24-5-76.

De kernitemmethode: een onaanvaardbare methode voor het bepalen van de zakslaag grens.

Ben Wilbrink.

De docent die voor de opgave staat om voor zijn tentamen de grens tussen voldoende en onvoldoende te trekken, wordt daarbij tot op de dag van vandaag niet of nauwelijks geholpen door goede technieken. Technieken, die wel beschikbaar zijn voor allerlei andere zaken zoals toetsconstructie, analyse van toetsresultaten, en selectieprocedures. Een aantal oorzaken voor dat ontbreken van cesuurbepalingstechnieken lijken wel aangewezen te kunnen worden, zoals:

(a)Het grootste deel van technieken voor het meten en beoordelen in het onderwijs stamt uit Amerikaanse literatuur. Omdat in het Amerikaanse onderwijs al vrij vroeg op aanzienlijke schaal overgegaan werd op studiepuntenstelsels (zie o.a. Bevers 1975), deed het zakslaag probleem zoals dat in het Nederlandse onderwijs zich op uitgebreide schaal voordoet zich in de Verenigde Staten niet in een de aandacht vragende vorm voor.

Voor bet oplossen van de cesuurbepalingsproblematiek hebben we dus in forse mate op eigen benen te staan, maar jammer genoeg wordt nog steeds erg weinig aandacht aan het probleem geschonken (De Groot 1964 a en b, Wijnen 1972, en een aantal publicaties van Van Naerssen over tentamenmodellen, Van Naerssen 1974).

(b) Het bepalen van de zakslaag grens is een beslissingsprobleem, en moet in principe langs besliskundige lijnen tot een oplossing worden gevoerd. Besliskundige technieken zijn echter tamelijk recent van ontwikkeling, en vooral toepassingen in complexe praktische situaties leveren nog forse problemen op.

De enige systematische methoden voor het vaststellen van de

grens voldoendeonvoldoende die in gebruik zijn, zijn de kernitem methode en de methode Wijnen. (En varianten daarop, zoals o.a. de kernitemmethode voorgesteld door Van Naerssen 1974). Over de methode Wijnen kunnen we hier erg kort zijn: in de huidige situatie, waarin geen behoorlijke cesuurbepalingstechnieken ontwikkeld en toepasbaar zijn, is de methode Wijnen het enige aanvaardbare alternatief. De methode erkent expliciet dat we over de onderwijs situatie erg weinig met zekerheid kunnen zeggen, en dat in afwezigheid van methoden die dat kleine beetje wél bestaande zekerheid kunnen uitbuiten in een rationele beslissingsprocedure t6ch op aanvaardbare (De Groot 1972) wijze beslissingen over personen genomen moeten worden. Een uit armoede geboren methode als het ware.

Uit bovenstaande uitspraak dat de methode Wijnen op dit moment de enige behoorlijke cesuurbepalingstechniek is, valt af te leiden dat de kernitem methode, en haar varianten, door de schrijver als niet aanvaardbaar worden beschouwd. De argumenten voor deze uitgangs stelling zullen in het volgende gegeven worden.

Het is allemaal begonnen in 1964 met een tweetal publicaties van De Groot. Om voldoende greep op de problematiek rond die kernitem. methoden te krijgen, zal begonnen worden met de belangrijkste bronnen uitgebreid te becommentariëren, evenals een enkele kritische bespreking zoals van Wijnen (1972).

A.D. De Groot: Wááraan voldoet een 'onvoldoende' prestatie niet? Paedagogische Studieën 1964, 39-44.

Het artikel bevat een aantal stellingen met summiere toelichting, die de voorbereiding vormen voor de behandeling van de kern item methode in een hierop volgend artikel (straks te bespreken).,Een belangrijke stelling waaraan 6ók de kernitem methode getoetst zou moeten worden, is zijn stelling

vijf:

"Als een leraar de door twee willekeurige leerlingen geleverde prestatie verschillend heeft beoordeeld, behoort hij dit verschil in

cijfer o£ objectieve, feitelijke gronden te kunnen rechtvaardigen; deze eis geldt met nog meer klem, als het ene cijfer een voldoende en het andere een onvoldoende is."

