

[Dit stuk is een vingeroefening. Of beter: een worsteling met de thematiek. Alle begin is moeilijk. Dit stuk is natuurlijk volkomen ongeschikt om aan docenten voor te leggen. Het is de worsteling van de beginnende expert met zijn vak. Het is goed dat het stuk bewaard is gebleven, en dat dit gepruts kan laten zien dat ik mijn loopbaan ook met de nodige misvattingen ben begonnen. Interessant is dat aan het eind van het stuk uitdraag dat in beginsel alle studenten moeten kunnen slagen. De dagelijkse praktijk in het onderwijs is een heel andere. Het thema zal mij blijven bezighouden, in meerdere gedaanten. (1992 ECER; blogs in de ;20-er jaren; bw 2025]

(C.O.W.O.) tel. 5252835.

11 november 1973

De richting waarin naar verbetering van beoordelingsmethoden gezocht kan worden.

Ben Wilbrink.

Beoordeling van studieprestaties is het onderwerp van dit informatieve stuk. Zoals zal blijken, is beoordeling van studieprestaties een nogal rekbaar begrip; geprobeerd is om de verschillende te onderscheiden vormen van beoordeling ook afzonderlijk, zij het wel in hun eventuele onderlinge samenhang, te bespreken. Daarbij zullen aan de orde komen:
de beoordeling zoals die aan het eind van een stuk onderwijs plaats kan vinden (ook wel summatieve beoordeling genoemd), de beoordeling die tijdens het onderwijs gegeven wordt, wat kan gebeuren d.m.v. tussentijdse toetsen, werkbeprekingen, e.d. (ook wel formatieve beoordeling genoemd), het probleem van de bepaling van de grens slagenzakken.

Geen systematische bespreking zal gewijd worden aan het fenomeen selectie, in de zin van methode om aantallen studenten te beperken. De (sub) faculteit heeft niet als taak om een dergelijk soort selectie, ook niet in verkapte vorm, uit te voeren. Daarover is in de wet niets vastgelegd. Zelfs de minister heeft een machtigingswet nodig om een dergelijke selectie te kunnen uitvoeren. Ook in de voorstellen voor herstructurering wordt uitsluitend gesproken over

een selectie met het oog op geschiktheid, en niet niet het oog op beperken van de aantallen. Enige vorm van selectie zal in dit stuk dus niet aan de orde zijn.

Alleen bespreken van de beoordeling van studietoestanden is natuurlijk eenzijdig, niet alleen de student, ook het, onderwijs en de docent worden beoordeeld of zouden dat moeten worden. Deze laatste vormen van beoordeling, die betrekking hebben op onderwijsbeoordeling, zijn anders van aard en opzet dan de beoordeling van studietoestanden (hoewel er natuurlijk veel raakvlakken zijn). In dit stuk zal de aandacht voornamelijk op de beoordeling van studietoestanden gericht zijn.

korte motivatie.

In het huidige universitaire onderwijs functioneert de beoordeling van studietoestanden als een systeem van beloning en straf. Anderen zullen ititischien meer geneigd zijn te spreken over motiverende betekenis van tentamens e.d. op de studieactiviteiten van de student. Hoe dat ook zij, de wijze waarop studietoestanden beoordeeld worden, en de consequenties verbonden aan hoge of lage cijfers, zijn in hoge mate bepalend voor de aard en de kwaliteit van het onderwijs. Vandaar dat juist de beoordeling van studietoestanden een belangrijk aangrijpingspunt is voor verbetering van het onderwijs. Dat wordt door studenten en docenten ook beseft, getuige bijvoorbeeld de vaak verhitte discussies over de vraag; wel of geen meerkeuze toetsen.

Ook door onderwijskundigen wordt de laatste jaren meer aandacht besteed aan de betekenis die beoordeling voor het onderwijs zelf heeft. Dat dringt zelfs door tot onze landelijke pers, waarin discussies plaats vinden over de mogelijkheden en wenselijkheden van het 'learning for mastery' model van Bloom c.s. Maar ook de beoordeling op zich wordt steeds meer bekeken vanuit oogpunten van relevantie voor het onderwijs, billijkheid voor de betrokken studenten, aanvaardbaarheid voor alle betrokkenen, e.d., waarvoor De Groot met name tekent.

