De subtoetsen voor stillezen van de Amsterdamse schooltoets 1969

E. F. Vroom

Bespreking van de beide stilleestoetsen van de A'damse schooltoets (basisonderwijs) 1969. Een aantal kwantitatieve gegevens. Bij het grote stuk vrij veel items met afwijkende p-waarden. Het bijzonder karakter (inhoudelijk) van tekstbegriptoetsen, problemen die voor de leerling daaruit voortvloeien. Kritiek op de objectiviteit van een aantal items. Relatie tussen p- en ris-waarden? Voldoen de toetsen aan de eisen van relevantie, specificiteit en evenwichtigheid? De beide toetsen zijn eerder als tekstbegriptoetsen te beschouwen, dan als stilleestoetsen. Voorstel tot ontwikkeling van diagnostische toetsen.

Hoewel 'objective tests' nog geen gemeengoed zijn in de verschillende sectoren van onderwijs, kan men zeker stellen dat het gebruik ervan een stijgende lijn vertoont, althans bij het basisonderwijs. Zo is de Amsterdamse schooltoets 1970 afgenomen van ca. 44.000 zesde-klassertjes, tegenover 6.000 in 1966. Constructie, afname en verwerking ressorteren nu onder het C.I.T.O., in samenwerking met het Nutsseminarie te Amsterdam. In zekere zin markeert deze taakoverdracht, dat een periode van experiment (toen het R.I.T.P.[1] zich met deze materie onledig hield) als afgesloten kan worden beschouwd. Of misschien is het beter om van een periode van terreinverkenning en voorbereiding te spreken, want het kan nog wel enkele jaren aanlopen voordat de toetsen inhoudelijk en qua vormgeving een definitieve gestalte hebben gekregen.

Wel ziet het er naar uit, dat, als de huidige omzetstijging doorzet, het jaarlijks toetsingsfestijn binnenkort tot de evenementen van nationale allure gerekend kan worden.

Een der voordelen van dergelijke, op grote schaal verrichte metingen is de verwerving van gegevens met een hoge statistische betrouwbaarheid. Daar de Samenwerkende Instituten deze kwantitatieve gegevens opsloegen in rapporten voor intern gebruik — publiceren kost geld en wat zou het nut geweest zijn? — ontstond als vanzelf de ietwat ongelukkige situatie dat de deelnemende scholen, mede door gebrek aan psychometrische kennis, alleen wat marginale opmerkingen wisten te plaatsen en niet tot een weerwoord kwamen dat zowel op het punt van toetsingstechniek als onderwijsdoelen tot een verantwoorde standpuntbepaling had kunnen leiden.

De bespreking van enkele subtoetsen, welke hieronder volgt, kan men beschouwen als een poging het gesprek over de waarde en de functie van dergelijke toetsen op gang te brengen in de (brede) kring van betrokkenen. De beperking tot de beide stilleestoetsen heeft het voordeel dat uitvoeriger kan worden ingegaan op detailkwesties (itemanalyse), omdat daardoor, beter dan bij een andere benadering (exemplarisch) de problemen van toetsconstructie worden belicht.

Tevens geldt dat kennis van de basisproblemen van toetsen voor tekstbegrip van belang is voor onderwijzenden in *alle sectoren van onderwijs, daar dergelijke toetsen voornamelijk in moeilijkheidsgraad* verschillen, hetgeen van reken- of kennistoetsen niet gezegd kan worden. Het zijn bovendien de toetsen die een voor het later schoolsucces zo belangrijk vermogen als de receptieve taalvaardigheid tot subject hebben en dan niet gesplitst in allerhande subvaar-

digheden (woordenschat, zinsbouw) maar als een *geheel* van activiteiten die slechts tot het vereiste resultaat voeren op voorwaarde dat de leerling ze op de juiste wijze *gecoördineerd* weet te ontplooien. Deze eis wordt door de overige taaltoetsen niet gesteld.

De beschouwing hieronder is geconcentreerd rondom twee aspecten, t.w. de plaats die de beide subtoetsen innemen in het geheel van de Amsterdamse Schooltoets 1969 (toegelicht aan de hand van een aantal kwantitatieve gegevens[2] en de beantwoording van de toetsen aan de geldende kriteria. (relevantie e.d.). Tenslotte worden enkele conclusies getrokken en suggesties gedaan.

1) Samen met het Nutsseminarie.

2) Ontleend aan De Amsterdamse Schooltoets, afname 1969 door Dr. E. Warries (Research Instituut voor Toegepaste Psychologie te Amsterdam, een intern rapport).

I. De plaats der beide subtoetsen voor Stillezen in het geheel

De toets omvatte 11 subtoetsen, nl. Hoofdrekenen I en II (70 items) en Rekenvraagstukken (25 items), *totaal Rekenen = 95 items.*

Algemene Kennis I en II, totaal *70 items.*

Extra vraagstukken, *35 items.*

Spelling (35 items), Woordbenoeming en Zinsontleding (35 items), Gemengde taalopgaven (35 items), Stillezen (groot stuk) (30 items), Stillezen (kleine stukken) (30 items) *Totaal taal 165 items.*

Totaal gehele toets *365 items.* Daarvan werden er enkele niet gescoord.

In de Extra Vraagstukken kwamen geen vragen van taalkundige aard voor, ze stelden uitsluitend rekenkundige problemen aan de orde. De Gemengde taalopgaven stelden de leerlingen voor de taken: de juiste aanvulling te leveren op een gedeelte van een vaste uitdrukking (item 1 t/m 7); een ontbrekend woord in te vullen in een zin (voorzetsels, werkwoorden) (item 8 t/m 17); een spreekwoord of staande uitdrukking te combineren met een tekstueel aanloopje (item 17, 18, 19), uit vier subzinnen met verschillende woordorde die te kiezen welke op een gegeven hoofdzin aansloot (item 20), een viertal samenhangende zinnen in de juiste volgorde te plaatsen (item 21 t/m 25), het tegengestelde te kiezen van een in een zin onderstreept woord (item 26, 27), het juiste voegwoord te plaatsen in een samengestelde zin (item 28 t/m 30).

in de beide stilleestoetsen kwamen geen hiermee vergelijkbare items voor.

