

[Dit stuk voor eigen gebruik legde de kiem voor het ORD-paper 1977 over subjectiviteit van zogenaamd objectieve keuzevragen.] [Ik had in 1976 nog niet in de gaten hoe sterk ideologisch 'Studietoetsen' van De Groot en Van Naerssen is. <https://benwilbrink.wordpress.com/2022/01/26/de-ideologie-van-1969-vierkeuzetoetsen-de-groot-van-naerssen/>]

open eind versus meerkeuze

ben wilbrink 1-3-76.

De Groot, A. D., & Van Naerssen, R. F. (red.) Studietoetsen, construeren, afnemen, analyseren. Den Haag: Mouton, 1973

Dit boek is uitsluitend gericht op toetsen bestaande uit meerkeuze items. Een nogal forse inperking van toetsingsmogelijkheden, die de auteurs willen rechtvaardigen door de voordelen die verbonden zijn aan deze toetsvorm. Daarom een kritische bespreking van de voordelen die in dit boek genoemd worden. Mellenbergh: Waarom meerkeuzevragen? (Hoofdstuk 1.) Mellenbergh gooit opstel en aanvulvragen op één hoop, de hoop van de open vragen'. O.a. zeker met betrekking tot beoordelingsbetrouwbaarheid is het onderscheid tussen opstel en aanvul vragen dermate groot dat ze niet gemakshalve over één kam geschoren kunnen worden. Zonder de door Mellenbergh als argument gehanteerde '90 onderzoeken' in detail te bekijken lijkt het niet mogelijk te overzien wat de gevolgen zijn t.a.v. de conclusies die Mellenbergh uit zijn onderzoek meent te moeten trekken: of een en ander wél of misschien ook niet betrekking heeft op die categorie van open vragen die van het aanvul type zijn, verder gewoon 'open eind' vragen te noemen.

Vergelijking open en geprecodeerde vragen. Een soort samenvatting van het desbetreffende hoofdstuk in de dissertatie van Mellenbergh.

Onderverdeeld in de volgende vier paragrafen:

1. het beoordelen van open vragen. Hier worden alleen opstellen besproken en het bekende resultaat dat deze beoordelingen weinig betrouwbaar zijn (bedoeld wordt dat beoordelaars het met elkaar niet erg eens zijn).

2. het voorspellen van schoolsucces. Geprecodeerde vragen zouden de beste voorspellers zijn. Bedacht moet bij die conclusie worden dat Mellenbergh niet duidelijk maakt of de open vragen in de

opstelvorm of in de open eind vorm zijn; dat meerkeuze vragen in deze onderzoeken van zeer hoge kwaliteit zijn (géén teacher made tests), en dat niet duidelijk is of bijv. slechtere beoordelaar betrouwbaarheid van opstel toetsen en eventueel open eind vragen niet op eenvoudige wijze te corrigeren is (door meerdere personen te laten nakijken, door méér vragen in de toets op te nemen, e.d.)

3. meten beide vraagvormen hetzelfde? De conclusie van Mellenbergh is een beetje triviaal: 'De tendens van deze onderzoeken is, dat het in vele gevallen mogelijk is met geprecodeerde vragen hetzelfde te meten als met open vragen, mits de geprecodeerde vragen goed geconstrueerd zijn.'

4. effecten op het onderwijs. "Waarschijnlijk zal het type vragen in de meeste gevallen weinig invloed uitoefenen op de wijze van voorbereiding door de leerlingen." (Waarvoor Mellenbergh als bewijsmateriaal een onderzoek van French gebruikt). Punt is echter niet of in bestaande situaties. en met de huidige primitieve of intuïtieve toetsconstructieprocedures die leiden tot ondoorzichtige toetsingssituaties voor de student, of in die situaties studenten anders gerangordend zouden worden wanneer zij getoetst zouden worden op een wijze waarop zij zich niet geprepareerd hadden. De vraag is nu juist hoe het mogelijk is de toetssituatie volkomen transparant te maken, zodat de studenten zich optimaal kunnen voorbereiden op de te stellen vragen, dus zowel wat de vorm van de vragen als wat de inhoud betreft (voorzover je die al zinvol kunt scheiden.) Met andere woorden: de kwestie is niet of het type vragen invloed uitoefent op de voorbereiding door de leerlingen, maar de vraag is welk type vragen de toetssituatie het meest transparant maakt.

'Meerkeuze vragen hebben echter het voordeel dat men er beter mee kan rekenen; de zogenaamde statistische analyse is bij meerkeuze vragen eenvoudiger uit te voeren.' (blz. 8)

Dat is natuurlijk nog maar de vraag, en dat hangt er ook van af of het gaat om opstel vragen of om open eind vragen. En wat is 'statistische analyse'. Toelichting:

Bij opstel vragen (maar daar gaat onze belangstelling in dit rapportje niet direct naar uit) is het vaststellen van de kwaliteit

van de toets, daar gaat het uiteindelijk om, ietsje ingewikkelder. Het valt echter te voorzien dat er procedures ontwikkeld kunnen worden die de docent goed in staat stellen om ook voor opstel vragen goede gegevens over de kwaliteit van zijn toets te verkrijgen (zeker de generaliseerbaarheidstheorie van Cronbach c.s. biedt hier grote praktische mogelijkheden). Daarnaast zijn er natuurlijk nog een belangrijke categorie 'opstelt vragen die zich min of meer laten behandelen als open eind vragen: te denken valt aan alle opgaven waarin gerekend, bewezen, geconstrueerd e.d. moet worden.

