

[Deze aantekeningen zijn onderdeel van een aanloop naar de mijn eerste docentencursus, over cesuurbepaling, in 1976. Er zijn dan nog een aantal jaren te gaan, gelegenheid tot scherper krijgen van de thematiek. Eenvoudige oplossingen zijn er niet, ik ben er tot heden op verschillende manieren mee bezig gebleven. bw 2025]

A. De zakslaag beslissing

16 april 1973

De zak/slaag beslissing is een belangrijke beslissing. Voor de student, omdat er voor hem vaak enorm veel van af hangt, in het beste geval kost het hem/haar de nodige extra tijd voor een herhaling van het tentamen. Voor de docentenstaf, omdat hij verantwoordelijk is voor de kwaliteit van het gegeven onderwijs, een kwaliteit die enerzijds tot uiting komt in de resultaten (o.a. aantallen geslaagden en gezakten), en die men anderzijds juist ook wil waarborgen of verhogen mede door het toetsingsbeleid.

Aan zak/slaag beslissingen, willen ze niet willekeurig zijn, ligt een bepaald beleid ten grondslag. Veelal is dat beleid niet expliciet, maar berust op stilzwijgende voortzetting van traditionele tentamen procedures. Dergelijke traditionele werkwijzen zouden misschien goede procedures kunnen zijn om te volgen, maar erg waarschijnlijk is het niet omdat slechts zelden bevredigende argumentaties gegeven worden voor het volgen van juist die, en geen andere, procedures. Natuurlijk kan gesteld worden dat het bestaande tentamen systeem niet meer en niet minder is dan een stel afspraken. waar een ieder weet waar hij zich aan te houden heeft. Er zijn situaties waar met dergelijke "afspraken" gewerkt kan worden. Ten aanzien van toetsingssituaties in het onderwijs is het echter mogelijk aan te tonen dat dit traditionele systeem van "afspraken" leidt tot ongewenste situaties. en dat het vervangen kan worden door andere overwegingen en procedures waarvan in principe aantoonbaar is dat ze niet tot bepaalde ongewenste situaties leiden. Een voorbeeld van een ongewenste situatie is dat er te grote aantallen foutieve zakbeslissingen genomen worden. Van

goede procedures bij het nemen van zak/slaag beslissingen mag geëist worden dat aannemelijk gemaakt kan worden dat aantallen foutieve beslissingen beperkt blijven tot wat aanvaardbaar is voor de betrokkenen.

Verschillende beleidslijnen als grondslag voor de beslissingsprocedure.

We zullen enkele van de meest voorkomende beleidslijnen voor zakslaag beslissingsprocedures bespreken.

1. De minst gecompliceerde is die, waar men een bepaald aantal kan laten "slagen". Een dergelijk beleid zal vrijwel uit sluitend gevolgd worden bij toelatingsselectie waar het aantal beschikbare plaatsen bekend is, en kleiner is dan het aantal gegadigden. Meestal is dat een dergelijk beleid voor toetsingen tijdens de cursus niet aanvaardbaar (en in het geval van een propedeutische periode een omstreden zaak). Meestal zal het quotum gevuld worden met "de besten". waarbij de examenprocedure zo opgezet moet worden dat de nodige garantie bestaat dat diegenen die de beste resultaten behalen ook behoren tot de groep van wat men "de besten" vindt. Hoe vanzelfsprekend dit laatste ook lijkt, toch zijn er omstandigheden waarin een dergelijke procedure niet gevolgd kan worden, bijvoorbeeld omdat zij onrechtvaardig kan zijn t.a.v. bepaalde groepen kandidaten die zich onderscheiden naar vooropleiding, sexe, of ze de weekend opleiding volgen of niet, en dergelijke.

2. Een veel voorkomend beleid is dat waar een getalsverhouding tussen geslaagden en gezakten bepaald wordt door bestaan de "tradities" binnen de onderwijsinstelling, of door de "ervaring" van de betrokken docenten, die "weten" dat van iedere jaargroep altijd een ongeveer even grote groep onvoldoende zal blijken te halen. Dit is een beleid dat soms beschaamd verzwegen wordt, maar vaak ook heel expliciet gesteld, verdedigd, en aan andere docenten als "norm" opgelegd. Het behoeft geen uitgebreid betoog, dat dit een rationeel

onverdedigbaar beleid is, en wel omdat de vooronderstellingen ervan veelal aantoonbaar onjuist zijn. (vgl. blz. 3 e.v. van dit stuk)

1*. Een geheel ander beleid is dat waar men werkt met goed omschreven criteria van de mate van beheersing over de onderwezen stof die als een voldoende prestatie zal gelden.