Moeilijkheid en (meet)betrouwbaarheid

Het eerste wordt uitgedrukt door het percentage van de goede antwoorden.

De meetbetrouwbaarheid van een toets geven we aan door een betrouwbaarheidsindex, die maximaal 1,00 kan bedragen. In het algemeen streeft men naar een index waarvan de waarde ligt boven 0,80. Een lagere index hoeft niet direct te leiden tot het verwerpen van de testuitslagen. Wel dient men dan te rekenen met een grotere "meetfout"; deze maakt de uit-

subtoets	*moeilijkheid*	*betrouwbaarheid*
Alg. Kennis I	67 %	.67
Alg. Kennis II	68 %	.71
Hoofdrekenen I	69 %	.86
Hoofdrekenen II	69 %	.88
Rekenvr. stukken	64 %	.88
Gem. Taalopg.	79 %	.78
Woordben./Zinsontl.	78 %	.84
Spelling	65 %	.74
Stellezen gr. st.	75 %	.67
Stillezen kl. st.	67 %	.77
Extra vr. stukken	62 %	.83

slag per leerling aanzienlijk onzekerder dan bij een betrouwbare score het geval is.'
(De Amsterdamse Schooltoets, afname 1967 — R.I.T.P.).
De betrouwbaarheid van een toets zal groter zijn naarmate de toets meer items telt en deze niet te veel verschillen in gemiddelde moeilijkheidsgraad.

N.B. Alle betrouwbaarheidsindices zijn berekend op basis van 35 items per subtoets. In een aantal gevallen moest daarom de betrouwbaarheid volgens een schattingsformule worden bepaald.

Alg. Kennis I en *Stillezen gr.st.* bleken het minst betrouwbaar. In het laatste geval zal ongetwijfeld ook het verschil in moeilijkheidsgraad per item een rol hebben gespeeld. Zie onder.

Itemkwaliteit
De moeilijkheidsgraad van een item wordt aangeduid door het percentage goede antwoorden (p-waarde). In het algemeen streeft men naar waarden die iets boven 50 liggen, in het rapport van 1967 wordt echter gesproken over 75% goede antwoorden als ideaal.
Ook wanneer men de grens vrij ruim trekt ($55 < p < 85$) kan het aantal „minder geslaagde" items per (sub)toets nog vrij groot zijn. Zo telde ik van de laatste er in Hoofdrekenen-I 9 (op 35), in Hoofdrekenen-II 12 (op 33). In Stillezen gr.st. vond ik echter *20* items (op 30) die niet aan *deze* norm voldeden, in Still. kl. st. waren het er 9 (op 29, één item was niet gescoord). Daarbij hadden de te gemakkelijke items de overhand: 14 stuks in Still. gr. st. en 3 in Stillezen kl. st.

De r is-waarde is een betrouwbaarheidsaanduiding waarmee correlatief wordt uitgedrukt in hoeverre een item bijdraagt aan de betrouwbaarheid van de hele subtoets. Deze zal hoger zijn naarmate het item de homogeniteit van de toets bevordert. Dit laatste is echter afhankelijk van de overeenkomst en de samenhang tussen de vaardigheden (psychisch en intellectueel) waarop de items een beroep doen. Het *gemiddelde* der ris-waarden der items van een toets X kan dus opgevat worden als een indicatie van deze samenhang. Uiteraard kunnen hiervoor ook de „gewone" betrouwbaarheidsindices dienst doen, doch daar deze mede afhankelijk zijn van het *itemaantal*, komt het me methodisch gezien juister voor van de ris-gemiddelden uit te gaan, wanneer we ons tot de homogeniteit willen beperken.

In onderstaande tabel zijn de 11 subtoetsen gerangschikt naar de aflopende waarden der ris-gemiddelden.

Subtoets	I gem. rff	II betrouwbaarheid
Rekenvr. stukken	.48	.88
Hoofdrekenen II	.46	.88
Hoofdrekenen I	.45	.86
Woordben./Zinsontl.	.43	.84
Extra vr. stukken	.40	.83
Gemengde Taalopg.	.38	.78
Spelling	.37	.74
Stillezen kl. stukken	.37	.77
Stellezen gr. stukken	.35	.67
Alg. Kennis II	.31	.71
Alg. Kennis I	.31	.67

Ter vergelijking zijn in kolom II ook de normale betrouwbaarheidsindices opgenomen. Zoals de tabel toont, geven de beide kolommen een duidelijke parallellie in het verloop der cijfers te zien, wat ook te verwachten viel. Voor ons is echter allereerst de rangschikking der subtoetsen van belang. We zien dat de toetsen die zich occuperen met een complex van afgebakende en logisch-systematisch geordende materie het best uit de bus komen (naast rekenen ook woordbenoeming/zinsontleding). De stilleestoetsen hebben het echter niet verder weten te brengen dan een bescheiden plaatsje tussen Algemene Kennis (vragen naar losse, onsamenhangende weetjes) en spelling (weetjes die soms wel — bijv. werkwoordsvormen —, vaker niet geordend samenhangen). Frappant is de relatief grote homogeniteit (t.o.v. de overige taaltoetsen) van de Gemengde (!) Taalopgaven.