In de visie van De Groot heeft de leerling recht op deze uitleg; voor de docent is de moeilijkheid om een dergelijke rechtvaardiging te vinden: daarvoor heeft hij ten minste een behoorlijke techniek voor bet bepalen van de zakslaag grens nodig. Een ander probleem is hier de rechtspositie van de leerling (student): hoe wordt hij beschermd tegen onzorgvuldige praktijken rond de cijfer en cesuur bepaling?

De Groot komt ook niet verder dan het aanstippen van een aantal praktijken die in ieder geval niet toelaatbaar zijn ("voorbeelden van hoe het niet moet"):

"(a) Examenwerk wordt dikwijls na de correctie niet teruggegeven, zelfs niet ter inzage.

(b) Er is vaak geen beroep op beslissingen.

(e) Men kan soms zelfs geen toelichting op de beoordeling krijgen.

(d) Er wordt dikwijls geen mededeling vooraf gedaan over de wijze van scoren, tellen en wegen van de onderdelen, zodat de leerling onnodig het risico loopt zijn uitslag door een verkeerde proefwerk- of examenstrategie te bederven.

(e) Er is geen praktisch bruikbare vorm van verweer (beroep) tegen éénmansbeoordelingen (en vraagmethoden!), met name op mondelingen tentamina, enz., enz.

Merkwaardigerwijs verklaart De Groot hier de student (althans, de leerling) onmondig waar het gaat om de bepaling van de zak-slaaggrens: men kan niet stellen dat de leerling daarover rechtens zou moeten kunnen meepraten."

In stelling 6 wordt nog eens uitdrukkelijk gewezen op de

twee geheel verschillende doelstellingen van beoordelingen in het onderwijs: de ene het betrouwbaar meten van verschillen tussen prestatieniveaus van studenten; de andere het leggen van een grens tussen zakken en slagen, of liever misschien tussen door laten gaan of het eisen van een herhaling. Een aantal primitieve cesuurbepalingsmethoden maken dit onderscheid niet ('grading on the curve', het hanteren van een bepaald afwijzings-percentage (althans geen al te grote afwijking van 'voorgaande jaren', e.d.; terwijl meer in het algemeen waarschijnlijk alle relatieve methoden (zie Wijnen 1972) voor deze beschuldiging vatbaar zijn. Of de kernitem methode het onderscheid wél maakt? We hopen te kunnen aantonen dat met name ook de kernitem. methode een duistere verwarring van beide doelstellingen is.

In stelling zeven wordt de invloed van een sterke Nederlandse traditie aangestipt: het gebruik van een tienpunts cijferschaal gekoppeld aan de gewoonte om de grens tussen voldoende en onvoldoende te leggen tussen de vijf en de zes. Deze traditie werkt absurde cesuurbepalingen in de hand. De Groot:

"Onvoldoende' zou moeten betekenen: Bij deze leerling heeft mijn onderwijs gefaald maar deze (absolute) betekenis is haast niet te handhaven als de officiële cijferschaal suggereert dat een onvoldoende net zo'n 'gewoon' cijfer is als een voldoende en dat onvoldoendes in liefst 5 verschillende gradaties kunnen voorkomen!"

"Stelling 10.

Om tot een gezond cesuurbeleid per vak te kunnen komen, moeten de volgende voorwaarden vervuld zijn:

A. Het doel van het onderwijs in kwestie en de functie van het vak daarin moeten duidelijk omschreven. zijn.
B. Uit A moeten duidelijk omschreven minimumeisen van kennis en kunnen afgeleid zijn.
C. De in B bedoelde minimumeisen moeten operationeel gedefinieerd zijn in termen van een reeks opgaventypen, met

objectieve beoordelingsnormen voor wat acceptabel (voldoende) is en wat niet (onvoldoende).

D. De beslissing voldoende/onvoldoende moet onmiddellijke praktische consequenties hebben, zowel voor de leerling als voor de leraar."

