TOETSEN TEKSTBEGRIPTOETSEN TEKSTBEGRIP?

Maarten van Gils

WOORD VOORAF

'Het is algemeen bekend dat bij het toetsen van tekstbegrip met open vragen meer getoetst wordt dan tekstbegrip alleen. Wij beoordelen immers ook de wijze waarop de leerling zijn antwoord formuleert. Wanneer het ons echter alleen gaat om het begrijpen van een tekst, zijn meerkeuzevragen hiervoor misschien beter geschikt. Zij bieden ook meer mogelijkheden tot een objectieve en snelle correctie. Men kan verwachten dat deze overwegingen er toe zullen leiden dat in de toekomst een deel van het examen in de meerkeuze-vorm zal worden afgenomen. Het is daarom wellicht verstandig leerlingen reeds nu met deze wijze van vragen vertrouwd te maken.' Tegen deze gedachtengang - die ik ontleen aan het 'Woord vooraf' bij een bekende toetsenverzameling[1] wil ik me in onderstaand artikeltje verzetten.

Toetsenmakers vinden het wetenschappelijk onverantwoord twee of meer vaardigheden tegelijk te meten. In hun ogen is bijvoorbeeld het stilleesstukje met open vragen onbetrouwbaar en 'besmet', omdat daarin het *tekstbegrip* niet voldoende is *geïsoleerd.* Zij prefereren de professionele multiple-choicetest, die één en niet meer dan één variabele probeert te meten: *luistervaardigheid* (of wat daarvoor doorgaat), *tekstbegrip* (of enkele deelvaardigheden daarvan), *schrijfvaardigheid* (of iets wat daarmee korreleert). En daarom noemen de konstrukteurs hun eigen toetsen *specifiek.*

Het is zeer de vraag of de multiple-choicetest wel verenigbaar is met de doelstellingen van het moedertaalonderwijs. De nadelen van deze vraagvorm zijn overigens al zo vaak gesignaleerd[2] dat ik er nu niet nader op hoef in te gaan. Ik beperk me hier tot een aanval op de *specificiteitstheorie* en tot de toetsing van de receptieve vaardigheden *lezen* en *luisteren.*

Mijn *boodschap* is: *De tekstbegriptoets meet voornamelijk testhandigheid en toevalsfaktoren.* Ik ga deze stelling met slechts één voorbeeld 'bewijzen'; het gebruikte techniekje is echter zo simpel, dat elke liefhebber het belastend materiaal gemakkelijk kan uitbreiden.

Als een tekstbegriptoets alleen tekstbegrip meet, dan moeten de ruwe skores van proefpersonen die niet over de tekst beschikken, zeer dicht in de buurt van de *raadkans* liggen. *Als* daarentegen die skores het raadpercentage aanzienlijk overschrijden, dan zullen we moeten aannemen dat er (ook) andere 'vaardigheden' worden gemeten dan tekstbegrip. Op deze redenering baseerde ik het nu volgende proefje.

PUZZEL

De onderstaande elf multiple-choice-items zijn te vinden in de toetsenverzameling waaruit ik zoëven ook een deel van het voorwoord heb geciteerd: *'Meerkeuze-vragen bij een aantal teksten uit Schrijvenderwijs'* van E. Pistoor. Ik nodig de lezer uit eerst zelf zijn krachten te beproeven op dit elftal (zonder de tekst erbij te halen) en pas daarna mijn relaas verder te volgen.

1. Een copywriter (r. 3) is een
   a. journalist die cursiefjes schrijft.
   b. schrijver van literaire teksten.
   c. schrijver van reclameteksten.
   d. tekstschrijver voor radio en t.v.

2. Wat is de taak van een copywriter? (eerste alinea)
   a. Contact-intensief bezig zijn met de moedertaal.
   b. De taal voor grote groepen toegankelijk maken.
   c. Gebruik maken van de zwakheden van de vrouwelijke natuur.
   d. Mensen beïnvloeden door middel van taaluitingen.

3. Welke 'brokken' (r. 9) maakt de reclametaal?
   a. Zij beïnvloedt de taal ongunstig.
   b. Zij beïnvloedt het koopgedrag van de meeste vrouwen ongunstig.
   c. Zij heeft de reclame zelf minder lonend gemaakt.
   d. Zij heeft een aantal 'forse' tekstschrijvers brodeloos gemaakt.

4. Tekstschrijvers gebruiken zo vaak Engels of verengelste uitdrukkingen (r. 12), omdat
   a. Engelse uitdrukkingen een bron van inspiratie vormen voor woordvernieuwing.
   b. het publiek ontvankelijker is voor een Engelse uitdrukking dan voor de gelijkwaardige term in het Nederlands.
   c. Nederlandse woorden vrijwel geen betekenis meer hebben.
   d. zij in het Engels veel beter contact-intensief met de woorden kunnen zijn dan in het Nederlands.