Bij open eind vragen is de betrouwbaarheid van beoordling naar verwachting toch tot grote hoogte op te voeren, in ieder geval tot een hoogte die zich gunstig onderscheidt van de betrouwbaarheid van meerkeuze items, waar toch bij de betrouwbaarheidskwestie ook de onbetrouwbaarheid voortkomend uit toevallige misleunen bij de constructie van de items of de keuze van de onderwerpen en de wijze waarop de vraag erover gesteld zal worden, mee te rekenen zijn. (We hebben in het vervolg nog uitgebreider gelegenheid om op deze kwestie terug te komen). Scores voor open eind vragen zijn dus perfect duidelijk, en laten nog makkelijker met zich rekenen dan scores op meerkeuze items omdat geen correcties voor raden en dat soort onbevredigende trucs hoeven te worden toegepast. In die zin is de statistische analyse van open eind vragen ook eenvoudiger, omdat er minder vooronderstellingen voor die analyse hoeven te worden gemaakt. (Wil men ook voor open eind vragen een item analyse op de computer laten uitvoeren, dan zal waarschijnlijk degene die het werk nakijkt een IBM formulier moeten invullen, waarbij een eenvoudige procedure kan zijn alleen streepjes te zetten voor vragen die fout beantwoord zijn.)

Kortom, het 'voordeel' dat Mellenbergh hier zo tussen neus en lippen noemt, blijkt geen duidelijke bestaansgrond te hebben. We zullen echter nog even afwachten wat andere auteurs in dit boek over de 'voordelen van de meerkeuze vorm' te zeggen hebben.

De Groot: Een definitie en verdere kenmerken.

Ook de Groot heeft de neiging de open eind vraag op dezelfde

grote hoop te deponeren waar ook opstel vragen en andere vormen met uitgebreide antwoorden al bijeengeveegd zijn. Nadat hij heeft duidelijk gemaakt dat objectieve scorbaarheid toch niet als hetzelfde als de betrouwbaarheid van vragen mag worden gezien, maakt hij vervolgens in het onderscheid tussen openeind vragen en meerkeuze vragen precies deze fout, namelijk dat minder objectieve scorbaarheid van open eind vragen ze ipso facto ook slechter, d.i. bijvoorbeeld minder betrouwbaar, maakt.

De Groot: Gebruik en nut van studietoetsen.(hoofdstuk 3)

Het automatiseringsvoordeel bij zeer grote aantallen studenten:

'Inderdaad zijn werkelijk ingrijpende vereenvoudigingen in de uitvoering (van beoordelingen, b.w.) alleen te bereiken als men tot gebruik van studietoetsen overgaat zij het tegen de prijs van veel moeizame voorbereidingsarbeid.' Opmerking: bij werkelijk zeer grote aantallen personen die getoetst moeten worden, denk aan eindexamens van middelbare scholen waar landelijke toetsen voor gemaakt worden, weegt de extra benodigde voorbereidingstijd (er worden zeer hoge eisen aan afzonderlijke vragen en de toets als geheel gesteld waaraan individuele docenten vrijwel nooit zouden kunnen voldoen waar het gaat om zelf gemaakte toetsen voor hun eigen onderwijs) weinig kost vergeleken met de kostenbesparing die voortvloeit uit de mogelijkheid van geautomatiseerde scoring, hebben we met een regel voordeel van meerkeuze vragen óók t.o.v. open eind vragen te maken. In vrijwel alle toetsingssituaties in het hoger onderwijs geldt echter dat de door automatiseerbaarheid te bereiken besparingen niet opwegen tegen de item constructie moeite en controle op de kwaliteit van de toets en items die zeer tijdrovend en kostbaar kunnen zijn. Tenzij het ook voor meerkeuzevragen mogelijk blijkt om constructieregels te formuleren die direct leiden tot onderwijskundig zinvolle vraagstellingen (dergelijke constructieregels zijn in de literatuur nog niet gegeven, hoewel een enkele poging daartoe o.a. door Bormuth gedaan is. Het is echter mogelijk om praktisch zeer bruikbare, zij het wetenschappelijk niet zo interessante, constructieregels te maken die leiden tot meer zinvolle vraag constructie óók voor meerkeuze vragen. Op het COWO worden werkzaamheden

in die richting uitgevoerd, die hopelijk zullen leiden tot een onderdeel van een cursus voor docenten). Zolang het gaat om hooguit enkele honderden studenten is de open eind vraag vorm direct concurrerend met de meerkeuze vorm voor wat betreft, snelheid en kwaliteit van verwerking, en waarschijnlijk zelfs, maar dat zal dit overzicht moeten uitwijzen, is de open eind vorm te prefereren vanuit onderwijskundige en andere overwegingen. Het is dus wat teleurstellend dat De Groot die directe vergelijking tussen meerkeuze vragen en open eind vragen niet maakt, waar toch dit boek bij uitstek geschreven is voor docenten die voor hun eigen stukje onderwijs beoordelingsinstrumenten willen construeren (landelijke toetsprogramma's voorzien zich van de nodige deskundige begeleiding en hoeven niet op de aanwijzingen in het boek van De Groot en Van Naerssen te drijven).

