# CITO-commentaar op Van Gils': 'De onbetrouwbaarheid van selectieve tekstbegriptoetsen''*

K. D. Thio
*CITO, Arnhem*

*De heer Van Gils is, in overleg met de redactie van Pedagogische Studiën, zo vriendelijk geweest ons de tekst van zijn artikel te voren toe te zenden. Zodoende heeft hij ons de gelegenheid tot een weerwoord in ditzelfde nummer gegeven. Daarvoor dank.*

*Het opstellen van zo'n weerwoord is evenwel moeilijk geworden door de overladenheid van het artikel met bijthema's: (1) aanmerkingen op de CITO-examenverslaggeving, (2) uithalen naar de psychometrie i.h.a., (3) eigen vingeroefeningen met psychometrische formules, (4) opmerkingen over zaken van examenbeleid, (5) toespelingen op validiteitsproblemen. Dit leidde tot een overdaad aan bijzonderheden die het beoogde hoofdthema: de betrouwbaarheid van tekstbegriptoetsen in het gebruikskader: eindexamens, dreigen te overschaduwen.*

*Wij beperken ons tot het hoofdthema en tot zaken in verband met de CITO-verslaggeving; daarnaast beperkte aandacht schenkend aan de bijthema's en aan zaken die wij als technische misverstanden zien.*

*Ad: 'Samenvatting'*

Omdat een samenvatting door zijn bondigheid een bijzondere indruk kan achterlaten, gaan we daar even op in.

Het artikel pleegt beslist géén 'aanslag op de axioma's van de neo-Ebeliaanse psychometrie'. Ten eerste bestaat die zó niet. Er is wel een 'testtheorie' (door Van Gils 'toetswiskunde' genoemd), die alle testvaklieden als leidraad dient. Ten tweede raakt het artikel de axioma's van die theorie (die er wel zijn) niet; geeft er zelfs geen blijk van dat de auteur ze uitdrukkelijk onder ogen heeft gezien.

De aanhaling plus toevoeging: a test must be reliable to be valid . . . but reliability does not guarantee validity' is op zich juist, maar kan licht de indruk achterlaten dat betrouwbaarheid van geen enkele betekenis is voor de validiteit en die indruk is nu weer niet juist en zelfs gevaarlijk. Over het gewicht van de 'theoretisch en empirisch gefundeerde bezwaren tegen de validiteit van multiple choice tests' valt te twisten. Dat 'alle indices van het homogeniteitsnetwerk als functies van elkaar gedefinieerd kunnen worden' is maar tot op zekere hoogte juist wegens de statistische aard van de verbanden.

Over de CITO-voorlichting aan de buitenwacht: men kan twisten over de hoeveelheid aan te bieden informatie. Zelfs als sommige indices algebraïsch uit elkaar af te leiden zijn; moet men dat zijn lezers, meestal technisch leken, maar zelf laten doen? Wel zijn we het met Van Gils eens dat het geen zin heeft om alles wat er maar uit te rekenen valt over de hoofden van de buitenwacht uit te storten. Maar aan zo'n overdaad menen wij ons niet schuldig te maken.

Op de 'kwetsbaarheid van meerkeuzetoetsen op hun sterkste punt' en de weinige 'indrukwekkendheid' van de betrouwbaarheid van de tekstbegripexamentoetsen komen we nog terug.

*Ad par. 1: 'Betrouwbaarheid als compensatie voor invaliditeit'*

In het algemeen zal geen testtheoreticus beweren dat betrouwbaarheid een vergoeding is voor 'invaliditeit'. Natuurlijk zul je in de eerste plaats moeten proberen een testinhoud te vinden die zo goed mogelijk voldoet aan het doel van de testprocedure. Maar is die testinhoud eenmaal omschreven, zo goed en kwaad als dat gaat, dan zul je je daarna ook moeten bekommeren om de betrouwbaarheid, als maat voor de verfijning waarmee je verschillen tussen geteste personen kunt vaststellen (onderscheidingsvermogen). Dit is dan de 'praktische' betekenis van testbetrouwbaarheid. Maar die betrouwbaarheid is ook van principiëler belang dan uit Van Gils' artikel blijkt. Dit blijkt uit de 'attenuatieformule' van de testtheorie:

*Dit artikel is een reactie op het artikel van Van Gils in dit nummer.

