PROEFWERKEN MET VIERKEUZE-VRAGEN VOOR HET BASISONDERWIJS.

E. WARRIES

Nadat in 1965 de regeling voor toelating tot het v.h.m.o. bij Koninklijk Besluit herzien was, zijn in de daarop volgende jaren in verscheidene plaatsen in ons land gemeenschappelijke proefwerken afgenomen aan de leerlingen van de zesde klas Basisonderwijs. Deze proefwerken, proeven, toetsen of schoolvorderingentests dienden er toe „kennis en inzicht" van betrekkelijk grote groepen leerlingen na te gaan. De uitslagen van de proeven werden bij de schoolkeuze c.q. toelating tot het vervolgonderwijs benut – in overeenstemming met de nieuwe wettelijke regeling – als tweede gegeven náást het geschiktheidsoordeel van het Hoofd der School.

Deze gemeenschappelijke proefwerken vormen blijkens hun herhaalde toepassing in enkele grote gemeenten (in objectief scoorbare vorm) een nuttig middel om grote groepen leerlingen tegelijkertijd en vergelijkenderwijs op hun prestaties te toetsen. De doelmatigheid van dit toetsmiddel is volgens de meeste test-deskundigen optimaal bij het gebruik van vragen en opgaven in de „meerkeuze-vorm". In de praktijk van het testen van school- en studieprestaties zal dit meestal betekenen: 4 of 5-keuze-items. Zo bestaat de laatste „Amsterdamse Schooltoets" (1968) geheel uit (ruim 300) opgaven in vierkeuze-vorm. *

De constructie van school- en studietoetsen is langzamerhand ook in Nederland een specialisme aan het worden: in ons land zijn het tot nu in het voortgezet onderwijs voornamelijk psychologen en vakleraren, die zich systematisch hierop hebben toegelegd.

Voor het Basisonderwijs en met name de vakken, die bij de Amsterdamse Schooltoets worden behandeld, zijn onlangs te Amsterdam permanente werkgroepen opgericht, die zich bezighouden met het schrijven van opgaven in vierkeuze-vorm. Op een eerste plenaire bijeenkomst van deze werkgroepen hielden wij over het schrijven van items een inleiding – die in dit artikel is uitgewerkt. Wij hopen met het onderstaande enig inzicht te geven in de problemen, die men noodzakelijk onder ogen moet zien als men zich bezig houdt met het construeren van gemeenschappelijke proeven. Het spreekt vanzelf, dat hier niet alle aspecten van „de toetsconstructie-techniek" ter sprake komen. Met name de toepassing van de „rekenregels" bij de toetsverbetering

moet hier onbesproken blijven. In het onderstaande wordt allereerst gesteld dat in de toepassing van schooltoetsen een zogvuldige en uitvoerige voorbereiding nodig is. Vervolgens worden zeer kort enige vraagvormen aangeduid. Daarna wordt uitvoerig gewezen op de eisen die gesteld moeten worden aan de vormgeving van de items en tenslotte worden enige richtlijnen aangegeven die men kan volgen bij het schrijven van de opgaven voor een toets.

1. DE NOODZAAK VAN ITEMSCHRIJFGROEPEN

Een opvallend kenmerk van toetsen bestaande uit opgaven in de meerkeuzevorm is dat het belangrijkste werk *vooraf* gaat en dat dit werk omvangrijker is dan het voorbereiden van een „gewoon" proefwerk. De voornaamste taak bij een gewoon proefwerk – dat immers bijna altijd bestaat uit „open vragen" waarbij de leerling vrij wordt gelaten in de formulering van zijn antwoord – komt pas achteraf. Het beoordelen van een gemaakt proefwerk is een taak, die hoge eisen stelt aan de vaktechnische bekwaamheid van de beoordelaar. Het is diezelfde vaktechnische bekwaamheid, die nu bij het maken van meerkeuzevragen vooraf een rol moet spelen. Bovendien moet hier nog een ander soort technisch kunnen mee gaan spelen en dat is de techniek van de toetsconstructie.

Bij de multiple choice opgave is het voorafgaande werk omvangrijker dan bij een gewoon proefwerk. Dit komt doordat men nu als vraagensteller niet alleen komt te staan voor het probleem *wat* te vragen, maar ook *hoe* het te vragen. Het is nu namelijk niet meer voldoende te volstaan met het stellen van vragen „zonder meer" aan de leerlingen over bijvoorbeeld rekenkundige problemen. Men moet nu ook gaan anticiperen op de oplossingsprocessen, die de leerling zou kunnen doorlopen, en op de fouten, die de leerling zou kunnen maken. Dit vereist vrij veel inzicht in het oplossingsproces en misschien ook in het leerproces van de kinderen. Daar komt nog iets anders bij: zodra we gedwongen worden precies te formuleren *hoe* we onze vragen gaan stellen, dus hoe we onze items gaan maken, moeten we ook heel duidelijk worden over datgene wat we willen meten. Wanneer we bijvoorbeeld de kennis „dat Montevideo de hoofdstad is van Uruguay" willen toetsen * dan kunnen we dat – alleen al door de keuze van onze vier alternatieven – op verschillend niveau doen. Of we voor de vier alternatieven vier Zuid-Amerikaanse landen nemen, of dat we één Zuid-Amerikaans en drie Europese landen nemen, of één Aziatisch, één Europees en één Afrikaans land, maakt erg veel verschil voor het te

testen niveau van kennis. Het eerste is bijvoorbeeld een vraag, die tamelijk gespecialiseerde kennis van de Zuid-Amerikaanse aardrijkskunde vraagt, terwijl de beide andere formuleringen minder eisen stellen ten aanzien van die kennis. Uit dit voorbeeld wordt dus duidelijk dat het probleem „hoe je gaat vragen" bij de objectieve studietoetsen een belangrijk (en vaak een moeilijk oplosbaar) probleem is.