Met punt D. bedoelt De Groot o.a. dat het hebben van een. voldoende of een onvoldoende voor een bepaald vak niet op onduidelijke wijze een rol mag gaan spelen bij overgangs of examen regelingen. Een onvoldoende betekent voor De Groot dat de studieprestaties van de leerling (student) op een niet aanvaardbaar laag niveau liggen, zodat het vak overgedaan moet worden, het liefst onder bijzondere begeleiding (het kan best aan het onderwijs gelegen hebben dat deze leerling een onvoldoende prestatie leverde). De Groot is van mening dat het hoger onderwijs op dit punt al bevredigend functioneert: een onvoldoende voor een vak betekent immers volgens hem dat het vak gewoon overgedaan moet worden. Deze rijkelijk optimistische opvatting ligt waarschijnlijk mede ten grondslag aan het (te vroeg) propageren van zijn kernitem methode als oplossing voor het cesuur probleem. Over het voldaan hebben aan eis D. kan men, voor wat het Hoger Onderwijs betreft, fors met De Groot van mening verschillen: de praktische consequenties die onvoldoendes hebben voor de docent zijn geheel afwezig. Nog maar al te vaak wordt de oorzaak van slechte prestaties (ook wanneer het gaat om relatief grote groepen studenten), bij de student gelegd, en niet bij het eigen tekortschieten bij inrichting van onderwijs en toetsing. Voor de student zijn de praktische consequenties van een onvoldoende vaak verre van duidelijk: overdoen, ja dat wél. Maar hoe? Waarom men een onvoldoende heeft gekregen is vaak niet duidelijk, zodat de student slecht geprepareerd is op het voorbereiden van zijn herkansing; die herkansing kost tijd, tijd die anders aan voorbereiding van andere vakken besteed had kunnen worden: daaruit vloeit voort dat een onvoldoende voor een bepaald vak de student in moeilijkheden kan brengen waar het gaat om de gestelde exameneisen, of zelfs alleen maar de eisen gekoppeld aan het recht op een studietoelage. Het zijn juist deze consequenties van zakslaag beslissingen die moeilijk te overzien, maar van groot belang voor

de student zijn. Op de één of andere wijze moet met de ernst van dit soort consequenties bij de vaststelling van cesuren rekening gehouden worden. In de praktijk zal dat waarschijnlijk betekenen dat de (sub)faculteit moet komen tot een regeling van de doorstroming waarin de cesuurproblematiek voor afzonderlijke vakken a.h.w. overkoepeld wordt. Voorbeeld van een dergelijk systeem is het Grade Point Average stelsel in de Verenigde Staten (Bevers 1975), als alternatief voor het Nederlandse hoger onderwijs voorgesteld door Wilbrink (1974).

Resumerend: een aantal eisen waaraan cesuurbepalingsmethoden moeten kunnen voldoen worden door De Groot genoemd:

(a) het verschil tussen voldoende en onvoldoende aangemerkte prestaties moet op 'objectieve, feitelijke gronden' gerechtvaardigd kunnen worden.

(b) die rechtvaardiging is in het bijzonder ook noodzakelijk jegens de betrokken studenten.

(c) het geven van een onvoldoende is een niet licht op te vatten zaak, omdat de docent er mee te kennen geeft dat zijn onderwijs bij de betrokken leerling (student) gefaald heeft. Een methode voor cesuurbepaling moet dan ook expliciet gericht zijn op het opsporen van tekorten in dat onderwijs.

(d) een methode tot cesuurbepaling moet, in het verlengde van punt c), gebaseerd zijn op operationeel gedefiniëerde minimum eisen van kennis en kunnen.

(e)uit d) volgt dat een methode tot cesuurbepaling onafhankelijk moet zijn van methoden gericht op het zo goed mogelijk onderscheiden tussen studieprestaties van verschillende studenten (Allais c.s. 1975 zouden zeggen dat het face of differentiation bij cesuurbepaling hoort te zijn: de onderwijsdoelstellingen, en bij meten van verschillen in prestatieniveaus: de studenten). Klassieke toetsconstructieprocedures zijn gericht op het optimaal differentiëren

tussen studenten, en ipso facto zijn op een dergelijke wijze geconstrueerde toetsen niet geschikt voor cesuurbepaling in de betekenis die De Groot daaraan geeft blijkens punt c).