$r(X, Y) = r(X_t, Y_t) \cdot \sqrt{r(X, X) \cdot r(Y, Y)}$

waarbij $r(X_t, Y_t)$ de 'ware' correlatie is tussen twee variabelen X en Y die je zou vinden als je X en Y beide foutvrij zou kunnen meten; r(X, X) en r(Y, Y) de meetbetrouwbaarheden zijn van de procedures (tests) waarmee je X en Y in feite meet, en r(X, Y) de correlatie die je in feite kunt krijgen. Stel eens dat Y een 'ideale' (volmaakt valide en meetfoutvrije) test van tekstbegrip is: dus $r(Y, Y) = 1$. Laat verder X een concrete tekstbegriptest zijn, die we tegen het ideaal willen houden. In het gunstigste geval is X in beginsel (d.w.z. afgezien van meetfouten) geheel valide. Dan is $r(X_t, Y_t) = 1$. Invullen in de attenuatieformule geeft dan:

$r(X, Y) = 1 \cdot \sqrt{r(X, X) \cdot 1} = \sqrt{r(X, X)}$.

Hieruit blijkt dat de validiteit van een concrete test, opgevat als correlatie met een 'ideale' test, *aan de bovenzijde begrensd* wordt door (de wortel uit) zijn meetbetrouwbaarheidscoëfficiënt! Dit is van belang voor een nuchtere waardering van beoordelingsprocedures waaraan 'men' veel validiteit toedicht, maar die een lage betrouwbaarheid blijken te hebben.

De mogelijke gevolgen van dit alles zijn met een hypothetisch voorbeeld verder te verduidelijken. Laat Y weer de ideale tekstbegriptest zijn, zodat weer $r(Y, Y) = 1$. Laat X een in beginsel geheel valide, maar onbetrouwbare test zijn, zodat $r(X_t, Y_t) = 1$, maar r(X, X) bijvoorbeeld 0,40. Laat tenslotte Z een minder valide, maar betrouwbare test zijn, zodat bijvoorbeeld $r(Z_t, Y_t) = 0{,}80$ en $r(Z, Z) = 0{,}90$. Toepassing van de attenuatieformule geeft dan voor X een validiteitscoëfficiënt van 0,63 en voor Z één van 0,76! Meetbetrouwbaarheid kàn dus wel degelijk een zeker gebrek aan validiteit vergoeden, in die zin dat een wat minder goed dekkende maar betrouwbare beoordelingsprocedure in een feitelijke toepassing méér 'meeneemt' van het ideaal dan een (in beginsel) beter dekkende, maar minder betrouwbare beoordelingsprocedure. Deze wellicht wat verrassende, testtheoretische uitkomst verdient wel eens aandacht gezien de vaak ongenuanceerde tegenoverelkaarstelling van betrouwbaarheid en validiteit. We willen overigens niet beweren dat inhoud en opzet van toetsen er niet toe doen, zolang ze maar betrouwbaar zijn.

*Ad par. 2: 'Op zoek naar een neutraal referentiekader'*

De KR 20 is niet zozeer een 'gemiddelde test–paralleltest correlatie', maar een ondergrens, 'onderschatting', van de testbetrouwbaarheid zelf. Tenslotte: de termen 'soortgenoot validiteit' en 'underachievers' worden door Van Gils op kleurrijke wijze opgevoerd, maar door ons in examenverslagen niet gebruikt.

*Ad par. 3: 'Bijdragen aan de toetswiskunde'*

Deze paragraaf bevat, naast eigen psychometrische vingeroefeningen van Van Gils, een aantal uithalen naar de psychometrie. We zijn het met de voorstelling van zaken in vele gevallen niet eens, maar daarop ingaan zou leiden tot uitweidingen over technische bijzonderheden, die teveel ruimte zouden vragen en voor het hoofdthema o.i. van geen wezenlijk belang zijn. Zulke aangelegenheden lenen zich meer voor rechtstreekse gedachtenwisseling met de auteur. We beperken ons tot enkele (hoofd-)zaken.