De relatieve beoordeling zoals onder (1) beschreven, en de "traditionele" beoordeling zoals onder (2) afgedaan, houden met het onderwijs zèlf niet veel verband. Dat ligt anders bij de aanpak onder (3) genoemd, reden om een dergelijk beleid iets uitgebreider te behandelen. (Voor selectievormen als onder (1) genoemd zij verwezen naar de betreffende, voornamelijk psychometrische, literatuur, voor (2) naar het boek 'Wijven en Zessen¹¹ van A.D. de Groot).

Procedure die leidt tot weloverwogen zakslaag beslissingen

Een makkelijk in de mond genomen term als "voldoende prestatie" blijkt bij nadere analyse een begrip met veel facetten te zijn. In hetgeen volgt zal een beknopt maar zo volledig mogelijk overzicht gegeven worden van de zaken waaraan aandacht besteed moet worden om tot een te verantwoorden zak/slaag beslissingsbeleid en procedure te kunnen komen.

1. Allereerst is het gewenst de doelstellingen van het betreffende vak te omschrijven, niet slechts in enkele globale uitspraken, maar in een tamelijk gedetailleerde vorm. Aanwijzingen voor hoe men bij het formuleren te werk kan gaan zijn bijv. te vinden in De Groot, van Naerssen e.a. "Studietoetsen". Hoofdzaak is dat een voor alle betrokkenen inzichtelijk overzicht van de inhoud en de opzet van het betreffende stuk onderwijs ontstaat.

2. Voor zover mogelijk wordt aangegeven hoe goed de onderwerpen uit het doelstellingenpakket door de student beheerst moeten worden (in termen van kennis, te blijken inzicht e.d.). Een eerste formulering van dergelijke maatstaven moet dan beschouwd worden als zeer voorlopig, in verband met hieronder te bespreken overwegingen die aanwijzingen op kunnen leveren dat deze maatstaven in bepaalde zin gewijzigd moeten worden.

(3. is weggefallen)

4. Een doelstellingen en criteriaformulering kan niet voorbijgaan aan de mogelijkheden die er zijn om de doelstellingen ook te realiseren in de mate zoals men die zich wenst. Het kiezen van geschikte onderwijsvormen speelt daarbij een rol, evenals de beschikbare middelen en de tijd. De belangrijkste factor vormt ongetwijfeld de student. Het onderwijsprogramma moet zo opgezet worden dat "redelijkerwijze" van de student verwacht kan worden dat hij aan de aan hem gestelde verwachtingen kan vol doen. Wat in ieder geval "redelijk" is, is niet zonder meer duidelijk, rekening zal gehouden moeten worden met persoonlijke verschillen in stijl van studeren, studie tempo, al eerder in deze studie opgedane vaardigheden en inzichten etc. Het lijkt niet juist om uit te gaan van wat "de gemiddelde student" aankan. het kan niet de bedoeling zijn om "de gemiddelde student" een goede oplei ding te geven en de anderen in de kou te laten staan.

5. Wanneer op grond van de in (4) genoemde overwegingen de doelstellingen en criteria formulering "bijge schaafd" is, moet een onderscheid gemaakt worden tussen criteria die aangeven wat de student in het ideale geval van dit studieonderdeel zal "meenemen", en maatstaven die aangeven wat de grens tussen "onvoldoende" en "voldoende", zou moeten zijn. Gedeeltelijk kan voor een bevredigende omschrijving van het

"Juist voldoende" niveau steun gezocht worden bij overwegingen als onder (4) genoemd. Voor een ander deel zal een stuk onderwijs "filosofie" expliciet gemaakt moeten worden. Want het ligt er duimen dik bovenop dat hier uitspraken gedaan zullen worden die vaak al een aanwijzing geven voor aantallen afgewezenen. In een onderwijssysteem waar het "gewoon" is dat tamelijk grote aantallen studenten voor afzonderlijke tentamens afgewezen worden, zal de neiging bestaan "reëel" en "ideaal" erg dicht bij elkaar te kiezen. Waar het echter om gaat is dat op dit punt duidelijk wordt dat het afwijzen van grote aantallen studenten, of juist het afwijzen van kleine aantallen studenten, mede een beleidskeuze is, en niet iets dat automatisch mag voortvloeien uit definities van doelstellingen en criteria, of uit bestaande gewoonten. Wanneer de maatstaven voor een nog juist voldoende prestatie zó hoog gekozen worden dat veel studenten daar niet aan zullen kunnen voldoen, dan moeten er voor een dergelijke keuze zwaarwegende argumenten gegeven kunnen worden. Het omgekeerde geldt natuurlijk ook, hoewel daar de zaken iets anders liggen, een onterecht gegeven onvoldoende weegt vaak zwaarder dan een onterecht gegeven voldoende; selectie tijdens het onderwijs is in velerlei opzicht ongewenst, en komt het onderwijs niet ten goede.