5. Wat is de overeenkomst tussen de verschillende soorten schrijvers die in de eerste alinea worden genoemd?
   a. Zij beïnvloeden bewust het publiek.
   b. Zij hanteren hetzelfde instrument.
   c. Zij willen persé gelezen worden.
   d. Zij zijn verantwoordelijk voor woorddevaluatie.

6. Het blijkt dat schrijvers van reclameteksten zeer contact-intensief kunnen zijn met de taal. Dit lezen we in de zin die begint met
   a. De superlatievenziekte (r. 34)
   b. De woorddynamiek (r. 25)
   c. Juist in de reclame (r. 22)
   d. Wij komen in de reclame (r. 30)

7. Een 'forse' tekstschrijver (r. 32) is een tekstschrijver die
   a. altijd het geschiktste woord gebruikt.
   b. een voorkeur heeft voor superlatieven.
   c. nooit kleine advertenties verzorgt.
   d. weinig waarde hecht aan de tekst.

8. Waarom worden in de reclame steeds minder superlatieven gebruikt? (derde alinea, r. 34-36)
   a. Superlatieven gebruiken is niet zindelijk.
   b. Superlatieven hebben geen zeggingskracht meer.
   c. Superlatieven horen niet bij dichterlijk taalgebruik.
   d. Superlatieven verschijnen ook op sportpagina's.

9. Welk bezwaar kun je tegen de stijl van de schrijver aanvoeren?
   a. Hij gebruikt te vaak hetzelfde woord.
   b. Hij gebruikt te veel dichterlijke woorden.
   c. Hij maakt nieuwe woorden die onbegrijpelijk zijn.
   d. Hij schrijft lange, ingewikkelde zinnen.

10. Uit welk(e) woord(en) blijkt dat de schrijver tracht de taal creatief te gebruiken?
    a. beroepscopywriter (r. 13)
    b. nylon petticoat (r. 15)
    c. superlatievenziekte (r. 34)
    d. taalconstructies (r. 27-28)

11. Wat is de kern van het betoog van Joop Roomer?
    a. De kritiek op de reclametaal is ongefundeerd, want de tekstschrijver geeft de superlatieven waar het publiek om vraagt.
    b. Nu de reclametaal op zindelijke wijze wordt gebruikt, blijkt dat de poëtische woordcombinaties opvoedende waarden bezitten.
    c. Tekstschrijvers zijn contact-intensief bezig met de taal en lijden minder aan de superlatievenziekte dan vroeger.
    d. Zowel copywriters als sportjournalisten hebben er een handje van niets-zeggende clichés te gebruiken.

OPLOSSING

Bij vele van deze items zal de identifikatie van het 'beste' alternatief u niet moeilijk zijn gevallen. U *wist* bijvoorbeeld al wat een copywriter doet (vraag 1 en 2), u *elimineerde* enkele onlogische afleiders (vraag 3, 4, 5 en 10), u *konstrueerde* al puzzelende de onbekende tekst met behulp van de informatie die in de itemformuleringen verstopt zat (vraag 7, 8 en 11). Slechts in een enkel geval (vraag 6 en 9) was het gemis van een tekst hinderlijk en dan moest u echt gaan *raden:* met nog altijd 25% kans op een goed antwoord. Dankzij uw algemene ontwikkeling, uw gezond verstand, uw vermogen tot kombineren en uw goede gesternte hebt u vermoedelijk een fraaie skore tussen 7 en 11 punten weten te behalen.

Gebruikers van *Schrijvenderwijs* zullen intussen hebben ontdekt, om welke tekst het hier gaat. Het is *Reclame en taal* van Joop Roomer (pag 122-123 van de achtste gewijzigde druk). Met de tekst erbij kunnen we een lijstje maken van de bedoelde antwoorden: 1 C, 2 D, 3 A, 4 B, 5 B, 6 C, 7 B, 8 B, 9 A (?), 10 C, 11 C.
Hoeveel had u er goed?

Deze elf vragen blijken niet specifiek genoeg om het tekstbegrip van *Moer-lezers* te meten. Dat is niet verwonderlijk, want de toets is bestemd voor *leerlingen.* Gezien de plaats van de tekst in het schoolboek (no. 17 op een totaal van 76 teksten en acht examenopgaven) mogen we aannemen, dat klas 4 VWO tot de doelgroep behoort. In hoeverre zijn ook hier onder merkwaardige omstandigheden hoge skores haalbaar voor leerlingen die over enige testhandigheid beschikken?