Een en ander zal In dit stuk uitgebreid aar, de orde komen, zij het dat

geen uitgebreide argumentaties en literatuurverwijzingen in een kort bestek als dit gegeven kunnen, worden. Voor verdere informatie kan men zich echter altijd tot het COWO wenden.

Summatieve beoordeling.

De term 'summatief' is ontleend aan Benjamin Bloom, en lijkt ook buiten een strict 'learning for mastery' model bruikbaar. Met summatieve beoordeling wordt bedoeld iedere vorm van beoordeling die plaats vindt aan het eind of als beëindiging van een integraal stuk onderwijs. Het klassieke tentamen is een voorbeeld van summatieve beoordeling.

Summatieve beoordeling kan op vele manieren plaats vinden: mondeling of schriftelijk, of m.b.v. meerkeuze toetsen. Ook werkstukken (van groepen of individuen) kunnen van een eindoordeel voorzien worden. Ons interesseert de vorm echter niet zozeer, wel de eisen die vanuit diverse gezichtspunten aan dergelijke vormen van toetsing gesteld mogen en zouden moeten worden.

Doelstellingen.

In het volgende zal telkens weer gerefereerd worden aan de bedoeling van de beoordeling, het doel waarvoor de beoordeling gehanteerd wordt. Wat de docent ook als doel voor ogen mag staan bij zijn beoordelingen, altijd zullen zijn onderwijs doelstellingen daarin een belangrijke rol moeten spelen: zij vormen als het ware het kader waarbinnen de verschillende onderwijs activiteiten, ook de beoordeling, zich afspelen. Hoewel dit stuk gaat over beoordelen, zullen we er niet omheen kunnen in het kort iets over doelstellingen te zeggen.

Het formuleren van expliciete en gedetailleerde doelstellingen is in ons onderwijs nog ongebruikelijk. Zolang ze nog niet geformuleerd zijn moet er naar gegist worden: een goede gissing kan vaak gemaakt worden door na te gaan wat er aan het eind van het onderwijs van de student gevraagd wordt in termen van

studieprestaties (en vaak heeft de student dat beter door dan de docent); de vragen op het tentamen zijn als het ware een operationalisatie van de (niet expliciet geformuleerde) doelstellingen. Uiteraard zou de docent zich teleurstellingen kunnen besparen door zijn doelstellingen wél te formuleren, waarmee hij tegelijk het leven van zijn studenten aanzienlijk zou kunnen veraangename.

Het formuleren van een doelstelling gaat ongeveer zo:

- geef aan om welke kennis, inzichten of vaardigheden het je te doen is (in globale bewoordingen)

- maak de globale doelstellingen 'harder' door bewoordingen te kiezen die aangeven wat de student moet kunnen doen wanneer hij over de bedoelde kennis, inzichten of vaardigheden zou beschikken
- geef aan onder welke omstandigheden (beschikbaar materiaal, literatuur, aard van de opgaven, in stress situatie of niet) de student van zijn kunnen blijk moet geven

- en tenslotte: specificeer welke prestatie aangemerkt kan worden als blijk van beheersing van de doelstelling.

Dit is alles, en natuurlijk veel te weinig, dat hier over formulering van doelstellingen gezegd kan worden. Voor verdere informatie zie bijv. De Cecco: Educational Psychology, of het onderdeel 'doelstellingen' van de docenten cursus die het COWO op dit moment nog in ontwikkeling heeft.

Wanneer in het volgende over doelstellingen gesproken wordt, dan is daarmee bedoeld een pakket doelstellingen, geformuleerd op de wijze als hierboven aangegeven. Wanneer over 'impliciete' doelen die de docent met zijn heeft, of zou kunnen hebben, gesproken wordt dan wordt dat ook zo aangegeven.

objectiviteit van de beoordeling.