'Daarmee komt een reeks problemen aan de orde objectiviteit, betrouwbaarheid, cesuurbepaling, constantie van normen, waarvan duidelijk is dat zij met studietoetsen beter kunnen worden opgelost." Opmerking: er is geen enkele reden om op voorhand te veronderstellen dat deze problemen niet of minder of zelfs minder efficiënt oplosbaar zouden zijn wanneer de toets bestaat itt. openeind vragen, of zelfs uit opstel vragen.

Wat evaluatie betreft, ziet De Groot grote mogelijkheden om 'objectieve vergelijkingen' te maken, juist met behulp van de studietoetsmethode. Daar valt hetzelfde bij op te merken als bij het laatste citaat: bij andere toetsvormen kan dat in principe even goed (minder objectief misschien, maar De Groot heeft juist zélf aan het begrip 'objectief' een ondergeschikte betekenis toegekend t.o.v. begrippen als betrouwbaarheid, validiteit, e.d.)

'Gebruik van studietoetsen heeft echter minstens twee belangrijke voordelen boven gewone 'proefwerken'. Ten eerste, de objectieve vergelijkbaarheid; ten tweede de hanteerbare vorm waarin de informatie binnen komt. Vierkeuzeitems kunnen met name zeer gemakkelijk worden gebruikt voor het maken van fouten analyses.' Opmerkingen: 'objectieve' vergelijkbaarheid van scores van verschillende studenten, of van dezelfde student bij verschillende gelegenheden: De Groot gaat hier het begrip 'objectief' zeer ver

uitbreiden boven zijn eerder gegeven beperkte definitie. Volgens die definitie hebben we aan die 'objectiviteit' weinig of niets bij vergelijkingen. Het gaat bij vergelijkingen juist om betrouwbaarheids-aspecten, die bij open eind vragen even goed aan de orde kunnen komen als bij meerkeuze vragen. Het gaat ook om validiteits-aspecten, waar open eind vragen mogelijk toch op belangrijke punten een streepje voor hebben op meerkeuze vragen. En ja, wat de hanteerbare vorm voor verdere rekenarij betreft, ook bij eenvoudig scoorbare open eindvragen kunnen we in principe een heel eind weg gaan rekenen. Foutenanalyses kunnen bij open eind vragen beter gemaakt worden, omdat we daar met raadkansen niets te maken hebben, en de mogelijke fouten door keuze van bepaalde alternatief-formuleringen al niet van te voren ver ingeperkt zijn.

De Groot: Onderwijskundige aspecten van de toetsvorm.(hoofdstuk 4)

'Items zijn objectief (geprecodeerd) zodat men met behulp van de item vorm tot onaanvechtbaar rechtvaardige relatieve prestatiebeoordelingen kan komen.' Een pertinent misleidende uitspraak: objectieve items en toetsen kunnen nog volstrekt onbetrouwbaar zijn, en daarmee een volstrekt onrechtvaardig middel om tot relatieve prestatie beoordelingen te komen. (vergelijk maar de uitspraken van de Groot over 'betrouwbaarheid' versus 'objectiviteit' in ditzelfde boek in hoofdstuk 2).

]Docenten 'worden ertoe gedwongen hun onderwijsdoelen te concretiseren en te operationaliseren.' Dit is gewoon lariekoek: in de praktijk is die noodzaak in het geheel niet aanwezig; anderzijds geldt natuurlijk evenzeer voor andere toetsvormen dat zij enige overdenking vragen m.b.t. waar het in het onderwijs om te doen was. Waar De Groot vervolgens over 'open opgaven' spreekt heeft hij het over opstel achtige opgaven, en niet over de eenvoudige openeind vragen. Telkens en telkens weer zien we een tegenstelling geconstrueerd worden tussen de beste vormen van meerkeuze toetsen, en de slechtste vormen van open vragen, geen wonder dat dan de meerkeuze vragen er als de 'redders van het onderwijs' uit komen. De vergelijking waar we in dit rapport in geïntereeserd zijn is echter die tussen open-eindvragen

en meerkeuzevragen, waar beide soorten vragen door de docent ten behoeve van beoordeling in zijn eigen stukje onderwijs, geïnstrueerd worden.