(f) de beslissing onvoldoende/voldoende moet zowel voor de betrokken student, als voor de docent onmiddellijke praktische consequenties hebben.

Er zijn waarschijnlijk nog wel meer eisen te formuleren, maar de bovenstaande zijn door De Groot zélf gestipuleerde eisen, en toetsing van de kernitem methode tegen déze eisen is in eerste instantie het meest interessant. Men mag immers aannemen dat De Groot zijn voorstel bedoeld heeft als zijnde in overeenstemming met de genoemde eisen.

A. D. De Groot De kernitemmethode voor de bepaling van de cesuur voldoende/onvoldoende. Paedagogische Studiën 1964, 425-44.

"Definitie: Een kernitem is een item, dat, naar het weloverwogen en liefst gedocumenteerde oordeel van de verantwoordelijke leraar, aan de volgende eisen voldoet:

1. (validiteit): het item vraagt naar iets (kennis, inzicht, evt. aangeleerde 'instelling'), dat in verband met de doelstellingen van het onderwijs en de behandelde stof van essentieel belang is (tot de 'kern' behoort, van belang is voor theoretisch inzicht of latere toepassing, hetzij bij het verdere onderwijs, hetzij in de praktijk; het is geen detailkwestie, en dgl.);

opmerkingen. Dat een test of toets (en daarmee ook de items waaruit zij samengesteld is) valide moet zijn, is een algemene eis (zie bijv. Standards for educational and psychological tests, 1974). Welke byzondere validiteitseisen daarboven nog gesteld moeten worden waar het gaat om 'kernitems', blijft erg in het vage bij De Groot.

De toelichting van De Groot is kernachtig: er valt uit af

te leiden dat hier geen sprake is van inhoudsvaliditeit, maar ofwel van construct validiteit (theoretisch inzicht), ofwel van predictieve validiteit (relevant voor later praktisch handelen). Voor de rechtvaardiging van de cesuur is een noodzakelijk onderdeel dan ook de demonstratie van construct- en/of predictieve validiteit. Op deze validiteitsproblematiek zal in het volgende nog uitgebreid ingegaan kunnen worden.

2. (betrouwbaarheid): het item discrimineert betrouwbaar tussen 'goede' en 'foutieve' antwoordmogelijkheden.

opmerking: De Groot bedoelt dat het item moet voldoen aan de gebruikelijke eisen m.b.t. item formulering en constructie. Alweer, niets bijzonders of specifieks voor juist kernitems.

Een heel andere kwestie is de eis van betrouwbaarheid die aan de verzameling te gebruiken kernitems als geheel gesteld moet worden, maar wat dit punt betreft zal eerst nog even de verdere uiteenzetting van De Groot gevolgd worden.

3. (geschatte moeilijkheid l): Het item mag niet te moeilijk zijn (het moet zo zijn dat 'ieder die een voldoende verdient het eigenlijk goed moet maken; dus geen vraag van het type:'alleen voor de goeden').

opmerking: vandaag de dag zou een dergelijke eis niet meer in bovenstaand vaag proza omschreven worden, maar er zou gezegd worden dat de meting criterium gerefereerd moet zijn. Of ook: het gaat hier om mastery scores. De problematiek van de constructie van criterium gerefereerde toetsen, is omvangrijk genoeg om a priori te verwachten dat de kernitem methode aan eis nr. 3 waarschijnlijk tot nog toe niet voldaan heeft. Een analyse van de kernitem methode, opgevat als procedure voor criterium gerefereerde meting, zal ongetwijfeld zeer verhelderend werken. Misschien heeft Van Naerssen (1974 a, 1974 b) ook al enkele gedachten in die richting geopperd, gezien zijn gelijktijdige publicatie van een artikel over de kernitem methode en een artikel over criterium gerefereerde meting. Het is duidelijk dat een kritische analyse van de kernitem methode

in het volgende vooral plaats zal vinden met gebruikmaking van de resultaten van theoretisch. en empirisch onderzoek naar methoden voor criterium gerefereerde meting.

