

U7CB
A 28

december 1969

MEER KEUZE TOETSEN

R.J. van Alphen de Veer
drs. P. van Weeren

Alle rechten voorbehouden.

Niets uit deze uitgave mag worden vermenigvuldigd en/of openbaar gemaakt door middel van druk, fotokopie, microfilm of op welke andere wijze dan ook, zonder voorafgaande toestemming van TNO.

Indien dit rapport in opdracht werd uitgebracht, wordt voor de rechten en verplichtingen van opdrachtgever en opdrachtnemer verwezen naar de Algemene Voorwaarden voor opdrachten aan TNO, dan wel de betreffende terzake tussen de partijen gesloten overeenkomst. Het ter inzage geven van het TNO-rapport aan direct belanghebbenden is toegestaan.

© TNO

NEDERLANDS INSTITUUT VOOR PRAEVENTIEVE GENEESKUNDE TNO

I N H O U D

blz.

1	INLEIDING	3
	1.1 Toetsen	3
	1.2 Geprecodeerde toetsen	4
	1.3 Meer keuze toets	5
	1.4 Voor- en nadelen van de meer keuze toets	5
2	VORMEN VAN MEER KEUZE-ITEMS	7
	2.1 Met betrekking tot de stam	7
	2.2 Met betrekking tot de alternatieven	8
	2.3 Met betrekking tot het item als geheel	9
3	AAN WELKE EISEN MOET EEN GOED M.K.-ITEM VOLDOEN?	11
	3.1 Relevantie	11
	3.2 Objektiviteit	11
	3.3 Specificiteit	11
	3.4 Efficiëntie	12
	3.5 Optimale moeilijkheidsgraad en differentiatie- mogelijkheden	14
4	DE KONSTRUKTIE VAN MEER KEUZE TOETSEN	15
	4.1 De leerstof omzetten in meer keuze items	15
	4.2 Lay-out	16
5	HET AFNEMEN	17
	5.1 Werktijd	17
	5.2 Instructie	17
	5.3 Raden	17
	5.4 Afkijken	17
6	NAKIJKEN EN CIJFERS BEPALEN	18
	6.1 Nakijken	18
	6.2 Kern-itemmethode van A.D. de Groot ter bepaling van de grens voldoende/onvoldoende	18
	6.3 Het omzetten van "ruwe scores" in cijfers	22
7	VOORBEELD	23
8	LITERATUUR	26

1 INLEIDING

1.1 Toetsen

Naast het onderwijzen zijn de meeste docenten ook belast met het controleren in hoeverre bepaalde stof "erin zit". De motieven voor deze controle kunnen verschillend zijn:

- controle op "het bestuderen" (studeert een leerling? studeert hij op de goede wijze?)
- controle op "het onderwijzen" (wordt de stof goed gebracht? is de stof niet te moeilijk? te gemakkelijk? juist gesteld?)
- als "maatstaf" (kan een leerling over? of een examen halen? hoe is het algemene peil van de leerlingen?)

Wát men in het algemeen te weten wil komen van de studieresultaten van de leerlingen kan men ruwweg verdelen in:

- kennis (is de stof geleerd en opgenomen?)
- inzicht en toepassing (is de stof begrepen en het geleerde bruikbaar? praktisch toepasbaar?)

De vragen die gebruikt worden om het onderwijsresultaat te meten kunnen open zijn of geprecodeerd (essay-type, opstelvragen). Een open vraag verlangt dat men zelf het antwoord bedenkt en formuleert. Een geprecodeerde vraag stelt de leerling voor een keuze uit reeds gegeven antwoorden.

Uit een verzameling van ± 70 empirische onderzoeken kan A.D. de Groot (1969) het volgende konkluderen:

- 1) Bij de open vragen worden zéér grote verschillen gekonstateerd tussen de beoordeling door verschillende docenten.
- 2) De geprecodeerde vragen zijn in het algemeen betere voorspellers van "sukses op school" (beoordeeld door de docenten) dan dat open vragen dat zijn.

- 3) Goed gekonstrueerde geprecodeerde vragen zijn in staat hetzelfde te meten als open vragen behoudens enkele specifieke vaardigheden: om te weten of iemand een opstel kan schrijven is het enige wat men doen kan hem een opstel te laten schrijven.
- 4) In de meeste gevallen zal de leerling zich niet anders voorbereiden op een toets of proefwerk met open of geprecodeerde vragen. Dit is uit onderzoek gebleken.

1.2 Geprecodeerde toetsen

Er zijn verschillende soorten en als een leraar of onderwijzer besluit deze vorm van toetsen te gebruiken, dan moet een keuze worden gemaakt. Men kent o.a.:

- 1) Het "waar/onwaar" type: Is bewering X waar of onwaar?

Het nadeel van dit soort vragen is in de eerste plaats dat deze vorm niet voor elke stof geschikt is, omdat er een absoluut juist of absoluut onjuist antwoord wordt verlangd. Verder zal er bij de ene leerling de tendens zijn meer vragen met ja te beantwoorden en bij de ander meer met nee, onafhankelijk van wat de juiste oplossing is. Bovendien is de kans op een juist antwoord bij gokken vrij hoog (namelijk 1 op 2).

- 2) "Matching": Het bij elkaar zoeken van gegevens.