Ook m,b,t, de moeilijkheid van vragen maakt de Groot de vergelijking tussen globale opstelvragen (qua moeilijkheid inderdaad moeilijk in te schatten) en langs perfect doorzichtige weg geconstrueerde toets items van het meerkeuze type waarbij alles en alles onderwijskundig verantwoord is (terwijl in de praktijk het veeleer zo is dat de docent de items voor zijn toets min of meer intuïtief zit te bedenken, en er weinig of geen kwaliteitscontrole op de items gedaan, bij empirisch onderzoek blijkt dat docenten de moeilijkheid van hun eigen items maar erg moeilijk kunnen schatten etc.)

"Kortom, psychometrische methoden van itemanalyse en testanalyse kunnen van veel belang zijn om de begripsvaliditeit van een toets onder controle te houden, d.w.z. te helpen bepalen en op te voeren. Vele van deze methoden nu waarover hieronder meer zijn alleen toepasbaar dankzij de toetsvorm, d.w.z. dankzij de samenstelling van een toets uit tientallen objectieve items." Opmerking: inderdaad is het zo dat een groot deel van de psychometrische technieken betrekking heeft op analyse van tests bestaande uit meerkeuze items. Het is echter misleidend om te stellen dat onderzoek naar de begripsvaliditeit van beoordelingsprocedures alleen maar mogelijk zou zijn wanneer de beoordeling is gebeurd met behulp van meerkeuze vragen. Ook opstel vragen, om maar eens een betrekkelijk ongunstig voorbeeld te noemen, lenen zich voor bepaling van begripsvaliditeit. Maar zeker de open eind vorm van vragen is zeer efficiënt waar het gaat om kwaliteits onderzoek zoals dat naar betrouwbaarheden en validiteiten.

De Groot: Psychometrische aspecten: itemanalyse. (hoofdstuk 5)

"Indien wij in één zin de psychometrische en daarmee natuurlijk ook de onderwijskundige merites van de item- en toetsvorm moeten omschrijven, dan kunnen wij niet beter doen dan te wijzen op de goed geregelde 'feedback', die de leraar bij toepassing ervan kan verkrijgen. Hij kan met weinig extra inspanning een duidelijk overzicht krijgen van wat wel en wat niet

goed is overgekomen bij zijn leerlingen en daardoor van zijn eigen fouten leren. Wordt het gemaakte werk gecorrigeerd aan de leerlingen teruggegeven, dan geldt dit voordeel ook voor hun leerproces: ook zij kunnen onmiddellijk zien waar hun tekorten liggen." Opmerkingen: in principe kan dit alles natuurlijk ook uitstekend gerealiseerd worden met opstelvragen. Zeker wanneer voor voortgangstoetsen tijdens het onderwijs geen selectieve beoordeling gepleegd wordt en de kwestie van betrouwbaarheid van de toets dan niet zo'n grote rol meer speelt. Het name voor de open-eindvragen geldt hetzelfde vleiende proza als door De Groot hier voorbehouden aan de meerkeuze vraagvorm.

Thorndike, R. L. Educational measurement for the seventies. In Thorndike, R. L. (Editor) **Educational measurement**. Washington, D. C. : American Council on Education, 1971.

"The first impact of technology on testing tended to stereotype item construction in one or another variation of the multiplechoice format. The focus on multiplechoice items was not solely a response to the IBM 805 test scoring machine; this type of item has many inherent virtues as a mode of testing application and understanding as well as information, But the readiness with which the format fitted into the commercially available answer sheets tended to focus test makers' attention upon this format to the exclusion of all others,"

"As character readers are developed, he (de student, b.w.) will be able to produce the answer by writing a set of numerals or letters. It can be anticipated that during the next twenty years these developments will result in a considerable increase in the diversity of test tasks including the greater use of constructed rather than recognition type of items."

"Concerns about testing's impact have been of two kinds. One type of concern is that there is too much testing and that too much weight is being given to it. The other is that objective tests of the sort that are widely used tend to pervert the learning process, either because they are poorly constructed or because they inherently measure limited and trivial educational outcomes,

Concern about the quality of tests and their wise use has been as prominent within the professions of educational and psychological measurement as among their critics. This concern has expressed itself concretely in several ways. Critical test reviews have continued to appear in successive editions of the Mental Measurements Yearbooks appearing in 1953, 1959, and 1965. In addition, the American Psychological Association and the National Council on Measurement in Education have prepared, and subsequently revised, statements of standards for the preparation and publication of psychological and educational tests (American Psychological Association. **Standards for educational and psychological tests and manuals.** Washington: APA, 1974). These publications are the most tangible signs of a concern for sound construction and wise use that is also expressed in textbooks and teaching, in articles and symposia, and in all the media by which test makers communicate with one another and with the public."