Met objectiviteit wordt door sommigen (o.a. De Groot) niet meer bedoeld dan de mate waarin beoordeling van antwoorden door een computer gedaan zou kunnen worden. waarmee subjectiviteit van menselijke beoordeling uitgeschakeld zou kunnen worden. Dergelijke subjectiviteit heeft namelijk een verstorende invloed op de juistheid van het oordeel. Maar niet alleen beoordeling van

antwoorden kan subjectief zijn, ook de keuze van problemen, de keuze van afleiders bij studietoetsen, de formulering van problemen e.d., die tot resultaat kunnen hebben dat de scores voor verschillende studenten op verschillende wijze beïnvloed worden. [In 1977 zou ik op dit thema een licht sensationele presentatie op de ORD geven: keuzevragen zijn subjectief, in tegenstelling tot de dogmatiek van het Cito <https://benwilbrink.nl/publicaties/77KeuzevragenORD.htm> bw 2025] Daarom willen we het begrip objectiviteit de volgende uitbreiding geven:

De objectiviteit van een beoordelingsprocedure, of een oordeel, heeft betrekking op de mate waarin allerlei irrelevante factoren die op het oordeel van invloed zouden kunnen zijn, geen rol kunnen spelen, of ten minste onder controle gehouden zouden kunnen worden.

Wat objectiviteit betreft zou je een beoordelingsprocedure dan ook kunnen vergelijken met een experimental design waarin men de invloed van allerlei variabelen waarin men geen direct belang stelt, tracht uit te schakelen of door statistische of experimentele methoden onder controle te houden.

In het volgende zullen we een aantal maatregelen noemen als voorbeelden van de wijze waarop men kan streven naar een grotere objectiviteit. Daarmee is nog niet gezegd dat onder alle omstandigheden dergelijke maatregelen genomen zouden moeten worden. Het is bijvoorbeeld inzichtelijk dat de eis van objectiviteit belangrijker wordt, naarmate de consequenties van de beoordeling ernstiger kunnen zijn voor de betrokkene. De moeite die gedaan zou moeten worden om een grotere objectiviteit te bereiken moet dan ook afgewogen worden tegen het belang van de beoordeling, maar ook bijv. tegen eventuele minder gewenste gevolgen die een grotere objectiviteit met zich mee kan brengen. Een volmaakt objectieve scoring vereist bijvoorbeeld dat met een of andere vorm van studietoets gewerkt wordt, wat in bepaalde onderwijs situaties, of voor betrokken studenten of docenten op zich niet gewenst kan zijn.

maatregelen.

Vragen die gesteld worden, kunnen op een van te voren

afgesproken, en ook aan studenten bekende, manier uit de stof gekozen worden. Hierbij zou een pakket doelstellingen uitstekende diensten kunnen bewijzen, daarbij zou bijv. aangegeven kunnen worden hoeveel vragen over ieder van de doelstellingen verwacht kunnen worden, welke de relatieve belangrijkheid van de doelstellingen is m.b.t. de toetsing, e.d.

Een stap op deze weg is ook het aangeven over welke onderwerper. in de stof welke soort vragen verwacht mogen worden.

Bij het gebruik van studietoetsen is het mogelijk om het hele itembestand uit te geven aan studenten, en dan op het tentamen de studenten een lijst met op random wijze uit datzelfde bestand getrokken items voor te leggen. (de lezer die het begrip 'transparantie', dat verderop nog expliciet aan de orde komt, zal inzien dat dergelijke maatregelen ook de transparantie van de beoordeling verhogen.) [Bob van Naerssen heeft dit inderdaad in praktijk gebracht, maar moest constateren dat er een eenvoudige truc mogelijk is die studenten een hoge score geeft zonder enige inhoudelijke kennis. Bob gebruikte ja/nee-vragen; studenten legden eenvoudig de ja-vragen bij elkaar, lazen die wel aandachtig door maar de nee-vragen lazen ze niet. Op het tentamen: iedere herkende vraag is een ja-vraag. bw 2025]

Zorg er voor dat alle studenten in ongeveer gelijke mate duidelijk is welke de beste strategie bij het beantwoorden van de vragen is. Wanneer onder studenten geruchten de ronde doen over het beste tentamen gedrag, dan is er op dit punt iets aan het handje. Vooral mondelinge tentamens zijn wat dit betreft erg kwetsbaar.