De psychometrische vingeroefeningen van de auteur, hoe leuk ook gevonden, zijn ten eerste geen aanval op de axioma's van de testtheorie. Ze lopen uit op algebraïsche identiteiten tussen psychometrische kengetallen, bereikt door sterker vereenvoudigende aannamen in te voeren dan de testtheorie gewoonlijk doet. Langs deze weg is geen weerlegging van axioma's (bijv. onafhankelijkheid van ware en fout-scores; homoscedasticiteit van meetfoutvarianties, etc.) te verkrijgen. Ten tweede is het nut van de formules van Van Gils voor toetsverslaggeving zeer betwistbaar. Het 'oerwoud' wordt er voor het lezerspubliek alleen nog warriger door. Ten derde leggen die formules soms betrekkingen vast, die – hoewel juist onder de ingevoerde aannamen – geen belangwekkende interpretatiemogelijkheden bieden, zeker niet voor de toetspraktijk. Wat heeft het onderwijspubliek aan een omrekening van bijvoorbeeld phi naar $RIP^2$?

Verder suggereert de hele wijze van behandelen en vergelijken van psychometrische kengetallen ('weinig vergoelijkende phi', het citaat uit Freudenthal, e.d.) een toekenning van morele eigenschappen aan puur vaktechnische indices, die o.i. niet terzake doet. Zo kun je ook zeggen dat de Fahrenheitschaal de 'echte' temperatuur teveel flatteert, vergeleken bijv. met de Celsiusschaal. De KR 20 is technisch de meest zinnige maat voor betrouwbaarheid binnen de huidige examenprocedures, die alleen psychometrische uitspraken toelaten over het onderscheidingsvermogen van déze toets binnen déze jaargang van examenkandidaten. Dat phi's en RIT's lager zijn dan bijv. de KR 20 is voor de statistisch ingevoerde lezer 'nogal wiedes'.

*Ad par. 4: 'Zwevende homogeniteitskriteria'*

Van Gils vindt kennelijk dat er een hard getal (norm?) gegeven zou moeten worden voor de betrouwbaarheid, resp. homogeniteit, van een examentoets en verwijt de 'toetswiskunde' zo'n hard getal niet te geven. Deze voorstelling is misleidend. De testtheorie is een geformaliseerd model van de werkelijkheid waarmee je allerlei betrekkingen kunt doorrekenen en in dit opzicht exact genoeg. Zo'n theorie kan vaak helpen vaststellen wat *on*mogelijk is (denk bijv. aan de attenuatieformule), maar kan niet volledig voorschrijven wat je in een bepaalde praktische situatie *moet* doen of eisen. Wie dat meent, verwart normatief met descriptief denken. De in CITO-verslagen gegeven 'normen' voor homogeniteitswaarden, RIT's e.d. kunnen niet anders dan enigszins willekeurig zijn; het zijn vuistregels, aan de testliteratuur ontleend. We willen nog eens nagaan of ze in deze vorm zinvol genoeg zijn om ze in verslagen op te nemen.

*Ad par. 5: 'Selektierelevante betrouwbaarheid'*

In deze paragraaf komen allerlei moeilijkheden aan de orde, die deels samenhangen met het norm(cesuur-)vaststellingsprobleem, deels met de betrekkelijke willekeurigheid in de omrekening van ruwe scores naar de officiële cijferschalen, deels met technische kwesties rondom 'ware' scores e.d. Veel daarvan heeft te maken met 'examenbeleid'. Deze paragraaf biedt, ondanks de titel, geen aanknopingspunten voor discussie over wat dan 'selektierelevante betrouwbaarheid' mag heten. Een opmerking: een deel van de interpretatie- en normeringsproblemen bij de examens zou verholpen kunnen worden door invoering van een 'geëquivaleerd' testsysteem; wat dan wel een verandering van examenorganisatie zou meebrengen.

