De onbetrouwbaarheid van selektieve tekstbegriptoetsen

MAARTEN VAN GILS

Samenvatting

Dit artikel draagt een wat provocerende titel. Het pleegt een aanslag op de axioma's van de neo-Ebeliaanse psychometrie en kant zich met name tegen rigoreus selektief gebruik van objektieve tekstbegriptoetsen in A.V.O.-eindexamens.

Het bekende adagium: 'A test must be reliable to be valid . . .' (Ebel 1965: 343) is niet kompleet zonder de belangrijke toevoeging: '. . . but reliability does not guarantee validity.' Desondanks handelen diverse toetswiskundige publikaties voornamelijk over de 'betrouwbaarheid', terwijl theoretisch en empirisch gefundeerde bezwaren tegen de validiteit van multiple-choicetests naar verhouding weinig aandacht krijgen. De kwantitatieve informatie die in de C.I.T.O.-examenverslagen wordt verstrekt, heeft zelfs uitsluitend betrekking op het homogeniteitsaspekt; en bovendien maakt de begeleidende tekst lang niet altijd scherp onderscheid tussen validiteit en betrouwbaarheid.

Een nader onderzoek van het gekompliceerde homogeniteitsnetwerk wijst uit, dat alle daartoe behorende indices kunnen worden gedefinieerd als funkties van elkaar en van het aantal toetsitems. De CITO-voorlichting aan de buitenwacht kan zonder enig informatieverlies beperkt blijven tot slechts enkele gegevens: de gemiddelde item-testkorrelatie en de standaardmeetfout (uitgedrukt in tienpuntenschaal-intervallen) geven een onvertekend beeld van de testhomogeniteit en van de onzekerheidsmarges rond de cijfers die uiteindelijk aan individuele leerlingen worden toegekend.

Tenslotte kan worden gekonkludeerd, dat de multiple-choicetest ook op zijn 'sterkste' punt kwetsbaar is. Uit enkele meettechnische manipulaties met recent cijfermateriaal blijkt althans, dat in 1974 en 1975 de selektierelevante betrouwbaarheid van de eindexamentoetsen voor de vakken Frans, Duits en Engels niet indrukwekkend hoog was.

1. *Betrouwbaarheid als kompensatie voor invaliditeit*

De toets-wiskunde speelt tegenwoordig een belangrijke rol in de centraal-schriftelijke eindexamens A.V.O. Bij de vakken Frans, Duits en Engels meet men sinds enkele jaren het tekstbegrip van de kandidaten met behulp van multiple-choicetests. Deze wijze van examineren laat weinig ruimte voor menselijke beoordelingsfouten, omdat een groot deel van de korrektieprocedure objektief-mechanisch kan worden afgewikkeld. Toch heerst onder de moderne talendocenten geen algemene tevredenheid.

Uit een enquêteverslag dat opent met de zin: 'Studietoetsen hebben hun beperkingen' (Wesdorp 1975: 346), blijkt duidelijk, wat er volgens het 'veld' aan de nieuwe examens mankeert. Enerzijds zijn de vragen niet altijd voldoende *objektief*: de leraren verschillen dan van mening met de officiële korrektiesleutel (item no. 1 van de enquête). Anderzijds zijn ze vaak te weinig *specifiek* en kunnen de leerlingen schertsafleiders ontmaskeren zonder veel begrip van de aangeboden teksten (item no. 6, 7 en 10). Het valt inderdaad niet mee toetsitems te vervaardigen die zowel objektief als specifiek zijn (Ebel 1965: 296–300), vooral niet als het gaat om tekstbegrip op de hoogste leesniveaus (Van Gils 1976: 17). Op dit punt biedt het enquêteverslag van Wesdorp (1975: 354) eigenlijk geen verrassing: de *validiteit* van de talen-toetsen wordt door de docenten in twijfel getrokken, en volgens de rapporteur op goede gronden.

Van sociolinguïstische zijde is tegen de objektieve taalvaardigheidstest aangevoerd, dat deze de *communicative competence* geweld aandoet (Bonset 1975). Ook in dat licht bezien zijn Wesdorps konstateringen geloofwaardig. Als de rapporteur gelijk heeft, dan schieten de examentoetsen tekort op het meest wezenlijke punt: hun reden van bestaan is dubieus. Dat nadeel kan natuurlijk nooit worden gekompenseerd door eventuele kwaliteiten op andere gebieden.

Het 'tegenargument' van Wesdorp (t.a.p.) dat er

nauwelijks klachten zijn over de betrouwbaarheid, is dan ook moeilijk te plaatsen. Vooreerst is het maar de vraag, in hoeverre akkurate metingen van een zo gekompliceerde vaardigheid als *tekstbegrip* noodzakelijk, wenselijk en mogelijk zijn: het klassieke ideaal, 'a collection of items which has a high average correlation with total scores and is dominated by one factor only' (Nunnally 1967: 255) lijkt hier niet gemakkelijk te realiseren. Maar bovendien: als een tekstbegriptoets iets anders toetst dan tekstbegrip, dan doet het weinig meer ter zake, hoe akkuraat de metingen zijn van het attribuut dat de tekst wèl bestrijkt.

In dit artikel zal de *validiteit* van de talen-examens pas aan het slot weer ter sprake komen. Wij koncentreren ons in de hoofdstukken 2 tot en met 5 op de toetseigenschap waarover aanzienlijk meer meettechnische informatie voor handen is: de 'betrouwbaarheid' of *homogeniteit*. In het voetspoor van Ebel presenteren sommige psychometrici deze eigenschap nogal nadrukkelijk als een speciale kwaliteit van multiple-choicetests. Ik hoop hierna aannemelijk te maken, dat de praktische betekenis van de zogenaamde betrouwbaarheid niet mag worden overschat.