Mondelinge of schriftelijke antwoorden altijd door meerdere docenten en/of studenten laten beoordelen, en van tevoren afspreken op welke wijze de verschillende oordelen gecombineerd zullen worden tot een eindoordeel, Daarbij moet men zich wel blijven realiseren dat gedeelde subjectiviteit nog geen uitgeschakelde subjectiviteit is. alleen toevallige bias kan zo enigszins gecorrigeerd worden, maar de systematische blijft nog van invloed. Daarom ook is het eigen oordeel van de student over de vraag, zijn antwoord, en de beoordeling daarvan, zeer relevant (daar komen we later nog op terug).

Stel voor de beoordeling van antwoorden voorschriften op, waarin vermeld kunnen worden de belangrijke punten die in ieder geval in een goed antwoord voor zouden moeten komen, in welke specifieke gevallen een antwoord als onvoldoende aangemerkt moet worden, e.d. Het spreekt vanzelf dat dergelijke voorschriften door de student ook geanticipeerd moeten kunnen worden, m.a.w. de student moet op de hoogte zijn van de wijze waarop zijn vragen en antwoorden beoordeeld zullen worden, en hij moet daar in de voorbereiding op de beoordeling rekening mee kunnen houden. Ook hier zou een pakket doelstelling formuleringen weer goede diensten kunnen bewijzen: enerzijds zijn daar door de docent eenvoudig beoordelingsvoorschriften uit af te leiden, anderzijds kan de student dat ook doen en aldus zijn studiestrategie goed kiezen.

De meerkeuze studietoets heeft als voordeel dat de beantwoording voor iedere student dezelfde vorm heeft: een aantal streepjes of cirkeltjes. Bij andere vormen van beoordeling wordt de beoordeling beïnvloed door stijlfactoren, netheid van schrijven, gelaatsexpressie, uiterlijk voorkomen, zorgvuldigheid van formuleren, bekendheid van de docent met de betrokken student, e.d. (voor uitvoerige bespreking zie 'Methodologie' van A, De De Groot). Deze beïnvloeding is niet alleen vaak niet uit te schakelen, maar de docent komt bovendien erg makkelijk in de verleiding om te veronderstellen dat zijn oordeel daar toch zeker niet door beïnvloed kan zijn. Er kan dan ook niet genoeg op gewezen worden dat onderzoek telkens weer de invloed van dergelijke factoren aantoont. De les die daar tenminste uit getrokken kan worden, is om het uiteindelijk oordeel veel minder als een vaststaand gegeven te beschouwen dan vaak nog gebruikelijk is. Bij de studietoets heeft men de uitgerekende standaard meetfout als indicatie voor de mate van 'onzekerheid' in de door de student behaalde score, maar hoewel voor andere vormen van beoordeling een dergelijke index moeilijk te verkrijgen is, moet ook hier terdege beseft worden dat de verkregen score als een schatting beschouwd moet worden, en niet als een precieze maat. Dit introduceert het probleem wie het voordeel van de twijfel moet krijgen: de student of de docent, op die problematiek zullen we in het hoofdstuk over bepaling van de grens slagenzakken nog terugkomen.

liet is mogelijk dat de beoordeling in de loop van de tijd verandert in

strengheid, of in spreiding van de scores. Dit probleem doet zich voor bij het nakijken van schriftelijke werken, en het afnemen van mondelinge tentamens, dan kan het voorkomen dat de eerste strenger beoordeeld worden dan de laatsten, of omgekeerd. Een zekere controle hierop is mogelijk wanneer de cijfers, in de volgorde waarin ze gegeven zijn, bestudeerd worden (bij voorkeur moeten de werken in gerandomiseerde volgorde nagekeken worden, of de studenten in gerandomiseerde en zeker niet in volgorde van aanmelding of alfabet opgeroepen worden).