4. (geschatte moeilijkheid 2): het item mag niet te gemakkelijk zijn."

opmerking: de eis die hier gesteld wordt is wederom een eis die ook ten aanzien van andere dan kern items gesteld pleegt te worden: het gaat weinig aan om als afsluiting van je onderwijs vragen te stellen die ook zonder dat onderwijs gevolgd te hebben, goed beantwoord kunnen worden. Het is dan natuurlijk wél zaak in de gaten te houden of het gegeven onderwijs inderdaad relevant was: items kunnen goed beantwoord worden door personen die bet onderwijs niet gevolgd hebben doordat ofwel de formulering van de vraag aanwijzingen over de juiste beantwoording bevat (zonder meer slechte items dus) ofwel omdat het onderwijs niet veel toevoegde aan de kennis en inzichten die de studenten aan het begin van dat onderwijs al had.den (een reële mogelijkheid). Maar goed, dit heeft niets met deze cesuurbepalings methode als zodanig te maken.

Wat we aan deze definitie van wat een kernitem zou moeten zijn, overhouden is kort gezegd:

constructvalidering van de items (afzonderlijk, of als groep ?) moet plaats hebben, in sommige gevallen aan te vullen met predictieve validering;
de onderwijsdoelstelling is 100 % mastery van de kernitems (anders een 'onvoldoende')

De eis van 100% mastery is absurd, maar wordt hier door De Groot wél met zoveel woorden aangegeven. Zijn praktische uitwerking van de kernitem methode houdt uiteraard niet aan die eis vaat (er zou vrijwel nooit meer iemand een voldoende behalen). De kloof tussen uitgangspunten er. praktijk is hier dermate groot dat het artikel van De Groot zijn geloofwaardigheid er door moet verliezen.

Wat meer welwillend zou je kunnen zeggen dat De Groot van de docent vraagt om het niveau van beheersing dat gewenst is, vaat te leggen in termen van de soort kennis en inzichten die op het tentamen moeten blijken, in plaats van in termen van waarschijnlijk'leid voor het goed malken van bepaalde soorten vragen (o'Lwel percentage goed te iiaken vragen). De veronderstelling hierbij is dat de docent in. staat is de leerstof (althans, vragen over die leerstof) hierarc',iisch te o,derien in termen van moeilijkheid, in J.eder geval rond de 'cesuur'. Of een dergelijke ordening dezelfde is als een ordening van de vragen naar 'niveau', is maar helemaal de vraag. De ordening uaar 'niveau' is louter afhankelijk van de inzichten van de docent; de ordening naar 'moeilijkheid' is empirisch toetsbaar (zoals bij het eerste tentamen al gebeurt). 'Moeilijkheid' en 'niveau' a priori gelijk stellen lijkt niet gerechtvaardigd. Of dit een serieus bezwaar is, en of de kernitem methode als cesuurbepalingsmethode onder eer dergelijk manco te 'leiden' heeft, zou nog uitgezocht moeten worden.

Als er op een of andere wijze bij de kerniteni methode inderdaad (zij het in de presentatie van De Groot nog impliciet) sprakex is van een of andere hiërarchische ordening van items, dan zou daarvan bij het aanwijzen van de I'kernitems' gebruik gemaakt moeten worden. Op zijn minst zou daardoor de methode van zijn allerergste subjectieve elementen ontdaan kunnen worden. (Van Naerssen 1974 doet een voorstel voor het aanwijzen van kernitems waarbij een stap in deze richting gezet wordt).