2. *Op zoek naar een neutraal referentiekader*

De verdiensten van objektieve studietoetsen worden veelal in een eufemistentaal beschreven. Sommige 'synoniemen' verdoezelen een wezenlijk onderscheid: *soortgenoot-validiteit* en *parallelbetrouwbaarheid* bijvoorbeeld. Andere termen hebben een Newspeak-funktie: zo stellen *underachievers* en *face-validity* de *subjektieve* lerarencijfers al bij voorbaat in een minder gunstig daglicht. Een tegengesteld jargon is best op te bouwen: men kan dan bijvoorbeeld spreken van *over-scorers* op *child-gradable multiple-guess tests* (Hoffmann 1967). Ik maak liever zo lang mogelijk gebruik van een 'neutraal' referentiekader: een wiskundig instrumentarium vergemakkelijkt technische operaties en kan tot falsifieerbare konklusies leiden. Men mag zulke konklusies nog altijd vanuit verschillende ideologieën interpreteren, maar zelfs dan moet daarover zakelijk worden gediskussieerd.

In de tabellen en in de begeleidende tekst van dit artikel komen diverse lettersymbolen en afkortingen voor. Sommige daarvan zijn in de literatuur gebruikelijk, andere werden speciaal om druktechnische redenen voor de gelegenheid ontworpen. Hun betekenis is als volgt.

Variantiegegevens

- k: aantal items. De te bespreken examentoetsen bestaan uit vijftig items, zodat steeds k = 50.
- SD: standaarddeviatie van een gegeven toets
- SD^2: testvariantie
- SE: standaardmeetfout
- SE^2: foutenvariantie
- Spq: som van de k itemvarianties

Homogeniteitsindices

- phi: (gemiddelde) item-itemkorrelatie
- RIR: (gemiddelde) item-restkorrelatie
- RIP: (gemiddelde)item-paralleltestkorrelatie
- RIT: (gemiddelde) item-testkorrelatie
- KR: (gemiddelde) test-paralleltestkorrelatie of homogeniteit (KR–20)
- F: signaal-ruisverhouding

Overige aanduidingen

- it.: aantal 'psychometrisch geslaagde items' (theoretisch maximum: it. = k = 50). Volgens CITO-normen wordt dit predikaat toegekend aan items met een p-waarde boven 0,50 en een RIT boven 0,30.
- M: gemiddelde testskore (theoretisch maximum: M = k = 50. In de examenpraktijk varieert M tussen 28 en 39)
- C: theoretische cesuurskore (varieert gewoonlijk tussen 22,5 en 33,5)
- % onv.: percentage examenkandidaten dat voor een gegeven toets een lager decimaal cijfer behaalde dan 5,45 (in de praktijk 10 à 25 %)
- M ± 2 SE: 95 % betrouwbaarheidsinterval rond de gemiddelde testskore

3. *Bijdragen aan de toetswiskunde*

3.1 *De Paramaribo-formules*

Jaarlijks na afloop van de eindexamens A.V.O. publiceert het CITO een groot aantal getallen, ook over de talen-toetsen. Elk afzonderlijk verslag bevat onder meer het aantal psychometrisch geslaagde items (it.), de homogeniteitskoëfficiënt die geacht wordt de parallelbetrouwbaarheid aan te geven (KR), de gemiddelde item-testkorrelatie (RIT), de standaardmeetfout (SE) en de standaarddeviatie (SD). Al deze gegevens, over 24 examentoetsen ontleend aan Meijering e.a. (1974; 1975), vinden we straks verzameld in tabel 2; de toetsen zijn daar gerangschikt naar hun homogeniteit, zodat in de eerste drie getallenrijtjes een regelmatige stijging zichtbaar is.

Hoewel de CITO-rapporteurs hun examens als

wijnjaren beschrijven, negeren ze de *relaties* tussen de diverse kwaliteitsaanduidingen. Vergelijkingen met de vorige oogst wijzen telkens uit, dat KR en RIT beide gestegen of beide gedaald zijn. In het laatste geval kan het kommentaar luiden: 'De gemiddelde RIT van de gehele toets (i.c. het examen VWO Duits 1974, M.v.G.) bedroeg 0,22, een lichte daling vergeleken bij vorig jaar (0,24). De betrouwbaarheid van het examen, uitgedrukt in de KR-20-formule, vertoonde eveneens een lichte daling: van 0,66 tot 0,62.' (Meijering e.a. 1974: 15). Dit soort informatie is nogal redundant, omdat KR en RIT in wezen hetzelfde meedelen (vgl. tabel 1 en 2). Het hechte verband tussen deze beide indices is bijvoorbeeld zichtbaar te maken met behulp van multipele-korrelatietechnieken.

Over de relaties tussen de gemiddelde RIT en de overige homogeniteitsindices zijn de meeste psychometrische handboeken opmerkelijk zwijgzaam. Zodoende presenteren ze de toetsalgebra als een geheimzinnig oerwoud van korrelatiekoëfficiënten, waarin alleen ervaren kaartlezers zouden kunnen doordringen. Toch is het plaatselijke wegennet minder ingewikkeld dan de reisgidsen suggereren. De voornaamste verbindingen zijn:

$$\text{phi} = \text{RIP} \times \text{RIT} = \text{RIP}^2 : \text{KR} = \text{KR} \times \text{RIT}^2$$
$$\text{KR} = \textit{RIP} : \textit{RIT} = \textit{RIP}^2 : \textit{phi} = \text{phi} : \text{RIT}^2$$

Deze eenvoudig te verifiëren formules (vgl. Gulliksen 1950: 88–107) berusten op de gebruikelijke vooronderstelling, dat items als parallelle subtests mogen worden behandeld (vgl. Guilford 1973: 414–416). Met behulp van deze basisrelaties kunnen we in principe alle gewenste homogeniteitsindices omschrijven als funkties van elkaar en van het aantal toetsitems. Bij wijze van voorbeeld volgen hier drie definities die voorzover ik weet in deze vorm nog niet zijn gepubliceerd. Ik heb ze uitgeschreven tijdens een verblijf in Suriname en presenteer ze daarom als de *Paramaribo-formules*.