Veronderstel dat een groep studenten beoordeeld is, en dat de cijfers vastgesteld zijn. Dan is een geweldig interessante vraag hoe die verzameling cijfers er uit zou hebben gezien wanneer voor die beoordeling andere maar vergelijkbare vragen waren gebruikt, Immers, wanneer de cijferverdeling er dan in belangrijke mate anders zou uitzien, zou dat betekenen dat we aan die cijfers maar bar weinig betekenis kunnen toekennen. Bij studietoetsen is het soms mogelijk, met erg veel voorzichtigheid en voorzorgen, in deze vraag enig inzicht te krijgen door de verzameling items in tweeën te splitsen, zodat vergelijkbare toetsen maar dan met de helft van het aantal vragen ontstaan, en dan scores per persoon voor ieder van de helften met elkaar te correleren. Zo kan onder bepaalde omstandigheden de bij de meeste computerbewerkingen van studietoetsen uitgerekende KR 20 beschouwd worden als een (onder)schatting van de bedoelde correlatie (een betrouwbaarheid, voor degenen die met die term beter vertrouwd zijn).

Bij mondelinge en schriftelijke procedures is het alleen met behulp van speciale maatregelen mogelijk enig inzicht in deze vorm van betrouwbaarheid van de beoordelingsprocedure te verkrijgen. in de meeste gevallen lijkt het niet onredelijk de schatting te maken dat de betrouwbaarheid in de buurt van .5 of .6 zal liggen, een fors stuk beneden de betrouwbaarheid van studietoetsen die veelal in de buurt van .8 kan liggen (en voor door testbureaus vervaardigde toetsen bij .90 of .95) Hier ontbreekt de ruimte voor een uitvoerige uiteenzetting over betrouwbaarheid, welke factoren daar zoal op van invloed kunnen zijn, en welke de consequenties van gebrek aan betrouwbaarheid zijn op de wijze waarop cijfers beïnvloed worden door de toevallige keuze van vragen in het tentamen. De Groot heeft een en ander op niet mis te verstane wijze uiteengezet in zijn 'Vijven

en Zessen (1966).

Ook hier is de les die de docent kan trekken: wees uiterst voorzichtig bij het omgaan met cijfers, het zijn niet meer dan indicaties, en vaak nog erg zwakke bovendien, van het 'werkelijke' niveau van de student of studenten.

Voor degenen die werken met sterk geformaliseerde toetsingsprocedures, zoals meerkeuze studietoetsen, is de boven gegeven korte bespreking waarschijnlijk niet bevredigend. Zij kunnen het best raadplegen het artikel over betrouwbaarheid, geschreven door Julian C. Stanley, en verschenen in het ook overigens meest aanbevelenswaardige handboek over beoordelingsprocedures, samengesteld door Robert L. Thorndikel getiteld "Educational Measurement", second edition, 1971. Mij zijn geen Nederlandstalige publicaties bekend waarin het betrouwbaarheid⁶ begrip niet op eenzijdige wijze behandeld wordt, ook de handleiding over toetsconstructie etc. van De Groot, Van Naerssen, e.a. niet uitgezonderd. Gedeeltelijk valt dat toe te schrijven aan de snelle ontwikkeling de laatste jaren in het denken over eisen die aan psychologische tests en onderwijs toetsen gesteld mogen en moeten worden, Zie daarvoor met name ook het artikel van Lee J. Cronbach in het genoemde handboek, dat handelt over validiteit (geldigheid).

geldigheid van de beoordeling.

Voor de bespreking van eigenschappen van beoordelingsprocedures zijn er een aantal kernbegrippen, waarvan de belangrijkste zijn: de op de vorige bladzijde summier besproken betrouwbaarheid, en de in het volgende te behandelen vormen van geldigheid (validiteit). De geldigheid waar het om gaat is de geldigheid van de interpretatie die gegeven wordt aan de gegevens die de beoordelingsprocedure oplevert.