Voorbeeld: Geef aan bij welke hoofdstad welk land hoort:

Engeland	Berlijn
Frankrijk	Londen
Italië	Madrid
Spanje	Rome
	Parijs
	Wenen

Deze vorm is in principe gelijk aan:

- 3) De "Multiple Choice" of meer keuze vorm

Bovenstaande vraag zou dan luiden:

Rome is de hoofdstad van:

- a) Engeland
- b) Frankrijk
- c) Italië
- d) Spanje

Parijs is de hoofdstad van:

- a) Engeland
- b) Frankrijk
- c) Italië
- d) Spanje

1.3 Meer keuze toets

Meestal wordt gekozen voor de meer keuze vorm (M.K.) boven matching, omdat M.K. algemener toe te passen en de scoring eenvoudiger uit te voeren is. Aan een M.K.-item onderscheidt men de volgende onderdelen:

- a) de stam, het gedeelte dat het algemene deel van het item uitmaakt en betrekking heeft op alle alternatieven;
- b) de alternatieven, de antwoordmogelijkheden op (of bijvoorbeeld de juiste aanvulling van) de stam. Van deze alternatieven is één het juiste antwoord. De overige zijn:
- c) de afleiders.

Het aantal alternatieven zou theoretisch zo groot mogelijk moeten zijn om de kans op een goed antwoord door gokken laag te houden. Echter omwille van de overzichtelijkheid en meer nog om het feit dat het vaak moeilijk is veel passende alternatieven te vinden, besluit men meestal tot het vier-alternatieven systeem.

Voorbeeld: Springvloed kan ontstaan:

- a) alleen bij nieuwe maan
- b) alleen bij volle maan
- c) bij nieuwe of bij volle maan
- d) bij geen van beide

1.4 Voordelen en nadelen van de M.K.-toets

Voordelen van een M.K.-toets zijn (in het kort):

- a) Gelijkheid van eisen, bijvoorbeeld ten opzichte van andere klassen en andere jaren.
- b) Objektieve beoordeling. Met de term objectief wordt hier bedoeld dat na de testkonstruktie, dus bij afname, correctie en de bepaling van de scores, de procedure zo is dat geen der leerlingen ten opzichte van een ander bevoor- of benadeeld kan zijn.

- c) Exakte beoordeling. De beoordeling is vooraf vastgesteld en verloopt volgens nauwkeurige wiskundige formules.
- d) Efficiënte beoordeling. Het belangrijkste en tijdrovendste werk is bij de open vorm de beoordeling en bij de M.K.-vorm de konstruktie van de items. De konstruktie behoeft één maal te worden gedaan door een groep docenten en de items zijn meerdere malen bruikbaar.
- e) Representatieve beoordeling. Over een groter gedeelte van de stof kan worden gevraagd door het groter aantal vragen.
- f) De mogelijkheid van systematische analyse van de resultaten.

Wat men mist bij M.K.-toetsen is de dialoog tussen docent en leerling en de toetsing van o.a. creativiteit, en persoonlijke stellingname. De vraag is echter in hoeverre het belangrijk is deze factoren bij de beoordeling een rol te laten spelen. Natuurlijk zal men gebieden houden die niet voor een M.K.-toetsing in aanmerking komen, zoals bijvoorbeeld het reeds genoemde opstel schrijven. De mogelijkheid echter blijft dat ook de vaardigheid in het opstel schrijven, omdat deze samenhangt met andere vaardigheden, toch met M.K. vragen meetbaar zal blijken te zijn.

2 VORMEN VAN M.K.-ITEMS

2.1 Met betrekking tot de stam

- a) Eenvoudige structuur, waarbij men in het oog houdt dat de vraag wel duidelijk is en éénvoudig.

Voorbeeld: Montevideo is de hoofdstad van

- A. Argentinië
- B. Bolivia
- C. Paraguay
- D. Uruguay

- b) Komplexe structuren. Het lezen en begrijpen van de stam zal in dit geval veel meer tijd kosten. Dit kan een nadeel zijn, omdat men bij voorkeur veel items per test heeft om een zo groot mogelijk deel van de stof te bestrijken; terwijl men aan de andere kant de test niet te lang kan laten duren. Een complexe stam mag niet tekort doen aan de duidelijkheid en de overzichtelijkheid van het item.

Voorbeeld: Een onderzoeker beweert in een artikel, dat de uitkomsten van een door hem gehouden sociaal experiment met een aantal volwassen proefpersonen geldig zullen zijn voor het gedrag van alle volwassenen in een soortgelijke sociale situatie. De door hem gemaakte denkstap is een voorbeeld van

- A. deductie
- B. inductie
- C. interpretatie
- D. operationalisatie

- c) Positief gesteld: Welk alternatief is juist? De meest gangbare vorm van vraagstellen.

Zie voorbeeld bij a) en b)

- d) Negatief gesteld: Welk alternatief is onjuist? Bij items met een dergelijke stam herhaalt men meestal na de alternatieven het feit dat de stam negatief gesteld was en onderstreept men (of drukt kursief) bovendien bijvoorbeeld het woordje "niet" of "onjuist". Dit om te voorkomen dat het over het hoofd gezien wordt.

Voorbeeld: Welk van de volgende vier woorden is niet juist gespeld?

- A. bureau
- B. directeur
- C. onmiddelijk
- D. weiland

N.B. Gevraagd werd welk woord niet goed gespeld is.