Het is van belang, dat het vele omvangrijke voorbereidende werk dat aan een studietoets verbonden is, plaatsvindt in daartoe samengestelde werkgroepen. In teamwork immers kan nu eenmaal méér gebeuren dan wanneer enkele en slechts weinige mensen elk afzonderlijk werken aan een stuk toets. Er is echter nog een andere overweging die ervoor pleit het itemschrijven in teams te beoefenen. Deze overweging betreft de strenge eisen die juist aan meerkeuze-opgaven gesteld mogen worden. De aard van het soort opgaven en de technologie die zich op het gebied daaromtrent heeft ontwikkeld, is zover, dat men aan dit soort items zeer hoge eisen kan en mag stellen. Om aan die nog te noemen strenge eisen te voldoen, is het blijkens onze ervaringen nodig in een groep te werken. Het werken met een groep, vooral als die van heterogene samenstelling is, ook naar geografische herkomst der leden, heeft grote voordelen. We mogen hierbij denken aan bijvoorbeeld de stokpaardjes, die item-schrijvers, of in het algemeen vragenstellers, noodzakelijkerwijze gaan berijden. Ik noem twee voorbeelden: één itemschrijver blijkt bij het kiezen van zijn opgaven de neiging te hebben zijn opgaven altijd zó te redigeren, dat er ongelukken of misdaden bij te pas komen. Een andere itemschrijver blijkt in zijn taalgebruik te poëtisch te zijn en vogels bijvoorbeeld altijd „zangertjes" te noemen. Het spreekt vanzelf dat zulke en minder onschuldige eigenaardigheden, die een toets ongunstig kunnen beïnvloeden, gemakkelijk kunnen worden gecorrigeerd in een groep van itemschrijvers, tot voordeel van de leerlingen en van de betreffende schrijver.

Dat het vormen van itemschrijfgroepen bij gemeenschappelijke proeven geen onvervulbare wens hoeft te blijven, mag ook gemotiveerd worden door te wijzen op de vaak zeer grote aantallen te toetsen leerlingen. De Amsterdamse Schooltoets bijvoorbeeld werd dit jaar door meer dan 16.000 kinderen gemaakt. Als we even denken aan de hoeveelheid werk, die doorgaans gestoken wordt in een proefwerk voor zeg 100 leerlingen, dan zouden we hier dus 160 maal zoveel energie in mogen steken. Met andere woorden: bij een zo groot aantal mag eigenlijk geen moeite teveel zijn om de voorbereidingen goed te doen.

2. DE VORM VAN EEN ITEM

Alvorens te komen tot een behandeling van de hierboven gemelde strenge eisen, eerst nog iets over een zeer elementaire zaak: Hoe ziet een vierkeuze-item eruit?

Aan het item onderscheiden we de „stam" en de vier „alternatieven". De stam bevat de vraagstelling, de probleemstelling of de opdracht aan de leerling. In de Amsterdamse Schooltoets hebben we in vorige jaren een aantal itemtypen leren kennen, waarbij stammen van verschillende vorm voorkomen. We hadden daarbij *incomplete zinnen,* waarbij het goede complement stond in één van de vier alternatieven.

Een voorbeeld is: *

Bij vroedschap denken we aan . . .

A een jonkvrouw
B een raadsman
x C een stadsbestuur
D godsdiensttwisten

Behalve dit soort incomplete zinnen, die op drie puntjes eindigen, bestaat er ook een itemtype, waarbij midden in de zin iets is weggelaten, bijvoorbeeld:

10 Miljoen schrijft men als een 1 met . . . nullen.

A 5
B 6
x C 7
D 8

Bij invul-items wordt de vraag ook wel aldus gesteld: „Welk woord past het beste op de open plek?" Bij de subtoets Gemengde Taalopgaven waar het ging om „het stellen", is deze vorm dikwijls toegepast: We gaan naar een plaats toe, waar we allemaal (. . .) onze trekken komen

Welk woordje past het best op de open plek?

x A aan
B bij
C in
D om

Er zijn nog meer vormen denkbaar, waarbij een *weggelaten stuk van de stam* moet worden aangevuld met één van de alternatieven. In het algemeen echter is er onder ervaren testconstructeurs een voorkeur

voor de directe vraag. De directe vraagstelling heeft wat betreft duidelijkheid – voor zowel de itemschrijf-groep als de leerlingen – grote voordelen boven andere vormen.

De directe vraag met een *ontkenning* erin heeft gevaren. Het denkproces, dat van een leerling wordt gevraagd in zo'n geval, wijkt waarschijnlijk enigszins af van wat hij bij andere opgaven moet doen.

Een voorbeeld:

Welke van deze verschijnselen is *niet* vulkanisch?

A asregen
x B fossielen
C geisers
D lava

Omdat de leerling hier in plaats van „het goede" nu „het foute" moet zoeken, wordt in het algemeen afgeraden dit soort vragen te stellen. Als we het toch willen doen, dan dienen we de leerling zeer nadrukkelijk op het negatieve karakter van de vraag attent te maken. De enige plaats waar we in de Amsterdamse Schooltoets wèl konsekwent de leerling „het foute" laten zoeken is in de subtoets „spelling". Daar is van de vier onderstreepte woorden er één verkeerd geschreven: dat woord moet opgezocht worden:

A B x C D

In de *zomer* rond het *middaguur* kan de zon *vel branden* op de bloemen.

Bij de spellingstoets wordt aangenomen dat de manier waarop we onze eigen of andermans schriftuur bekijken om die te verbeteren, óók neerkomt op het min of meer automatisch zoeken naar fouten. Bij de spellingsopgaven wordt dit dus eveneens van de leerlingen gevraagd.

Behalve de hier behandelde stamvormen kennen we ook nog de stam, waarin aan de eigenlijke vraagstelling een *inleiding* voorafgaat.

Die kan, soms noodgedwongen, vrij lang uitvallen:

In Engeland is het 1 uur vroeger dan in Nederland.
De boot van Hoek van Holland naar Harwich (Engeland) vertrekt om 7.00 uit Hoek van Holland. Hij komt precies 10 uren later in Harwich aan. Hoe laat is het dan in Harwich?