Opmerking: het laatste citaat heeft betrekking op gestandaardiseerde tests die op letterlijk fantastisch grote schaal (oplagen van miljoenen toetsformulieren voor bepaalde tests per jaar) afgenomen worden. Daar bleek allereerst de harde noodzaak van vaststelling van criteria voor de kwaliteit van deze toetsen. Dezelfde criteria zijn natuurlijk relevant voor toetsen die de docent uitsluitend ten behoeve van zijn eigen onderwijs maakt, zij het dat in die situatie aan vele kwaliteitseisen niet kan worden voldaan omdat de daarvoor benodigde research inspanning in geen enkele verhouding staat tot het aantal te testen studenten. Het is echter van groot belang om heel expliciet te signaleren aan welke kwaliteitseisen de teacher-made test moeilijk of niet kan voldoen, zodat met die niet te vermijden gebreken rekening gehouden kan worden op het moment dat op basis van de individuele toets scores beslissingen ten aanzien van personen genomen gaan worden.

Wesman, A. G. Writing the test item. In Thorndike, R. L. (Editor) Educational Measurement. Washington, D. C. : American Council on Education, 1971.

"It has often been held that items, such as the shortanswer type, that require the examinee to supply an answer are inherently more

valid as measures of important educational functions than are items which require 'merely' the selection of an answer from among a small number of options (which is the characteristic of such types as the multiplechoice and the matching format). (Note: an item is valid to the degree it measures what it is intended to measure. Validity depends on proper representation of the idea (i.e. content) to be appraised as well as on the skill with which the item is prepared.) The argument generally is put forth that by its very essence the ability to produce the correct answer represents a higher level of competence than does the ability to recognize which of several responses offered by the printed page is best or correct. The argument is sufficiently persuasive so that it is often accepted at its face value, unchallenged and deemed unassailable. The truth of the matter is not, however, quite so simple.

It perhaps is reasonable to assume, at least in theory, that a greater degree of competence is evidenced by production of a correct response than by identification of a correct response from a list of stated options. The examinee who knows a fact and can produce it without clues should also be able to recognize the fact when it is ' presented in multiplechoice form. The reverse need not be true a student who could not have produced the free response may be able to select the correct when faced with it as one of a small number.of options. Between the two response modes, then, the superiority of freely produced responses might be acknowledged if other considerations were equal. But other considerations are not equal the nature of the function being measured and the skill with which the items are constructed in practice usually swing the advantage from one measurement approach to the other.

Simply put, it is more difficult to write good shortanswer freeresponse items than it is to write good multiplechoice selectresponse items, especially if one wants to test for something other than simple fact, The essential problem in preparing freeresponse items is so to state the problem that the student knows exactly what response is expected of him. If the response is truly to be short answer, it should be a single word, a name, a symbol, a formula, a sign, or the like. Relatively few areas in which testing is done offer instances in which answers to questions calling for generalizations, principles or applications can be stated adequately by the examinee with a word

or two. a sign, etc. The result is that the large preponderance of shortanswer items measure no higher level of mental process than memory or fact."

Met permissie: de laatste opmerkingen van Wesman zijn volslagen onterecht: allereerst vervalt hij in de bekende drogreden goede meerkeuze items te vergelijken met middelmatige open-eindvragen. Ten tweede geldt juist en met name voor meerkeuze items dat de alternatieven ook bestaan uit één of twee woorden, een getal, e.d., zodat uit dien hoofde zijn bezwaren tegen open-eindvragen evenzeer gelden ten aanzien van meerkeuze vragen, zodat beide vraagsoorten elkaar in dit opzicht gelijk blijven. Maar terug naar Wesman:

Hij noemt als bezwaren tegen open eind vragen verder dat het vaak niet mogelijk blijkt de vraag z6 te formuleren dat alle studenten die de vraag kunnen beantwoorden hetzelfde antwoord gebruiken. Dit is echter pas een regel bezwaar wanneer het grote aantal studenten verhindert dat de docent zélf alle toetsen nakijkt. Kijkt de docent (bijvoorkeur samen met enkele collega's) zelf de toetsen na (in sommige situaties kunnen de studenten best hun eigen werk, of het werk van andere studenten nakijken), dan is een probleem dat interpretatieverschillen de betrouwbaarheid van de beoordeling kunnen verlagen. Dat wil zeggen, dat verschillende docenten het met elkaar niet eens zijn over de vraag of een bepaald afwijkend antwoord dat toch niet duidelijk fout is, dan ook goed gerekend moet worden. Dit bezwaar dat tegen openeind vragen aangevoerd zou kunnen worden is slechts een schijnbaar bezwaar: er wordt dan gezegd dat in vergelijking met meerkeuze vragen waar de scoring van te voren geheel vastgelegd wordt, de problemen die bij beoordeling van open eind vragen zich voordoen een extra bron van onbetrouwbaarheid zouden vormen. Dat hoeft echter in het geheel niet zo te zijn: objectieve scorbaarheid heeft geen enkele garantie te bieden voor betrouwbaarheid van de vragen, wanneer een vraag slecht geformuleerd is, of zelfs het verkeerde alternatief als 'goed' in de scorings sleutel terecht komt, wordt er weliswaar perfect betrouwbaar gescoord, maar dan wel perfect betrouwbaar verkeerd. Korter gezegd: het gaat niet aan om volstrekt willekeurige aspecten uit de gehele beoordelingsprocedure te lichten, en te zeggen dat op die punten een vraagsoort meer of minder

betrouwbaar zou zijn. Betrouwbaarheid van beoordeling is niet een éénduidig begrip (zie o.a. Stanley, J. C. Reliability, in Thorndike, R. L. (Editor) Educational Measurement. Washington, D. C. : American Council on Education, 1971)