Een hoge subtoetshomogeniteit impliceert overigens niet, dat de door de betrokken toets aan de orde gestelde materie en/of vaardigheden representatief zou(den) zijn voor de hele taalmaterie en/of totale taalvaardigheid. Voor deze externe relaties raadplege men de correlatiematrix van het genoemde rapport.[3]

Tenslotte nog een *inhoudsaspect*, waardoor niet alleen déze twee toetsen maar toetsen voor tekstbegrip zich in een batterij altijd onderscheiden van de andere. Het bijzondere van dergelijke toetsen is immers dat zij niet uit 'nur' items bestaan, maar uit items en nog iets, nl. de tekst of teksten. Het verschil is niet zonder belang, dunkt me. Voor de leerling betekent het, dat hij voor de beantwoording van de stamvraag het „normale" typografische kader van de stam + nevengedrukte antwoorden moet doorbreken om elders de gegevens op te diepen die hij nodig heeft. Is dit gebeurd (bij Stillezen gr. st. pas na terugbladeren), dan kan hij zich aan de vergelijking van de keuze-antwoorden zetten maar door de plausibiliteit der alternatieven zal het keuzeproces (soms? vaak?) leiden tot een zodanige concentratie op het item dat de tekstgegevens weer ongemerkt achter de kim van het bewustzijnsveld wegzakken. De uiteindelijke beslissing is dan het resultaat van een proces van *kortsluiting* tussen stam en antwoordmogelijkheden en is genomen op grond van overwegingen waarin de tekstgegevens geen rol hebben gespeeld.

Een tweede onderscheid, eveneens samenhangend met de aard van dergelijke toetsen, is te vinden in de omstandigheid dat, in tegenstelling tot bijv. rekentoetsen, de keuze-antwoorden elkaar zelden uitsluiten. A is plausibel, maar voor B valt ook wat te zeggen. Wanneer de leerling niet voldoende op de toets is voorbereid en niet weet dat, bij een dilemma, altijd de tekst de doorslag geeft, zal hij eenvoudig bij zijn algemene kennis of gezond verstand te rade gaan. Een aantal leerlingen zal bovendien in het kinderlijk vertrouwen niet bedrogen te worden *alle* alternatieven als min of meer aannemelijk beschouwen, ook al staan ze in lijnrechte tegenspraak tot de tekst. Zij zien de keuze-antwoorden eenvoudig als aanvullingen op de tekst. Deze zegt wel dat de Hoangh-Ho traag stroomt omdat hij breed is, maar bij de antwoorden staat duidelijk: omdat hij was ingedamd. En dat zou er toch niet staan, als hij *niet* was ingedamd..

II. *Beantwoorden de toetsen aan de geldende kriteria?*

Zoals in de inleiding werd gesteld, kan met de overdracht van werkzaamheden aan het C.I.-T.O. de experimentele fase als afgesloten worden beschouwd. Daaraan dient echter nadrukkelijk de beperking te worden toegevoegd: voor wat de technisch-organisatorische kant van constructie, afname en verwerking betreft. In dit opzicht demonstreert de schooltoets *als geheel* een zeker vakmanschap. Toch vallen er ook op dit punt bij de toetsen voor Stillezen wel enkele kritische opmerkingen te plaatsen.

a. *gemiddelde moeilijkheidsgraad.*

Zie de eerder gemaakte opmerkingen. Vooral in Stillezen gr.st. is het aantal „minder ge-

slaagde" items hoog, te hoog.

b. *objectiviteit.*

Alternatieven dienen zodanig geredigeerd te zijn dat er *geen twijfel* kan bestaan aan de juistheid van het goede antwoord. Dit ter beoordeling van deskundigen, waartoe zich in dit geval elke ontwikkelde Nederlander mag rekenen. Die twijfel bestaat wel bij de volgende items:[4]

G8. De tekst zegt dat de draken wreed en bloeddorstig waren. Dit suggereert sadisme en moordlust, maar in de stam wordt alleen over verslinden gesproken. Antwoord B is dus zeker te verdedigen, m.i. is het zelfs beter dan C.

G14. Antwoord B is gelijkwaardig aan D. De gouverneur is de man die de soldaten tot zijn beschikking heeft. Dat wijsheid in deze zaak een middel voor een oplossing zou zijn, moet voor de onderdanen toch een vergezochte gedachte zijn.

K1. B is *onbetwijfelbaar beter* dan C. Men lette op het tijdstip waarop de vraag gesteld wordt
nl. nadat het *hele* stukje gelezen is. Het slimme antwoord van Bendoel geeft ons de zekerheid van zijn slimheid, die eerst nog slechts verondersteld werd (zo slim was als men zei).

N.B. Dit is ook het item met de laagste riswaarde. Misschien was deze waarde hoog geweest als niet de constructeur de fout had begaan het verkeerde alternatief als juist aan de computer toe te voeren.

K5. C en D zijn gelijkwaardig aan A. De beide antwoorden zijn volkomen plausibel, *op basis van de tekst.*

K6. Ook D is goed te verdedigen. Zou de brandweer (met groot materieel) uitrukken als het vuur al gedoofd was?

K16. A is niet onbetwijfelbaar juist omdat de tekst niet vermeldt wat de dierenbescherming *overal* (dus ook buiten Nederland en België) doet. Integendeel: de tekst suggereert min of meer dat de vergelijking tussen fauna- en dierenbescherming tot beide landen beperkt wordt.

Hierbij aansluitend nog de volgende kritiek. (formeel, redactioneel).

G7. Wanneer *vliegen* door de leerlingen wordt opgevat als *zich m.b.v. vleugels voortbewegen,* zal hij niet B maar C kiezen. De vraag lijkt me op deze wijze tot onnodige verwarring te leiden.