Om dat te achterhalen legde ik de elf items *twee maal* voor aan een proefgroep van 71 leerlingen uit drie verschillende klassen (5 HAVO, 4 atheneum en 5 atheneum): de eerste keer *zonder tekst* en meteen daarna *met de tekst erbij.* Tabel I geeft een overzicht van de resultaten, berekend over alle deelnemers. Per item vermeld ik het percentage goede antwoorden (de P-WAARDE) èn de percentages leerlingen die een afleider aanstreepten (de A-WAARDEN)[3]. U vindt telkens eerst de skores *zonder* en daaronder de skores *met* tekst; de p-waarden zijn gekursiveerd.

Tabel I P- EN A-WAARDEN PER ITEM

ITEM NO.	*PERCENTAGES (P- EN A-WAARDEN)*				
	A	B	C	D	(niet ingevuld)
1.	11	6	*72*	11	(–)
	1	3	*96*	–	
2.	4	26	–	*70*	(–)
	20	4	–	*76*	
3.	*68*	15	14	1	(1)
	83	8	8	–	
4.	7	*65*	–	28	(–)
	1	*87*	–	11	
5.	64	*14*	5	14	(3)
	34	*51*	–	15	
6.	26	47	*22*	3	(1)
	7	21	*65*	7	
7.	33	*46*	10	8	(3)
	10	*70*	6	14	
8.	3	*89*	7	1	(–)
	–	*96*	3	1	
9.	*28*	4	28	36	(4)
	45	1	25	28	
10.	5	21	*46*	26	(1)
	1	11	*59*	28	
11.	5	7	*54*	23	(10)
	4	13	*75*	8	

Welke voorlopige konklusies kunnen we trekken uit deze empirische gegevens? Twee wil ik er hier presenteren: de eerste heeft betrekking op de A-WAARDEN, de tweede op de P-WAARDEN.

1 De helft van de afleiders funktioneert niet of nauwelijks onder normale omstandigheden; de meeste van die afleiders zijn zelfs onaantrekkelijk voor leerlingen die *niet* over de tekst beschikken (zie: 1 B, 2 C, 3 D, 4 A en C, 5 C, 6 D, 8 A, C en D, 9 B, 10 A en 11 A). Bij tekstloze toetsing zou een alternatievenverhouding

25 : 25 : 25 : 25 theoretisch ideaal zijn; geen enkel item slaagt er echter in om die verhouding zelfs maar te benaderen.

2 Het feit dat zó veel schertsafleiders ook voor leerlingen gemakkelijk herkenbaar zijn, draagt natuurlijk fors bij aan de hoge p-waarden. De invloed van 'raadvaardigheid' op 'tekstbegrip' is hier veel groter dan de officiële ideologie van de meettechnici toelaat. Over de hele toets bedraagt de gemiddelde p-waarde *zonder* tekst .55 en *met* tekst .76. Gesplitst naar de drie nivo's in mijn proefgroep gedraagt die gemiddelde p-waarde zich als volgt (zie tabel II).

Tabel II GEDRAG VAN DE GEMIDDELDE P-WAARDE

	5 HAVO	*4 ath.*	*5 ath.*	*totaal*
met tekst	.69	.77	.81	.76
zonder tekst	.46	.53	.65	.55
verschil	.23	.24	.16	.21

Het *verschil* tussen de beide p-waarden (*met* en *zonder* tekst) is mijns inziens de enig juiste maatstaf voor specifiek tekstbegrip: invloeden van raadvaardigheid en algemeenverbale intelligentie zijn daarin zorgvuldig uitgeschakeld. Dat verschil is schrikbarend laag, zeker als we het vergelijken met de hogé p-waarden die er vlak boven staan.

Een tijdje geleden heb ik wat gestoeid met een objektieve *luisteroefening* van het CITO die ik - zonder band - liet maken door een VWO-examenklas. De gemiddelde p-waarde bedroeg bij die gelegenheid .47 en op de twaalf vierkeuze-items kwamen 24 schertsafleiders[4] voor. Ik wees er toen op, dat bij dergelijke toetsen de leerlingskores altijd worden vertekend door de invloed van *voorkennis* en *kombinatievermogen*: niet alleen in 6 VWO, maar ook op lagere nivo's. Dat geldt evenzeer voor de *tekstbegriptoets* die nu aan de orde is. Tabel II wijst uit, dat er *over de hele linie* meer raadvaardigheid wordt gemeten dan tekstbegrip, ook al varieert de verhouding tussen die beide variabelen per subgroep.

Scholieren die - al dan niet in een examensituatie - met moderne objektieve tekstbegriptoetsen te maken krijgen, kunnen uit het bovenstaande iets heel belangrijks leren. Zij zouden zich vooral niet moeten storen aan het ouderwetse advies: 'Lees de tekst eerst goed door voor je met de beantwoording van de vragen begint.' Hun enige doel is immers een hoge testskore en daarbij past geen tijdrovende *leesstrategie*. 'Wat je in je kop hebt telt toch ook mee!'[5] Veel doeltreffender lijkt de *invulstrategie* die sommige eindexamenkandidaten volgen wanneer ze de multiple-choicetests Frans, Duits en Engels krijgen voorgelegd: ze analyseren eerst de toetsitems en raadplegen ter kontrole de bijbehorende (zinnen uit de) tekst. Op die manier werken ze methodologisch

verantwoord vanuit een konkrete probleemstelling via werkhypothese en toetsing naar een oplossing toe.