Nog eens anders geformuleerd: bij open eind vragen kunnen inderdaad interpretatie problemen ontstaan op het moment dat de toetsformulieren gescoord moeten worden. Bij meerkeuze items hebben we natuurlijk met dezelfde problemen te maken: de docent kan daar op nogal subjectieve wijze vraagstellingen genereren en antwoorden als de enig juiste of de beste aanwijzen, waarover hij onenigheid zou krijgen met collega's wanneer hij hen de toets zou voorleggen. Bij een behoorlijke voorbereiding van een meerkeuze toets hoort dan ook altijd de controle bestaande uit het laten maken van de opgaven door op het vakgebied deskundige collega's, om dubieuze vragen te kunnen signaleren. Welnu, een dergelijke controle moet uiteraard ook met open vragen uitgevoerd wordt, zodat ook daar op dezelfde wijze vragen verwijderd kunnen worden die op meerdere manieren goed beantwoord kunnen worden (op onverwachte manieren goed beantwoord kunnen worden). Bij de openeind vragen blijft het dan altijd nog mogelijk dat zich een enkel probleem. voordoet b bij de werkelijke toets scoring; maar daar staat tegenover dat eventuele onvolkomenheden in meerkeuze items bedenkt worden met de mantel van de objectieve scorings liefde, wat wil zeggen dat de beoordeling weliswaar geen onbetrouwbare variantie meer toevoegt, maar dat die onbetrouwbare variantie &r desalniettemin toch maar in zit: onduidelijke, slechte, dubieuze items zijn ipso facto onbetrouwbaar, hoewel perfect betrouwbaar scoorbaar, en het vervelende is dat item analyse procedures dergelijke items niet altijd (vaak niet ?) boven water halen (en wat dan, moet je dergelijke slechte items alsnog, nadat de toets is afgenomen, verwijderen? Dan kun je hetzelfde doen met open eind vragen die niet eenduidig beantwoord worden).

Ook in dit opzicht: betrouwbaarheid, lijken openeind vragen niet onder te doen voor meerkeuze vragen, zijn beide in wezen aan elkaar gewaagd. Ongetwijfeld wanneer aan beide vraagsoorten evenveel zorg besteed wordt bij de formulering en het uitproberen,

Stanley, J. C., & Hopkins, K. D. Educational and psychological measurement and evaluation. Englewood Cliffs, N. J.: Prentice Hall, 1972.

Een reminder m.b.t. meerkeuze items en de min of meer vanzelfsprekende associatie die vaak gelegd wordt tussen deze vraagvorm en de herkenning van het goede antwoord als de activiteit van de student:

"Multiple choice items were formerly referred to simply as 'Recognition' items, as if they were capable of testing only rote knowledge. Obviously, they can be as highlevel as the effort and ingenuity of an able writer can make them. Even a simple arithmetic reasoning item such as 'How many apples can I buy for 42 when each apple costs 7?' involves application. A teacher can leave a blank at the end of such a question, or he can provide several options and let the student choose the correct option. In neither case is 'recall' alone involved, as it might be if we asked what year England entered World War I."

"The free response test has the obvious advantage of familiarity and 'naturalness'. It may almost completely eliminate guessing, because the student chooses among the (usually large) number of options he can recall. The freeresponse test is particularly valuable for mathematics and the physical sciences when the stimulus appears in the form of a problem requiring computation. It also has wide application to test situations when it is presented in the form of maps, charts, and diagrams for which the pupil is required to supply, in spaces provided, the names of parts keyed by numbers or letters."

"A limitation of the freeresponse test is that it tends to measure highly factual knowledge, consisting of isolated bits of information, because questions with unrelated short answers are used. The isolatedknowledge level is avoidable if the writer is ingenious.¹¹ Opmerking: dit bezwaar wordt door de ene tekstboek schrijver na de andere gemaakt, altijd zonder verwijzing naar wat voor empirisch onderzoek of plausibele onderwijskundige redenering dan

ook, kennelijk uitsluitend op basis van 'ervaring?'. Die 'ervaring' is in vele gevallen bovendien z6 dat openeind vragen vaak door de docent zelf gemaakt worden, en meer gebreken vertonen, dan meerkeuze items uit gestandaardiseerde tests die door gespecialiseerde itemschrijvers ontwikkeld en door andere specialisten op hun kwaliteiten onderzocht zijn. Terecht maakt Stanley & Hopkins hier dan ook de opmerking dat aan genoemd gevaar, dat evenzeer geldt voor meerkeuze items (zie citaat boven aan de bladzijde) alleen maar tegemoet gekomen kan worden door aan de formulering van openeind vragen dezelfde zorg te besteden als aan de formulering van meerkeuze vragen.