- e) De vragende vorm: Bij deze vorm moet men eerst alle alternatieven lezen, voordat het juiste antwoord gegeven kan worden.

Zie voorbeeld bij d)

- f) De stellende vorm: De alternatieven zijn het logisch vervolg van de stam. Na het lezen van de stam kan men het antwoord reeds bedenken zonder de alternatieven eerst gelezen te hebben.

Zie voorbeeld bij a)

2.2 Met betrekking tot de alternatieven

- a) Een absoluut juist (of onjuist) antwoord. Bij deze items kan men eventueel gebruik maken van de uitspraken "geen van deze drie" of "alle drie" als alternatief.

Voorbeeld: Wat is de juiste vertaling van "development"?

- A. enveloppe
- B. ontering
- C. ontwikkeling
- D. geen van deze drie

- b) Geen van de alternatieven is absoluut juist of onjuist. Hier is één alternatief (dat door deskundigen algemeen als het beste antwoord aanvaard zou worden). Bijvoorbeeld indien men naar een definitie vraagt zal het moeilijk zijn in korte bewoordingen een absoluut juist antwoord te vinden. In bovengenoemde vorm is dat dat niet vereist.

Voorbeeld: Welke van de volgende uitspraken komt het best overeen met Jefferson's begrip van democratie?

- A. Democratie is de staatsvorm zoals de bijbel die voorschrijft
- B. Democratie vereist een krachtig nationaal bewind
- C. Het doel van regeren is de welvaart van de mensen te verhogen
- D. Het doel van regeren is de mensen te beschermen tegen radikale of opstandige minderheden

c) De alternatieven vormen een schaal: Een reeks alternatieven die een kwalitatieve of kwantitatieve schaal vormen. Hierbij dient men om verwarring te voorkomen de alternatieven in een volgorde te plaatsen (bijvoorbeeld van hoog naar laag, van sterk naar zwak, etc.). Vaak zal het niet gaan om een uiterst nauwkeurig antwoord, maar meer om de kennis omtrent een algemeen niveau van bijvoorbeeld een bepaalde grootheid. In andere gevallen rangschikt men de alternatieven volgens alfabetische volgorde.

Voorbeeld: In Nederland lag het gemiddelde aantal kinderen per ouderpaar in 1965 het dichtst bij

- A. 1
- B. 2
- C. 3
- D. 4

2.3 Met betrekking tot het item als geheel

Behalve het gangbare item met één vraag, één antwoord en drie afleiders ziet men ook: items waarin meerdere vragen gekombineerd worden. Het is vaak erg moeilijk om vier passende alternatieven te vinden. Een oplossing kan dan zijn om bijvoorbeeld twee vragen en twee alternatieven te combineren tot één item met één stam (waarin twee vragen) en vier alternatieven; of bijvoorbeeld een combinatie van een vraag en uitleg (wat deed ? en waarom?)

Voorbeeld: Wanneer men waterstof en zwavel een verbinding laat aangaan door zwavelijzer met water te overgieten, dan ontstaat er

- A. geen warmte, maar er treedt wel verbranding op
- B. geen warmte, omdat de zwavel niet verbrandt
- C. warmte, maar er treedt geen verbranding van de zwavel op
- D. warmte, omdat de zwavel verbrandt

Voorbeeld: Wordt de gemiddelde grootte van de boerderijen in de Verenigde Staten de laatste jaren groter? Waarom?

- A. Ja, omdat de grond zijn natuurlijke vruchtbaarheid verliest, moet er meer land ontgonnen worden
- B. Ja, omdat het gebruik van landbouwmachines grotere bedrijven meer rendabel maakt dan kleinere
- C. Nee, want de moeilijkheden om werkkrachten te vinden dwingen veel boeren hun bedrijf in te krimpen
- D. Nee, want grote familiebedrijven worden steeds bij erfenissen in kleinere gedeeltes

3 AAN WELKE EISEN MOET EEN GOED M.K.-ITEM VOLDOEN?

3.1 Relevantie

Het item moet relevant zijn ten opzichte van wat het meten wil.

- geen te lange, ingewikkelde verhalen vooraf; hierbij meet men eerder stillezen dan wat men wil meten.
- geen strikvraag alternatieven, waarbij de leerling door één woordje over het hoofd te zien de zin totaal anders kan begrijpen.
- liefst geen item waarbij men moet weten wat letterlijk in de tekst wordt vermeld.

Het is dus van belang dat men zich steeds afvraagt: wat meet dit item en wat wil men meten!

3.2 Objektiviteit

- alleen het juiste antwoord moet als juist te verdedigen zijn. Er mogen bijvoorbeeld geen twee afleiders ook op een bepaalde wijze als juist te verdedigen zijn.
- In de stam mag niet te weinig informatie staan. Er mag geen twijfel bestaan over wat precies bedoeld is met de vraag.

3.3 Specificiteit

De specificiteit is gehandhaafd, indien slechts die leerlingen het item kunnen oplossen die de stof bestudeerd hebben. Er mogen geen zogenaamde indicatoren in het item aanwezig zijn, waardoor men ook zonder kennis van de stof een item kan oplossen. Men moet géén stereotype zinnen uit de stof herhalen, men meet dan de mate waarin iemand de stof uit het hoofd kent en dat is meestal niet de bedoeling.