Wat de procedure betreft doet De Groot de aanbeveling om kernitems met een lage pwaarde nog eens kritisch te bekijken: of bijv. op dit punt het onderwijs wel voldoende is geweest; of het item wel behoorlijk geformuleerd was, ed. IJet zwakke van dáze 'raad' is dat deze controle volstrekt subjectief is, terwijl juist op dit punt de docent zou moeten kunnen beschikken over meer objectieve of meer valide toetsings procedures. Eén van de mogelijkheden daartoe zou misschien zijn een uitwerking van jánpiaxáa de generaliseerbaarheidstheorie naar items of doelstellingen als 'face of differentiation' (Allais et aliis 1975). Op zijn minst is beoordeling van de items door onafhankelijke 'deskundigen' (vakbroeders) een

noodzakelijk onderdeel van de procedure.

Iets fundamenteler is de vraag of van docenten en leraren wel verwacht mag worden dat zij toets items kunnen schrijven die aan minimum eisen van kwaliteit voldoen (Standards 1974). Mijns inziens, maar ik geef toe dat ik daar misschien eens wat onderzoek over moet opsnorren, is die verwachting sterk overspannen. Een en ander hangt samen met het ontbreken van een behoorlijk theoretisch kader waarbinnen vraagconstructie kan plaats vinden (Wilbrink, maart 1976).

De gemiddelde pwaarde van de kern items wordt bepaald. Vraag daarbij, is of de mate van spreiding van de pwaarden nog var. belang is: het is toch duidelijk dat er ergens een grens moet zijn m.b.t. wat nog toelaatbaar is aan het uiteenlopen van pwaarden van de aangewezen kernitems. Daarover moet iets concreets te zeggen zijn, aangezien de aanname van De Groot is dat ieder kernitem in principe gelijke pwaarde moet hebben: fluctuaties van pwaarden mogen dan niet groter zijn dan te verwachten steekproef fluctuaties rond één bepaalde 'ware' proportie (nl. proportie van studenten die de stof op het gewenste niveau, begrepen heeft). Te vermoeden valt dat ook op dit laatste puntje de psychometrische moeilijkheden zich weer opstapelen, omdat er tevens een onvolledige correlatie tussen de items zal zijn (nog afgezien van verlaging van correlaties die volgt uit ongelijke pwaarder.). Je zou zeggen dat, althans in principe, een toetsing mogelijk moet zijn van de veronderstelling dat de verzameling kernitems betrekking hebben op hetzelfde 'niveau' zullen we maar zeggen. Bij verwerping var, de veronderstelling zou eerst gezocht moeten worden naar een nieuwe of een veranderde verzameling kernitems die wél aan de veronderstelling voldoet. We zien op deze manier dat een aantal veronderstellingen die aan de kern item methode ten grondslag liggen, niet door De Groot (of later door anderen) tot empirische toetsen zijn uitgewerkt, waar dat in principe heel wel mogelijk lijkt te zijn, en waarschijnlijk ook in praktische uitvoering geen grote problemen hoeft te geven (door het opstellen van nomogrammen bijv. waarmee de individuele docent toetsing kan uitvoeren).

Doet zich vervolgens het probleem voor dat er een correctie voor raden toegepast wordt om een schatting te krijgen van de 'ware' proportie studenten die de stof op het bedoelde 'niveau' beheersen. Die gecorrigeerde pwaarde gebruikt de Groot dan als ondergrens, en de ongecorrigeerdeals bovengrens, waarbij de aftestgrens dan tussen beide in gekozen wordt. (kunnen zich wat probleempjes bij voordoen afhankelijk van de frequentieverdeling, maar dat is niet essenti~el: je kunt altijd de studenten de benefit of the doubt geven en de laagste aftest grens kiezen.). Je kunt je afvragen welke beweegreden De Groot heeft om de aftestgrens niet op de gecorrigeerde pwaarde vast te stellen (wat in overeenstemming zou zijn met de uitgangspunten van de kernitem methode), maar daarboven. Ook dit is een ingebouwde 'benefit of the doubt', die op zich niets met de kernitem methode z~lf te maken heeft. Argumenten voor die afronding naar boven geeft De Groot niet; je zou kunnen vermoeden dat dit een correctie a priori is op de in de praktijk volgens de kern:item methode te streng uitkomende cesuur.