*Ad par. 6: 'Selektief gebruik van tekstbegriptoetsen'*

Eerst een terminologische zaak. Van Gils stelt in deze paragraaf 'selektief' gebruik van tekstbegriptoetsen tegenover gebruik in 'strikt evaluatieve, nivobewakende funktie'. Dit is echter in het raam van de AVO/VWO-examens een onjuist gebruik van psychometrische terminologie, dat overigens nogal eens voorkomt. Het begrip selektie heeft te maken met de toepassing van tests voor het uitzoeken van de 'meest geschikte' sollicitanten (voor een bepaalde bedrijfsfunktie), kandidaten (voor een bepaalde opleiding), e.d. De psychometrische theorie over selektie is opgehangen aan predictieve validiteiten, d.w.z. aan correlaties van de selektietests met uitwendige slaagmaatstaven.

Bij de huidige examens is een dergelijke selektieprobleemstelling formeel niet aan de orde. Ze gaan in beginsel uit van de gedachte dat de kandidaten beoordeeld moeten worden op het (in meer of mindere mate) bereikt hebben van de onderwijsdoelen: meer een (onderwijskundig) 'beheersingsgezichtspunt' dan een (psychotechnisch) selektiegezichtspunt. Maar zodoende hebben de examens bij uitstek een 'nivobewakende evaluatieve funktie'! Vergelijking van meerkeuze-tekstbegriptoetsen met andere beoordelingsmethoden op het stuk van selektie is hier niet terzake en blijft onberedeneerd. Het hele artikel overziende, krijgt men de indruk dat Van Gils met 'selektierelevante betrouwbaarheid' eigenlijk 'onderscheidingsvermogen' bedoelt, een testeigenschap die meer met betrouwbaarheid dan met validiteit te maken heeft en voor de 'nivobewakende funktie' van examens niet onbelangrijk is. Alleen wordt deze zaak hier binnen een verkeerde terminologie ingevoerd.

Ten tweede Van Gils' betwisting van het betrouwbaarheidsvoordeel van tekstbegriptoetsen. Deze zaak is in zijn algemeenheid niet zo eenvoudig; wij komen er in de slotbeschouwing op terug. Hier zij alleen de vinger op een technische misvatting gelegd. Van Gils stut zijn geringe waardering voor de betrouwbaarheidswinst van de tekstbegriptoetsen op een vergelijking van ( in 10-cijferschaal uitgedrukte) standaardmeetfouten: 0,6 bij de gemiddelde tekstbegriptoets en 0,8 bij een opstelbeoordelingsonderzoek door Jansen en Wesdorp, aangehaald in een ander artikel van zijn hand (Van Gils, 1975). Maar je mag betrouwbaarheden niet vergelijken op grond van standaardmeetfouten alléén! Er is nl. een analytisch verband tussen betrouwbaarheid, standaardmeetfout en standaardafwijking.
Het aangehaalde artikel raadplegend ( blz. 415 daarvan) vindt men een gemiddelde standaardafwijking van 1.04 (in 10-cijferschaal uitgedrukt) en een gemiddelde standaardmeetfout van 0,78 of 0,79 (in dezelfde schaal). Met de kleinste van die twee geeft de formule

$$r(X, X) = \frac{s^2(T)}{s^2(X)} = 1 - \frac{s^2(E)}{s^2(X)};$$

waarbij r(X, X) de betrouwbaarheid; $s^2(T)$ de 'ware'