Soorten indicatoren zijn onder andere:

- herhaling van een woord uit de stam in het juiste alternatief.
- één of meer foute alternatieven sluiten taalkundig niet aan bij de stam; bij het lezen "passen" sommige alternatieven al niet.

Voorbeeld: De verschillen die optreden tussen organismen van dezelfde soort worden genoemd

- A. adaptie
- B. erfelijkheid
- C. natuurlijke selectie
- D. variaties

- één of meer alternatieven zijn zo duidelijk als afleiders gemaakt, dat ze door de leerling direkt als fout gezien worden, zonder kennis van de stof.
- het juiste is te herkennen doordat het véél langer of korter is dan de andere.
- het feit dat de juiste alternatieven systematisch van plaats verwisselen. Dit voorkomt men door de alternatieven in alfabetische volgorde te zetten. Bijvoorbeeld: komt nu het juiste antwoord van drie op elkaar volgende items steeds op de tweede plaats, dan is het toeval en niet systematisch.
- "nooit", "altijd", "meestal" en "zeker" zullen weinig voorkomen in het juiste alternatief, dit kan dus een indikator zijn.
- de afleiders overlappen elkaar of sluiten elkaar in (als katoen niet van een dier komt, komt het ook niet van een schaap).

Voorbeeld: Katoen komt van een

- A. dier
- B. mineraal
- C. plant
- D. schaap

Zo'n item heeft dus eigenlijk maar twee alternatieven, want er is altijd slechts één juist antwoord, dus A en D zijn fout.

3.4 Efficiëntie

Efficiënt is een item wanneer het relevante informatie verschaft en in korte tijd gelezen en begrepen wordt. Een item moet dus liefst kort, duidelijk, eenvoudig gesteld en grammatikaal juist zijn.

- Vaak kan een item duidelijker worden door bijvoorbeeld in de alternatieven niet steeds dezelfde woorden te herhalen, maar deze bij de stam te trekken.

Voorbeeld: Wat is de beste definitie van een ader?

- A. een bloedvat waardoor bloed met veel afvalstoffen stroomt
- B. een bloedvat waardoor bloed naar het hart toe stroomt
- C. een bloedvat waardoor bloed van het hart af stroomt
- D. een bloedvat waardoor zuurstofarm bloed stroomt

duidelijker: Een ader is een bloedvat, waardoor

- A. bloed met veel afvalstoffen stroomt
- B. bloed naar het hart toe stroomt
- C. bloed van het hart af stroomt
- D. zuurstofarm bloed stroomt

- Een ontkenning werkt gauw verwarrend en vertragend! Het moet daarom **zeer** duidelijk worden aangegeven in een N.B. onder het item. Een dubbele ontkenning moet vermeden worden.

Voorbeeld: Het is niet juist om te beweren dat de klassieke psychoanalyse zich niet bezighoudt met:

- A. gekonditioneerde reflexen
- B. mathematische gedragsmodellen
- C. testkonstruktie
- D. theorievorming

N.B.: Gevraagd wordt wat niet juist is

beter: De klassieke psychoanalyse houdt zich onder meer bezig met:

- A. gekonditioneerde reflexen
- B. mathematische gedragsmodellen
- C. testkonstruktie
- D. theorievorming

- Een lange inleiding bij een item is alleen gerechtvaardigd, indien deze bijdraagt tot de duidelijkheid van de vraagstelling. Kan de inleiding worden gemist zonder dat de vraag onbegrijpelijk wordt: men late hem weg.

3.5 Optimale moeilijkheidsgraad en differentiatiemogelijkheden

De moeilijkheidsgraad van een item wordt aangeduid met de letter p:

$$p = \frac{\text{aantal leerlingen dat het juiste antwoord gaf}}{\text{totaal aantal leerlingen}}$$

De ideale moeilijkheidsgraad ligt afhankelijk van het doel van de toets bij $\pm .90$ of tussen $.40$ en $.70$. In het eerste geval meten we of de stof begrepen en overgekomen is, in het tweede geval willen we differentiëren tussen betere en slechtere leerlingen. Zelfs een geroutineerd "items-maker" heeft deze moeilijkheidsgraad niet in de hand. Men kan echter moeilijker en gemakkelijker items maken door:

- te hoeden voor indicatoren, maar te manipuleren met de afleiders. Hoe dichter deze bij het juiste alternatief liggen, des te moeilijker is het item.

4 DE KONSTRUKTIE VAN M.K.-TOETSEN

4.1 De leerstof omzetten in M.K.-items

Als zeer belangrijk moet gezien worden het kontakt met kollega's. Op de volgende punten is kontakt zeer gewenst:

- 1) Bij het bepalen van wat belangrijk is en getoetst moet worden en wat niet zo belangrijk is en niet getoetst behoeft te worden.
- 2) Bij het bepalen in welke verhouding het aantal items over de onderdelen van de stof verdeeld moet zijn.
- 3) Bij het vaststellen: welk soort vragen stelt men over welke onderdelen (kennis? begrip? inzicht?)
- 4) Bij de uiteindelijke beslissing over:
 - a. of het item voldoet;
 - b. welke items gebruikt worden voor een bepaald proefwerk. De overige goedgekeurde blijven bewaard om bij een volgende gelegenheid dienst te doen.