"Obvious advantages of the fixedresponse test (meerkeuze toets) are its apparent ease of construction (maar zie citaat boven aan de blz.), applicability to a wide range of subject matter (wat ook voor andere toetsvormen geldt), objectivity of scoring (waar ook nadelen aan verbonden zijn die door de 'objectiviteit' van de scoring versluierd worden, waar ze bij openeind vragen direct blijken), and wide sampling of knowledge tested per unit of working time (wat evenzeer geldt voor openeind vragen)."

"It is unfortunate that many persons assume the multiplechoice test to be popular because it is easy to score. Its major virtue is that it requires the examinee to discriminate among alternatives. The 'bestanswer' type of multiple choice item can measure the degree of understanding of abstract concepts." Opmerking: Stanley & Hopkins begeven zich hier buiten het kader waarin zij de toetsingsproblematiek in hun boek behandelen. Dat kader is vrijwel uitsluitend de meettechnische kant van de zaak, inclusief behandeling van allerlei praktische probleempjes waarvoor de docent komt te staan, en ethische problemen. De geciteerde opmerking heeft betrekking op de aard van de cognitieve operaties die 'getoetst(worden, een onderwerp waaraan zij, behalve een summiere bespreking van de cognitieve taxonomie van onderwijsdoelstellingen van Bloom c.s., geen aandacht besteden. De vraag blijft dan ook volledig open of wat Stanley en Hopkins hier stellen wel houdbaar is (zij geven zelf geen verdere argumentatie, behalve een op zich wel interessant voorbeeld). In het door het COWO te ontwikkelen docentencursusonderdeel 'constructie van toetsvragen' komt juist aan de orde welke 'cognitieve operaties'

bij uitstek in onderwijssituaties aan studenten bijgebracht moeten worden, en ook getoetst moeten worden. Met name komt daar aan de orde de problematiek rond beheersing van abstracte begrippen, en de wijze waarop dergelijke beheersing van abstracte begrippen toetsbaar gemaakt kan worden. Het ziet er op dit ogenblik niet naar uit dat het meerkeuze item daarbij grotere mogelijkheden biedt dan vragen die korte antwoorden vereisen (openeind vragen). Ook wie het door Stanley en Hopkins gegeven voorbeeld analyseert zal ontdekken dat in één item een groot aantal verschillende 'bits' kennis tegelijk getoetst worden, wat je met evenveel recht als Stanley en Hopkins dit a priori 'het begrijpen, van een abstract begrip' noemen, het paraat hebben van een groot aantal stukjes 'kennis' kunt noemen. Een dergelijk item zou je om wille van een duidelijke toetsing, interpreteerbare resultaten, feedback naar student zowel als docent, e.d. juist bij voorkeur opknippen in een aantal kleinere items. Wat betreft toetsing van echt gecompliceerde cognitieve operaties: het is in veel gevallen zeer de vraag of dat wel moet gebeuren in toetssituaties zoals wij die in ons hoger onderwijs kennen; veeleer moeten dergelijke activiteiten in het onderwijs zélf als 'Handelingsonderdelen' (De Groot) aan de orde gesteld worden. Echter, ook dit is een probleem dat op deze plaats niet verder besproken kan worden.

Hofstee, W. K. B. Participatie controle door 'onbenullige' toets items. Nederlands Tijdschrift voor de Psychologie, 1973, 28, 189-198. [niet online?]

"Intussen is aan de universiteiten de hausse van deze discussie achter de rug (klassiek tentamen of meerkeuze toets) en het pleit is wel ongeveer beslecht in het voordeel van de keuzetoets. Dat wil overigens niet zeggen dat de strijd op relevante, b.v. onderwijskundige en psychometrische argumenten is gewonnen. Want terwijl bij voorstanders van keuzetoetsen het objectiviteitsargument centraal stond, heeft naar alle waarschijnlijkheid het efficiëntieargument de doorslag gegeven: we veronderstellen tenminste dat de toetsvorm (keuze, essay, mondeling) voornamelijk voorspelbaar is vanuit de oplage, en dat lijkt een plausibele veronderstelling."

"Een tweede hypothese ter verklaring van de huidige populariteit van keuzetoetsen is er een die lastiger zou zijn waar te maken maar die echter te belangrijk is om buiten discussie te blijven. Wanneer men nl. begripsvragen definiëert als vragen waarop slechts weinige uitverkorenen het goede antwoord vinden, dan voldoen veel studietoetsen aan de eis dat ze begrip meten. Wie dus bevreesd was dat ieder die zijn lesje geleerd heeft voor een keuzetoets zal slagen, merkte al spoedig dat hij daar althans niet bang voor behoefde te zijn."