6. Nu is in principe duidelijk waar de grens tussen voldoen de en onvoldoende ligt. "In principe", want het volgende probleem is de inrichting van een toetsingsprocedure die de gegevens oplevert op grond waarvan uiteindelijk de zak/slaag beslissingen genomen moeten worden. De toetsing kan vele vormen hebben: de meerkeuze studietoets ligt erg voor de hand, maar ook essay-toetsen, mondelinge toetsen e.d. zijn vaak, met de nodige waarborgen, bruikbaar. De toets moet inhoudelijk een goede afspiegeling zijn van de inhoud van het onderwijs, of van de geformuleerde doelstellingen. Dit is niet een principe dat slechts met de mond beleden hoeft te worden, het kan heel expliciet als leidraad bij het maken van de toets

fungeren, en van de uit eindelijk geconstrueerde toets moet ook aannemelijk gemaakt kunnen worden dat ze inderdaad een goede afspiegeling is van het eraan voorafgaande onderwijs (of van de daarvan ten grond slag liggende doelstellingen).

Eventueel kan ook hier weer blijken, dat het nodig is de aanvankelijke doelstellingen omschrijving te wijzigen, omdat het vaak weinig zin heeft te werken met doelstellingen die op geen enkele wijze (direct of indirect) getoetst worden. Voor de constructie van meerkeuze studietoetsen bestaan uitvoerige handleidingen, zoals De Groot & Van Naerr»sen. Voor andere toetsvormen gelden vergelijkbare constructieprincipes.

Welke toetsingsvorm ook gekozen wordt, altijd zal men zich af moeten vragen of onderdaad de toetsingsresultaten datgene weer geven (weer kúnnen geven), wat men zich voorstelde te toetsen (te onderwijzen).

7. Het onder (6) genoemde is nog niet voldoende om een goede toets te maken. Er zijn nog allerlei toevallig heden die roet in het eten kunnen gooien. Er zijn vele, van zeer toevallige tot min of meer systematische fac toren

die de hoogte van den score mede bepalen. Een toevallige factor is bijvoorbeeld de migraine waarmee student X op de tentamendag rondloopt een systematische factor is tentamenvrees die sommige studenten meer plaagt dan anderen, en onvermijdelijk in de tentamenscores tot uitdrukking komt. Het is mogelijk een min of meer com pleet

overzicht van dergelijke "storende" invloeden te geven, wat we op deze plaats niet zullen doen. Wat be langrijk is, is dat enige zekerheid moet bestaan dat de tentamenscores in geen al te grote mate door dergelijke niet bedoelde toevalligheden beïnvloed worden (bijv.: een betrouwbaarheidscoëfficiënt als K.R. 20 die voor meerkeuze studietoetsen vaak uitgerekend wordt, geeft enig inzicht in de mate waarin een bepaald soort toevalligheden invloed op de toetsscores heeft). Op grond van onderzoekgegevens bekend uit de literatuur kan van sommige toetsprocedures al apriori gesteld worden dat ze op dit punt

ernstig ontoereikend zijn. Een voorbeeld daar van is het mondelinge tentamen onder vierogen, waar een groot deel van de score bepaald wordt door beoordelingsfouten van de docent, en een ander groot deel door de toevallige keuze van een te klein aantal onderwerpen. Ook hier geldt, evenals bij (6), dat enig inzicht moet bestaan in de mate waarin allerlei toevalligheden een invloed op de scores kunnen hebben gehad, of in positieve zin: zekerheid dat sommige van deze onbedoelde invloeden door de keuze van de toetsingsopzet geen rol van betekenis meer kánnen spelen. (Zowel (6) als (7) hebben betrekking op nogal technische, maar zeer belangrijke zaken. Het is niet mogelijk deze onderwerpen in dit korte verband voldoende recht te doen, verwezen zij naar de psychologische test theorie, of naar advisering op deze punten door onderwijsresearchers).

8. Zoals de toets er nu uit ziet representeert hij goed de doelstellingen (6), en geeft hij scores waar we vertrouwen in kunnen hebben (de betrouwbaarheid, besproken in (7)). Dit zijn echter niet de enige eisen die aan een toetsingsprocedure gesteld kunnen worden. Enige andere eisen zijn:

a. transparantie

De eisen die voor de toets gesteld worden moeten van te voren aan de student volledig duidelijk, met inbegrip van het niveau dat nog net goed is voor een "voldoende" prestatie. De student moet zich zo goed mogelijk op de toets voor kunnen bereiden, met zo weinig mogelijk onnodige en overbodige studieactiviteiten. Hij moet weten wat er gevraagd wordt (kan worden), hoe daarover gevraagd zal worden, hoe goed hij de vragen zal moeten kunnen beantwoorden om een voldoende te krijgen.