A 15.00
x B 16.00
C 17.00
D 18.00

Zoals uit dit voorbeeld blijkt, is het soms nodig de vraagstelling te la-

ten voorafgaan door een tekst, waarin het probleem wordt gesteld. Bij tekstbegrip- of stillees-items brengt de aard van het werk mee, dat een stuk tekst in lengte variërend van 100 tot 1500 woorden door de leerling wordt gelezen, waarna meer items worden beantwoord. Hier behoren dus meerdere items altijd bij elkaar; ze kunnen niet afzonderlijk gesteld worden of worden overgeheveld naar een andere toets. In het algemeen geldt echter als regel, dat een item één afzonderlijke meeteenheid moet zijn, die als afzonderlijk item in een kaartenbak kan staan. Stilleestoetsen vormen op deze regel dus een uitzondering.

3. EISEN, WAARAAN EEN ITEM OF EEN TOETS MOET VOLDOEN

In de loop van de tijd is het construeren van tests en van school- of studietoetsen een specialistisch vak geworden. Dat heeft meegebracht, dat testconstructeurs langzamerhand vrij hoge eisen zijn gaan stellen aan het eindproduct, zowel het afzonderlijk item als de hele toets. We spreken hier uitsluitend over enkele eisen die reeds tevóren zonder rekenwerk aan de items gesteld kunnen worden en we laten de zeer belangrijke psychometrische eisen hier onbesproken. We volstaan met te constateren, dat de analyse achteràf van de antwoorden der leerlingen informatief en onontbeerlijk is.

Hieronder worden zes eisen behandeld die op het ogenblik voor de gemeenschappelijke proefwerken het belangrijkste lijken te zijn. Daarbij is gebruik gemaakt van twee boeken. Allereerst Ebel: „Measuring educational chievement", een standaardwerk op het gebied van studietoetsen. Verder uit het reeds genoemde en nog te verschijnen boek „Studietoetsen" het hoofdstuk van G. J. Mellenbergh en mej. W. Lans over itemschrijven.

3.1 De eis van relevantie

Hierbij stelt de itemschrijver zich de vraag of hij inderdaad bezig is de vragen te stellen, die hij met de toets had willen stellen. Voortdurend moet de schrijver zich ervan bewust zijn, dat zijn items relevant zijn ten aanzien van de bedoelingen van de toetsgebruiker. Wat verwacht de gebruiker van de toets? Waartoe dient de toets? Bij het stellen van deze vragen blijkt dikwijls dat de bedoelingen van de toets niet duidelijk geformuleerd zijn en kost het alleen al om die reden moeite steeds in het achterhoofd te houden waar de opgaven toe dienen. Misschien zou het goed zijn om de doeleinden van elke toets tevoren goed met elkaar af te spreken. Ebel (p. 284) adviseert een korte doel-omschrijving, die voor onze schooltoets er bijvoorbeeld als volgt zou kunnen gaan uitzien:

Het doel van de schooltoets is na te gaan in hoeverre het basisonderwijs is overgekomen in het zesde leerjaar wat met het basisonderwijs is bereikt einde maart – begin april. Om dat na te gaan, willen we vragen stellen, die:

1. in meerkeuze-vorm zijn
2. kennis en inzicht in de stof vragen
3. betrekking hebben op de „basisstof" uit de Proeve voor een Leerplan van het Nutsseminarium voor Pedagogiek
4. door vakexperts op dezelfde manier worden beantwoord
5. geen stereotype zinswendingen of „verbalismen" meten.

De bovenstaande doelomschrijving is hier overigens geïmproviseerd ingevoegd. De bedoeling van deze improvisatie is duidelijk: aangeven wat voor soort kenmerken zo'n doelomschrijving kàn hebben en wat voor soort relevante aspecten men zou kùnnen noemen. Er staat bijvoorbeeld in genoemd, dat het gaat om einddoelstellingen van het basisonderwijs en er wordt gezwegen over toelatingsselectie voor het v.h.m.o. Als we dat element van selectie voor de „moeilijke" scholen in het voortgezet onderwijs wèl relevant voor de toets achten, dienen we dat tevoren vast te stellen en zo mogelijk vast te leggen. Hetzelfde geldt voor de stofomschrijving in de „Proeve" en voor de uitdrukking „kennis en inzicht". Als we de bedoelde stofomschrijving te ruim of te beperkt achten of als we „inzicht" niet relevant vinden maar bijvoorbeeld liever spreken over „toepassing van kennis", dan kunnen we en moeten we dat doen. Tevoren! Alleen immers als we tevoren omschreven hebben wat we relevant achten, kunnen we later tijdens het schrijven of bij de controle op het geschrevene, aan de eis van relevantie tegemoetkomen. Relevantie is uiteraard een essentieel kenmerk van een school- of studietoets. Hoe boeiend het itemschrijven als bezigheid ook mag zijn, nooit mag de testconstructeur zich laten fascineren door zijn eigen opgaven. De vraag „waar het om gaat" moet altijd gesteld blijven.

3.2 De eis van evenwichtigheid

Hierbij gaat het erom dat we ons bij de uiteindelijke samenstelling van de toets realiseren in welk evenwicht de verschillende delen van de toets met elkaar staan. De Amsterdamse Schooltoets 1968 bijvoorbeeld bevatte ongeveer 160 taal-items, 85 reken-items en 70 algemene kennisitems. Is dit inderdaad in overeenstemming met onze bedoelingen? Zijn we ons ervan bewust, dat de taalscore, in vergelijking met de rekenscore, tot stand komt door bijna twee maal zoveel opgaven? Denken we er ook aan, dat aldus in de Totaalscore „Rekenen plus Taal" de taalopgaven zwaarder vertegenwoordigd zijn dan de rekenopgaven? De

zwaarte waarin de verschillende toetsen (en delen daarvan) aanwezig zijn, is van groot belang bij de scoring. *

Niet alleen voor de scoring is het onderlinge evenwicht van de verschillende delen van de hele toets van belang; ook voor de „sturende" invloed die de toets op het onderwijs kan hebben, is de evenwichtigheid een factor, die in het oog moet worden gehouden. Het is nu eenmaal mogelijk, dat op datgene wat niet gevraagd wordt in een examen (of in een toets die afgenomen wordt aan het eind van een onderwijs-periode) in een volgend schooljaar door sommige docenten minder nadruk zou worden gelegd. Het lesprogramma is vol, de tijd is kort, de voorbereiding kost moeite: hoe verleidelijk is het dan niet om aan de onderwijsdoeleinden, die *niet* in de toets gemeten worden, wat minder aandacht te schenken.

