaties, dat wil zeggen uit situaties waarin het taalgebruik functioneel is. De score op een functionele taaltoets laat een uitspraak toe over het taalgebruik in de situaties die tot het domein behoren. Een voorbeeld: laat het domein (globaal) omschreven zijn als - voor jezelf notities of aantekeningen maken - dan kan men op basis van de prestatie op een functionele taaltoets die bij dit domein ontwikkeld is, aangeven hoe goed iemand zich in de werkelijke situatie kan gedragen.

De scores op een functionele taaltoets horen daarvoor een zo rechtstreeks mogelijke relatie met het domein te hebben. Deze eis heeft consequenties voor de constructiewijze. Gegeven een domein van taaltoetsen, is de constructie van een functionele taaltoets simpel, althans in theorie. Men trekt een steekproef van taken uit het domein en legt deze aan de leerlingen voor. Als het domein welomschreven is en de steekproef representatief, is generalisatie op basis van de toetsprestatie naar het domein zonder meer mogelijk. De nauwkeurigheid waarmee generalisatie mogelijk is, wordt in de regel *inhoudsvaliditeit* genoemd. Deze soort validiteit is voor een functionele taaltoets een essentiële eigenschap.

In verband met criteriumtoetsen is van meet af aan veel aandacht besteed aan het vraagstuk van de inhoudsvaliditeit. Bij het merendeel van de criteriumtoetsen is evaluatie van de inhoudsvaliditeit helemaal geen simpele zaak. Dit vindt zijn oorzaak in het feit dat het zelden mogelijk is het domein volledig te beschrijven. Veelal, en dat geldt ook in ons geval, is het slechts mogelijk het domein te preciseren via een zogeheten 'table of content specifications' of *toetsmatrijs*. In zo'n situatie is aselecte trekking van opgaven onmogelijk, waardoor bepaalde statistische aannamen onbruikbaar zijn. De mogelijkheden om de inhoudsvaliditeit te evalueren zijn dan beperkter en ook veel minder stringent. Men maakt dan wel gebruik van oordelen van experts of inhoudelijke specialisten over de relatie tussen toets en domein. Wij hebben ons evenwel van een tweede mogelijkheid bediend: uitgaande van dezelfde toetsmatrijs hebben twee teams van constructeurs parallelversies van de functionele taaltoetsen ontwikkeld. De mate waarin de twee versies equivalent zijn, is een indicatie voor de inhoudsvaliditeit.

Onze toetsmatrijs is een beschrijving van de taalgebruikssituaties waarop de taaltoetsen betrekking dienen te hebben. In het voorafgaande onderzoek, waarin oud-leerlingen hun oordelen over de taalgebruikssituaties gaven,

hebben we tamelijk brede omschrijvingen van het taalgebruik toegepast. Voor de toetsconstructiewerkzaamheden zijn deze omschrijvingen te globaal. Het is dus lang niet zeker dat de beide programma's straks tot dezelfde conclusie zullen leiden.

Nu gaat het bij evaluatie van de inhoudsvaliditeit niet alleen om de vraag of de taaltoetsen het gehele domein (of een steekproef daaruit) bestrijken. Van belang is óók de vraag naar de relatie tussen het volbrengen van de taaltoetsen onder de afnameconditie en het volbrengen van de werkelijke taalgebruikssituatie. Idealiter is deze relatie perfect, dat wil zeggen dat leerlingen de taaltoetsen beter volbrengen naarmate hun taalbeheersing ook buiten de testsituatie beter is. Naarmate de taaltoetsen en de condities waaronder ze afgenomen worden meer gelijk zijn aan werkelijk taalgebruik onder werkelijke omstandigheden, zal de inhoudsvaliditeit groter zijn. Om dit te bereiken maken we gebruik van *simulatie*.