NIEUWE OPGAVEN

Wie snel de kwaliteit van een tekstbegrip- of luistertoets wil schatten, beschikt in de *specificiteitsproef* over een praktisch hulpmiddel om het ergste kaf van het koren te scheiden. Zelf ga ik alle *nieuwe opgaven* op dit gebied *eerst zonder tekst of band* te lijf. Slaag ik er dan in meer dan de helft van de vragen korrekt te beantwoorden, dan staat voor mij vast dat de toets in kwestie ongeschikt is voor selektief of diagnostisch gebruik in de klas. Pas wanneer mijn eigen 'nulskore' in de buurt van het raadpercentage ligt (p = .25 bij vierkeuzevragen), is een diepgaander onderzoek nodig: naar de 'onjuistheid' van de aangegeven afleiders bijvoorbeeld of naar de relevantie van de gestelde vragen. Er zijn mij *slechtere* vierkeuze-items bekend dan die bij *Reclame en taal.* Vooreerst is het een pluspunt, dat nergens (behalve misschien bij vraag 9) redelijke twijfel mogelijk is aan de juistheid van het aangegeven 'beste' antwoord. Vervolgens: als we achter elk item de vier antwoordmogelijkheden schrappen, resteert er een aantal goede open vragen. En tenslotte kan uit tabel I worden afgeleid, dat bij de meeste items het gebruik van de tekst leidt tot een significante stijging van de p-waarde: ook dát is niet bij elke objektieve tekstbegriptoets gegarandeerd. Maar dat alles kompenseert niet het gebrek aan specificiteit: *in deze vorm toetst deze tekstbegriptoets te weinig tekstbegrip.*

Als reeds een oppervlakkig onderzoek aantoont dat een bepaalde multiple-choicetest evident ondeugdelijk is, dan zal een *goede* meettechnische analyse die eerste indruk bevestigen en zeker niet tegenspreken. Geruststellende mededelingen over item-testkorrelaties (*R.I.T.*) en betrouwbaarheid (*KR 20*) missen in zo'n geval iedere bewijskracht; dat soort informatie is trouwens zelden terzake en altijd onvolledig[6]. De psychometrie is gelukkig geen toverkunst waarmee men recht kan praten wat eigenlijk krom is.

Tegen mijn specificiteitsproefje met *Reclame en Taal* is wel wat in te brengen: kritiek op deze ene toets tast 'het systeem' nog niet aan, ik had de 'raadvariantie' nader moeten specificeren etc. Desondanks wens ik voor mijn konklusies een tamelijk algemene geldigheid te claimen. De door mij gevolgde methode is gemakkelijk *herhaalbaar*: op grotere schaal, op diverse nivo's en met andere opgaven[7]. Ik nodig de lezers - en zeker ook de toetsenmakers - graag uit via replikatie-onderzoek mijn hypotheses te verifiëren, resp. te falsifiëren.

NOTEN

1 E. Pistoor: *Meerkeuze-vragen bij een aantal teksten uit Schrijvenderwijs,* Meulenhoff Educatief Amsterdam 1973

2 P.J.M. Groot: *Het toetsen van taalvaardigheid,* Wolters-Noordhoff, Groningen 1973, p. 17-25.
H. Bonset: *Notities t.a.v. meerkeuze-tekstbegriptoetsen op het eindexamen,* MOER 1974: 6, p. 305-306.
H. Bonset: *Enkele sociolinguïstische bezwaren tegen objektieve toetsen voor taalvaardigheid,* Levende Talen no. 313 (aug. 1975), p. 335-345.

3 P.J.M. Groot, a.w., p. 86-87.

4 M. van Gils: *Geachte toehoorders,* Levende Talen no. 312 (juni 1975), p. 225-227.

5 K. Hennephof: *Leesonderwijs,* MOER 1975: 4, p. 214 vv.

6 H.W. Feltkamp: *Kwantificatie van tekstbegrip,* Levende Talen no. 304 (jan.-febr. 1974), p. 35-36.
M. van Gils: *De middelmatige opsteljury,* Levende Talen no. 314 (okt. 1975), p. 416 vv.

7 Speciaal aanbevolen materiaal voor nader onderzoek: de eindexamenopgaven Nederlands tekstbegrip die sinds enkele jaren worden voorgelegd aan de overwegend theoretisch opgeleide leerlingen van het L.T.O.