Interessant is dat De Groot de mogelijkheid niet vergeet, dat studenten die het wél weten, toch eeri fout antwoord geven (hoewel hij in de gegeven uitwerking die proportie toch maar op nul stelt, wat ook in overeenstemming is met de nogal ideale definitie van het kernitem). Hij waarschuwt wél dat bij items met zeer hoge pwaarden de proportie studenten die het antwoord niet goed hebben terwijl ze het wél weten groter kan worden dan de proportie nietweters die goed raden. Maar daar past dan als aantekening bij dat beide proporties in dat geval niet groot kunnen zijn.

De cesuur wordt dan zodanig gelegd dat de relatieve frequentie van geslaagden op het hele tentamen ligt tussen de boven en ondergrens voor de gemiddelde p waarde van de kernitems. Op het eerste gezicht ziet deze regel er volstrekt willekeurig uit. Waarom?

er wordt slechts met één bron van onbetrouwbaarheid (van het zeer grote aantal mogelijke en relevante bronnen van onbetrouwbaarheid) rekening gehouden: de raadkans. De manier

waarop dat gebeurt leent zich ook nog wel voor enige kritische kanttekeningen.

er wordt doorgeredeneerd van de proportie personen die de kernitems beheersen (hoewel, is dat wel de proportie personen? is de gemiddélde pwaarde bij niet perfect gecorreleerde items gelijk aan bedoelde proportie? In ieder geval lijken een aantal begrippen hier niet of onvoldoende gedefinieerd te zijn.)(bij aanname van homogeniteit van de kernitems, een aanname die zich ook best voor empirische toetsing leent, is het zo dat je inderdaad de items als perfect gecorreleerd kunt beschouwen (mits gecorrigeerd voor onbetrouwbaarheid). Zodat dit probleem zich misschien heel eenvoudig laat oplossen. De draad weer oppakkend: er wordt doorgeredeneerd van de proportie personen die de kernitems beheersen, naar de proportie (de proportie met de hoogste scores uiteraard) studenten die voor het tentamen als geheel dient te slagen. Een punt waarop De Groot uiteraard de kritiek wel anticipeerde. Probleem is en blijft echter waarom in vredesnaam niet de cesuur gelegd wordt uitgaande van de scores op alleen de kernitems: zolang kernitems en andere items niet perfect correleren (in ruwe scores ditmaal) blijft het moeilijk te verdedigen studenten een onvoldoende te geven die wél de gevraagde proportie kernitems goed hebben, terwijl er misschien ook wel eens bezwaren zullen zijn, zeker in die gevallen waarin de kernitem methode met recht en rede toegepast kan worden, tegen het doorlaten van studenten die niet de gevraagde proportie kernitems goed hebben. Het is na de uitleg van De Groot voor mij nog een open vraag of het toevoegen van nietkernitems op zich een bijdrage kan leveren aan het met groter betrouwbaarheid nemen van beslissingen rond de door de keuze van kernitems vastgelegde niveau grens (wat niet dezelfde grens hoeft te zijn als in de door De Groot aangegeven procedure resulteert in de cesuur op basis van álle tentamenvragen). Op zijn minst is hier een meer formele grondslag voor deze praktijk een vereiste. (De Groot geeft als andere redenen dan de verhoging van de betrouwbaarheid nog de moeilijkheid van het vinden van goede kernitems, en het ook nog willen differentiëren van studenten boven de aftestgrens. Als kernitems zo moeilijk te maken zijn, is de methode niet praktisch

uitvoerbaar, tenzij andere procedures als vervanging geconstrueerd kunnen worden wat op voorhand niet onmogelijk lijkt. Het willen differentiëren van studenten over de hele cijferschaal is een streven dat zich in principe niet laat verenigen met optimale procedures voor cesuurbepaling (liever: met studietoetsconstructieregels gericht op het zo aanvaardbaar mogelijk nemen van zak-slaagbeslissingen). Grote verwarring zaait De Groot door er op te wijzen dat compensatie van kernitems en nietkernitems in feite door de docent geaccepteerd is op het moment dat hij dit tentamen in zijn geheel geschikt acht. Dit is een perfecte cirkelredenering.