(G11, 12, 13) Het antwoord op de vraag van G12 is in de items G11 en G13 te vinden, wat toch niet de bedoeling zal zijn.

G18. A ontkent de zinnigheid van de gestelde vraag. De vertrouwde vraag-antwoordrelatie wordt daardoor wel erg verrassend doorbroken. Waren de leerlingen op dit soort verrassingen voorbereid?

G19. Op welk tekst-tijdstip is de vraag gesteld? Voor of na het moment dat de gouverneur zijn inval krijgt?

G24. Als de draken de brief tijdens de schemering ontdekken is hij waarschijnlijk eerder (zie A en B) aangeslagen. T*ezamen* zijn A en B dus wel plausibel. D (Dat kun je niet weten) geldt echter alleen als men tussen A en B moet kiezen. Bedoeld is echter: dat kun je niet weten, omdat de tekst niets zegt van het tijdstip van aflevering van de brief. Zal de leerling de bedoeling van D echter hebben doorgehad?

(K2) zie voor antwoord A de opmerking bij G19. (A is echter hier niet het juiste antw.).

K22. Waarschijnlijk hebben een aantal leerlingen — zelf kinderen — moeite gehad met de gelijkstelling: in zijn kinderjaren = de tijd waarin je op handen en voeten kruipt. Met de ochtend van zijn leven wordt dan ook eerder de kleuter-, nog beter *peutertijd* bedoeld.

(K27, 28, 30 K4) Volgens de handboeken verdienen direct gestelde vragen de voorkeur boven dergelijke halfaffe zinnen in de stam.

Voorzover de vorige aanmerkingen terecht zijn gemaakt, bewijzen ze dat ook puur vaktechnisch op de beide Stilleestoetsen nog wel iets valt af te dingen. Oorzaak daarvan zal wel zijn dat aan de eindredactie te weinig zorg is besteed. Wanneer er een proefafname

is gehouden, moet toch ook het grote aantal items met een van het gemiddelde ver afwijkende moeilijkheid zijn gesignaleerd.[5]

Ter illustratie van de stelling dat dit soort vaktechnische tekortkomingen de homogeniteit (en daardoor ook de betrouwbaarheid) van een toets schaden, geef ik hier de lezer een lijstje van 59 items der beide toetsen, tezamen ondergebracht in één reeks en geordend naar hun oplopende ris-waarden. De manipulatie lijkt me excusabel, (iets anders dan gerechtvaardigd!), omdat de toetsen in gemiddelde ris-waarde niet beduidend verschillen en bovendien voor hetzelfde doel zijn geconstrueerd.
kolom I — ris-waarde; kolom II itemaanduiding; kolom III — p-waarde;
kolom IV — items met afwijkingen van gemiddelde p-waarde ($55 < p < 85$) = meest linkse kruisje; items die niet aan de eis van objectiviteit voldoen of anderszins slecht zijn geredigeerd = meest rechtse kruisje.

N.B. de items die in de bespreking tussen haakjes zijn geplaatst, werden niet in kolom IV (rechtse kruisjes) opgenomen. Waarschijnlijk werd bij deze items de antwoordkeus door de slechte redactie nauwelijks beinvloed.
Vooral bij minder betrouwbare subtoetsen lijkt me de rangschikking der items naar hun ris-waarden een handzame methode om aan de analytische nabeschouwing een richtsnoer te bieden. Zoals uit de tabel blijkt, bestaat er verbond tussen lage klassering op de ris-index en zekere constructiefouten. In dit gebied vindt men ook een naar verhouding groot aantal moeilijke items, die daarom al verdacht zijn omdat men ze eerder bij de hoog geklasseerde zou verwachten. Men mag immers aannemen dat leerlingen die correcte maar moeilijke items goed weten te beantwooren, ook in staat zijn het merendeel der gemakkelijke items op te lossen, hetgeen ipsco facto tot een *hoge* ris-waarde zou leiden voor de moeilijke items. Men ziet deze veronderstelling globaal bevestigd voor de items met een ris-waarde van 30 en hoger. Middelt men bijv. de p-

I	II	III	IV	I	II	III	IV	I	II	III	IV
.05	K 1	43	x x	.35	K20	73		.43	K12	60	
.13	K29	32	x	.35	G16	96	x	.43	K14	85	
.14	G24	23	x x	.36	G25	95	x	.43	G 9	61	
.17	K30	13	x	.36	K 4	83		.43	G12	87	x
.18	G 8	12	x x	.36	K 6	75	x	.43	G27	60	
.20	K 5	91	x x	.36	K15	79		.44	K10	77	
.22	G28	96	x	.36	K19	78		.44	G30	64	
.24	G18	16	x x	.37	K24	89	x	.45	G 1	51	x
.24	G 7	48	x x	.37	G26	91	x	.45	G22	83	
.25	G 2	79		.38	G15	92	x	.45	K17	73	
.27	G21	98	x	.38	K22	62	x	.46	K28	78	
.29	K16	68	x	.39	G 6	94	x	.47	K 2	40	x
.29	K25	90	x	.39	G23	93	x	.48	K11	45	x
.29	K 7	79		.40	G 5	88	x	.48	K18	57	
.30	G17	93	x	.40	K13	51	x	.48	G10	82	
.30	G29	84		.41	K 3	82		.48	G13	80	
.31	G11	64		.41	G 4	92	x	.50	K23	63	
.33	G 3	95	x	.42	G19	69	x	.51	K 9	72	
.34	G14	86	x x	.42	G20	88	x	.54	K27	60	
.34	K26	76		.42	K 8	71		één item niet gescoord.			