$$1.\ \text{phi} = \frac{F}{F+k} = \frac{k(RIT^2)-1}{k-1}$$

$$2.\ \text{RIP} = \frac{F}{\sqrt{(F+1)(F+k)}} = \frac{k(RIT^2)-1}{(k-1)\,RIT}$$

$$3.\ \text{KR} = \frac{F}{F+1} = \frac{k(RIT^2)-1}{(k-1)\,RIT^2}$$

Bij het opstellen van deze formules, waarin RIT, RIP en phi opnieuw *gemiddelde* korrelaties representeren, kon ik gebruik maken van het feit dat KR zich tot F verhoudt als een gekwadrateerde sinus tot een gekwadrateerde tangens. Strikt genomen gelden de formules alleen voor toetsen die zijn opgebouwd uit parallelitems. Empirisch kan echter worden aangetoond, dat ze voor examentoetsen even goed voldoen als bijvoorbeeld de testverlengingsformule van Spearman en Brown. In de praktijk zijn RIP en RIR nagenoeg identiek, zodat de tweede Paramaribo-formule doorgaans bruikbaar is als definitie van de gemiddelde RIR. Iets nauwkeuriger is echter de volgende variant.

$$2a.\ \text{RIR} = \frac{F}{\sqrt{\{F + k/(k-1)\}\,(F+k)}} = \frac{k(RIT^2)-1}{\sqrt{k(k-2)\,RIT^2+1}}$$

3.2 *Het homogeniteitsnetwerk*

De Paramariboformules ontsluiten de *hoofdwegen* door het psychometrische oerwoud en tonen bovendien aan, dat het praktisch nut van de kronkelige *zijpaden* dubieus is. Een examentoetsanalyse die van elk item niet alleen de RIT vermeldt maar ook de RIR (en soms nog alternatief-test- en alternatief-restkorrelaties), is weinig meer dan een imposante verzameling tautologieën.

Het netwerk van homogeniteitsindices kan ad infinitum worden uitgebreid, bijvoorbeeld met de 'coefficient of alienation' en de 'index of forecasting efficiency' (Guilford 1973: 344–349). De termen suggereren enigszins dat het gaat om validiteitsaanduidingen, maar niets is minder waar. De bijbehorende formules berusten eenvoudig op het axioma dat elke toets zijn eigen kriterium is. Deze misvatting verdient naar mijn oordeel in voorkomende gevallen een krachtige bestrijding (Van Gils 1975: 418–426). In dit verband mag er nog wel eens tegen gewaarschuwd worden, dat CITO-publikaties de KR-20 presenteren als 'een getal dat aangeeft hoe constant de toets *de* vaardigheid van de leerlingen meet' (bijv. Meijering e.a., 1975: 14); het – door mij gekursiveerde – bepaalde lidwoord wekt de indruk dat er geen validiteitsvraag meer hoeft te worden gesteld.

In konkrete gevallen berekent het CITO de testhomogeniteit uit de p-waarden der items en de

testvariantie. De toegepaste methodes zijn rijkelijk ingewikkeld. Een efficiënt programma gaat uit van de gemiddelde item-testkorrelatie; heeft men hierna nog behoefte aan andere indices, dan kan de hoogte daarvan worden geschat met behulp van Paramaribo-formules.

1. $RIT^2 = SD^2 : k(Spq)$
2. phi $= \{k(RIT^2) - 1\} : (k - 1)$
3. RIP = phi : RIT
4. RIR = RIP (±)
5. KR = RIP : RIT
6. F = RIP : (RIT — RIP)

Omdat deze zes homogeniteitsaanduidingen transformaties van elkaar zijn, kan men ze desgewenst op vele andere manieren uitrekenen. Voor de nauwkeurigheid van de uitkomsten maakt het weinig uit, aan welke formules uit het netwerk men de voorkeur geeft; het lijkt raadzaam niet de moeilijkste te kiezen. Tabel 1 toont enkele getalswaarden die de indices konkreet kunnen aannemen voor 50-itemtoetsen. De onderlinge verschillen zijn goeddeels te verklaren uit het feit dat telkens in een andere schaal gewerkt wordt; daarbij moeten we overigens aantekenen, dat negatieve indexwaarden in de praktijk zelden voorkomen. De bovenste en onderste rijen van tabel 1 geven een aardige indruk van de theoretisch mogelijke minima en maxima; maar bij examentoetsen liggen de F's meestal tussen 2 en 7, zodat de tussenliggende rijen nu voor ons het interessantst zijn.

Tabel 1 wijst dan duidelijk uit, dat KR het best geschikt is om een hoge toetskwaliteit te suggereren. De zogenaamde betrouwbaarheid, eens getypeerd door Freudenthal (1975: 46) als 'een vergoelijkende functie van de toets-steekproeffout', komt al snel in de buurt van zijn maximum en onderscheidt zich des te gunstiger van RIT en phi naarmate het aantal items groter is. In de Ebeliaanse psychometrie staat de KR-20 zeer hoog aangeschreven, terwijl de rechtlijnige F en de zeer weinig 'vergoelijkende' phi in toetsanalyses zelden of nooit vermeld worden.

De *itemhomogeniteit* verschijnt in de CITO-verslagen als RIT en dus is de getalswaarde meestal positief. Bij de interpretatie van zo'n index moeten we echter rekening houden met de systematische fout die erin verstopt zit: als 50 parallel-items gemiddeld nihil interkorreleren, dan is elke RIT toch nog 0,1414, namelijk de reciproke vande wortel uit het aantal toetsitems (vgl. Guilford 1973: 321). Treffen we dus in een *voorbeeld*tabel van Van der Hoeven (1975: 150–151) negen items aan met een RIT van 0,14 of minder, dan is de konklusie onontkoombaar: in de desbetreffende herexamentoets zullen heel wat negatieve interkorrelaties zijn voorgekomen.