b. billijkheid

De toetsprocedure moet verdedigbaar zijn tegenover de student, met name ook de afwijzende beslissing moet met

argumenten gestaafd kunnen worden. Liefst moet daarover van te voren al de nodige informatie verstrekt zijn. Wanneer een toetsprocedure opgezet wordt volgens de lijnen die in dit stuk beschreven zijn, bestaat de rechtvaardiging uit de beschrijving van de gehele ontwikkelingsprocedure, inclusief de redenen waarom deze procedure gevolgd is.

c. doelmatigheid

De gehele toetsprocedure, d.w.z. de toetsing zelf, de voorbereiding door de docenten, en de voorbereiding van de studenten op de toets, plus de consequenties van de beslissingen die op grond van de toetsresultaten genomen worden, dit hele complex moet "de moeite waard" zijn, in juiste relatie staan tot het doel dat men ermee heeft. Daartoe moet de toets voldoen aan de onder (6) en (7) genoemde voorwaarden, maar ook moet de tijd die alle betrokkenen er aan besteden, de kosten die er aan verbonden zijn, de nadelen die het toetsen met zich mee brengt, dit alles moet ruimschoots overtroffen worden door de positieve functies van de toetsing. Voor zover het zak/slaag beslissingen betreft moeten alle moeite en inspanningen, en de kosten of nadelige effecten die verbonden zijn aan de foute beslissingen, in een dergelijke verhouding staan tot de voordelen van deze procedure, dat het geheel een te rechtvaardigen zaak is. Eventueel kunnen bij overwegingen als de boven genoemde, alternatieve procedures in hun voor en nadelen vergeleken worden met de in feite gehanteerde of nog te hanteren procedure* Een heel eenvoudig alternatief is bijvoorbeeld om iedereen te laten slagen; deze procedure zou bepaalde nadelen met zich mee kunnen brengen, zoals verslapping van de motivatie en studieinzet, hoewel dit nu juist nadelen zijn die waarschijnlijk met een aangepaste opzet van het studieprogramma te ondervangen zijn.

Een ander aspect van de doelmatigheid van een selectieve tentamenprocedure heeft betrekking op de mate waarin in de

voorafgaande studie de betreffende groep studenten al sterk geselecteerd is. Het betreft hier een problematiek die in de testliteratuur bekend staat als "restriction of range". Voor een universitaire studie zal het ongetwijfeld zo zijn dat in de eerste twee jaren van de studie de groep studenten sterk geselecteerd wordt, met als gevolg dat verschillen in prestaties in de rest van de studie in belangrijke mate toevallig zijn, kunstmatig gecreëerd door de gewoonte om onder alle omstandigheden met een 10-punts cijferschaal te werken (of door de stilzwijgende veronderstelling dat iedere groep studenten altijd een belangrijk aantal onvoldoende presterenden zal bevatten, een veronderstelling die pertinent onjuist is wanneer er sprake is van sterke "restriction of range").

9. Hoe kun je een "ideale" grens zakken/slagen vaststellen? Met "ideaal" is dan bedoeld dat geen rekening gehouden wordt met factoren als onbetrouwbaarheid van de toetsingsprocedure etc., die natuurlijk later wel moeten dienen voor een bijstelling van de "ideale" grens tot een "praktische" grens slagen/zakken. Op de een of andere wijze moet een nog juist voldoende niveau van beheersing van de stof, zoals we dat in de doelstellingen formulering hebben Vastgelegd, overbrengen naar de toetsingssituatie: wat is een nog juist voldoende (ongecorrigeerd voor onbetrouwbaarheid e.d.) prestatie in de toetsingsprocedure. Als de doelstellingenformulering duidelijk en gedetailleerd is, enals de maatstaven daarbij aangegeven ook voldoende eenduidig zijn, dan zal daaruit op vrij eenvoudige wijze (maarwel met de nodige zorgvuldigheid) uit afgeleid kunnen worden welke toetsprestatie nog juist voldoende zou zijn.

En nu die onbetrouwbaarheid: een toetsingsprocedure levert een aantal scores op, die hopelijk in belangrijke mate bepaald worden door de mate waarin de betrokkenen de stof beheersen. Daarnaast echter zijn er een groot aantal min of meer "toevallige" factoren die invloed hebben op de hoogte van