waarden van de 9 laagst geklasseerde items, daarna die van het laagste tweede tot en met tiende item, daarna het derde t/m elfde enz., dan komen de volgende cijfers te voorschijn (in horizontale volgorde).
(Opm. Elk getal dient nog door 9 gedeeld te worden).
374 - 410 - 478 - 520 - 597 - 664 - 666 - 654 - 702 - 749 - 756 - 734 - 739 - 746 - 762 - 752 - 743 - 758 - 741 - 743 - 758 - 777 - 743 - 743 - 752 - 765 - 737 - 741 - 744 - 766 - 762 - 769 - 735 - 728 - 701 - 737 - 715 - 700 - 695 - 658 - 670 - 683 - 676 - 655 - 613 - 610 - 615 - 631 - 639 - 628 - 615.
De gecursiveerde waarde geeft de som van 9 items, waarvan het centrale item een riswaarde van .30 heeft. Men ziet dat daarna de som van de p-waarden een tijd lang vrij constant blijft, om tegen het eind te zakken tot een minimum van ca. 620:9=69 (gemiddelde laagste p-waarde).[7]

Relevantie, specificiteit, evenwichtigheid
Was de hierboven geleverde kritiek (objectiviteit, moeilijkheidsgraad) allereerst bestemd voor de toetsconstructeurs, we komen nu op een terrein waar de problemen slechts in samenspraak met het onderwijs tot een oplossing kunnen worden gebracht. Problemen als: wat is eigenlijk tekstbegrip, wat zijn daarvan de componenten, hoe hoog mogen de eisen zijn die wij aan de zesde-klassertjes in dit opzicht kunnen stellen, enz. Onnodig te zeggen dat het gesprek hierover nog helemaal op gang moet komen. Een gesprek ook dat hopelijk tot een verscherpt inzicht leidt op het punt van taalleerdoelen, omdat de toetsen deze leerdoelen dermate concretiseren dat ook wij gedwongen zullen worden ons eenduidig uit te drukken.

'Afgezien van de praktische verdienste van de toets mag als fundamentele verdienste van de schooltoets het volgende genoemd worden.

De toets verschaft de mogelijkheid onderwijsdoelstellingen te concretiseren in de vorm van opgaven en schept aldus de mogelijkheid de discussie over leerdoelen concreter en naar wij hopen ook effectiever te maken. (....). Het C.I.T.O. hoopt, mede dankzij de respons die het uit het onderwijs hoopt te verkrijgen op de toets, een volgende serie nog beter aan te doen sluiten bij de huidige ontwikkelingen op leerplangebied.'[7]

c. Relevantie
Daar het leerdoel 'tekstbegrp' ook in de Proeve[8] maar vaag is omschreven, moesten de constructeurs voor een groot deel zelf bepalen wat relevant geacht kon worden. Een verantwoording van hun keuze hebben zij echter niet verstrekt. Door deze nalatigheid is het vrijwel onmogelijk het eindresultaat te evalueren, daar er geen gegevens zijn waartegen men dit resultaat zou kunnen afzetten. Probeert men, uitgaande van de items, door terugredeneren te achterhalen wat de constructeurs met elk item hebben 'voorgehad', dan ziet men zich in het merendeel der gevallen geplaatst tegenover de mogelijkheid van een *aantal* intenties. De oorzaak hiervan is dat door de vraagstelling in de vierkeuze-vorm niet slechts een bepaalde 'zuivere tekstbegripvaardigheid' wordt gemeten, maar *bovendien* het vermogen van de leerling deze vaardigheid te effectueren in een itemitisch bepaalde situatie. De schijnplausibiliteit van de afleiders kan het oplossingsproces gemakkelijk een zodanige wending geven, dat in feite een andere vaardigheid dan de (mogelijk) bedoelde de uiteindelijke keuze bepaalt. Ook andere factoren kunnen echter tot (misschien onbedoelde) verschuivingen leiden.
Enkele voorbeelden.
De items G 21 en G 13 zijn beide reproductievragen. Zo zegt de tekst dat Tsjoe Tsjou zijn karakters met een penseel tekende en G 21 vraagt: Waarmee schreef Tsjoe Tsjou? De alternatieven luiden: A. met een ganzeveer, B. met een pen, C. met een penhouder, D. met een penseel. De afleiders leveren hier voor de leerling geen enkel probleem op. Over geen van de genoemde schrijfinstrumenten wordt in de tekst gesproken, zij zijn duidelijk 'verzonnen'. Resultaat p = 98, ris .27.
G 13 vraagt: Tsjoe Tsjou schreef met een