Het psychometrische netwerk is nodeloos gekompliceerd en bovendien vrijwel onbegrensd: publiekgericht werkende voorlichters zullen zich bij het hanteren van termen en begrippen enige beperking moeten opleggen. Wie aan talendocenten iets wil meedelen over test- of itemhomogeniteit, kan gelukkig volstaan met één redelijk doorzichtige, efficiënte rekeneenheid: daarvoor is zo op het oog de (gemiddelde) RIT nog het best geschikt.

Een lastig multiple choicevraagje tot slot van deze paragraaf. Wie het goed oplost, heeft het netwerk doorzien. Welke homogeniteitsindex kan als volgt worden gedefinieerd?

$$\{(k - 1)\,RIP + \sqrt{(k - 1)^2\,RIP^2 + 4k}\} : 2k$$

Vul in: a. KR *b.* RIT *c.* RIR *d.* phi.

Tabel 1: Samenhangende waarden van homogeniteitsindices (berekend voor 50-itemtoetsen)

F	KR	RIT	RIP	RIR	phi
— 1	min oneindig	0	min oneindig	—1	—0,0204
0	0	0,1414	0	0	0
1	0,5000	0,1980	0,0990	0,0985	0,0196
2	0,6666	0,2401	0,1601	0,1595	0,0384
3	0,7500	0,2747	0,2060	0,2055	0,0566
4	0,8000	0,3042	0,2434	0,2429	0,0740
5	0,8333	0,3302	0,2752	0,2747	0,0909
6	0,8571	0,3535	0,3030	0,3026	0,1071
7	0,8750	0,3746	0,3278	0,3273	0,1228
50	0,9803	0,7141	0,7001	0,7000	0,5000
100	0,9900	0,8205	0,8124	0,8123	0,6666
150	0,9933	0,8689	0,8631	0,8630	0,7500
oneindig	1	1	1	1	1

3.3 *Maximale item-testkorrelaties*

De in de toetspraktijk bereikte RIT-waarden liggen gewoonlijk aanzienlijk lager dan 1 en dat kan worden gezien als een verontrustend verschijnsel. CITO-medewerkster M. van der Linden-Mulder (1972: 10) neemt aan, dat de RIT's niet kunnen uitgaan boven bepaalde maxima die afhankelijk zouden zijn van de p-waarden der items. Zij brengt deze opvatting als volgt in kaart:

P–WAARDE	*MAXIMALE RIT*
100 en 0	0
90 en 10	0,57
80 en 20	0,70
70 en 30	0,77
60 en 40	0,78
50	0,80

Dit zonderlinge tabelletje berust vermoedelijk op een grafiek van Nunnally (1967: 132–133), die de maximale korrelatie uitbeeldt tussen een dichotome variabele met p-waarden tussen 0 en 1 en een normaal verdeelde variabele. Vertaalt men deze grafiek echter zonder meer naar de in Nederland gebruikelijke vierkeuzetoets, dan heeft dat consequenties voor bijvoorbeeld de KR-20: en men kan dan moeilijk blijven beweren, dat 'de betrouwbaarheid wordt uitgedrukt in een getal tussen 0 en 1' (Van der Linden-Mulder 1972: 5). Als namelijk een perfekt homogene toets theoretisch denkbaar is, dan volgt daaruit noodzakelijk, dat zo'n toets ongeacht de p-waarde der items met elk van die items perfekt positief zal korreleren. De vraag is dus maar, welke vooronderstellingen hier wèl voor RIT, maar nìet voor KR zijn ingebouwd.

Het is een van de basisassumpties van Nunnally (1967: 26), dat testskores altijd normaal zijn verdeeld. Echter: realistische frequentieverdelingen zijn moeilijk verenigbaar met maximale homogeniteit (Wijnen 1971: 106–113). Nunnally's assumptie impliceert dus, dat ook het KR-maximum aanzienlijk lager ligt dan 1. Voor een gegeven studietoets van 36 items berekende Wijnen (t.a.p.) een theoretisch maximale homogeniteit van 0,88; op grond van de formules uit paragraaf 3.1 zou daarmee een gemiddelde RIT van 0,44 overeenkomen. Homogene testverlenging tot 50 items kan desnoods een KR van 0,91 en RIT's van gemiddeld 0,43 opleveren. Binnen zo'n toets zouden naast RIT's van 0,57 à 0,80 (vgl. Van der Linden-Mulder) ter kompensatie veel lagere RIT's moeten voorkomen om het gemiddelde van 0,43 veilig te stellen.

Dit cijfervoorbeeldje bevestigt nog eens, dat KR stijgt bij homogene testverlenging, maar dat de RIT's dan geneigd zijn te dalen. Het laatste is niet moeilijk te verklaren. Elk item draagt zelf bij aan de testvariantie, maar deze bijdrage is van groter gewicht naarmate de toets korter is: in het meest extreme geval korreleert een toets die uit slechts één item bestaat, perfekt met zichzelf. Kortom, een definitie van de maximale RIT mag de faktor k niet verwaarlozen en moet verder berusten op dezelfde axioma's als die welke aan een definitie van de maximale KR ten grondslag liggen.

4. *Zwevende homogeniteitskriteria*

Wanneer is een selektieve toets homogeen genoeg om een rol te mogen spelen in het eindexamengebeuren? Het schijnt lastig te zijn deze beleidsrelevante vraag met een hard getal te beantwoorden: exaktheid van normen is bepaald geen kenmerk van de toets-wiskunde. Van der Hoeven (1975: 48) vindt een KR van 0,70 voor talen-toetsen 'vrij goed', maar Meijering e.a. (1974; 1975) vermelden herhaaldelijk, dat een KR van 0,75 'wenselijk is voor het nemen van beslissingen'. In de handboeken worden veel hogere maatstaven genoemd (KR's van 0,80 à 0,90), waaraan echter blijkens tabel 2 alleen de toetsen voor de heterogene MAVO-3-populaties weten te voldoen. Een en ander doet vermoeden, dat de CITO-normen meer op een empirische dan op een theoretische basis zijn gefundeerd.