variantie; $s^2(E)$ de foutenvariantie en $s^2(X)$ de testvariantie is, aan dat

$$r(X, X) = 1 - \frac{(0,78)^2}{(1,04)^2} = 0,44$$

moet zijn. Deze waarde geeft Van Gils in dat artikel zelf ook; het is de betrouwbaarheid van de gemiddelde, enkele beoordelaar in dat onderzoek. Maar we gaven de formule boven, omdat deze ons ook vertelt dat de opstelbeoordelingsprocedure met één beoordelaar per opstel, globaal maar 44 % 'ware' variantie (variantie toe te schrijven aan 'ware' scores, aan 'echte' verschillen tussen getesten) bevat. De tekstbegriptoetsen met gemiddeld KR 20 = 0,78 mogen evenwel minstens 78 % ware variantie op hun rekening schrijven en dat scheelt nogal wat. Het onderscheidingsvermogen van de tekstbegriptoetsen is echt veel groter dan dat van de opstelbeoordelingsprocedure met één beoordelaar per opstel. Tenslotte een opmerking over Van Gils' vaststelling dat de schoolonderzoekcijfers, hoewel misschien wat minder betrouwbaar, 'door de vakdocenten in het algemeen valide worden geacht' en dat (daarom!?) de subjektief oordelende examinatoren in plaats van 'het objektieve toeval' de laatste verantwoordelijkheid voor de individuele eindcijfers moeten hebben. Deze redenering onderstelt een, door niets in het artikel gesteunde, grotere validiteit van het schoolonderzoek en miskent eisen die de openbare legitimeringsfunktie van diploma's aan examens stelt: betrouwbaarheid, uniformiteit e.d.

*Slotbeschouwing*

1. *De CITO-verslaggeving*

De kritiek daarop omvat aanmerkingen op meer of minder gelukkige zinsneden in de CITO-examenverslaggeving, op een vermeende overdaad (redundantie) aan psychometrisch gedoe en op een verwaarlozing van validiteit ten gunste van betrouwbaarheid.

Over de wijze van aanbieding in onze verslaggeving willen wij best verder nadenken; wij zijn daar trouwens steeds mee bezig. De beschuldiging van overdaad wijzen wij af; zeker de laatste jaren is het cijfermateriaal (nog) soberder geworden. De vele vaktermen die Van Gils in dit artikel bij elkaar heeft gesleept geven, door de hele aanbiedingswijze, de lezer in dat wij dit alles in onze examenverslaggeving ook elke keer overhoop halen; niets is minder waar. Blijft het voor eeuwige twist vatbare vraagstuk hoever je met bijv. psychometrische gegevens in bijzonderheden moet afdalen. Je moet evenwel rekening houden met vele lezers en iets meer bieden dan waar de minst belangstellende om vraagt. In dit verband is het wel aardig te noemen dat het door Van Gils aangehaalde enquêteverslag van Wesdorp gemiddeld 67 % positieve reacties aangeeft (enquêtevraag nr. 70) op de CITO-verslaggeving. We zouden het trouwens de docenten–lezers niet graag aandoen om hun de theoretisch kleinstmogelijke verzameling gegevens te bieden, waarna ze dan, bijv. met behulp van Van Gils' formules, de rest maar zelf moeten uitrekenen. We houden verder staande dat de KR 20 in de gegeven omstandigheden het meest toepasselijke en eenvoudige kengetal is voor de betrouwbaarheid (onderscheidingsvermogen) van de toetsen. Tenslotte dan over het verwijt van teveel aandacht voor 'betrouwbaarheid' ten koste van 'validiteit': verrichten van psychometrische analyse is bij elke examenafname gewenst, als algemene nacontrole en om opduikende problemen (bijv. een niet goed afgestemde moeilijkheidsgraad van toetsen) zo vlug mogelijk te onderkennen. Uitgebreid ingaan op het 'validiteitsvraagstuk' heeft alleen zin als er iets nieuws – ingrijpende wijziging van toetsopzet – te melden valt en dat is zelden het geval.

In alle nuchterheid moet worden gezegd dat snelle en dramatische ontwikkelingen (vooruitgang) op validiteitsgebied eenvoudig niet te verwachten zijn.

2. *De (on-)betrouwbaarheid van tekstbegriptoetsen*

Welke betrouwbaarheidswinst hebben de tekstbegriptoetsen feitelijk gebracht? Deze vraag vraagt op zijn beurt weer naar een vergelijking met open vraagvormen voor hetzelfde doel. Rechtstreekse vergelijkingen binnen de situatie waar het hier om gaat zijn helaas niet mogelijk bij gebrek aan bijv. cijfermatige analyses van de betrouwbaarheid van vóór het tekstbegriptoetstijdperk gebruikte beoordelingsmethoden. Zouden we de betrouwbaarheid van bijv. opstelbeoordeling met één beoordeling per opstel als aanknopingspunt beschouwen, dan was de kous snel af: de tekstbegriptoetsen winnen het dan met vele lengten (zie ons commentaar bij par. 6). Maar opstelbeoordeling is een bijzonder berucht vraagstuk en misschien niet helemaal eerlijk als vergelijkingspunt. Dan blijven alleen de algemene onderzoeksbevindingen over open vraagvormen over als aanknopingspunten.