penseel de karakters duidelijk en fors, maar tegelijkertijd ook:
A. duidelijk en sierlijk, B. fors en vlot, C. sierlijk en fors, D. vlot en sierlijk.
De tekst geeft: fors en duidelijk maar tegelijk sierlijk en vlot. Hoewel het item op het eerste gezicht een nogal onbenullige indruk maakt, blijken de leerlingen moeite te hebben met de opgave door vergelijking van *elk* alternatief met de tekst tot een juiste keus te komen. (p = 80). Hoe zou de uitslag echter geweest zijn wanneer er geen afleiders in het spel waren geweest om verwarring te stichten? De hoge ris-waarde (.48) wekt het vermoeden dat het vermogen om in dit soort situaties het hoofd koel te houden van aanzienlijke invloed is op de totaalscore.
Er zijn meer van dergelijke itemparen aan te halen. 'Ik eet nooit mensenvlees meer', besloot de Groene draak. G 28 vraagt: Wilde de Groene Draak na de brief nog mensen verslinden? Alternatieven A, B, C, zijn pure verzinsels, die nergens door de tekst worden gesteund en D zegt: 'Neen, hij zou dat nooit meer doen.' Resultaat p = 96, ris. 22.
Tekst: . . snikte de Rode, die ontroerd was door zijn eigen goedheid.
G 27: Waardoor was de Rode Draak ontroerd. A. door de mooie letters van Tsjoe Tsjou, B. door de wijze woorden van Tsjoe Tsjou, C. door medelijden met de mensen, D. door zijn eigen goedheid. Laten we aannemen dat ook dit item als een reproductievraag is bedoeld. Maar nu is het verschil in waarde tussen de afleiders enerzijds en het juiste alternatief anderzijds veel geringer. D. (juist) wordt expliciet door de tekst verstrekt, A en B niet maar zijn plausibel op grond van de tekst, C misschien niet, hoewel aan de plausibiliteit sec niet valt te twijfelen. Resultaat p = 60, ris .43.
De conclusie uit het vorige is, dat een omschrijving van 'zuivere' tekstbegripsvaardigheden, hoe gedetailleerd ook, nog niet garandeert dat daarop afgestemde items noodzakelijk die vaardigheden op identieke wijze aan de orde stellen. Edoch, daar kan men nog vrede mee hebben. Bedenkelijker wordt de zaak echter als de vierkeuze-vorm ertoe leidt dat de leerling tot foutieve beslissingen komt omdat hij over de *bedoeling* van het item in het onzekere verkeert. Hij effectueert dan een subvaardigheid waarop dit item toevallig *niet* mikte. Bijv. hij trekt een conclusie, of demonstreert zijn inzicht, terwijl hij slechts tot reproduceren werd uitgenodigd.
Ook hiervan enkele voorbeelden.
K7. Was Johan Tippel met een taxi naar de telefooncel gekomen?
Tekst: De heer Tippel verliet de telefooncel en rende naar de nog op hem wachtende taxi, die hem naar de IJhaven moest terugbrengen.
A. Ja, de taxi stond er nog, B. Ja, want hij kwam van het IJ, C. Nee, bij de telefooncel stond toevallig een taxi, D. Nee, want bij het IJ is ook een telefooncel.
Indien ook dit item slechts om reproductie vraagt had de vraag beter kunnen luiden: Waaruit blijkt dat J. T. met een taxi naar de telefooncel was gekomen? Antwoord: de taxi stond er nog (of wachtte nog op hem). De reden waarom niet deze vraag gesteld is, zal duidelijk zijn: men had met het probleem van de afleiders in de maag gezeten. Vermoedelijk hadden ze geen van alle hun afleidende taak naar behoren vervuld en was het item veel te gemakkelijk geworden.

In dit geval koos 13% der leerlingen het antwoord B.
De kans bestaat, dat er een aantal geweest zijn die antwoord A eenvoudig te simpel gevonden hebben. Ze gingen liever een stapje verder: ja, want met de taxi die daar nog op hem stond te wachten was hij van het IJ gekomen.

Ook dit geval ('onder'-vraging) staat niet op zichzelf.
Een zeer illustratief voorbeeld is al eerder aangehaald, nl. K1 (zie objectiviteitsbspreking). Ik noem verder, zonder commentaar, G2-B, G7-C, G9-D, G14-B, G18-B en C, G24-A en C, G30-A, K5-C en D, K6-C en D, K13-C en D; dit lijstje is mogelijk niet compleet en vatbaar voor kritiek. Een discussie over deze

kwesties zou echter ongetwijfeld aan het licht brengen dat items voor tekstbegrip vaak zo complex van aard zijn, dat een adequate afstemming op de te toetsen subvaardigheden vooralsnog een kwestie schijnt van goed geluk.

Tenslotte nog enkele andere zaken die direct of indirect met het probleem van de relevantie in verband staan.

Slechts in drie items werd uitdrukkelijk gevraagd waarom of hoe iets was *volgens de schrijver* of *het verhaal?* Was het bij de andere items dan niet de bedoeling dat de leerlingen een keus maakten *op grond van de tekst*? Zo ja, was hun dit in voldoende mate bekend?

In sommige gevallen geven de bewoordingen van de stamvraag een zeer duidelijke localisering van het tekstgedeelte waar het antwoord is te vinden, in andere gevallen niet of nauwelijks. Hetzelfde geldt voor de alternatieven. Bij gebrek aan een verantwoording valt niet uit de toets op te maken of in de keuze van deze itemtypen methodisch te werk is gegaan.

Soms is het tekstgedeelte waarnaar vraag en juiste antwoord verwijzen zeer beperkt, dan weer beslaat het vele regels. Opzet, toeval?

In het grote stuk komen geen samenvattingsvragen voor die op het *geheel* slaan. Dienen dergelijke opgaven niet tot de leerdoelen gerekend te worden?

Het zelf hoofdzaken van bijzaken kunnen onderscheiden mag men toch ook als een relevante vaardigheid beschouwen. Deze komt nu niet aan bod door de gerichte vragen van de itemschrijver die daardoor zèlf bepaalt wat hij van belang vindt.

De kleine stukken zijn alle zakelijk van inhoud. Teksten met een meer lyrische inslag (sfeer, stemming) ontbreken, alsmede teksten van een humoristisch, ironisch of badinerend karakter. Hoort het begrijpen of aanvoelen van dergelijk proza dan niet tot de doelstellingen van het taalonderwijs?

Uit hoofde van de eis van *specificiteit* dient een toets geen kwesties aan te roeren, die in *deze* toets niet op hun plaats zijn. In een grammaticatoets moeten geen spellingsvragen worden gesteld, en omgekeerd. De items behoren qua intentie a.h.w. te convergeren naar het doel dat vóór de constructie is gedefinieerd, nl. als het complex van kennis en/of vaardigheden waarvan men het bezit wil onderzoeken.