Uit het vorige hoofdstuk valt onder meer te konkluderen, dat de 'normen' voor aanvaardbare hoogten van KR en RIT met elkaar zouden moeten harmoniëren. De KR-norm van 0,75 korrespondeert (bij 50 items) met een gemiddelde RIT van 0,27. Daarentegen verwachten Meijering e.a. van hun 'psychometrisch geslaagde items' een RIT boven 0,30; zouden per toets 50 items nog net aan die voorwaarde voldoen, dan mochten we een KR van 0,79 verwachten. Zelfs binnen één en hetzelfde normenstelsel is de speelruimte mijns inziens nogal groot.

Uit tabel 2 kunnen we aflezen, dat de examentoetsen in het algemeen voldoen aan de KR-norm van Van der Hoeven. Ongunstige uitzonderingen zijn alleen de VWO- en HAVO-opgaven voor het vak Duits, waarover enquêteur Wesdorp (1975: 353–354) vele klachten kreeg. We moeten overigens konstateren, dat ook deze toetsen in 1974 en 1975 onverkort werden gebruikt 'voor het nemen van beslissingen'. Als het er op aan komt, zijn de kriteria geheel vrijblijvend.

Niet anders is het gesteld met de normen voor de 'psychometrisch geslaagde items' (vgl. tabel 2). Wie

op dit punt een hard empirisch kriterium wenst, kan uit het resultatenlijstje wellicht de eis distilleren dat er van elke 50 examenvragen minstens 22 meettechnisch in orde moeten zijn. Maar het predikaat (of het ontbreken daarvan) heeft geen enkele consequentie: itemselektie à la Nunnally (Wijnen 1971: 114–119) blijft achterwege en alle items, psychometrisch geslaagd of niet, tellen gewoon mee 'voor het nemen van beslissingen'.

Het ontbreken van vastomlijnde homogeniteitsnormen is op zichzelf nog niet zo bezwaarlijk. Men beschikt immers altijd over de mogelijkheid om in de laatste fase van de skoreverwerking de meetvaardigheid van een toets in de normen te verdiskonteren. Zo zou bijvoorbeeld een lage homogeniteit automatisch een lage aftestgrens en weinig onvoldoendes kunnen opleveren, indien de cesuurbepaling zou geschieden volgens de methode van Wijnen (1971). Een duidelijk zwakke toets hoeft dan niet al te veel schade aan te richten, omdat met de spreiding van de decimale cijfers redelijk te manipuleren valt (vgl. Rommes 1974).

Uit tabel 2 zijn echter weinig duidelijke relaties te extrapoleren tussen KR en RIT enerzijds en het toetsgemiddelde (M), de cesuur (C) en het percentage onvoldoendes anderzijds. Hoogstens kan uit tabel 3 worden gekonkludeerd, dat de verschillen in moeilijkheidsgraad tussen de toetsen ten dele in verband staan met schooltype en vak. Het ziet er dus naar uit, dat van de hier geschetste mogelijkheden tot kompensatie van een gebrekkige meetbetrouwbaarheid in de praktijk weinig gebruik wordt gemaakt. Dat vermoeden zal in het volgende hoofdstuk worden bevestigd.

5. *Selektierelevante betrouwbaarheid*

5.1 *Standaardmeetfouten in verschillende schalen*

De zes homogeniteitsindices die in paragraaf 3.2 ter

Tabel 2: CITO-gegevens betreffende de tekstbegriptoetsen Duits, Engels en Frans (Centraal Schriftelijk Eindexamen MAVO, HAVO en VWO, 1974 en 1975)

No.	*Examen*	*vak*	*jaar*	*it.*	KR	RIT	SE	SD	M	C	% onv.
1	VWO	D	74	9	0,62	0,22	2,56	4,09	38,81	33,5	10
2	HAVO	D	75	10	0,66	0,24	2,98	5,23	33,32	28,5	18
3	VWO	D	75	10	0,68	0,24	2,70	4,87	36,14	31,5	17
4	HAVO	D	74	12	0,72	0,26	3,08	5,70	30,27	24,5	16
5	HAVO	E	75	20	0,73	0,27	3,11	6,10	30,85	26,5	24
6	MAVO-4	E	75	17	0,74	0,27	3,04	6,03	30,62	26,5	24
7	MAVO-4	E	74	16	0,76	0,28	2,86	5,83	32,77	29,5	27
8	VWO	E	74	18	0,76	0,28	2,85	5,85	36,15	30,5	17
9	VWO	E	75	19	0,76	0,28	2,84	5,93	36,31	30,5	17
10	HAVO	E	74	21	0,77	0,29	3,05	6,30	32,24	27,5	23
11	MAVO-4	D	74	15	0,77	0,29	3,01	6,28	34,19	28,5	18
12	HAVO	F	75	17	0,77	0,29	2,86	6,05	34,85	29,5	19
13	HAVO	F	74	22	0,77	0,29	2,82	6,04	36,76	30,5	16
14	MAVO-4	D	75	23	0,78	0,29	2,85	6,17	34,19	28,5	18
15	VWO	F	75	22	0,79	0,30	2,83	6,29	36,17	30,5	19
16	VWO	F	74	25	0,81	0,30	2,67	5,95	38,45	32,5	17
17	MAVO-3	E	74	26	0,81	0,31	2,97	6,80	32,12	27,5	24
18	MAVO-3	D	74	30	0,82	0,32	2,97	7,11	33,25	27,5	21
19	MAVO-3	E	75	33	0,82	0,33	2,91	6,94	32,12	27,5	24
20	MAVO-4	F	74	31	0,84	0,34	2,73	6,82	35,25	29,5	20
21	MAVO-3	D	75	35	0,85	0,35	2,77	6,96	36,23	30,5	21
22	MAVO-4	F	75	35	0,86	0,36	3,07	8,10	31,87	24,5	20
23	MAVO-3	F	74	30	0,87	0,37	3,05	8,42	28,56	22,5	24
24	MAVO-3	F	75	28	0,88	0,38	3,10	8,88	29,46	22,5	25
Gemiddeld (24 examens)				22	0,78	0,30	2,90	6,36	33,79	28,5	20