de scores. Zo kan het dus gebeuren dat iemand die de stof juist voldoende beheerst, en dus een 6 zou verdienen, door toevallige omstandigheden een 4 krijgt, of misschien een 7. Een bepaalde score is dus niet meer dan een benadering van het werkelijke cijfer dat de student zou verdienen. Een en ander betekent dat de grens zakken/slagen zoals die hierboven "idealiter" bepaald was, gecorrigeerd moet worden voor onnauwkeurigheid van de scores die de toetsingsprocedures oplevert. Deze correctie moet zó plaats vinden dat aangenomen mag worden dat vrijwel allen, die in wezen tenminste een 6 zouden verdienen, die ook krijgen. Voor het geval van een studietoets ligt het dan in de rede om uit te gaan van de score die "idealiter" als juist voldoende was aangemerkt, laten we zeggen dat dit een ruwe score van 30 is, en daarvan trekken we af twee maal de "standaard meetfout" van de toets (De standaardmeetfout is een maat voor de onnauwkeurigheid van scores). Tweemaal deze grootte aftrekken betekent dat alleen diegenen worden afgewezen van wie de feitelijke prestatie nauwelijks meer kan worden opgevat als een toevalsafwijking van de "ideale" juist voldoende score. (Anders gezegd: iedereen wordt afgewezen, van wie met een waarschijnlijkheid van 95% vaststaat, dat de voor deze onderwijssituatie juist voldoende prestatie niet gehaald wordt door de getrokken student), (parafrasering van Wijnen, blz. 54)., Het 95% kans interval. dat correspondeert aan twee standaard meetfouten boven en onder de "ideale" score, is een in de sociale wetenschappen gebruikelijk niveau van zekerheid, en lijkt voorlopig ook het beste uitgangspunt voor deze zak/slaag beslissingen. De standaard meetfout is een grootte die routinematig meestal uitgerekend wordt bij computer verwerking van toetsresultaten. Stel dat deze uitkomt op 2,30. Dan vinden we als praktisch te hanteren grens voor zak/slaag beslissingen: $30 - 2 \times 2,30 = 25,40$, scores van 25 of lager leveren een onvoldoende op. Voor toetsingsprocedures die anders zijn ingericht, zoals essaytoetsen, mondelinge tentamens e.d., is het vaak moeilijk of onmogelijk om concreet

inzicht in de betrouwbaarheid van de ruwe scores (in eerste instantie gegeven cijfers) te schatten. Meestal komt daar nog bij dat de onbetrouwbaarheid van deze toetsingsprocedures aanzienlijk veel groter is dan bij studietoetsen, doordat een groter aantal "toevalligheden" de score kunnen beïnvloeden, waarbij vooral de subjectieve beoordeling van prestaties zeer grote invloed op de scores kan hebben. Soms zal het mogelijk zijn door een bepaalde werkwijze enige van deze storende invloeden uit te schakelen. Voor situaties als hier geschetst ligt het voor de hand om alle in eerste instantie uitgedeelde onvoldoendes nog eens te bekijken, maar nu uitgaande van de vraag: hoe groot zou de kans kunnen zijn dat deze cijfers niet meer dan toevallig afwijken van wat we in eerste instantie als "juist voldoende" hebben beoordeeld? In overleg met het C.O.W.O. zou bijvoorbeeld gezocht kunnen worden naar aanwijzingen en argumenten die bij het beantwoorden van die vraag voor een bepaalde toetsingssituatie relevant kunnen zijn. Het ligt in ieder geval voor de hand dat een vijf, zoals die gewoonlijk becijferd wordt, al vrij makkelijk opgevat kan worden als niet meer dan toevallig afwijkend van een 6, wat zou betekenen dat de cijferschaal zoals die in eerste instantie gebruikt is bij de beoordeling, voor het nemen van zak/slaag beslissingen tenminste één punt opgeschoven wordt: alle vijven worden zessen. Wat dan met de andere cijfers moet gebeuren kan hier buiten beschouwing blijven omdat het ons hier alleen om de cesuurbepaling te doen is. (zie daarvoor bijlage B).

10. De beslissingen zijn nu gevallen, maar ze zijn nog niet definitief. Het is bijvoorbeeld mogelijk dat het aantal afwijzingen verrassend groot is. In dat geval zal overwogen moeten worden om alsnog de criteria voor wat een voldoende prestatie is enigszins te laten zakken, wanneer enige aanwijzing bestaat dat er geen sprake is van verzuim of gebrek aan inspanning van de studenten. Immers, ondanks alle zorgvuldige overwegingen en waarborgen, blijven altijd de mogelijkheden

bestaan dat het onderwijs tekort schiet, de toetsingssituatie te weinig transparant is, de toets te moeilijk is, de toets een aantal (psychometrisch) ongewenste eigenschappen bezit die resulteren in een te groot aantal afgewezenen.

11. De toets is achter de rug. de beslissingen zijn gevallen. Het geheel moet nu afgerond worden met een goede herkansingsprocedure voor degenen die afgewezen zijn. Tot een goede procedure kan behoren: intensieve voorlichting aan de studenten over hun specifieke tekorten in voorbereiding en/of op de toets gebleken vaardigheden; een zorgvuldig opgezette herkansingstoets op zo kort mogelijke termijn; alleen op die onderwerpen opnieuw toetsen waar de prestaties op de eerste toets veel te wensen overlieten, etc.