Welke zijn echter de vaardigheden die het stilleesonderwijs aan de leerlingen wil bijbrengen? En gaat het alleen maar om 'vaardigheden' en niet ook om de verwerving van een aantal attitudes, zoals nauwkeurigheid en concentratie. Laten we eens nagaan waartoe het leesonderwijs tenslotte voert.

a. De toename van de technische vaardigheid. Een snellere en met minder moeite verlopende visuele en semantische identificatie van de woordsymbolen en de combinaties daarvan. De opnamebreedte van het lezend oog neemt toe en dankzij de verworven routine kan de leerling mettertijd ook lange verhalen en dikke boeken 'aan', de leesperiode verlengt zich, soms tot een tijd van enkele aaneengesloten uren.

b. Het vorige leidt weer tot een geleidelijke functieverandering van het verwerkend bewustzijn. Hoe groter de technische vaardigheid, des te meer 'ruimte' komt er vrij voor processen op hogere niveaus. Het geheugen krijgt zijn kans, omdat de nieuwe inhouden de oude niet meer volledig verdringen, zodat bijvoorbeeld binnen het bestek van een alinea oud en nieuw geïntegreerd kunnen worden en als beeld- of begrips'gestalten' een groter tekstgeheel 'darstellen', dit ondanks het line-air-temporele verloop van het opnameproces. De leerling gaat passages 'overzien', ook meer ingewikkelde samengestelde zinnen blijken, zij het via zijwegen, tenslotte toch in een 'mededeling' uit te monden (zij het van een fijner gestructureerd karakter met hoewels, tenzijs, omdats, desondanks' e.d.).

c. De greep op de inhoud is intussen losser geworden èn meeromvattend. Het kind gaat a.h.w. op afstandlezen waardoor weer ruimte vrijkomt voor reflecties op de inhoud, zonder

dat daardoor de receptie wordt gestagneerd. De grotere distantie doet gemakkelijker de 'draad' van het verhaal zien en vasthouden en permitteert een voortdurende selectiviteit tijdens de voortgang, wat tot allerlei fluctuaties in de leesintensiteit kan voeren. 'Nou ja, dat is van minder belang, dat gaat weer over . . .', of: 'nu wordt het spannend zeg', enz.

d. Tenslotte kan het lezen een duidelijk doelbewust karakter krijgen; verschillende leeswijzen ontstaan. Studerend lezen, oriënterend lezen, splitsen zich af als specifieke vormen, soms spontaan, soms aangeleerd. Typografische signalen zoals cursivering, alineëring krijgen samen met ondertiteling, paragrafennummering en inhoudsopgaven een functionele waarde, de benutting ervan beïnvloedt tevens de leesefficiëntie, ook bij het ontspannend lezen. Jeugdboeken worden eerst even doorgesnuffeld, men wijst elkaar op de bladzijden die bijzonder leuk of spannend zijn. Naast de leesvaardigheid als geschetst onder *a* en *b* ontwikkelt zich een meer kritische instelling, tot uiting komend in een genuanceerder oordeel: 'een beetje langdradig, in het begin nogal moeilijk, hoofdstuk 3 had best weggelaten kunnen worden, enz.'

Op deze grove schets moge het een en ander zijn aan te merken, in grote lijnen zal het beeld toch wel juist zijn.

Hoe verhouden zich nu de beide stillees-(!)toetsen tot de door het onderwijs ontwikkelde vaardigheden? Het antwoord is duidelijk. De constructeurs gaan blijkbaar van de veronderstelling uit dat men het *hele* complex van *stillees*vaardigheden uit toetsttechnisch oogpunt gezien mag gelijkstellen aan de vaardigheden die in toetsen voor *tekstbegrip* worden geëffectueerd. Een bewijs voor de gerechtvaardigdheid van deze handelwijze ontbreekt echter. De gang van zaken doet denken aan een selectieproef voor tienkampers in de atletiek waarbij men op grond van de prestaties voor één onderdeel tot een geschiktheidsoordeel komt.

Als stilleestoetsen zijn de opgaven eenvoudig *te* specifiek. Het zijn tekstbegriptoetsen. Daar deze qualitate qua een zeker even groot beroep doen op de (algemene) intelligentie als op taalvaardigheid sec, zullen ze in de hele batterij maar in geringe mate bijdragen aan de onderzoekbreedte.

Tenslotte nog de eis van *evenwichtigheid*. We zullen, om hierover iets zinnigs te kunnen zeggen, de toetsen als tekstbegriptoetsen opvatten en hen vanuit dit gezichtspunt beschouwen.

Daar de theorievorming m.b.t. de subvaardigheden die samen de vaardigheid 'tekstbegrip' constitueren nog geheel in het stadium van het voorbereidende grondwerk verkeert, dient ook een uitspraak over de distributie van de items over de subvaardigheden te worden opgeschort, totdat door onderzoekingen een duidelijk beeld is verkregen van de *samenhang* tussen deze vaardigheden en het belang dat men aan elke categorie, afzonderlijk moet toekennen. Waarschijnlijk zal het nodig zijn voor dit onderzoek een aantal toetsen te ontwerpen die ieder een maximale homogeniteit vertonen, dus uitsluitend reproductievragen, of uitsluitend inzichtsvragen enz. bevatten. Het type true-false-item is daarvoor vermoedelijk beter geschikt dan het vierkeuze-item, gezien de complicaties die het laatste schept, hetgeen ze voor researchdoeleinden minder geschikt maakt. Ook de ondoorzichtigheid van de hier besproken toetsen is voornamelijk aan deze complicerende factor te wijten. Men dient steeds weer de afleiders buiten beschouwing te laten, als men de items naar hun intentie wil determineren, maar uiteraard heeft een dergelijke benaderingswijze geen enkele zin. De problemen dienen *vanuit de leerling* aan de orde gesteld te worden en voor hem fungeren de afleiders niet als de technische middelen die objectieve scoring mogelijk moeten maken, maar als in beginsel reële antwoordmogelijkheden. In een aantal gevallen vindt hij het goede antwoord dankzij de informatie die de foutmogelijkheden verschaffen. De reproductievraag of inzichtsvraag is een intelligentie-

vraag geworden door de rol die de afleiders hebben gespeeld in het oplossingsproces.