Tabel 3: Selektierelevante betrouwbaarheid van de tekstbegriptoetsen uit tabel 2

Rangno.	*Schooltype*	*Skoreschaal*		*Transformaties in de decimale schaal*		
vgl. tabel 2	vak en jaar	M	SE	Intervallen 5,45 — 10,0	SE	M ± 2 SE
	VWO					
1	Duits 74	38,81	2,56	16,5 × 0,2758	0,71	5,6 — 8,1
3	Duits 75	36,14	2,70	18,5 × 0,2459	0,66	5,3 — 7,8
8	Engels 74	36,15	2,85	19,5 × 0,2333	0,66	5,6 — 7,9
9	Engels 75	36,31	2,84	19,5 × 0,2333	0,66	5,6 — 8,1
16	Frans 74	38,45	2,67	17,5 × 0,2600	0,69	5,8 — 8,2
15	Frans 75	36,17	2,83	19,5 × 0,2333	0,66	5,6 — 8,1
	HAVO					
4	Duits 74	30,27	3,08	25,5 × 0,1784	0,55	5,5 — 7,5
2	Duits 75	33,32	2,98	21,5 × 0,2116	0,63	5,3 — 7,7
10	Engels 74	32,24	3,05	22,5 × 0,2022	0,62	5,3 — 7,6
5	Engels 75	30,85	3,11	23,5 × 0,1936	0,60	5,2 — 7,5
13	Frans 74	36,76	2,82	19,5 × 0,2333	0,66	5,8 — 8,1
12	Frans 75	34,85	2,86	20,5 × 0,2219	0,63	5,6 — 7,8
	MAVO-4					
11	Duits 74	34,19	3,01	21,5 × 0,2116	0,64	5,6 — 7,9
14	Duits 75	34,19	2,85	21,5 × 0,2116	0,60	5,6 — 7,9
7	Engels 74	32,77	2,86	20,5 × 0,2219	0,63	5,1 — 7,3
6	Engels 75	30,62	3,04	23,5 × 0,1936	0,59	5,2 — 7,3
20	Frans 74	35,25	2,73	20,5 × 0,2219	0,61	5,6 — 7,8
22	Frans 75	31,87	3,07	25,5 × 0,1784	0,55	5,7 — 7,9
	MAVO-3					
18	Duits 74	33,25	2,97	22,5 × 0,2022	0,60	5,6 — 7,8
21	Duits 75	36,23	2,77	19,5 × 0,2333	0,65	5,6 — 8,1
17	Engels 74	32,12	2,97	22,5 × 0,2022	0,60	5,3 — 7,6
19	Engels 75	32,12	2,91	22,5 × 0,2022	0,59	5,3 — 7,6
23	Frans 74	28,56	3,05	27,5 × 0,1655	0,50	5,5 — 7,4
24	Frans 75	29,46	3,10	27,5 × 0,1655	0,51	5,7 — 7,5

sprake zijn gekomen, delen iets mee over de verhouding tussen standaardmeetfout en standaarddeviatie: bij een KR van 0,75 (een van de CITO-normen) is SE =½SD en moeten we dus rekenen op 50 % steekproef-fouten. Het is helaas nooit te achterhalen, welke individuele testskores het meest van meetfouten te lijden hebben. Wel is SE een nuttig gegeven om de onzekerheidsmarges enigszins te schatten. Gerekend over verschillende toetsen is SE bovendien een tamelijk konstante grootheid, terwijl daarentegen SD in grootte meer varieert (zie tabel 2). Relatief de hoogste SD's komen (natuurlijk) voor bij de examens MAVO-3 en bij het vak Frans. Zijn k, SD en Spq bekend, dan kan SE rechtstreeks worden berekend met behulp van een formule die tot het homogeniteits-netwerk behoort:

$$SE^2 = SD^2(1 - KR) = \frac{SD^2(1 - RIT^2)}{(k - 1)RIT^2} = \frac{k(Spq) - SD^2}{k - 1}$$

Voor examenkandidaten en hun docenten is SE een veel interessanter gegeven dan KR of RIT. SE kan bovendien worden vertaald in de geijkte tienpuntenschaal, waarmee alle betrokkenen vertrouwd zijn. Hoe de transformatie van ruwe testskores in examencijfers verloopt, wordt onder meer uiteengezet door Van der Hoeven (1975: 151–153). Uit diens verslag blijkt, dat behalve het aantal items, de standaarddeviatie en de som van de itemvarianties nog een vierde inputgegeven van belang is, namelijk

de *theoretische cesuurskore*.

De theoretische cesuurskore C (zie tabel 2) ligt midden tussen de hoogst mogelijke onvoldoende en de laagst mogelijke voldoende prestatie, en komt overeen met het decimale cijfer 5,45. De bijbehorende hypothetische testskore wordt aangewezen via een 'absoluut' normeringssysteem dat volgens skeptici neerkomt op een toepassing van de Wet van Posthumus. Als tweede vaste punt fungeert altijd het decimale cijfer 10,0, dat toegekend wordt aan een kandidaat die alle 50 items korrekt heeft beantwoord. De plaats waar men de cesuur vastlegt, is dus beslissend voor de feitelijke grootte van een iteminterval en daarmee voor de feitelijke grootte van SE. De formule voor de berekening van standaardmeetfouten kan dus enigszins worden uitgebreid en toegespitst op de uiteindelijke examencijfers:

$$SE = \frac{10 - 5{,}45}{k - C}\sqrt{\frac{k\,(Spq) - SD^2}{k - 1}} = \frac{\sqrt{21{,}125(Spq) - 0{,}4225\,SD^2}}{50 - C}$$