In het kort nu een procedure die men kan volgen om van de stof te komen tot een serie M.K.-items waaruit men een toets kan samenstellen:

- 1 Een voorlopige lijst van onderwerpen, begrippen, namen, principes, e.d. die de inhoud van de leerstof zo volledig mogelijk beschrijft.
- 2 Een definitieve lijst van onderdelen die aangeeft welke stof in aanmerking komt voor het maken van items.
- 3 Achter de in lijst 2 vermelde onderdelen aantekeningen omtrent soort items (kennis, begrip of toepassingsvragen) en hoeveel per onderdeel
- 4 Aan de hand van lijst 2 concept-items maken. Men begint met de stam en - tenminste - het juiste antwoord. Ziet men direkt reeds één of meer geschikte afleiders, dan schrijft men die erbij; anders komt dat later. Tracht niet eerst een essay vraag te formuleren om dan tot de M.K.-vorm te komen, dat werkt onnatuurlijk verwrongen items in de hand.

- 5 Bedenken van goede afleiders. Ga na:
- a) Wat is een voor de hand liggende fout?
 - b) Wat is een gangbare fout? (vroegere proefwerken, mondelinge beurten, vragen in de les)

Lukt dit niet, tracht dan een vervangend item te bedenken, maar laat het onafgemaakte item niet verloren gaan.

Misschien weten kollega's goede alternatieven of komt men er zelf later op.

Kontroleer de items nu ook:

- voldoet het aan de eisen? (zie hoofdstuk 3)
- is het afgestemd op het doel?
- niet te moeilijk?
- is het te verbeteren?

Twijfelt men, verzin dan een alternatief-item; hoe meer keus, hoe beter straks het gehalte van de toets.

- 6 Is het item te vermenigvuldigen, dan variatie en analoge vraagstelling, zo mogelijk met vaste alternatieven; dan voegt u deze nieuwe items bij de lijst voor de groepsbespreking.

4.2 Lay-out

Belangrijk is dat de regels niet te lang en niet te kort zijn en dat de items duidelijk van elkaar te onderscheiden zijn doordat ze niet te dicht op elkaar staan. Ieder alternatief dient op een nieuwe regel te beginnen.

5 HET AFNEMEN

5.1 Werktijd

De leerlingen moeten zoveel tijd krijgen, dat de meesten binnen de tijd klaar komen. Als, nadat de tijd verstreken is, blijkt dat bijvoorbeeld nog slechts 10% klaar is met de toets, dan is deze te lang of te moeilijk geweest. Deze toets is nog niet waardeloos, maar niet alle onderwerpen zijn nu getoetst. Dat moet vermeden worden. Bovendien kunnen vooral jeugdige leerlingen hierdoor worden ontmoedigd en de toets minder serieus gaan nemen.

5.2 Instructie

De instructie moet duidelijk zijn. Zelfs de zwakste leerling uit de klas moet begrijpen wat er van hem wordt verwacht, hoewel hij de toets zelf misschien helemaal niet kan maken. Is de situatie onduidelijk, dan krijgen intelligente leerlingen een voorsprong.

5.3 Raden

Maakt een fout antwoord of "niet ingevuld" geen verschil, dan kunnen de leerlingen raden. Men dient zich er echter van te overtuigen dat alle leerlingen dan weten dat ze beter kunnen raden dan overslaan! Indien alle leerlingen door dezelfde instructie de items die ze niet weten raden, en het grootste deel komt klaar, dan is een korrektie voor het raden niet nodig, omdat iedereen hetzelfde voordeel van het raden had.

5.4 Afkijken

Het afkijken is bij M.K.-proefwerken wat gemakkelijker dan bij normale proefwerken. Men kan de leerling nauwelijks verantwoordelijk achten voor het feit dat de verleiding soms te groot wordt. Het afkijken kan preventief bestreden worden door:

- a) intensieve, doch de leerling niet storende surveillance.
- b) leerlingen zover uit elkaar zetten dat het onmogelijk is af te kijken, bijvoorbeeld in een aula of gymnastieklokaal.
- c) een procedure waarbij de leerling op een antwoordblad het juiste alternatief moeten invullen. Hierbij staan de even nummers links en de oneven rechts. Men vouwt nu het blad overlangs dubbel. Degenen die links zitten beginnen met de even nummers en degenen die rechts zitten met de oneven. Halverwege wisselt men.

6 NAKIJKEN EN CIJFERS BEPALEN

6.1 Nakijken

Het nakijken kan men doen door een juist en zeer duidelijk ingevulde test te vergelijken met een door een leerling ingevulde test. Ook kan men van schablonen gebruikmaken (dit is een blad waarin gaten zitten op de plaatsen waar de juiste kruisjes horen). Deze methode gaat snel, maar men kan erbij over het hoofd zien, dat er twee kruisjes zijn gezet of verbeteringen gemaakt. Men noteert het aantal juiste antwoorden op het formulier en bovendien maakt men een matrix waarmee later de p-waarden (moeilijkheidsgraad) kunnen worden bepaald:

		vraag no:								totaal score
		1	2	3	4	5	6	7	8	
leerling no:	1	1	1	1		1	1		1	6
	2		1	1		1		1	1	5
	3	1	1	1	1	1		1	1	7
	4		1	1		1	1	1		5
	5	1	1			1	1		1	5
	totaal aantal juist per vraag	3	5	4	1	5	3	3	4	

$$p = \frac{3}{5} = .60 \text{ voor vraag 1}$$

Met deze gegevens kan nu volgens de methode van A.D. de Groot de grens tussen voldoende en onvoldoende worden bepaald.