III. *Samenvatting, conclusies, suggesties*

In de Amsterdamse Schooltoets 1969 blijken de beide subtoetsen voor Stillezen (tezamengenomen) qua betrouwbaarheid en homogeniteit acher te blijven bij de rekentoetsen. Vooral het grote stilleesstuk steekt ongunstig af. Bij de laatstgenoemde toets bleken 20 items (op een geheel van 30) af te wijken van de gestelde 'norm' voor de gemiddelde moeilijkheidsgraad. De distributie van dergelijke items schijnt een zekere samenhang te vertonen met de plaats die zij innemen volgens de rangschikking naar hun oplopende ris-waarden.
Misschien kan het zelfde gezegd worden van de items met een betwijfelbare objectiviteit.

De beide toetsen werden tevens geconfronteerd met de eisen van relevantie, specificiteit en evenwichtigheid. Daar de toetsmakers nalieten een verantwoording aan hun werkstuk toe te voegen, kwam de bespreking min of meer in de lucht te hangen.
Wel kon gesteld worden, dat deze zgn. stilleestoetsen hoofdzakelijk als tekstbegriptoetsen functioneren, waarbij echter werd aangetekend dat de vierkeuzevorm voornamelijk een beroep doet op rationeel-discriminerend denken (alternatievenvergelijking). Is dit inderdaad de belangrijkste vaardigheid in het complex van vaardigheden dat men voor 'verwerken' van teksten nodig heeft? Ik weet het niet, ik weet ook niet of het C.I.T.O. het weet. Er zijn meer vragen op dit terrein. Bijvoorbeeld: hebben de toetsconstructeurs zich gebaseerd op researchresultaten t.a.v. de processen van receptieve taalverwerving en staat hen een hiërarchie van vaardigheden voor ogen welke samen de vaardigheid 'tekstbegrip' constitueren? Zo ja, dat men zijn geheime wetenschap den volke kond maken. Zo niet, wat kopen we dan in hemelsnaam voor al deze 'objectiviteit'? Een mogelijkheid leerlingen te vergelijken t.a.v. hun bekwaamheden in het maken van vierkeuzeproefwerken, van alternatieven vergelijking en -eliminatie, van rationeel-discriminerend denken? Is dat de einddoelstelling van het stilleesonderwijs op de basisschool?

Moge ik tenslotte, als leraar Nederlands, hier enige wensen op tafel leggen m.b.t. de toekomstige activiteiten van het C.I.T.O.-R.I.T.P.
Als gevolg van de democratisering van het onderwijs en de wenselijkheid het selectiemoment naar de 15-jarige leeftijd te verplaatsen, zullen de schooltoetsen voor het basisonderwijs steeds meer hun zin verliezen. Scholengemeenschappen met driejarige brugperioden en middenscholen zullen op een veel betrouwbaarder wijze de leerlingen zichzelf laten voorsorteren voor de opleidingen van hun keuze en begaafdheid. Deze te verwachten ontwikkeling stelt echter het probleem aan de orde van onderwijsmethodes voor zeer heterogene groepen waarbij elke individuele leerling krijgt wat hem toekomt, zoveel mogelijk aangepast aan zijn persoonlijke behoeften. Om die te kennen zal het onderwijs allereerst behoefte hebben aan *diagnostische* toetsen, waaruit de leraar leert wat er verkeerd is aan zijn didactiek en de leerling wat hèm ontbreekt. Dergelijke toetsen zullen dus niet gebruikt worden om te selecteren, maar als bezinningsmomenten functioneren tijdens het leerproces, om daaraan richting te geven. Het zal daarom ook niet mogelijk zijn ze onafhankelijk van de gevolgde didactische praktijk te ontwikkelen maar ze dienen samen daarmee één 'unit' te vormen voor een bepaalde, beperkte onderwijsperiode. (bijv. 6 weken). Dat bij de samenstelling van deze 'units' en unitsseries (vaak vertakt geschakeld) het proces van de taalverwerving (zoals ook het inlopen van achterstanden) optimale kansen moet krijgen, is duidelijk, maar evenzeer is duidelijk dat we onderzoeksvormen zullen moeten ontwikkelen die aan het licht brengen, waaróm we doen wat we doen, en wel of geen resultaat boeken.
Ik meen dat *deze* vorm van dienstverlening aan het onderwijs van onschatbaar meer waarde zal zijn.

Noten

1. samen met het Nutsseminarie.
2. Ontleend aan De Amsterdamse Schooltoets, afname 1969 door Dr. E. Warries (Research Instituut voor Toegepaste Psychologie te Amsterdam) een intern rapport')
3. Bijv. correlatie met het rapportcijfer. Spelling .46; Woordben./zinsontl. .44; Still. kl. st. .38; Gem. Taalop. .33; Still. gr. .32.
4. K voor itemnummer = Stillezen kleine stukken, G = groot stuk.
5. Door tijdgebrek werd er in 1970 in elk geval geen proefafname gehouden.
6. Wellicht geven de rekentoetsen een beter ideaalbeeld van de samenhang tussen ris – en p-waarden.
7. Zie Schooltoetsen basisonderwijs, verantwoording inhoud toets 1970 – pag. 5.
8. Proeve van een leerplan voor het basisonderwijs – Nutsseminarie, 1967. Deel B 1968.

Over de schrijver

Leraar Nederlands aan het Roncalli College te Bergen op Zoom.