De interpretatie die men wil geven aan de beoordelingsresultaten moet duidelijk zijn voordat aan de beoordelingsprocedure begonnen wordt. Deze raadselachtige uitspraak betekent slechts dat er bij het vaststellen van de beoordelingsprocedure, en het bedenken van de vragen, de wijze waarop antwoorden gescoord worden, e.d.,

duidelijk moet zijn welk soort gegevens men graag van de student wil hebben om die gewenste interpretaties terecht te kunnen maken. In navolging van Cronbach's genoemde artikel onderscheiden we het gebruik van toetsing om de student(en) te beschrijven en het hanteren van toetsing dat gericht is op het nemen van beslissingen over studenten. Beide bespreken we afzonderlijk.

geldigheid van beoordelingsprocedure als beschrijving van de student.

Allereerst gaan we kijken naar de relevantie van de beoordelingsprocedure met betrekking tot het gegeven onderwijs, en meer in het bijzonder voor wat de doelstellingen van het onderwijs betreft. Ieder stuk onderwijs, of het nu door een docent persoonlijk gegeven wordt of voornamelijk plaats vindt op de kamer van de student, is gericht op het bijbrengen van intellectuele en/of motorische vaardigheden. Bij de summatieve beoordeling wordt dan onderzocht in hoeverre de opzet geslaagd is, in welke mate de studenten zich die vaardigheden eigen hebben weten te maken. Het ligt voor de hand dat een allereerste vereiste, wil men nagaan of de beoordeling daaraan voldoet, is dat nauwkeurig omschreven is om welke vaardigheden het gaat, onder welke condities de studenten van hun vaardigheden blij moeten geven, en welke eisen aan de kwaliteit van beheersing van die vaardigheden gesteld worden. Iets andere woorden, de doelstellingen van het onderwijs moeten behoorlijk omschreven zijn. is dat het geval, dan moeten we nagaan in hoeverre de vragen en opgaven die deel uit maken van de beoordelingsprocedure, representatief zijn voor de doelstellingen van het onderwijs.

maatregelen

In de meeste gevallen zal eerst een begin gemaakt moeten worden met het expliciet formuleren van doelstellingen op de wijze zoals op blz. 2 kort aangegeven.

Zolang er nog geen doelstellingen geformuleerd zijn, moeten we ons op bijv. de volgende wijze behelpen:

Besprek zorgvuldig iedere afzonderlijke vraag of opgave die in de

beoordelingsprocedure voorkomt of kan komen. Dat kan een bespreking tussen docenten zijn, van de docent met vakgenoten die niet direct bij zijn onderwijs betrokken zijn, of bespreking tussen de docent en zijn studenten.

Vragen die gesteld kunnen worden zijn bijvoorbeeld:

- is de opgave belangrijk
- is het voorafgaande onderwijs gericht geweest op o.a. het kunnen voldoen aan dit soort opgave
- is tussen vakgenoten onenigheid mogelijk over juiste oplossing of beantwoording
- is de moeilijkheid 'representatief'?

Ook aan de verzameling vragen en opgaven als geheel moet aandacht besteed worden:

- is de keuze van vragen wat inhoud of vorm betreft misschien op bepaalde wijze eenzijdig geweest?
- hebben studenten (en ook de docent) de indruk dat dit pakket vragen hen in staat stelde een goed beeld van hun kunnen te geven?

Zijn wél expliciet geformuleerde doelstellingen beschikbaar, dan kunnen eenvoudig de vragen en opgaven in hun totaliteit vergeleken worden met het pakket doelstellingen. De vraag is dan of de vragen een goede afspiegeling van de doelstellingen vormen. Uiteraard hoeft hiermee niet gewacht te worden tot de beoordelingsprocedure achter de rug is, er zijn immers geen concrete gegevens over de door studenten gegeven antwoorden voor nodig. Deze controle kan vóóraf uitgevoerd worden.

Wanneer zou blijken dat de vragen niet geheel representatief zijn voor de doelstellingen, dan kan een ongelukkige doelstellingen formulering daar de oorzaak van zijn (herformuleren dus), maar meestal zal het gaan om discrepanties waar te weinig of te veel vragen met betrekking tot bepaalde doelstelling(en) gesteld worden en vergelijking tot het belang dat men aan de doelstelling hecht. In dergelijke gevallen is de remedie ook eenvoudig; enige vragen schrappen, of enige vragen extra formuleren.