De eis van evenwichtigheid is derhalve van groot belang voor zowel de interpretatie van de scores door de leerling behaald, als voor de invloed die misschien op het onderwijs zou kunnen uitgaan.

3.3 De eis van efficiëntie

Bij mondelinge universitaire examens komt het nogal eens voor, dat de hoogleraar over een stuk van de bestudeerde stof een monoloog van vijf à tien minuten houdt, eindigend met een vraag naar de zienswijze van de examinandus. Deze wordt dan geacht te antwoorden met een betoog, dat weliswaar niet zo lang behoeft te zijn als dat van de professor, maar dat in elk geval niet mag bestaan uit een korte mededeling over de mate waarin de student het eens is met het gestelde.

Een dergelijke wijze van ondervragen – hoe bruikbaar ook in de beschreven situatie – wordt bij objectieve school- en studietoetsen als inefficiënt beschouwd. De inleiding tot de vraag moet kort en duidelijk zijn en niet teveel leestijd vragen. Ook de alternatieven mogen niet zó geformuleerd zijn, dat de leerling veel tijd nodig heeft om te ontdekken en te begrijpen wat er staat. Lans en Mellenbergh geven het volgende voorbeeld van a) een lange en b) een bekorte inleiding.

a. Jan gaat naar de groenteboer om voor zijn moeder boodschappen te doen. Hij moet daarvoor eerst de grote weg oversteken. Daarbij moet hij goed uitkijken, eerst naar links en dan naar rechts. Hij koopt bij de groenteboer 9 kg. aardappelen van 48 cent per kilo. Hoeveel moet hij dan betalen?

A f 3,84
B f 3,92
x C f 4,32
D f 4,80

b. Als je 9 kg. aardappelen van 48 cent per kilo koopt, hoeveel moet je dan betalen?

A f 3,84
B f 3,92
C f 4,32
D f 4,82

Het spreekt vanzelf, dat het bekorte item uit het voorbeeld als een efficiënter item beschouwd moet worden dan het lange. Waarom is deze efficiëntie van zo groot belang? Omdat inefficiënte items teveel tijd van de leerling vergen en van dit soort items derhalve een gering aantal per lesuur kan worden voorgelegd. De tijd, die in de school voor toetsen kan worden vrijgemaakt is altijd beperkt; we moeten proberen zoveel mogelijk vragen te stellen als redelijkerwijze de leerlingen kunnen beantwoorden. Dat we véél vragen moeten stellen staat in verband met de psychometrische eis van meetbetrouwbaarheid. Een eindscore die gebaseerd is op 50 items is nu eenmaal „betrouwbaarder" dan één die na een test van 25 items is berekend. Bij een „betrouwbare" score is per definitie de toevalsfactor geringer. Elke uitslag kan een paar punten te hoog of te laag zijn uitgevallen. Net zomin als een IQ van 89 precies 89 is, is een schooltoets-uitslag exact juist. Er is een zekere marge omhoog en omlaag. Bij betrouwbare toetsen is deze marge gering. Dus: hoe meer (goede) items, hoe beter.

In de Schooltoets blijken het vooral de rekenvraagstukken te zijn, die op het punt van efficiëntie problemen geven. De aard van dit soort reken-items brengt mee, dat het vraagstuk moet worden ingeleid en dat daarna pas de vraag wordt gesteld. Dikwijls zien wij nu dat de inleiding informatie bevat, die op zichzelf niet nodig is om de gestelde vraag te beantwoorden. De itemschrijver noemt de knikkerende of winkelende kinderen bij naam en toenaam, voert naar believen ouderfiguren of winkeliers in en doet zijn best om een zo concreet mogelijke situatie te schetsen voor de leerlingen. Is dit nodig? Het antwoord op deze vraag moet gegeven worden door degenen, die de test samenstellen. Het is mogelijk hier verschillende zienswijzen over te hebben. Sommigen menen, dat de „echte" rekenvraag wordt versluierd door onnodige franje. Anderen hebben de indruk, dat het juist van belang is de leerling concrete problemen voor te leggen en dat „abstractie" uit den boze is. Het is moeilijk één oplossing te geven voor dit probleem van de inleiding in de stam van het item. Van groot belang is evenwel, dat het probleem niet over het hoofd wordt gezien en dat de efficiëntie-vraag

– eventueel telkens opnieuw – wordt gesteld.

3.4 De eis van objectiviteit

School- en studietoetsen worden ook wel „objecieve studietoetsen" genoemd. Het woord „objectief" geeft een enkele maal aanleiding tot misverstand, omdat de tegenstelling ervan – subjectief – wordt opgevat als „bevooroordeeld" of „unfair". Op zo'n manier zou dus elke niet-objectieve methode een ongunstige kwalificatie meekrijgen. Dat is niet in overeenstemming met de toetstechnische opvatting van de term. Objectief betekent niet anders dan dat – vanaf een bepaald moment –het door de leerling gemaakte werk volgens een uniforme, „automatische" procedure kan worden beoordeeld. Bij de meerkeuzevragen, die in de Schooltoets worden gebruikt, begint deze objectieve beoordelingsprocedure dus zodra de leerling de letter van het door hem gekozen alternatief heeft aangestreept op zijn antwoordblad. Vanaf dat moment kan een goed geïnstrueerde machine het werk overnemen als tenminste nog aan één voorwaarde is voldaan. Het goede antwoord moet ondubbelzinnig vaststaan. Het is opmerkelijk, dat hier vaak fouten voorkomen, doordat meer antwoorden goed zijn, doordat het goede antwoord er eigenlijk niet bij is of doordat de opgave om een andere reden onoplosbaar is. We geven twee voorbeelden van items waar meer dan één antwoord te verdedigen valt:

Waarom kwamen de Noormannen in ons land?