Alleen in de noemer van de breuk speelt C een rol. Nu zal de *selektierelevante standaardmeetfout* volgens de formule groter zijn, naarmate C dichter bij 50 ligt. Is dus de gemiddelde testskore hoog en wil men toch aan een groot aantal leerlingen onvoldoende cijfers toekennen, dan kan in de slotfase zelfs een hoge toetshomogeniteit volledig worden verknoeid. Tabel 3 verschaft enig inzicht in de werking van dit mechanisme: zo heeft bijvoorbeeld de VWO-toets Frans 1974 (no. 16) oorspronkelijk een lage SE, die echter na vertaling in de decimale schaal als een der hoogste uit de bus komt. Voor de kandidaat die zich door het uitgereikte absolute cijfer gedupeerd acht, is het vermoedelijk maar een schrale troost, dat zijn relatieve plaats in de groep 'betrouwbaar' is vastgesteld.

5.2 *Royale betrouwbaarheidsintervallen*

De onzekerheidsmarges rond de feitelijk toegekende examencijfers komen het best uit, wanneer niet de standaardmeetfout, maar het betrouwbaarheidsinterval onze aandacht krijgt. Een 95-procentsinterval is in de sociale wetenschappen gebruikelijk; dat omvat ongeveer vier standaardmeetfouten, zodat de marges aan beide zijden van een tekstbegrip-cijfer ruwweg 1 à 1¼ punt in beslag nemen (zie tabel 3, laatste kolom). Hierbij doet zich de moeilijkheid voor, dat betrouwbaarheidsintervallen niet symmetrisch zijn ten opzichte van de ruwe testskores, maar ten opzichte van de bijbehorende 'ware'skores, die jammer genoeg onbekend zijn. Deze moeilijkheid, die beschreven wordt door Nunnally (1967: 220–221) en Wijnen (1971: 97–100), heb ik hier omzeild door per toets het 95-procentsinterval rond de *gemiddelde* skore als vergelijkingsobjekt te kiezen. Vervolgens heb ik aan de hand van Meijering e.a. (1974; 1975) mogen vaststellen, welke decimale cijfers binnen de bij elke examentoets gehanteerde schaal nog juist vielen binnen het gebied 'M ± 2 SE'.

Uit de laatste kolom van tabel 3 blijkt nu, dat voor een gegeven onderwijssituatie 'typische' kandidaten per jaar en per vak met uiteenlopende cijfers beloond kunnen worden, soms zelfs met cijfers boven òf onder de cesuurskore 5,45. Het 95-procents betrouwbaarheidsinterval is niet alleen onbehoorlijk groot, het heeft bovendien geen vaste plaats op de tienpuntenschaal. Een en ander komt natuurlijk, doordat telkens tegen het eind van de relatieve meetprocedures 'absolute' normen zijn ingevoerd in de berekeningen. Mutatis mutandis geldt ook voor andere dan 'gemiddelde' prestaties, dat de daaraan toegekende cijfers over de jaren heen moeilijk met elkaar te vergelijken zijn.

Dit probleem wordt ook gesignaleerd door Prick (1976). Hij suggereert, gesteund door een bekende tabel van Van den Ende en De Groot, de volgende oplossing: stem in de zeventiger jaren de AVO-examennormen af op de cijferkonventies die omstreeks Kerstmis 1947 van kracht waren in het VHMO, zodat bijvoorbeeld 25 % van de kandidaten een 4 of een 5 krijgt en 30 % een 6. Dankzij deze systematische Posthumusering zou telkens opnieuw bevestigd kunnen worden, dat de opeenvolgende leerlingenpopulaties kwalitatief niet uiteenlopen en dat het algemene taalvaardigheidsniveau van jaar tot jaar konstant blijft (vgl. Van der Hoeven 1975: 147–148). Mijns inziens leidt het opvolgen van Pricks advies in het gunstigste geval tot een gering verlies aan selektierelevante betrouwbaarheid en tot grote winst aan 'faith validity'.

Men kan een betere vergelijkbaarheid van landelijke examencijfers ook op andere manieren nastreven. Wie – in tegenstelling tot Prick – wil vasthouden aan lineaire transformatie van ruwe testskores, handelt waarschijnlijk het meest rationeel als hij relatieve cesuren hanteert die uitsluitend worden bepaald door gemiddelde en standaardmeetfout (Wijnen 1971). Kent hij bovendien aan M-2 SE, M en M + 2 SE vaste plaatsen toe op de tienpuntenschaal (bijvoorbeeld resp. 5,5, 6 en 6,5), dan zal het *informatiegehalte* van de centrale beoordelingscijfers zonder twijfel sterk toenemen. De *spreiding* van die

cijfers zal in het algemeen gering zijn bij een lage testhomogeniteit, misschien zelfs zo gering dat de theoretische raadskore als 'bijna voldoende' moet worden gewaardeerd (vgl. Rommes 1974). Maar ook in dat extreme geval kan men tegen de gang van zaken moeilijk gefundeerde bezwaren aanvoeren.

6. *Selektief gebruik van de tekstbegriptoetsen*

De multiple-choicetests hebben minder winst aan selektie-relevante betrouwbaarheid opgeleverd dan de CITO-voorlichting doet vermoeden. Wij kunnen de grote hoeveelheid homogeniteitsinformatie uit de examenverslagen over 1974 en 1975 in één simpel gegeven samenvatten: de standaardmeetfout van de gemiddelde talen-toets bedraagt ongeveer 0,6 punt op de gebruikelijke tienpuntenschaal (vgl. tabel 3). En daarna is de vergelijking met een meer traditioneel examenonderdeel illustratief. Het opstel voor Nederlands, dat als een zeer ónbetrouwbaar meetinstrument bekend staat, hield er bij het HAVO-examen 1972 een SE van naar schatting 0,8 punt op na (Van Gils 1975: 415). Het relatief geringe verschil tussen 0,6 en 0,8 kan moeilijk worden opgevoerd als een krachtig empirisch bewijs voor de stelling dat vierkeuze-items in het algemeen akkurater meten dan minder gesloten vraagvormen. Zeker mag men er nooit een verwijzing op baseren naar mogelijke relaties tussen betrouwbaarheid en validiteit (vgl. Ebel 1965: 376–395), in antwoord op theoretische en intuïtieve bezwaren tegen aard en inhoud van konkrete toetsitems.