Ideaal zou zijn om een uitgebreide doelstellingen formulering op te zetten, waaruit volgens bepaalde spelregels vragen en opgaven voor

beoordelingsprocedures afgeleid kunnen worden. Op het ogenblik is op diverse plaatsen onderzoek gaande naar zeer geavanceerde methoden om dit te bewerkstelligen. Er zijn echter ook eenvoudige methoden, die binnen het handbereik van iedere docent liggen. Eén van de mogelijkheden is het maken van een doelstellingen tabel zoals door Bloom c.s. gehanteerd wordt. Een voorbeeld daarvan is te vinden In W. Meuwese "OnderwijsResearch", (Aula) blz 155.

In veel gevallen zal het mogelijk zijn een klein experiment uit te voeren; Prepareer voldoende vragen en opgaven voor twee beoordelingsprocedures. Verdeel de vragen in tweeën, en wel zó dat er twee reeksen vragen ontstaan die ongeveer even representatief voor de doelstellingen (of het onderwijs) geacht mogen worden. Vraag de medewerking van studenten om beide beoordelingsprocedures te ondergaan, en correleer de twee reeksen scores die aldus verkregen worden. Het spreekt vanzelf dat de nodige methodologische voorzorgen getroffen moeten worden, waarover hier niet verder gesproken kan worden (voor hulp bij het opzetten van een dergelijk experiment kan men zich tot het COWO wenden). Wanneer er weinig verband tussen beide reeksen scores zou blijken te bestaan, dan is er ook geen behoorlijke inhoudelijke geldigheid mogelijk (een alternatieve verklaring zou kunnen zijn dat de vragen onvoldoende betrouwbaar waren, onvoldoende betrouwbaarheid betekent echter altijd tegelijkertijd onvoldoende geldigheid, omdat tussen beide een voorwaardelijk verband bestaat (zie daarover bijv. De Groot en Van Naerssen)

Het genoemde experiment kan uitgevoerd worden op basis van een expliciet doelstellingenpakket met bijbehorende procedure voor het genereren van vragen voor de beoordeling.

Ook wanneer geen doelstellingen expliciet geformuleerd zijn, kan het experiment uitgevoerd worden. Het enige dat daarvoor nodig is dat de docent(en) op de hun vertrouwde wijze twee reeksen vragen opstellen die zij beide ongeveer even representatief vinden voor wat zij met hun onderwijs hebben willen bereiken; dit laatste oordeel berust in hun geval op moeilijk achterhaalbare overwegingen, wat echter in het kader van dit kleine experiment geen hinderpaal hoeft te zijn. Integendeel, uit dit experiment zou kunnen blijken dat het niet werken met expliciete doelstellingenformuleringen kan leiden tot een

tekort in de inhoudelijke geldigheid van de beoordelingsprocedure, en dat kan een stimulans vormen om tot expliciete doelstellingen formulering over te gaan.

Een andere vraag die relevant is in het kader van de geldigheid van beoordelingsprocedures voor beschrijving van de student, is of inderdaad in de scores de eigenschappen tot uitdrukking komen waarover we enig inzicht willen krijgen (de door de student verworven intellectuele vaardigheden bijv.). Het is namelijk niet denkbeeldig dat de scores tot stand komen onder invloed van geheel andere eigenschappen dan degene waarin we geïnteresseerd zijn.

Om daar een voorbeeld van te geven: bij beoordelingsprocedures zijn we niet geïnteresseerd in de intelligentie van de studenten (dan kunnen we ook maar beter een behoorlijke intelligentie test afnemen), het is echter mogelijk dat verschillen in scores in belangrijke mate terug te voeren zijn op verschillen in intelligentie. Zou dat bij onderzoek het geval blijken te zijn, dan zou onderzocht kunnen worden wat daarvan de oorzaak zijn. Oorzaken zouden kunnen zijn: duistere formulering van de vragen, specifieke keuze van vraagvormen, subjectiviteit in de beoordeling, een onvoldoende gestructureerde onderwijs situatie, e.d.