A Ze waren door stormen verdwaald.
B Ze wilden de Christenen doden.
C Ze wilden met ons vechten.
D Ze wilden veel buitmaken.

Bij deze opgave is niet voldaan aan de objectiviteitseis, want als de testconstructeur één van de vier alternatieven als het goede antwoord aanmerkt, is zijn scoring altijd aanvechtbaar. Evenzeer is in het onderstaande voorbeeld uit de Schooltoets (afhankelijk van het belangrijkheidscriterium) een alternatieve scoringswijze verdedigbaar, zeker voor bewoners van de mijnstreek.

Welke delfstof is op het ogenblik in ons land het belangrijkste?

x A aardgas
B aardolie
C steenkool
D zout

Men stelt als algemene regel doorgaans, dat een item aan de objectiviteits-eis beantwoordt, als deskundigen het over het juiste antwoord

eens zijn. Daar is nog aan toe te voegen, dat die deskundigen het „zonder discussie" eens moeten zijn. Overigens is het natuurlijk heel goed mogelijk, dat de experts discussiëren over de vraag of een bepaald item *voor de leerlingen* (die de stof beheersen) eenduidig is. Dàn is discussie uiteraard legitiem. Zodra de discussie evenwel gaat over het juiste antwoord – zonder meer – moet het item in deze vorm al verworpen zijn.

3.5 De eis van specificiteit

Meten we met rekenitems inderdaad rekenvaardigheden? Is een stilleestest een meting van kunnen lezen? Representeert de titel van de subtoets wezenlijk de opgaven die onder het opschrift volgen? Dekt de vlag de lading? Deze vragen hebben betrekking op de specifiekheid of specificiteit van de toets. Deze eis houdt in dat de opgaven niet iets ànders gaan meten dan ze pretenderen. Het gevolg zou namelijk zijn, dat de toets-uitslagen niet meer een indicatie zijn voor het al of niet beheersen van de gevraagde stof, maar dat ze met „iets anders" te maken hebben. Dat „iets anders" kan op zichzelf een verdienstelijke bekwaamheid vertegenwoordigen, maar het is *niet* specifiek voor wat we wilden meten.

Bij opgaven als in de schooltoets komt de eis van specificiteit zeker ook naar voren. Zo dient de itemschrijver zich er bij het maken van rekenvraagstukken rekenschap van te geven of het *lezen* van de opgaven niet teveel problemen stelt aan de leerling. In dat geval krijgt immers de rekentoets een „stillees"-element, dat we beter apart kunnen testen. Evenzeer vrage men zich af of ingewikkelde becijferingen nodig zijn om het inzicht van leerlingen in rekenkundige opgaven te toetsen; misschien is het beter om voor cijferen een aparte toets te maken en de getallen in de opgaven voor rekenkundig inzicht zo eenvoudig mogelijk te houden. Een andere toets waar de eis van specifiek toetsen blijkt is het stillezen. Als men bij het stillezen „slimme" vragen gaat stellen toetst men weliswaar leesvaardigheden, maar dan loopt men het risico meer iets als „(sociale) intelligentie" en minder het hebben leren lezen te meten.

Bij meerkeuze-items is de eis van specificiteit vooral daarom van belang, omdat bij deze opgaven zo gemakkelijk „indicatoren" voor het goede antwoord kunnen vóórkomen. Het talent van sommige leerlingen om — zonder over de vereiste kennis te beschikken — tòch het goede antwoord te vinden, brengt dan een element van *test-sophistication* in de toetsresultaten. Door dit element van test-slimheid wordt de specificiteit van de toets geschaad.

Indicatoren kunnen op allerlei manieren ontstaan. Zo kan het juiste antwoord soms herkend worden door de zorg waarmee het is geformuleerd; deze zorgvuldigheid kan bijvoorbeeld blijken uit de moeilijke woorden, die er in voorkomen of uit de lengte van het alternatief. Bij een stilleesstukje over een schipbreuk komt de volgende opgave voor:
Wat betekent hier „de achttien koppen tellende bemanning"?

A De bemanning was koppig geworden uit boosheid.
B Er waren achttien volwassen matrozen aan boord.
x C Met de kapitein, de stuurman en alle anderen meegerekend, waren er achttien mensen aan boord.
D Toen men ging tellen, waren er nog achttien overlevenden.

Het is duidelijk dat bij dit (ook in ander opzicht trouwens niet zo goede) item het juiste antwoord door zijn lengte er uitspringt.
De leerling wordt soms naar het juiste antwoord geleid door een niet eens bewust herkende indicator. Indien bijvoorbeeld een bepaalde term uit de stam letterlijk of als bekend synoniem herhaald wordt in het goede alternatief kan de oplossing van het item „intuïtief" goed worden gemaakt. Soms ook geeft de grammaticale structuur een aanwijzing voor het juiste antwoord. De incomplete zin „Een andere naam voor medicamenten is . . ." in het onderstaande voorbeeld ligt uitsluitend goed in het gehoor bij de keuze van alternatief B. Een dergelijke indicator ontsnapt gemakkelijk aan de aandacht van de itemschrijver terwijl sommige leerlingen erdoor op de juiste keuze komen. Soms kan de indicator door logisch redeneren worden gevonden. In het tweede voorbeeld is in zoverre het goede antwoord geïndiceerd, dat men door redeneren kan vaststellen dat òf B òf D goed is.
Een andere naam voor medicamenten is . . .

A apparatuur
x B geneesmiddelen
C operatie
D verdoving

Op welke grondsoort worden in ons land suikerbieten verbouwd?

A meestal op laagveen
B niet op klei
C vooral op hoogveen
x D op kleigronden

Er zijn nog vele andere indicatoren mogelijk voor het juiste antwoord. Sommige itemschrijvers hebben de neiging het juiste alternatief een

vaste, of juist een regelmatig wisselende, positie te geven. Wanneer leerlingen dit doorzien, is aan de eis van specificiteit niet voldaan, want de uitkomsten worden nu mede door andere niet-specifieke vaardigheden van de leerling – slimheid, testervaring – bepaald. Als oplossing wordt aanbevolen de alternatieven altijd in alfabetische of, bij getallen, in een logische volgorde te plaatsen.