Simulatie is een belangrijke techniek met veel toepassingsmogelijkheden. Bij de bouw van de dam in de Oosterschelde wordt op grote schaal gesimuleerd. Psychologen simuleren bij de bestudering van de omgang met complexe systemen als vliegruigen, schepen, wapens en kernenergiecentrales. Ook in de educatieve sfeer is simulatie geen onbekende. Denk aan de oude meesterproeven, waarbij gezellen als blijkt van hun kunnen een proefstuk of meesterwerk vervaardigden. Denk ook aan het praktische gedeelte van het rijvaardigheidsexamen waarbij de kandidaten gedurende een half uur aan het verkeer deelnemen onder het toezicht van de examinerator. Zelfs de spreekbeurt, een onderdeel van het spreekvaardigheidscurriculum, is te beschouwen als simulatie voor een werkelijke taalgebruikssituatie, het houden van een lezing.

Geen simulatie is perfect. Anders zou het geen simulatie zijn. Voor een hoge inhoudsvaliditeit is een nauwkeurige nabootsing ook geen voorwaarde. Sommige reducties zullen geen effect hebben op de inhoudsvaliditeit, andere daarentegen wel. Van elk verschil tussen de afnamecondities en de werkelijke condities zal overwogen moeten worden wat het meest waarschijnlijk effect daarvan is op de prestaties van leerlingen. Bijvoorbeeld, voor veel werkelijke taalgebruikssituaties geldt dat er een persoonlijk belang in het geding is. Wie een sollicitatiebrief schrijft, doet dat uit behoefte naar een baan. Is deze behoefte sterk, dan zal de schrijver er veel aan gelegen zijn een goede brief te schrijven. Zo'n persoonlijk belang is in testomstandigheden niet te simuleren. Het is evenwel de vraag of dat belang tot een betere brief leidt. Alleen als het persoonlijk belang effect heeft, gaat dat ten koste van de inhoudsvaliditeit van de testtaak.

Zolang er nog weinig bekend is over het effect van reducties in de afnamecondities, is het in het algemeen verstandig ernaar te streven deze condities op de werkelijke omstandigheden te laten lijken. Het behoeft geen betoog dat dit streven op gespannen voet staat met de principes van een efficiënte en gestandaardiseerde meting. Het construeren van functionele taaltoetsen heeft daarom iets van het laveren tussen Scylla en Charibdis. Aan de ene zijde

dreigt het gevaar van een toets met uitstekende meettechnische kwaliteiten die beperkt inhoudsvalide is. Aan de andere zijde loert een toets met een inhoudsvaliditeit die groot had kunnen zijn, als de meettechnische kwaliteit maar voldoende was geweest. In de volgende paragraaf zal ik laten zien hoe we aan beide maastromen trachten te ontkomen.

3. Telefoneren met Randstad

Bij wijze van voorbeeld zal ik één van de programma-onderdelen introduceren. Het *gedragsdomein* dat bij dit onderdeel behoort, is als volgt omschreven:

MONDELING INFORMATIE VRAGEN

Bijvoorbeeld eenvoudige informatie op een reisbureau, postkantoor of in een winkel. Maar ook ingewikkelder informatie over hypotheeken, verzekeringen, regels en voorschriften van de overheid.

In het onderdeel voert de leerling een telefoongesprek met een intercedent van Randstad, gespeeld overigens door een test-assistent. In dit gesprek moet de leerling over diverse arbeidsvoorwaarden voor vakantiewerkers informatie vragen. De opdracht tot het telefoongesprek is verpakt in een *situatiebeschrijving*. De beschreven situatie is een vervolg op een ander onderdeel, waarin de leerlingen een tijdelijke baan is aangeboden door Randstad. De situatie komt er op neer, dat de leerling zojuist een gesprek met een kennis, Kees, heeft gevoerd. Kees heeft veel problemen met z'n vakantiewerk gehad. In de situatiebeschrijving worden zes problemen aangeduid. De *opdracht* voor de leerling luidt dan als volgt:

'Je bent nogal geschrokken van Kees' z'n verhaal. Misschien is je vakantiebbaan toch niet zo aantrekkelijk als je eerst dacht. Wat zijn eigenlijk jouw rechten en plichten? Dat zou je nu wel eens willen weten.'

Daarom ga je zo meteen René Blaak van Randstad nog eens opbellen. Zorg dat je van René alles te weten komt wat voor jou van belang is in de kwesties die Kees aanvoerde. Meteen na het gesprek met Randstad krijg je schriftelijke vragen over de informatie die je van René gehad hebt.'