12. Op enkele bekende methoden voor de cesuurbepaling is in dit stuk niet of nauwelijks ingegaan: de kernitem methode van De Groot, de methode Nedelsky, de methode Wijnen. De reden voor dit verzuim is dat de beide eerste typisch betrekking hebben op meerkeuze studietoetsen, en dat ook Wijnen zijn methode aan de hand van studietoetsen heeft ontwikkeld en onderzocht. Afgezien daarvan zijn er nog talrijke min of meer ernstige bezwaren tegen bovengenoemde methoden in te brengen, die voor een belangrijk deel te maken hebben met de grote mate van willekeur waarmee af te wijzen percentages studenten door deze methoden (ook door de methode Wijnen) worden bepaald. In tegenstelling daarmee is in dit stuk er juist van uitgegaan dat alle mogelijke moeite gedaan moet worden om tot een glashelder beleid voor deze zak/slaag beslissingen te komen. Een goed voorbeeld daarvan is het mondelinge hoofdvak tentamen waarvoor per definitie iedereen slaagt die opkomt, een constructie die bijvoorbeeld binnen enkele psychologie subfaculteiten wel voorkomt. Dit is reëel te doen wanneer bijvoorbeeld als onderwerp van het tentamen naast de literatuurlijst een korte scriptie fungeert, die enige tijd vóór het tentamen ingeleverd wordt, en waar bij tekortschieten van

deze scriptie de opdracht voorlopig aan de student teruggesteerd wordt. Meestal zal een dergelijke scriptie geaccepteerd kunnen worden, in welk geval de kandidaat ervan verzekerd is dat zijn tentamen tenminste met een voldoende gehonoreerd zal worden.

A.D. de Groot & R.F. van Naerssen: "Studietoetsen: construeren. afnemen, analyseren". Monton, Den Haag, 1969
A.D. de Groot: "Vijven en zessen". Wolters, Groningen, 1966
W.H.F.W. Wijnen: "Onder of boven de maat, een methode voor het bepalen van de grens voldoende/onvoldoende bij studietoetsen¹¹. Swetz & Zeitlinger, Amsterdam, 1971.

B. Het bepalen van een cijferschaal

Met het oplossen van het cesuurprobleem is de grens tussen voldoende en onvoldoende bepaald, waarmee op de traditionele cijferschaal van 1 t/m 10 de punten 5 en 6 bepaald zijn. De vraag is nu of er een min of meer objectieve methode gevonden kan worden om ook de andere cijfers van deze schaal aan bepaalde prestaties te verankeren¹¹. Wanneer er gewerkt wordt met meerkeuze studietoetsen zijn er enkele praktische methoden om boven en ondergrenzen te bepalen. Voor het bepalen van een bovengrens, welke toetsprestatie correspondeert aan een negen bijvoorbeeld, kan de toets afgenomen worden aan enkele collegadocenten die bij opstelling van de toets niet betrokken zijn geweest. Ná afname van de toets aan de studenten kan geredeneerd worden dat in ieder geval de beste prestaties beloond moeten worden met een acht of negen. Of ook, en dat is meer in lijn met de benadering zoals die bij de cesurbepaling ook gehanteerd is, kan van te voren een poging gedaan worden om te definiëren welke toetsprestatie ongeveer overeen komt met wat men wil of kan beschouwen als een goede beheersing over de stof, zoals omschreven in de doelstellingen van het betreffende stuk

onderwijs. Voor het bepalen van een ondergrens, een minder belangrijk probleem, zou de studietoets afgenomen kunnen worden aan studenten die misschien wel enigszins thuis zijn op het betreffende vakgebied, maar nog geen cursus als die waar de toets betrekking op heeft, doorlopen hebben. In dit geval moet gewaarschuwd worden tegen een methode die wel eens gebruikt wordt: op grond van inspectie van de verkregen eindscores nagaan of er een groep studenten is die "duidelijk slechter is, zich ongunstig onderscheidt van de rest van de groep. Deze methode heeft bij onderzoek bewezen slechts te berusten op toevalsfluctuaties, en mag onder geen omstandigheid gebruikt worden (daarom komt zij ook niet voor in het hoofdstuk over de cesuurbepaling).

Voor schriftelijke tentamens kunnen dezelfde proefafnames worden gebruikt als voor studietoetsen, mits gewaarborgd kan worden dat bij het nakijken van deze werken de beoordelaars niet kunnen weten dat het werkjes van docenten, van studenten die zich niet op deze cursus hebben voorbereid, of van studenten die gewoon het tentamen doen, betreft. Voor mondelinge tentamens wordt een en ander heel erg moeilijk, en zal een cijferschaal op vaak nogal subjectieve wijze gebruikt worden. De nodige extra zorgvuldigheid is dan geboden. Enigszins kan aan al te grote subjectiviteit het hoofd geboden worden door zeer expliciete beoordelingsvoorschriften te maken. Dergelijke voorschriften moeten in intensieve besprekingen van docenten onderling, bij voorkeur bijgestaan door een onderwijskundige, opgesteld worden. Het is natuurlijk ook mogelijk om voor mondelinge tentamens per definitie iedereen die opkomt een acht te geven, waarbij eventueel voorwaarden gesteld kunnen worden aan het mogen opkomen (bijv. het gemaakt hebben van een korte scriptie, die als "voldoende" geaccepteerd is). Dit is misschien een beetje een vreemd voorstel. maar er kan op gewezen worden dat een strategie als hier genoemd vaak niet meer of minder willekeurig is dan traditionele procedures waarin het grootste deel van de

1-10 cijferschaal gebruikt wordt.