Bekend is verder, dat de algemene ontwikkeling van een leerling soms de toetsresultaten ten onrechte kan beïnvloeden.

In welke grote stad staat het Empire State Building?

A Hamburg
B Moskou
x C New York
D Stockholm

Dit item, bedoeld om gedetailleerde aardrijkskundige kennis te toetsen, kan door een leerling met enige bekendheid met de Engelse taal beter beantwoord worden, dan door anderen. Aan de eis van specificiteit is hier niet beantwoord.

3.6 De eis van bekende moeilijkheidsgraad

In sommige toetsen wil men uitsluitend zeer moeilijke, in andere slechts gemakkelijke vragen stellen. Soms wenst men binnen de toets de moeilijkheidsgraad af te wisselen of langzamerhand te veranderen. Het is noodzakelijk tevoren afspraken te maken over de gewenste moeilijkheid. Indien differentiatie tussen leerlingen die de stof wel en niet beheersen vooropstaat zal gestreefd worden naar een percentage goede antwoorden per item van iets minder dan 70 %. Dit percentage pleegt men – in tegenstelling tot wat logischerwijze verwacht mocht worden – de *moeilijkheids*graad te noemen. Indien uitsluitend gemeten moet worden of basis-doeleinden bereikt zijn, waarvan bekend is dat ze door bijna alle scholen en leerlingen inderdaad gerealiseerd worden, dan wordt een hoger antwoordpercentage, 80 % of 90 % nagestreefd. Heeft men te maken met leerlingen, waarvan men weet dat ze in de gegeven toets-situatie niet erg op hun gemak zullen zijn, dan kan men de eerste of meerdere subtoetsen met gemakkelijke items aanvangen. Andersom zijn er ook argumenten om – bijvoorbeeld wegens de vermoeidheidsfactor – de lastigste vragen aan het begin van de toetsdag te stellen.

Welk argument het zwaarst moet wegen, dient steeds opnieuw beslist te worden. Hoofdzaak is hier – en dat geldt voor al de gegeven technische eisen – dat men de problematiek onderkent en afweegt in hoeverre men aan de gestelde eisen redelijkerwijze kan beantwoorden.

Door M. Derksen-Mögelin is onlangs aangetoond dat het tevoren schatten van de moeilijkheidsgraad een te moeilijke opgave is. Wèl bleek uitvoerbaar een rangordening van items van moeilijk naar makkelijk. Wil men nochtans om bepaalde redenen exactere gegevens over de moeilijkheid dan dient men een proefafname (zie 4.8) te organiseren.

4. BEKNOPTE HANDLEIDING VOOR HET SCHRIJVEN VAN SCHOOLTOETSITEMS

Hiervóór zijn nu behandeld de eisen, waaraan eenmaal geschreven items moeten beantwoorden. Men kan zich afvragen of er ook manieren aan te geven zijn, waarop een team van itemschrijvers het beste komen kan tot het gewenste eindproduct: een verzameling items of een toets, die aan de genoemde eisen voldoet? Er zijn uiteraard verschillende werkwijzen mogelijk, maar het is wellicht nuttig hier de richtlijnen te verstrekken, die tot nu toe bij de toetsconstructie in het R.I.T.P. ongeveer aldus hebben gegolden. Daarbij zijn de volgende (acht) stappen te onderscheiden:

4.1 Bespreking over wat - hoe - hoeveel - wanneer

Eigenlijk spreekt het vanzelf, zeker gezien de hiervoor besproken relevantie-eis, dat het schrijven van de opgaven moet worden voorafgegaan door een bespreking over de te verrichten (schrijf-) werkzaamheden. In een dergelijke bespreking worden voor (soms ook door) de itemschrijvers richtlijnen voor de constructie vastgesteld.

Allereerst zal aangegeven worden over welke leerstof de toets of de verschillende subtoetsen zullen gaan en wat daarbij wèl en niet mag worden gevraagd. * Tevens komt aan de orde in welke vorm de opgaven worden aangeboden en hoeveel opgaven en subtoetsen uiteindelijk moeten overblijven. Gezien het feit dat van de allereerste producten (verzameling items), blijkens de ervaring, dikwijls de helft ongeschikt is, is het nodig tevoren vast te stellen hoeveel méér items in de beginvoorraad moeten zitten. Verder moet ook vastgesteld worden wanneer de toets in de klas moet zijn.

Het tijdstip waarop de toets moet worden afgenomen, is om minstens twee redenen van belang. Ten eerste omdat dit tijdstip aangeeft welke stof op school behandeld is en het derhalve de inhoud van de toets meebepaalt. Ten tweede omdat dit tijdstip de „deadline" aangeeft waar het tijdschema voor de hele constructiefaze naar toe moet werken. In het algemeen wordt de duur van de constructieperiode sterk onderschat. Een jaar vóór de toetsdag moet men beginnen.

In de eerste bespreking kan tenslotte ook van gedachten worden ge-

wisseld over de te volgen schrijfregels, zoals deze hieronder zijn aangegeven.

4.2 Het schrijven van de items

Aan de hand van de vastgestelde richtlijnen gaan itemschrijvers elk afzonderlijk hun „huiswerk" maken: het schrijven van de items. De werkwijze bij het schrijven van items is uiteraard sterk individueel. Aangeraden wordt wèl om elk item te beginnen met een „item-idee" dat de vraagstelling, of althans het soort voor te leggen probleem, vastlegt. Ht is van belang dit idee vast te houden, hetzij in het hoofd, hetzij op papier, tot na het schrijven, want de ervaring leert dat, wanneer het item eenmaal gestalte heeft gekregen, de oorspronkelijke vraagstelling soms verloren is gegaan en eigenlijk een nieuwe vraag is gecreëerd. Als men het item-idee vastgelegd heeft kan men tenminste – zonder de nieuwe ontstane vraag geheel en al te verwerpen – op zeker moment constateren dat het huiswerk op dit punt nog niet klaar is.