Na afloop van het telefoongesprek, tijdens welke de leerling aantekeningen mag maken, wordt hem of haar een serie van elf schriftelijke korte-antwoord vragen voorgelegd. De vragen gaan over de informatie die de leerling in heeft moeten winnen. Het is de leerling overigens bekend dat een test-assistent de rol van Randstad-intercedent speelt.

Het is de taak van de test-assistent om te antwoorden op de vragen van de leerlingen. Voor elk van de zes problemen is het antwoord van de assistent voorgeschreven in een *ideaal-antwoordmodel*. De manier waarop dit gedaan is, laat de assistent voldoende ruimte om het gesprek een natuurlijk verloop te kunnen geven. Essentieel is dat de aanwijzingen ook eisen bevatten voor de vraagwijze van de leerling in een *ideaal-vraagmodel*. De vragen van de leerling

moeten dus aan bepaalde eisen voldoen, voordat de test-assistent het ideaal-antwoord mag geven. Dit principe wordt wel 'triggering' genoemd: de leerling moet de trekker overhalen alvorens de informatie te ontvangen. Een voorbeeld: na de vraag - ik wil eens weten hoe dat allemaal nou zit als je ziek wordt - zal de test-assistent doorvragen en nog niet de informatie over ziektegeld geven. Daarvoor is nodig dat de vraag een combinatie is van A en B, waarbij



De vraag - als ik nou een been breek, hoe zit het dan, krijg ik dan nog geld? - voldoet aan de eis en zal dus door de in het ideaal-antwoordmodel vervatte informatie gevolgd worden.

Wij hebben dit onderdeel onlangs afgenomen aan ruim 200 leerlingen, die net een LBO- of MAVO-examen gedaan hadden. Per leerling kost dat gemiddeld ongeveer 20 minuten: 5 minuten voorbereidingstijd, 7 minuten voor het telefoongesprek en nog eens 7 minuten voor de schriftelijke vragen achteraf. Van alle gesprekken zijn audio-opnamen gemaakt. De bedoeling van de opnamen is tweërlei:

- De opnamen stellen ons in staat het vraaggedrag van de leerlingen te scoren. Per probleem onderscheiden we de volgende categorieën:
 - a. de wijze van vragen is in de eerste beurt conform het ideale vraagmodel;
 - b. de wijze van vragen is in de tweede of latere beurt conform het ideale vraagmodel;
 - c. de wijze van vragen is in geen enkele beurt conform het ideale vraagmodel;
 - d. het probleem is in het geheel niet aangevoerd.
- Door de opnamen kunnen we het gedrag van de test-assistent controleren: houdt de assistent zich aan de voorschriften?; krijgen alle leerlingen een gelijkwaardige behandeling?; etc.

We beschikken niet alleen over scores voor het vraaggedrag, ook de schriftelijke vragen achteraf worden gescoord. Zo hebben we twee soorten indicatoren voor de kwaliteit van het taalgebruik van de leerling in de gesimuleerde situatie. De afnamen zijn verricht, maar de scoring nog niet. Ik kan daarom niet ingaan op de kwaliteit van de metingen, de scores dus. Wel wil ik op één punt nog een toelichting geven, en wel over de confrontatie van de twee soorten scoringen. Want op kleine schaal, bij ongeveer 20 leerlingen, is daar inmiddels aandacht aan besteed.

Confrontatie van de twee soorten scoringen, van de schriftelijke antwoorden en van de wijze van vragen stellen is interessant. Niet onverwacht blijkt dat scoring van de twee verschillende typen gedrag tot verschillende resultaten kan leiden. De volgende twee typen discrepanties komen namelijk voor:

type 1: de leerling stelt in het gesprek de vraag op correcte wijze, maar mist de schriftelijke vraag die achteraf gesteld wordt;

type 2: de leerling stelt in het gesprek de vraag niet correct, maar beantwoordt de schriftelijke vraag achteraf correct.

De discrepantie van type 1 kan voorkomen wanneer de leerling na een correcte vraag:

- het gegeven antwoord niet begrijpt en vervolgens niet door- of navraagt;
- het gegeven antwoord noch in het hoofd noch op papier heeft vastgelegd, dan wel foutief heeft vastgelegd;
- bij het toepassen van de nieuw verworven informatie fouten maakt;
- het correcte antwoord niet is aangeboden door de test-assistent.