Overzichten als bijvoorbeeld in Coffman (1971) en Mellenbergh (1971) geven een wisselend beeld. Tussenbeoordelaarscorrelaties kunnen bijv. van

0,44 tot 0,96 lopen; zie tabel 1 op blz. 10, 11 bij Mellenbergh. De interpretatie van zulke literatuuroverzichten is niet eenvoudig. Er is een verband met het 'vak': bij exacte vakken kunnen vaker hogere betrouwbaarheden worden bereikt dan bij de talen. Verder zijn niet alle open vragen over één kam te scheren: kort-antwoord-vragen geven meestal hogere betrouwbaarheden dan vragen die uitgebreider stelarbeid van de kandidaat vergen. Van belang is ook *hoeveel* open vragen je binnen een bepaalde testtijd wilt of kunt stellen: hoe meer afzonderlijke vragen, hoe betrouwbaarder de procedure in het algemeen wordt. Verder zitten er technische en theoretische problemen aan het vergelijken van betrouwbaarheden van open en gesloten vraagvormen. De betrouwbaarheidstheorie voor procedures met menselijke beoordelaars is ingewikkelder dan die voor geprecodeerde toetsen; o.a. door de moeilijk grijpbare gedragsverschillen tussen beoordelaars. Betrouwbaarheidscoëfficiënten voor open werk, uitgedrukt in tussenbeoordelaarscorrelaties, *over*schatten de werkelijke tussenbeoordelaarsovereenstemming, omdat het mechaniek van de correlatieberekening verschillen in gestrengheid en variabiliteit tussen beoordelaars wegwerkt. Zulke correlaties geven in hoofdzaak een rangorde overeenstemming aan, en niet zonder meer de mate waarin beoordelaars als 'paralleltests' te beschouwen zijn. Maar verschillen in gestrengheid en variabiliteit tussen beoordelaars werken zich natuurlijk wel uit in de *feitelijke scores* die kandidaten krijgen!
De algemene slotsom uit onderzoeksbevindingen is en blijft dat objectieve testvormen hogere betrouwbaarheden bereiken dan procedures met 'opstelvragen'. Bovendien hebben objectieve testvormen een testeconomisch voordeel (goedkoper scoring en meestal méér 'meetpunten' per eenheid testtijd) dat groter wordt naarmate het om meer kandidaten gaat en naarmate men het open werk betrouwbaarder tracht te maken door meervoudige beoordeling.

Twee vragen blijven staan:

1. Is het betrouwbaarheidsvoordeel van de tekstbegriptoetsen belangwekkend genoeg?
2. Is hun betrouwbaarheid als zodanig hoog genoeg voor het doel: de AVO/VWO-examens?

Vraag 1. is bij gebrek aan rechtstreeks toepasselijke gegevens alleen schattenderwijs en algemeen te beantwoorden. Laten we de betrouwbaarheid van tekstbegriponderzoek met andere middelen dan geprecodeerde toetsen en met één beoordelaar per kandidaat, schatten op 0,70. Dat is het gemiddelde van de coëfficiënten in tabel 1 van Mellenbergh en, gezien de overwegingen boven, zeker vrijgevig genoeg. Dat is ook minder dan de gemiddelde KR 20 van de tekstbegriptoetsen, 0,78, maar het verschil is niet zo groot. Wel geeft de noodzakelijke tussenkomst van één beoordelaar per kandidaat bij de grote aantallen kandidaten in moderne talen examens nog altijd een testeconomisch voordeel voor de tekstbegriptoetsen. Stel nu dat we twee 'parallelle' beoordelaars per kandidaat invoeren. Dan geeft toepassing van de Spearman–Brown formule:

$$R(X, X) = \frac{Kr(X, X)}{1 + (K - 1)\, r(X, X)}$$

waarbij r(X, X) de oorspronkelijke betrouwbaarheid en R(X, X) de betrouwbaarheid na verlenging met factor K, aan dat de betrouwbaarheid van de *twee*voudige beoordelingsprocedure:

$$R(X, X) = \frac{2(0,70)}{1 + 0,70} = 0,82$$

wordt. Dit is hoger dan de gemiddelde KR 20, 0,78, van de tekstbegriptoetsen. Om te berekenen hoeveel de tekstbegriptoetsen verlengd zouden moeten worden om ook gemiddeld 0,82 te behalen is dezelfde formule te gebruiken, waarbij dan K opgelost moet worden:

$$0,82 = \frac{K(0,78)}{1 + (K -)(0,78)}$$ . Dit geeft K = 1,28.

De huidige tekstbegriptoetsen zouden dus met 28 %, d.w.z. met 14 items (ongeveer twee tekstfragmenten plus vragen) verlengd moeten worden. Hieruit blijkt o.i. dat de tekstbegriptoetsen, zowel wat betrouwbaarheid als testeconomie betreft, en zowel feitelijk als potentieel (bij het nastreven van verdere verbeteringen in de examens) nog altijd aanmerkelijke voordelen hebben. Daarbij komt dat geprecodeerde testvormen veel beter hanteerbaar zijn binnen een verfijnder examentechniek.

Vraag 2. – is de betrouwbaarheid van de tekstbegriptoetsen hoog genoeg voor het doel – vraagt naar een stellige uitspraak over het voor de huidige examens te eisen onderscheidingsvermogen. Zo'n vraag is niet gemakkelijk te beantwoorden omdat daarvoor een harde 'kostenschatting' nodig is

voor met name aantallen misclassificaties rondom de 'ware' cesuurscore, die van een bepaalde (on) betrouwbaarheidsgraad het gevolg zijn: je moet dan aangeven wat het individu en samenleving kost als een kandidaat een niet terechte voldoende dan wel een niet terechte onvoldoende krijgt. Zo'n schatting is, zeker bij de eindexamens, moeilijk te geven bij gebrek aan goede aanknopingspunten ervoor. (Het probleem is trouwens nog ingewikkelder door de compensatieregels voor groepen vakken.) Maar algemeen kan men dit stellen: bij een gegeven examenopzet en afgezien van validiteitsvragen, is een beoordelingsprocedure die betrouwbaarder en 'testeconomischer' is dan andere (voor hetzelfde doel) ook inderdaad verkieslijker. De tekstbegriptoetsen voldoen o.i. wel aan deze uitspraak, al is hun betrouwbaarheid zeker voor verbetering vatbaar. Daarvoor zou òf een verfijnder examentechniek nodig zijn, òf brute toetsverlenging. Maar in ieder geval verwachten wij dat verbetering van de betrouwbaarheid goedkoper zal zijn voor tekstbegriptoetsen dan voor tekstbegriponderzoek met open vragen. Ons voorbeeld: verlenging van de tekstbegriptoetsen met 28 % geeft ongeveer dezelfde betrouwbaarheid als tekstbegriponderzoek met open vragen onder *twee*voudige beoordeling.

Betrouwbaarheid en testeconomie overtuigen natuurlijk niet als men stelt dat de tekstbegriptoetsen zo weinig validiteit in zich dragen dat die voordelen dit gebrek niet meer kunnen compenseren. Hierop verder ingaan zou leiden tot bespreking van de 'theoretisch en empirisch gefundeerde bezwaren tegen tekstbegriptoetsen', maar dat wordt – gegeven het hoofdthema van het artikel – een ander verhaal.

*Literatuur*

Coffman W. E.: *Essay Examinations*. In: Thorndike, R. L. (ed.): Educational Measurement, ACE, Washington, 1971, Ch. 10.

Mellenbergh, G. J.: *Studies in Studietoetsen* (diss.), Psychol. Laboratorium, U.v.A., 1971.