"Enkele argumenten voor de bovenstaande, ietwat cynische hypothese, die er dus op neerkomt dat keuzetoetsen ingang vonden omdat ze niet zo makkelijk bleken als men eerst had gevreesd: ten eerste blijkt dat zowel docenten als studenten geneigd zijn moeilijke items begripsvragen te noemen (Vonk, 1967). (Voor literatuur verwijzingen zie Hofstee, b.w.) Verder blijkt de voor gissen gecorrigeerde moeilijkheidsgraad van de doorsnee toetsvraag bij ongeveer 0,60 te liggen (Wijnen 1971). Wanneer we aannemen dat de toetsscores symmetrisch verdeeld zijn, zoals doorgaans bij benadering het geval is, volgt hieruit dat gemiddeld ongeveer de helft van de mensen zou zakken als de meest natuurlijke (maar daarom nog niet goede, b.w.) beslissingsregel zou worden gehanteerd: nl. dat iemand na correctie voor gissen zonder verdere omrekening zestig percent goed, dus een 'zes' (nogmaals: in de praktijk blijken cijfers of deze idiote manier nogal eens toegekend te worden; onderwijskundige of psychometrische argumenten voor die handelwijze bestaan niet. Het is de bedoeling van Hofstee te beschrijven hoe het vaak gaat, niet hoe het beter zou moeten, b.w.) moet halen. De docent zal dus meestal 'teleurgesteld' worden door de toetsresultaten."

"Dat deze teleurstelling gemakkelijk omslaat in respect voor de studietoets als instrument is dan inderdaad een speculatie van schrijver dezes. Als argument daarvoor zou kunnen worden aangevoerd dat de neerwaartse druk op de slaagpercentages m.n. aan de universiteiten aanzienlijk is. Aangezien voorts motiveringen van zuiver numerieke aard (nl. dat veel mensen moeten zakken omdat de onderwijscapaciteit beperkt is) zeker niet algemeen geaccepteerd worden, is ieder 'onderwijskundig' 'objektief' argument voor hoge zakpercentages welkom. Zo kan de keuzetoets een hoogst twijfelachtige rol gaan spelen

(zie ook Warries 1970)"

"Natuurlijk is er geen enkele reden waarom een studietoets als zodanig moeilijk zou moeten zijn. Aangenomen moet worden dat de door Wijnen geobserveerde moeilijkheidsgraad de resultante is van eventueel impliciete onderwijskundige visies. Verder speelt misschien een rol dat de maker van keuzevragen uit beduchtheid voor te gemakkelijke items juist zijn toets te moeilijk maakt." [Dit zijn toch allemaal citaten uit Hofstee? bw]

"Wie de zin en het goed recht van de gesignaleerde ontwikkeling in twijfel trekt kan twee kanten op. Hij kan met relatieve normen gaan werken, dus geheel afzien van de moeilijkheidsgraad van de toets; de meest intelligente en eindwandfreie procedure is dan hetgeen momenteel in Groningse medische kringen de 'cesuur volgens Wijnen' wordt genoemd. Hij kan ook de onderwijsfilosofie die tot moeilijke toetsen leidt onder ogen zien en er een alternatief tegenover stellen."

Dit laatste doet Hofstee dan, en hij stelt voor om 'onbenullige items' te gebruiken, items die niet meer bedoelen dan participatie controle te zijn: degenen die het onderwijs hebben gevolgd, de literatuur gelezen, e.d., moeten de items goed kunnen maken. "De vragen zijn 'onbenullig' in de zin van: zonder veel nadenken te beantwoorden door ieder die kennis heeft genomen van de leerstof. Gestreefd wordt naar eenvoudige vragen over de hoofdzaken van de leerstof."

Hofstee werkt een en ander uit in de vorm van tweekeuze vragen, maar het zal duidelijk zijn dat ook vragen die een kort te formuleren antwoord eisen hier goed gebruikt kunnen worden.

Mijn bedoeling met de inbreng van deze aanpak van Hofstee is te laten zien dat open eind vragen óók buiten de sfeer van rekenopgaven e.d. juist bijzonder zinvol kunnen zijn. Het gevaar bestaat dat men uit herhaalde opmerkingen in de literatuur dat openeind vragen bijzonder geschikt zijn voor rekenopgaven e.d., en dat voor andere leerstof het risico bestaat dat ze te zeer op detailkennis ingaan (hoewel dat laatste argument in deze bespreking meermalen van tegenargumentatie is

voorzien), de conclusie trekt dat meerkeuze vragen misschien meer geschikt zijn dan open eind vragen wanneer het gaat om leerstof die zich niet laat toetsen door rekenopgaven of manipulaties met kaarten, diagrammen, e.d. Hofstee laat zien dat juist bij uitstek voor vakken waar met 'fuzzy concepts', uitgebreide literatuurlijsten, etcetera wordt gewerkt (sociale wetenschappen, letteren faculteit, en onderdelen uit programma's economie, medicijnen etc.) heel simpele vragen (open eind of meer of tweekeuze) goed in overeenstemming met de onderwijsdoelstellingen kunnen zijn.

Voor verdere bespreking van deze wijze van toetsen, en de manier waarop ze gerealiseerd kan worden, zie het artikel van Hofstee.

[projecten/76Open_vs_keuzevragen_1-3-76.pdf](#)