Mogelijkheid d. hebben we op grond van de audio-opnamen vrijwel steeds kunnen uitsluiten. De drie andere mogelijkheden lijken zich evenwel voor te doen.

De discrepantie van type 2 kan voorkomen wanneer de leerling na een foutieve vraag:

- het correcte antwoord aangeboden wordt;
- op grond van reeds verworven kennis de schriftelijke vraag correct weet te beantwoorden.

De mogelijkheden e. en f. zouden door een juiste inrichting van de testsituatie niet mogen voorkomen, maar voor mogelijkheid f. lijkt hierop geen waterdichte controle mogelijk.

Ondanks deze discrepanties lijkt het de moeite waard de beide soorten scoringen te handhaven en te verbeteren. Onze overweging daarbij is dat beide scoringen relevant zijn voor de betrokken taalgebruikssituatie. Men kan immers het verwerken, vastleggen en toepassen van de aangeboden informatie beschouwen als een essentieel onderdeel van het mondeling informatie inwinnen.

4. Tot besluit

We staan nog maar aan het begin van de ontwikkeling van functionele taaltoetsen. Het valt daarom nog niet te zeggen waartoe deze ontwikkelingen zullen leiden. Maar ik voorzie dat deze nieuwe toetsvorm een hoge vlucht zal nemen. Op diverse terreinen wordt aan de ontwikkeling ervan hard gewerkt. In de SLO-projecten *Gericht Schrijven* zien we aandacht voor het schrijven in functionele situaties en voor toetsvormen die daarbij passen. Op het CITO zijn de eerste ervaringen opgedaan met functionele schrijfvaardigheidsexamens (Rijlaarsdam, De Glopper & Kreeft 1984). Op de T.H. Twente gebruiken te men al enkele jaren geleden ten behoeve van het onderzoek naar

taalverkeersproblemen een functionele taaltoets (Jansen & Steehouder 1981). Op het SCO worden momenteel onder leiding van Westorp functionele taaltoetsen ontwikkeld ten behoeve van een proefonderzoek naar het niveau van het taalonderwijs in het basisonderwijs (Westorp, Van den Bergh & Hoeksma 1984). Op het gebied van het vreemde-talenonderwijs ontwikkelt het CITO al geruime tijd functionele spreektoetsen. Deze opsomming is zeker nog niet compleet.

Ik meen dat er nog veel aandacht moet worden aan de kwaliteit van functionele taaltoetsen als meetinstrument. Betrouwbaarheid en validiteit zijn fundamentele eigenschappen van welke meetinstrument dan ook. Zonder deze eigenschappen is elke meting zinloos. Voor de evaluatie van de betrouwbaarheid zijn inmiddels vele methoden ontwikkeld. Van groot belang lijkt mij de toepassing van de zogenaamde latente trekmodellen uit de *item-responsietheorie* (Hambleton 1983). Met deze modellen kan men schattingen verkrijgen van de vaardigheid van een leerling, onafhankelijk van de precieze taak die men voor hem of haar uit het domein gekozen heeft. Deze modellen zijn in statistisch opzicht nog niet af. En ook is de toepassing gebonden aan hele strenge voorwaarden, die in de praktijk van het taaltoetsonderzoek niet gauw vervuld zijn. Maar er zijn ook andere kwesties die aandacht eisen, problemen bij de evaluatie van de inhoudsvaliditeit. Enkele daarvan zijn in het voorgaande al aangestipt:

- het is onmogelijk het domein waarop de toets betrekking heeft exact vast te leggen;
 - over het effect van reducties in de afname-situatie hebben we nog bijzonder weinig inzicht.
- Het zal veel inventiviteit vergen en waarschijnlijk nog lang duren voordat deze problemen opgelost zijn. Ik hoop dat het lukt, want een goede functionele taaltoets heeft veel voordelen. Zo'n toets is motiverend voor de leerling; zo'n toets is belangwekkend voor het publiek als het om onderwijs-doelstellingsvraagstukken gaat; en op de korte termijn is de constructie van zo'n toets een uitdaging voor taalbeheersingsonderzoekers.