[NB Over correctievoorschriften, CV's, bestaat het misverstand dat ze nodig zijn voor eerlijke en/of betrouwbare beoordeling. Nadat ik in 1977 op de ORD had laten zien dat 'objectieve' meerkeuzevragen even subjectief zijn als korte open vragen, duurde het ook voor mij nog enkele decennia voordat ik in de gaten kreeg dat voor CV's hetzelfde geldt: ze zijn ook gewoon subjectief hoor. In de Amerikaanse literatuur bestaat er de term 'frozen subjectivity' voor. Het is een ernstig probleem bij onze eindexamens. bw 2025]

C. Het gebruik van beoordelingen

1. Beoordelingen kunnen gebruikt worden om de student informatie te geven over zijn vorderingen, en over zijn nog bestaande tekortkomingen. Een geprogrammeerde instructie is een onderwijsvorm waarin dit principe erg ver is doorgevoerd, ook computeraidedinstruction hoort hier thuis. Maar ook meer traditionele onderwijsvormen kunnen zo ingericht worden dat er frequente beoordelingen of toetsen gebruikt worden, bedoeld om de student te motiveren, te begeleiden bij zijn studie programmering, inzicht in de eigen vorderingen te geven etc. etc. Dergelijke beoordelingen kunnen gebeuren in de vorm van kortestudietoetsen, in de vorm van discussie (bijvoorbeeld met kleine groepjes medestudenten), etc. Het is niet de bedoeling om cijfers te geven, maar om informatie te geven die voor de student bruikbaar is bij zijn verdere studie.

2. Dezelfde beoordelingen en beoordelingsvormen als onder (1) genoemd, kunnen gebruikt worden om de docent inzicht te geven in de vorderingen van individuele studenten, zodat hij in staat is om, waar nodig, extra begeleiding en advies te geven. Het spreekt vanzelf dat deze beoordelingen louter een informatieve functie hebben, dat ze afgestemd zijn op de doelstellingen van het betreffende studieonderdeel, en dat ze een afspiegeling vormen van eisen die ook in de afsluitende toets gesteld zullen worden (wanneer er gewerkt wordt met afsluitende toetsen). Omdat deze "tussentoetsen", "diagnostische toetsen" of "formatieve toetsen" zoals ze ook wel genoemd worden, een louter informatieve functie hebben,

hoeven er geen zware eisen aan gesteld te worden wat betreft hun betrouwbaarheid, vraagstukkenkeuze, etc; het gaat er slechts om de toetsen zo op te stellen, of dusdanige procedures te gebruiken, dat er inderdaad een goede bijsturende werking op de studieactiviteiten van studenten van kan uitgaan.

3. Aan het einde van een cursus kan een afsluitende toets plaatsvinden, die als functie heeft de student en eventueel de docent, inzicht in het bereikte niveau van beheersing over het stuk leerstof te geven. Ook dit is een stuk informatie verschaffing waaraan eventueel enig advies gekoppeld kan worden wanneer het i.v.m. bepaalde gebleken zwakheden voor de student raadzaam zou zijn bij zijn verdere studie daar toch nog enige extra aandacht aan te schenken. Een louter afsluitende, dus niet selecterende, toets is kenmerkend voor onderwijsvormen zoals die op de betere bladzijden van hun rapporten door Posthumus en door A.D. de Groot zijn gedaan. Zie bijv. 3.3. "middelen ter realisering van een selectievrije opzet" in "Selectie voor en in het hoger onderwijs" van A.D. de Groot.