Bijkomend punt: waarom zou je op basis van de gevonden p-waarde voor de kernitems een afwijzingspercentage vaststellen dat bizar gelijk is of dicht in de buurt ligt van 1-p? Waarom bijv. zou je geen rekening houden met de onbetrouwbaarheid van die gemiddelde p-waarde? En een correctie aanbrengen die die onbetrouwbaarheid in rekening brengt? Andere vraag: waarom zou je bij de beslissing over individuele studenten gebruik maken van indices die betrekking hebben op de hele groep? Waarom zou je bij beslissingen over studenten in het 'grensgebied,' gebruik maker van informatie verkregen uit de hele groep studenten waarbij een groot aantal die juist zonder twijfel boven of onder de cesuur zitten? Is het bijv. in amerikaanse systemen van criterium-gerefereerde toetsing niet mogelijk gebleken om beslissingen over individuele studenten te nemen zonder van groepsgegevens gebruik te maken zoals in deze kernitem methode gebruikt worden?

In de laatste paragraaf bespreekt De Groot o.a. de vraag of de problematiek van de grensbepaling niet eenvoudig verschoven wordt naar de keuze van kernitems. Wat mijns inziens inderdaad het geval is. Het weerwoord van De Groot is dat je met de aanwijzing van kernitems in ieder geval wat meer houvast hebt, dat de cesuurbepaling wat meer bespreekbaar wordt. Hij ziet daarbij over liet hoofd het grote aantal problemen dat juist aan die kernitem vraag formulering vast zit, met name ook op het punt van de moeilijkheid van de vragen. Die moeilijkheid bijv. is toch heel eenvoudig door schijnbaar onbetekenende veranderingen in de vraagvorm

te manipuleren, en omgekeerd blijken er fluctuaties in pwaarden voor te komen als gevolg van wijzigingen in de formulering, die volkomen onverwacht zijn. (in het algemeen blijken pwaarden door docenten heel moeilijk van te voren te schatten). Weliswaar is altijd wel duidelijk over welk stukje van de stof een bepaalde kernitem vraag handelt, maar de wijze waarop die stof in de vraag aan de orde wordt gesteld hoeft in het geheel niet zo duidelijk te zijn. Met name allerlei handboeken over toetsconstructie behandelen het itemschrijven nog steeds als een 'kunst', een activiteit waarvoor geen goede regels gegeven kunnen worden (behalve voorschriften hoe je het allemaal niet moet doen), waar geen behoorlijke theoretische grondslagen voor gegeven zijn. En op dat drijfzand wil De Groot een cesuurbepalingsmethode bouwen! Dan zijn er eerst nog wel een aantal andere mogelijke aanknopingspunten waar veel meer mee te bereiken valt, zeker ook op het punt van het meer bespreekbaar maken van de procedure van grensbepaling. Het bespreekbaar maken van de cesuurbepaling is geen uniek voordeel van de kernitem methode. Evenmin als het een uniek voordeel van de kernitem methode is dat het 'dwingt' over de 'doelstellingen' na te denken. Kernitems zijn wel zeer concreet, zoals De Groot zegt, maar de kernitem methode bouwt nu juist op een aantal eigenschappen van items die in het geheel niet concreet zijn.

Het antwoord dat De Groot geeft op het bovengenoemde bezwaar: dat de kernitem methode de cesuurproblemen niet oplost maar slechts verplaatst, is niet terzake. De blote mededeling dat items toch 'concreet, zijn, is geen beantwoording van de vraag. Dat de procedure van zoeken naar en formuleren van kernitems dwingt tot reflectie over het onderwijs en voert tot een dieper inzicht in dat onderwijs en hoe het overkomt, blijven steken in het stadium van opini~rende uitspraken. De Groot sluit zijn pleidooi af met de afgrijselijke dooddoener dat natuurlijk iedere methode 'verkeerd' gebruikt kan worden.