Literatuur

- Blok, H.
1984 *Functionele taalvaardigheid: de ontwikkeling van evaluatieve taalprogramma's*. Amsterdam: SCO.
- Blok, H. en K. de Glopper
1983 *Taal voor het leven: meningen van oud-leerlingen LBO en MAVO over hun taalgebruik*. Harlingen: Flevodruk.
- Groot, A.D. de en R.F. van Naerssen
1975 *Studietoetsen: construeren, afnemen, analyseren, deel II*. 's Gravenhage: Mouton.
- Hambleton, R.K. (ed.)
1983 *Applications of item response theory*. Vancouver, British Columbia: Educational Research Institute of British Columbia.

- Jansen, C.J.M. en M. Steehouder
1981 'Een taalverkeersprobleem: de voorlichting over individuele huursubsidie'. In: M. Steehouder en C.J.M. Jansen (red.): *Taalbeheersing 1981*. Enschede: Vereniging Interuniversitair Overleg Taalbeheersing (V.I.O.T.)/T.H. Twente.
- Rijlaarsdam, G., K. de Glopper en H. Kreeft
1984 'Functionele schrijfvaardigheidstoetsen: een voorstel voor nieuwe examens stelvaardigheid LBO/MAVO'. *Levende talen*, juni.
- Wesdorp, H., H. van den Bergh en J.B. Hoeksma
1984 *De constructie van een aantal instrumenten ter meting van functionele taalvaardigheden ten behoeve van een periodiek peilingsonderzoek in het basisonderwijs*. Amsterdam: SCO.

Training van opstelbeoordelaars: effecten en problemen

Kees de Glopper

1. Inleiding

Voor de beoordeling van de produktief-schriftelijke taalvaardigheid zijn uiteenlopende evaluatiemethoden beschikbaar. Het meest gangbaar zijn de directe beoordelingsmethoden waarbij schrijfproducten door beoordeelaars op hun waarde worden geschat. Het langs directe weg beoordeelen van de stelvaardigheid plaatst moedertaaldocenten en taalbeheersingsonderzoekers keer op keer voor lastige problemen. De betrouwbaarheid van opsteloordeelen is immers vaak dubieus; het is eerder uitzondering dan regel dat een op één moment door één beoordeelaar gegeven oordeel over een opstel overeenstemt met een ander, onafhankelijk verkregen oordeel. Met de betrouwbaarheid is ook de validiteit van de beoordeling van de produktief-schriftelijke taalbeheersing in het geding. Het is lang niet altijd duidelijk wat oordelen over opstellen nu eigenlijk meten: welke kwaliteiten van schrijfproducten komen in de gevelde oordelen tot uitdrukking?

Verbeteringen van de kwaliteit van *opstelbeoordelingen* worden langs verschillende wegen nagestreefd. Vaak worden achteraf, ná beoordeling, maatregelen getroffen om de betrouwbaarheid van de beoordeling te verhogen. Dit gebeurt bijvoorbeeld door de oordelen van twee of meer onafhankelijk werkende beoordeelaars door optelling te combineren tot een *jury-oordeel*. Zonder problemen is deze aanpak niet: er zijn duidelijke aanwijzingen dat beoordeelaars consistent verschillend beoordelen; dit brengt met zich mee dat bij jury-vorming metingen van verschillende betekenis opgeteld worden.

In veel gevallen wordt ingegrepen in de beoordelingstaak, met de bedoeling de beoordelingstaak beter te omschrijven. Daarbij worden dan analytische beoordelingschema's verstrekt waarin te beoordelen deelaspecten van schrijfproducten onderscheiden en gespecificeerd worden. Ook wordt wel gewerkt met een beoordelingsschaal van in kwaliteit oplopende voorbeeldopstellen. Om te garanderen dat beoordeelaars analytische richtlijnen en/of schalen van voorbeeldopstellen op dezelfde en bedoelde wijze gebruiken wordt met name in taalbeheersingsonderzoek werk gemaakt van *training van beoordeelaars*.

Met het gebruik van analytische schema's, voorbeeldschalen en training is ervaring opgedaan in een onderzoek naar de aanpak en resultaten van het stelvaardigheidsonderwijs in het Nederlandse voortgezet onderwijs.¹ Op de opzet van de trainingsprocedure, de effecten van de training en op de tijdens