6.2 De kern-item methode van A.D. de Groot

(een uitgewerkt voorbeeld volgt in hoofdstuk 7)

Kern-items

Nadat men heeft bepaald uit welke items een toets zal bestaan, kiest men uit deze items een aantal kern-items. Een kern-item is een vraag die naar de mening van de docent juist moet kunnen worden beantwoord door die leerlingen die de stof goed hebben bestudeerd en begrepen.

De docent zal dus in grote trekken op de volgende punten letten bij de beoordeling van de items op hun geschiktheid als kern-item:

- 1) Het kern-item moet vragen naar een fundamenteel aspect van de stof. Een bepaalde hoofdzaak van de stof die men behoort te hebben begrepen of dient te worden gekend. Er wordt in kern-items niet naar detail kwesties of minder belangrijke feiten gevraagd.
- 2) Het item mikt op onderscheidend vermogen tussen hen die de stof voldoende of onvoldoende beheersen.
- 3) Een kern-item dient bij uitstek te voldoen aan de eisen voor een goed item, zoals in hoofdstuk 3 vermeld.
- 4) Het kern-item mag niet te moeilijk zijn. Niet alleen de goede leerlingen moeten het kunnen beantwoorden, maar ook zij "die een voldoende verdienen". Het mag ook niet te gemakkelijk zijn. Een leerling die de stof niet voldoende heeft bestudeerd of begrepen dient het antwoord niet te kunnen vinden op basis van oppervlakkige kennis of door handigheid in het oplossen van M.K.-toetsen.

Opmerking: Het verdient aanbeveling zich te oefenen in het schatten van de moeilijkheidsgraad door deze te voorspellen en deze voorspelling met de resultaten te vergelijken.

De p-waarde (moeilijkheidsgraad)

Praktisch verstaat men (zie hoofdstuk 3.5) onder de p-waarde van een vraag: het aantal juiste antwoorden op de vraag gedeeld door het aantal leerlingen dat de toets heeft gemaakt. Theoretisch is deze waarde echter niet geheel juist. De mogelijkheid is aanwezig dat, bij de beantwoording van een bepaalde vraag, enkele leerlingen geluk hebben (of juist pech). Deze leerlingen beantwoorden de vraag juist, hoewel ze het goede antwoord niet weten (of onjuist terwijl ze het wel weten). Er wordt uitgegaan van het principe dat de kans dat een leerling die het antwoord op de vraag wél weet, maar deze per ongeluk toch niet juist beantwoordt (dus pech heeft) zo klein mogelijk moet zijn. Dit geldt ook voor de kans op "geluk hebben". Er dient dus een korrektie voor geluk of pech te worden toegepast.

Met de volgende formule, gebaseerd op bovenstaand uitgangspunt, berekent men de gekorrigeerde p-waarde:

$$\frac{(A \times p) - 1}{A - 1}, \text{ waarin } A \text{ het aantal antwoordmogelijkheden per vraag is.}$$

Gebruikt men dus vier antwoordmogelijkheden (alternatieven), dan wordt de gekorrigeerde p-waarde:

$$\frac{(4 \times p) - 1}{3}$$

(Zie voor afleiding van de formule eventueel De Groot 1964)

De grens voldoende/onvoldoende

Om de grensscore voldoende/onvoldoende te vinden maakt de kern-itemmethode gebruik van de gemiddelden van beide p-waarden (de "normale" en de gekorrigeerde). De gemiddelde p-waarde (P1) berekent men over alle kern-items, dus men telt de p-waarden van de kern-items op en deelt dit getal door het aantal kern-items. De gemiddelde gekorrigeerde p-waarde (P2) berekent men met behulp van de gemiddelde p-waarde (P1) volgens de formule:

$$P_2 = \frac{(A \times P_1) - 1}{A - 1}, \text{ waarin } A \text{ weer het aantal alternatieven per vraag is.}$$

Voor vier alternatieven dus:

$P_2 = \frac{(4 \times P_1) - 1}{3}$

Deze beide waarden worden gebruikt als onder- en bovengrens van een interval, waarbinnen de score moet liggen die aangeeft welke leerlingen nog juist voldoende hebben.

Om dit interval nu te bepalen zet men eerst P1 en P2 om in procenten door de komma twee plaatsen naar rechts te zetten. (Dit is vermenigvuldigen met 100). Men zorgt nu dat het percentage leerlingen dat voldoende krijgt, ligt tussen de grenzen van het interval dat door deze percentages wordt afgegrensd. Dit is zeer eenvoudig af te lezen, indien men een tabel maakt als in het voorbeeld in tabel II (hoofdstuk 7) met een kumulatieve frekwentiekolom in procenten (kolom 3) (zie opmerking bij tabel II, pag. 25)

Het is mogelijk dat binnen deze grenzen twee percentages liggen of géén percentage. Men geeft de leerlingen dan "the benefit of the doubt", dat wil zeggen, men zoekt dat percentage dat het dichtst bij de ondergrens (= P2) ligt.