De wijze waarop men aan ideeën komt is niet voor te schrijven. Sommige itemschrijvers bestuderen leerboeken, anderen putten uit schriftelijk werk van hun leerlingen en weer anderen gaan „zonder meer" zitten denken tot de ideeën komen. De werkwijze bij het zoeken van ideeën is dus niet gemakkelijk te uniformeren. In zekere zin geldt dit ook voor het uitwerken van de ideeën. Toch is op het terrein van het uitwerken wel eens een aantal regels aangeraden. Aan het reeds enkele malen genoemde hoofdstuk in het boek „Studietoetsen" ontlenen we de volgende regels:

a. Schrijf elk item-idee op ook al weet men niet hoe tot een vierkeuzevorm te komen.
b. Probeer niet gewone, open, proefwerkvragen om te vormen tot meerkeuze-items.
c. Maak gebruik van reeds eerder afgenomen items. Zeker bij Schooltoets-items is het mogelijk goede opgaven in enigszins gewijzigde vorm te herhalen of minder goede items te verbeteren. Het is noodzakelijk de herkomst te vermelden – althans in de itemschrijfgroep.
d. Stuur concept-items in volgens de afgesproken lay out: Bijvoorbeeld links de stam, rechts de alternatieven onder elkaar in alfabetische volgorde en met de hoofdletters A, B, C en D.
e. Vermijd de fouten, die hiervóór besproken zijn:
   - te lange inleidingen, die meer dan de noodzakelijke informatie geven.
   - neem geen stereotype zinswendingen, rijtjes of cliché's over uit de leerboeken.

- het item moet over de gevraagde stof gaan en niet voornamelijk de intelligentie meten
- zorg dat er slechts één alternatief is, dat werkelijk goed is
- vermijd indicatoren
- formuleer duidelijk en kort, en grammaticaal en logisch juist
- maak er geen strikvraag van
- houd rekening met de afgesproken moeilijkheidsgraad.

4.3 Vermenigvuldigen en distribueren

Alle itemschrijvers van één team – meestal is er voor elk schoolvak een afzonderlijk team – zenden hun items naar een centraal punt. Daar worden de items (na, indien nodig leesbaar gemaakt te zijn, dus overgetikt) vermenigvuldigd en vervolgens ter beschikking gesteld aan alle leden van de betreffende schrijfgroep.

Zoals in alle constructiestappen is het ook hier noodzakelijk het tijdschema strikt te volgen. De leden van de schrijfgroepen kunnen slechts dàn het ontvangen werk amenderen of verwerpen als ze daar ruimschoots de tijd voor krijgen.

4.4 Bestudering en verbetering van andermans items

Het laten verbeteren of aanvullen van items door anderen is efficiënter dan zèlf streven naar perfecte items. Anderzijds is het amenderen van items van collega's een leerzame bezigheid, die op zichzelf weer tot nieuwe ideeën kan leiden. Bij de inzending van de items dient er al rekening mee te worden gehouden, dat de collega-lezer de gedachtengang van de itemschrijver moet kunnen volgen. Bij rekenopgaven bijvoorbeeld, zou men de veronderstelde verkeerde oplossingen van de leerlingen erbij kunnen geven * zodat de lezer niet noodzakelijkerwijs veel tijd in het narekenen gaat steken, maar zich op die veronderstelde oplossingen zelf kan concentreren. Bijvoorbeeld:

Jan en Rob hadden samen 150 knikkers. Jan verloor het derde deel van zijn knikkers en Rob won er 30. Toen hadden ze samen weer 150 knikkers. Hoeveel knikkers had Jan eerst?

a 30 knikkers (1/3 = 30, niet verder doorgedacht)
B 40 knikkers

$$\frac{150-30}{3}$$

C 50 knikkers (1/3 van 150)
x D 90 knikkers

Men kan ook de teamleden het probleem van een moeilijk realiseerbaar item-idee, al dan niet met een paar alternatieven uitgewerkt, voorleg-

gen. Anderen zien vaak oplossingen, waar de schrijver zelf niet op kan komen. Een voorbeeld van zo'n „niet compleet te krijgen" item:
Waardoor ontstaat een zonsverduistering?

x A De maan schuift voor de zon.
B Er hangen wolken voor de zon.
C ?
D?

Bij het verbeteren van andermans items worden uiteraard de reeds genoemde maatstaven aangelegd, waaraan een goed item dient te voldoen. Het is mogelijk dat een vrij goed item toch een lezer op een zijns inziens beter idee brengt. Zo'n idee moet dan onmiddellijk uitgewerkt worden en als alternatief worden genoteerd. Alle verbeteringen en aanvullingen moeten op een gezamenlijke bespreking weer worden doorgenomen. Het verdient daarom aanbeveling zulke amendementen in leesbare vorm in te dienen vóór de bespreking, zodat ze eventueel ter plaatse nog vermenigvuldigd kunnen worden.

Als algemene regel voor het bestuderen en verbeteren geldt: verwerpen òf een alternatief voorstel doen. Indien men het eens is met het item-idee, maar de uitwerking niet fraai vindt, moet men zelf andere voorstellen doen òf men moet adviseren het item te verwerpen. Critiek in de geest van „Dit is verwarrend" of „Is dit wel goed Nederlands?", kan op zichzelf terecht zijn, maar draagt in deze vorm niet bij tot een vruchtbare gedachtenwisseling. Als men zelf geen betere mogelijkheden kan vinden, dient men de critische opmerkingen te laten volgen (of kortweg te vervangen) door een verwerp-advies.

4.5 Bespreking over de items

Het allereerste doel van deze itembespreking is te komen tot een voorraad van items, waarover alle aanwezige teamleden een gunstig oordeel hebben uitgesproken. Daarnaast kan als tweede doelstelling nog gelden het doen van een uiteindelijke – evenwichtige – keuze over de te gebruiken items in de samen te stellen test. Ten derde kan de bespreking resulteren in het formuleren van noodzakelijk geachte verdere werkzaamheden, zoals het huiswerk aan voorlopig verworpen items, de aanvulling van de voorraad en het nogmaals bestuderen van andermans items.

Tijdens de itembespreking – of besprekingen als er veel items zijn ingediend – moet door de voorzitter een tamelijk „straf" beleid worden gevoerd, omdat een itembespreking gemakkelijk kan terugvallen naar een doeleindenbespreking – die al eerder gevoerd moet zijn.