4. Het ligt voor de hand om bij afsluitende toetsen zoals onder (3) omschreven, toch te zoeken naar criteria waaraan tenminste voldaan moet zijn om als "geslaagd" te kunnen door gaan. Dat verwijst naar het uitgebreide stuk over de cesuur bepaling onder (A) in deze bijlage opgenomen. Doel van het aanleggen van een minimum prestatieniveau kan zijn om te voorkomen dat degenen die evident niet hebben deelgenomen aan het onderwijs (of zelfstudie), bijv. in verband met langdurige ziekte, of ook door onwil etc. etc., op onjuiste gronden een studieonderdeel zouden kunnen afsluiten. Ander mogelijk doel is om de studie voor het betreffende vak voor de student te kunnen structureren door verwijzing naar de uiteinde lijk aangelegde norm, waarbij het dan wel zo moet zijn dat bij redelijke studieinspanning iedereen ruimschoots aan die

norm kan voldoen. Een apart geval is de eindtoets die een apert selectieve functie heeft:

5. Toetsen die bedoeld zijn als selectieinstrument, als onderdeel van toelatingsexamens bijv., moeten aan heel specifieke eisen voldoen, waarvoor verwezen kan worden naar de psychologische literatuur over selectie, of een concept rapport van Wilbrink "Levert selectie in het Hoger Onderwijs iets op?" C.O.W.O., 1971 <https://benwilbrink.nl/publicaties/71SelektieCOWO.htm> en <https://benwilbrink.nl/publicaties/71SelektieCOWO.pdf>

6. Een belangrijk doel waarvoor beoordelingen gebruikt kunnen worden, naast de feedbackfunctie voor de individuele student, is het signaleren van tekortkomingen in het onderwijs.

Voorzover het om een bepaald studieonderdeel gaat, is dit vaak niet meer dan de som van de informatie die de docent over alle individuele studenten krijgt. Immers, wanneer veel studenten met een bepaald onderwerp dezelfde moeilijkheden hebben, ligt het voor de hand te zoeken naar mogelijkheden in de onderwijssituatie om deze moeilijkheden bij voorbaat te voorkomen. Een eenvoudige zaak dus, zij het dan dat de docent zich moet aanwennen de tekortkomingen bij het eigen onderwijs te zoeken, en niet aan zijn studenten te wijten. Iets ingewikkelder is de vraagstelling of een stuk onderwijs zoals dat gegeven is, in overeenstemming is met zijn doelstellingen. Wordt inderdaad datgene bereikt, wat men met dat stuk onderwijs bereiken wilde? Antwoorden op deze vraag kunnen vinden, betekent dat doelstellingen en de maatstaven die men daarbij wil aanleggen, zó geformuleerd worden dat het mogelijk wordt informatie te verzamelen die een inzicht geeft in de mate waarin men met het onderwijs bereikt wat men wil bereiken. Deze informatie kan soms verkregen worden uit gangbare toetsingsprocedures, maar daarbij zijn ook andere middelen te gebruiken (volgens andere principes geconstrueerde toetsen, vragenlijsten, discussie. analyse van werkmateriaal). Een aardig voorbeeld is om na te gaan welke bladzijden uit

studieboeken de sporen van meest intensieve bestudering dragen, een gegeven waaruit zou kunnen blijken dat er discrepanties bestaan tussen wat de docent als belangrijkste doelen van zijn cursus ziet en hoe de student met de stof omgaat.

7. Mutatis mutandis gaat hetzelfde op voor een compleet onderwijsprogramma. Bij analyse van tentamenresultaten bijvoorbeeld zou kunnen blijken dat een bepaald vak fungeert als "bottle neck" voor het doctoraal. terwijl dit niet in overeenstemming is met het belang dat aan dat vak in het kader van de doctoraal opleiding toegekend wordt. Bij iets grondiger opgezet onderzoek zou kunnen blijken dat het voor vak A en vak B, die veel gemeenschappelijks hebben, vaak voor komt dat studenten die voor het ene vak zakken, bij het andere vak meteen goede cijfers halen. Dergelijke gegevens zijn aan leiding tot verder onderzoek, gericht op het verbeteren van de onderwijssituatie in een hele (sub)faculteit.

Hoe moet nu, in het licht van bovenstaand overzicht, bijvoorbeeld het mondelinge, afsluitende, tentamen in traditionele vorm beoordeeld worden? Het heeft geen informatieve functie voor de student die slaagt. De student die zakt kan zich vaak beroepen op een stuk onderwijskundig en psychometrisch inzicht. zoals omschreven in de bijlage over het bepalen van de grens zakken/slagen, en stellen dat de afwijzende beslissing onvoldoende gefundeerd is, m.a.w. dat hij in zijn belangen geschaad is, dat hem een stuk onrecht is aangedaan. De docent is meestal niet in staat op grond van de informatie uit deze tentamens verbeteringen in zijn onderwijs aan te brengen. Ook voor de faculteit zijn er weinig of geen aanknopingspunten voor onderwijsveranderingen, of een beter beleid, aan te ontleen (behalve simpele zaken, als het op de vingers tikken van een te scherp selecterende docent). Aan een aantal belangrijke eisen die aan toetsingsprocedures gesteld kunnen of moeten worden, wordt door de traditioneel

mondelinge toets vaak dus niet voldaan; wat een reden moet zijn om te zoeken naar andere, betere, meer op het onderwijs zelf betrokken toetsingvormen.

C.O.W.O., 16 april 1973 Ben Wilbrink