Basisgedachte van de kern-itemmethode

De basisgedachte is geheel in overeenstemming met de gangbare methode van toetsen die ervan uitgaat dat iedere perfecte vraag (hier kern-item) een geschikte toets is om te zien of een leerling de leerstof wel of niet voldoende beheerst. De kern-itemmethode onderscheidt zich op de volgende punten:

- 1) De definitie van wat een perfecte vraag is exakter bepaald. Natuurlijk blijft dit een kwestie ter beoordeling van de leraar, die verantwoordelijk is, maar de norm is expliciet en vooraf vastgesteld.
- 2) Om de kans op foutieve beoordelingen van leerlingen te verminderen worden meerdere perfecte vragen (kern-items) als criterium gebruikt ter bepaling van de grens voldoende/onvoldoende.
- 3) De kern-items bepalen niet zonder meer het individuele cijfer, maar spelen slechts een bijzondere rol bij de schatting van het percentage voldoende. Bij de bepaling van het cijfer tellen ook de overige items mee en spelen de kern-items geen bijzondere rol.

Waarom niet alleen maar kern-items?

Er zijn enkele redenen waarom men slechts van een toets eist dat er een aantal kern-items zijn en niet dat alle vragen kern-items zijn:

- 1) Kern-items, die werkelijk goed zijn, zijn zeer moeilijk in grote hoeveelheid te maken. Bovendien vallen veel in opzet geschikte kern-items als zodanig af, omdat ze achteraf te moeilijk of te gemakkelijk blijken te zijn.

- 2) Hoe meer vragen des te betrouwbaarder zal de toets meten. Een vergroting van de toets kan heel werkzaam ten aanzien van deze betrouwbaarheid zijn, zonder dat voor de aanvulling kern-items worden gebruikt.
- 3) Men wil méér weten dan wie boven of onder de grens voldoende/onvoldoende "ligt". Door wat moeilijkere en wat gemakkelijkere vragen in te lassen differentieert de toets, zodat ook tussen net voldoende en de maximum score verschillende cijfers gegeven kunnen worden.

6.3 Het omzetten van "ruwe scores" in cijfers

De score "aantal juiste antwoorden" noemt men: de ruwe score. Om deze scores te veranderen in normale cijfers van 1 tot en met 10 zijn verschillende methoden bekend. Bepaalde voordelen hebben methoden die op een nauwkeurige, doch ingewikkelde berekening zijn gebaseerd. Eenvoudiger methoden voldoen echter ook behoorlijk, mits zorgvuldig toegepast.

De kern-itemmethode levert slechts een punt op de cijferschaal van 1 - 10, namelijk de grens ~~5/6~~. De korresponderende ruwe score neemt men meestal als laagste ruwe score, waaraan men een "zes" toekent. Men werkt niet graag met halve punten, dit maakt de berekeningen moeilijker en het verschil wordt bovendien zeer klein. Een tweede punt kies men nu min of meer willekeurig. Hierbij let men erop dat er niet ten onrechte weinig (of juist veel) hoge cijfers gegeven worden. Dit kan gebeuren als een proefwerk erg moeilijk of erg gemakkelijk uitvalt of als de behaalde ruwe scores voor het grootste deel zeer dicht op elkaar liggen. Men kiest bijvoorbeeld de 10 voor de hoogst haalbare ruwe score (= "alle vragen goed"). De ruwe scores tussen de laagste score met 6 en de hoogst haalbare verdeelt men zodanig, dat bij elk cijfer ongeveer even veel ruwe scores behoren. Dit betekent in de praktijk dat zessen, zevens en achten het meest voorkomen.

Valt een proefwerk te moeilijk of te gemakkelijk uit, dan verplaatst men het punt waarop men "de 10" toekent. Dus men stelt bijvoorbeeld: de 10 valt bij een ruwe score lager dan de maximaal haalbare hoogste score (= "alle vragen goed"), als het proefwerk achteraf te moeilijk blijkt te zijn.

7 VOORBEELD

Proefwerk in een klas van 30 leerlingen met 25 vragen. Nadat de vragen waren nagekeken, kon het volgende overzicht in tabelvorm worden samengesteld (tabel I). Door het aantal goed per vraag te delen door het aantal leerlingen kreeg men de moeilijkheidsgraad ($= p$). Men telde de p 's van de kern-items op en deelde deze door het aantal kern-items om de gemiddelde p -waarde (P_1) te berekenen voor de bovengrens en door middel van de formule berekende men de ondergrens (P_2). Vooraf werden de kern-items 1, 2, 11, 16, 20 en 21 als kern-item geschrappt, omdat ze óf te moeilijk óf te gemakkelijk waren of omdat er om andere redenen aanmerkingen waren op het item als kern-item. De p 's van de kern-items lagen tussen .60 en .73. De volgende berekeningen werden dus uitgevoerd:

De gemiddelde moeilijkheidsgraad van de uiteindelijk aanvaarde kern-items is: $\frac{.67 + .60 + .73 + \dots + .63}{10} = .67$ (P_1). Dit is de "bovengrens" voor het percentage voldoende.

De ondergrens wordt $\frac{(4 \times .67) - 1}{4 - 1} = \frac{2.68 - 1}{3} = .56$ (P_2).

Tussen 67 en 56 procent ligt het percentage 63.33 (zie tabel II). Dit korrespondeert met een ruwe score van 16. Degenen die deze score haalden krijgen dus een 6.

Geven we nu aan hen die score 22 haalden een 10, dan wordt de verdeling van de cijfers als in kolom 4 van tabel II.