De eindexamens Frans, Duits en Engels bestaan thans uit twee onderdelen: schoolonderzoek en landelijke toets. Naar ik meen worden de schoolonderzoekcijfers door de vakdocenten in het algemeen valide geacht; dat ze ook betrouwbaarder (homogener) zouden zijn dan de landelijke testresultaten, lijkt voorshands minder aannemelijk. Toch bevat de examenwetgeving zeer strakke bepalingen inzake het middelen van de beide eindcijferkomponenten. De vigerende afrondingsregels zijn tot op procenten van punten bindend, zelfs binnen de grenzen van een halve standaardmeetfout: (5,3 + 5,6) : 2 = 5,45 = 5, maar (5,4 + 5,6) : 2 = 5,50 = 6. Zulk wiskundig purisme is onder de gegeven omstandigheden niet verdedigbaar. Als de CITO-toetsen een rol blijven spelen in de selektieprocedure, dan kan men mijns inziens nog maar het best terugkeren naar de toestand van vóór 1973: de laatste verantwoordelijkheid voor de definitieve individuele eindcijfers berustte toen bij subjektief oordelende examinatoren en niet bij het objektieve toeval.

In recente publikaties over de AVO-eindexamens heeft het betrouwbaarheidsaspekt onevenredig veel aandacht gekregen. Het is zaak, dat voortaan de validiteit (weer) centraal komt te staan en dat de selektieve funktie van de huidige examens daarmee in verband wordt gebracht. Mocht nader onderzoek uitwijzen dat de kritische opmerkingen van Bonset (1975) en van de geënquêteerde talen-leraren (Wesdorp 1975) terecht zijn (persoonlijk acht ik dat nader onderzoek niet eens meer nodig), dan is er in toekomstige selektieve examens geen plaats meer voor tekstbegriptoetsen met vierkeuzevragen. Dit soort opgaven kan dan waarschijnlijk nog wel worden ingezet voor een strikt evaluatieve, niveaubewakende funktie.

Literatuur

Bonset, H., Enkele sociolinguïstische bezwaren tegen objektieve toetsen voor taalvaardigheid. *Levende Talen*, no. 313, aug. 1975, p. 335–345.

Ebel, R. L., *Measuring Educational Achievement*. Englewood Cliffs, New Jersey, 1965.

Freudenthal, H., Een internationaal vergelijkend onderzoek over wiskundige studieprestaties. *Pedagogische Studiën*, jg. 52, no. 2, februari 1975, p. 43–55.

Gils, M. van, De middelmatige opsteljury. *Levende Talen*, no. 314, oktober 1975, p. 409–428.

Gils, M. van, Toetsen tekstbegriptoetsen tekstbegrip? *Moer*, jg. 1976 no. 1, p. 12–19.

Guilford, J. P. en B. Fruchter: *Fundamental Statistics in Psychology and Education* (fifth edition). Tokyo etc., 1973.

Gulliksen, H., *Theory of Mental Tests*. London-Sydney 1950.

Hoeven, H. C. van der, De normering van het centraal schriftelijk meerkeuzewerk. *Levende Talen*, no. 311, april 1975, p. 146–154.

Hoffmann, B., *The Tyranny of Testing*, A reasoned attack on the exaggerated claims of mass 'objective' testing and its damaging effects on the vitality of the nation (tweede druk als pocket). New York-London, 1967.

Linden-Mulder, M. van der, Toelichting bij de in de verslagen gebruikte statistische en psychometrische begrippen, in: *De eindexamens 1972 in de vorm van meerkeuze-toetsen* (CITO-publikatie no. 23). Arnhem, 1972.

Meijering e.a., P. H., *De eindexamens AVO–VWO 1974 in de vorm van meerkeuze-toetsen* (CITO-publikatie no. 32). Arnhem, 1975.

Meijering e.a., P. H., *De eindexamens AVO–VWO 1975 in de vorm van meerkeuze-toetsen* (CITO-publikatie no. 39). Arnhem 75.

Nunnaly, J. C., *Psychometric Theory*. New York etc., 1967.

Prick, L., Het normeren van examens. *Levende Talen*, no.

317, april 1976, p. 190–196.
Rommes, A. E. N., Van toetsscores naar tentamencijfers. *Onderzoek van Onderwijs*, jg. 3 no. 2, juni 1974, p. 8–13.
Wesdorp, H., De eindexamentoetsen Moderne Talen. Meningen van MAVO-, HAVO-, VWO-docenten over de kwaliteit van de toetsen en de invloed ervan op het onderwijs. *Levende Talen*, no. 313, augustus 1975, p. 346–355.
Wijnen, W. H. F. W., *Onder of boven de maat*. Een methode voor het bepalen van de grens voldoende–onvoldoende bij studietoetsen. Amsterdam, 1971.

Curriculum vitae

M. J. A. M. van Gils (geb. 1939) doorliep te Breda het gymnasium alpha en studeerde Nederlandse taal- en letterkunde te Nijmegen (doctoraalexamen in 1965, hoofdvak: diachronische taalkunde). Hij is leraar Nederlands aan het Henric van Veldekecollege en taalbeheerser-vakdidacticus aan de Katholieke Leergangen.

Privé-adres: Churchill-laan 40, Maastricht