Meestal worden de complete items stuk voor stuk aan de orde gesteld, waarbij ten eerste vastgesteld moet worden of één alternatief ondubbelzinnig het juiste antwoord bevat. Indien dit zo is, kan men verder gaan met dit item; indien dit niet zo is, moet het item verworpen worden. Als een item verworpen moet worden, kan het zonodig direct als „huiswerk" ter amendering worden toegewezen aan één of meer teamleden. meen zeldzaam en er is geen reden om aan te nemen, dat itemschrijvers op dit punt uitblinkers zijn. Eerder integendeel. Omdat nochtans een eindbeslissing moet worden getroffen over de uiteindelijke vormgeving van de schooltoets, is een redactie-commissie het aangewezen middel.

Het werk van de eindredacteuren kan zeer eenvoudig gestructureerd zijn: „ja" of „nee" zeggen tegen het eindproduct. Hun werk kan echter ook gecompliceerder zijn en bijvoorbeeld omvatten: het doen van voorstellen over de volgorde van de subtoetsen of van de items in een subtoets, het afwijzen van sommige items of het vragen van een andere vormgeving. De eindredacteuren kunnen ook tevoren vragen om een ruimere keus aan items of een reservevoorraad aan items, zodat onmiddellijk beslissingen kunnen worden genomen over de definitieve vorm.

Welke taak de redacteuren ook hebben, altijd is een duidelijke omschrijving daarvan noodzaak. Een andere noodzakelijke voorwaarde is: tijd. Redacteuren, die de semi-definitieve vorm van een Schooltoets in mogen zien, maar niet meer dan een week de tijd hebben om die – niet anders dan telefonisch – met andere redacteuren te bespreken, verrichten geen doorwrocht werk. Een maand is wel nodig.

4.8 Proefafname

Het vooraf beproeven van een Schooltoets of van een parallelvorm van een Schooltoets is noodzakelijk om een aantal in de handboeken beschreven redenen. *

De proefafname kan weinig of veel pretenderen en naar verhouding arbeidsuren vragen van de test-specialisten, wier taak het is de afname te organiseren en de analyse uit te voeren. In het eenvoudigste geval verzamelt men in een klein aantal klassen wat gegevens van de onderwijzers of proefleiders over „hoe't ging". In het meest uitgewerkte geval, begint men met een landelijke steekproef van leerlingen te trekken en eindigt men met een „normering vooraf", waardoor de uislagen van de Schooltoets snel – op de dag van afname (of eigenlijk van scoring) in gestandaardiseerde scores of in percentielscores kunnen worden omgezet.

Indien men tijd genoeg heeft, kan men in de besprekingsgroepen aan de uitkomsten van een proefafname de nodige aandacht besteden. Op

grond van bijv. de antwoordpercentages van de vier alternatieven bij één item, kunnen items verbeterd worden. Dit behoort evenwel tot de soort rekenkundige bewerkingen die wij hiervóór hebben aangekondigd niet te zullen bespreken.

Literatuur.

DERKSEN-MÖGELIN, M. E. *Beoordelingen van l.o.-toets items.* Doctoraal werkstuk. Amsterdam: Psychol. Lab. der Universiteit, 1968.

EBEL, R. L. *Measuring Educational Achievement.* New Jersey: Prentice Hall, 1965.

GROOT, A. D. DE. Het nut van een Schooltoets in de zesde klas l.o. In: *„Amsterdamse Schooltoetsen"*, Samenwerkende Instituten, Groningen: Wolters-Noordhoff, 1967.

GROOT, A. D. DE en NAERSSEN, R. F. VAN. (red.) *Studietoetsen,* Groningen: Wolters-Noordhoff, nog niet verschenen.

LINDQUST, E. F. (ed.) *Educational Measurement,* Washington: American Council on Education, 1951.

Voorbeelden uit drie jaren Schooltoets, Amsterdam: Nutsseminarium voor Pedagogiek, 1968.

Nutsseminarium voor Pedagogiek te Amsterdam. *Proeve van een Leerplan voor het Basisonderwijs.* Groningen: Wolters-Noordhoff, 1967.

* De Amsterdamse Schooltoets wordt elk jaar opnieuw samengesteld door het Nutsseminarium voor Pedagogiek en het Research Instituut voor de Toegepaste Psychologie (waar de schrijver werkzaam is) beide aan de Universiteit van Amsterdam. De toets bestaat uit opgaven voor Taal, Rekenen en „Algemene Kennis".

* Het voorbeeld is gekozen uit het hoofdstuk van G. J. Mellenbergh en mej. W. Lans in een binnenkort te publiceren boek *„Studietoetsen"*, Constructie, afname en analyse, dat onder redactie staat van Prof. Dr. A. D. de Groot en Dr. R. F. van Naerssen.

* Het juiste alternatief wordt in de voorbeelden met een kruisje aangegeven.

* Bij de berekening van de invloed van die zwaarte moet onder meer ook nog verdisconteerd worden de moeilijkheid van de opgaven.

* We nemen nu maar even aan dat aan het gesprek over de leerstof een bezinning over de doelstellingen van het onderwijs vooraf is gegaan. Die bezinning zal overigens — zie ook A. D. de Groot 1966 — des te vruchtbaarder zijn naarmate ze meer concreter opgavenmateriaal ter illustratie van beschouwingen gebruiken kan!

* Deze goede gewoonte namen wij over van mejuffrouw G. Boomsma van het Nutsseminarium voor Pedagogiek.

* Zie onder meer Lindquist p. 250 e.v.

Curriculum Vitae

E. Warries, geboren in 1926, studeerde van 1957 tot 1965 psychologie aan de Universiteit van Amsterdam. Hij promoveerde in 1968 bij professor dr. A. D. de Groot op een proefschrift over de effecten van vormingscursussen.

Als medewerker van het Research Instituut voor de Toegepaste Psychologie heeft hij zich de laatste drie jaren vooral bezig gehouden met de ontwikkeling van objectieve studietoetsen.