TABEL I

juiste antwoorden (een 1 = een juist antwoord)

leerling no.	item-nummer																									aantal juiste antwo per leerl
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	
1	1	1		1		1		1		1	1		1		1	1	1				1	1	1	1		16
2	1	1			1				1	1	1		1	1	1		1			1			1	1	1	14
3	1	1	1	1	1	1				1		1	1	1			1	1	1	1	1	1	1	1	1	19
4	1	1			1	1		1			1	1	1	1			1	1		1		1	1	1	1	16
5	1		1		1		1	1		1			1	1				1	1				1	1		12
6	1	1	1			1		1		1		1		1				1	1	1	1	1		1	1	14
7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1		1	1		1	1	1	1	22
8	1	1	1	1	1		1	1	1	1		1	1	1	1		1	1		1	1		1	1	1	20
9	1	1	1				1	1	1		1	1	1				1		1			1	1	1	1	15
10	1	1	1		1			1	1			1	1					1	1			1	1			12
11	1	1	1	1	1	1	1	1				1	1	1	1	1	1	1	1	1	1	1	1		1	20
12			1	1	1	1	1	1	1	1	1	1	1			1	1	1				1	1	1	1	18
13	1	1	1		1		1	1		1	1					1	1	1	1	1	1	1	1	1	1	19
14	1	1	1	1	1			1	1			1	1			1	1	1	1	1		1	1		1	17
15	1		1	1	1	1	1	1	1	1		1	1	1												12
16	1	1	1	1			1		1	1	1	1	1												1	11
17	1	1	1	1			1	1	1	1	1	1	1	1		1	1	1	1	1	1	1	1	1	1	22
18	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1		1		1		1	1	1	21
19	1	1	1	1			1	1	1					1	1		1	1	1	1	1	1	1	1	1	18
20	1	1	1	1				1	1	1	1	1	1			1		1		1		1		1	1	16
21	1	1		1	1		1	1		1	1	1	1				1	1	1		1		1	1	1	18
22		1	1				1	1	1	1		1	1		1			1	1	1	1	1	1	1	1	17
23	1	1	1	1	1	1	1	1	1	1			1	1			1	1	1			1	1	1	1	19
24	1	1	1	1				1	1			1	1	1	1	1	1	1	1	1	1	1	1	1	1	20
25	1		1					1	1	1			1	1	1				1			1	1		1	12
26	1	1		1									1				1					1		1		7
27	1	1	1	1	1	1	1	1	1				1				1	1	1	1		1		1		16
28	1	1	1	1	1	1	1	1	1	1			1	1	1		1				1	1	1			17
29	1	1	1	1	1		1	1	1	1			1	1		1	1	1		1	1		1	1	1	19
30	1	1	1				1		1		1			1		1		1		1		1	1		1	13

het aantal
leerlingen
dat een
vraag juist
beantwoordde

28 26 25 20 18 12 19 26 22 18 13 19 26 19 11 9 21 21 18 17 14 21 25 19 25

p = .93 .87 .83 .67 .60 .40 .63 .87 .73 .60 .43 .63 .87 .63 .37 .30 .70 .70 .60 .57 .47 .70 .83 .63 .83

↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑

Vooraf werden de vragen, met ↑ of ↓ aangeduid, als kern-item gekozen. Achteraf werden slechts degene met ↓ definitief als kern-item aangenomen. De overige (↑) vielen om een of andere reden uit. (Zie hoofdstuk 6-2 over kern-itemmethode).

N.B.: Het is natuurlijk ook geoorloofd de fouten te scoren als dit gezien het aantal gemakkelijker is. Als men bij de verdere berekeningen echter maar steeds van het aantal juiste antwoorden uitgaat.

TABEL II

1	2	3	4
ruwe score = aantal goede antwoorden	aantal leerlingen dat deze score behaalde	percentage leerlingen dat deze score of lager behaalde	Cijfer
22	2 $\frac{2}{30} =$	6.67	10
21	1 $\frac{3}{30} =$	10.00	9
20	3 $\frac{6}{30} =$	20.00	9
19	4 etc.	33.33	8
18	3	43.33	7
17	2	50.00	7
16	4	63.33	6
15	2	70.00	5
14	2	76.67	5
13	1	80.00	4
12	4	93.00	4
11	1	96.67	4
7	1	100.00	4
totaal aantal leerlingen dat de toets maakte:		30	

P2 = 56

P1 = 67

Opmerking: Kolom 3 is een kumulatieve frekwentiekolom in procenten, dat wil zeggen, het aantal leerlingen dat een bepaalde score behaalde of lager wordt gegeven in procenten.

8 LITERATUUR

Als literatuur kan worden aanbevolen het werk dat bij de samenstelling van de bovenstaande verhandeling over meer keuze toetsen uitvoerig werd geraadpleegd en waaraan ook voorbeelden werden ontleend:

A.D. DE GROOT & R.F. VAN NAERSSEN E.A.: Studietoetsen, konstrueren, afnemen, analyseren; Mouton en Co., Den Haag, 1969.

tevens:

ROBERT L. EBEL: Measuring educational achievement; Prentice-Hall inc., New Yersey, 1965.

en:

A.D. DE GROOT: De kern-itemmethode voor de bepaling van de caesuur voldoende/onvoldoende in Paedagogische Studiën, Wolters, Groningen, 1964, pag. 425 - 440.