Andere factoren die van invloed kunnen zijn op de scores, terwijl dat toch niet de bedoeling van de beoordelingsprocedure is: verbale begaafdheid, zoals die blijkt uit helderheid van formuleren, goed kunnen lezen en interpreteren van opgaven. e.d.; nauwkeurigheid van werken (wanneer nauwkeurigheid van werken van groot belang geacht wordt, dan is zij ongetwijfeld ook als object van onderwijs gekozen, en zal zij als zodanig terecht een plaats in de beoordeling kunnen krijgen; in andere gevallen mag nauwkeurigheid van werken, waar grote persoonlijke verschillen in bestaan, niet van invloed op de scores zijn) studiegewoonten (wanneer verschillen in de wijze van voorbereiding systematisch samenhangen met de hoogte van scores, dan moeten studenten over die samenhang volledig ingelicht worden, en het liefst ook geoefend worden in de 'beste' studiemethodiek; misschien moeten ook veranderingen aangebracht worden in de vorm van syllabi e.d., of de vorm van de beoordelingsprocedure zelf.). etc.etc.

maatregelen.

In het bovenstaande zijn al enige mogelijkheden genoemd van beïnvloedende factoren, en maatregelen die daartegen genomen zouden kunnen worden.

Het is meestal niet erg eenvoudig om er achter te komen of bepaalde factoren op de toetsscores van invloed zijn. Wanneer op welke gronden dan ook vermoedens rijzen over invloed van niet bedoelde factoren op de verdeling van toetsscores, zal vaak een op dat speciale geval toegesneden experiment uitgevoerd moeten worden om er achter te komen of de veronderstelde invloed inderdaad aanwezig is. In het geval van de invloed van bepaalde intelligentie verschillen bijvoorbeeld kan een intelligentie test afgenomen worden die de verbale intelligentie bijv. goed meet; blijkt er een aanzienlijk verband te bestaan tussen scores op de toets en die op de test voor verbale intelligentie, dan is er alle reden om de toets te herzien.

Een bekend verschijnsel in het onderwijs is dat scores op studietoetsen voor verschillende vakken binnen eenzelfde studierichting sterk met elkaar correleren. In die gevallen bestaat de gerede verdenking dat de tentamens een behoorlijk beeld geven van verschillen tussen studenten met betrekking tot de zorgvuldigheid waarop men zich op tentamens voorbereidt. Het spreekt vanzelf dat dan de scores niet erg veel inzicht kunnen verschaffen op het bereikt zijn van de voor ieder vak vastgestelde doelstellingen

De begripsmatige geldigheid zoals zij genoemd wordt (of: construct validiteit) is van belang zodra enigerlei specifieke interpretatie van beoordelingsgegevens gemaakt wordt. Bij iedere interpretatie moet dan ook de vraag gesteld, en zo mogelijk beantwoord, worden of de beoordelingsprocedure de interpretatie kan rechtvaardigen; of er geen alternatieve interpretatie mogelijkheden zijn die tot andere conclusies zouden kunnen leiden. Cronbach: *"And construct validation is needed to give meaning to the results to explain a poor performance, to forecast how it will handicap the student, to suggest how to teach more effectively in the future, etc."*

Data die mogelijk inzicht geven in de begripsmatige geldigheid van een bepaalde beoordelingsprocedure, zijn~de ook boven al genoemde correlaties met scores voor andere vakken. Interessant zijn dan met name correlaties met cijfers voor vakken die vergelijkbaar zijn in onderwerpen en soort intellectuele vaardigheden die vereist worden enerzijds, en geheel anderssoortige vakken anderzijds. Vindt men een geringe correlatie met cijfers behaald op gelijksoortige vakken, dan kan dat wijzen op een gebrek aan begripsmatige validiteit, evenals trouwens een grote correlatie met nogal onvergelijkbare vakken.

Let vooral op groepen studenten die op grond van de scores te onderscheiden zijn. Het is mogelijk dat een bepaalde groep studenten zich door andere onderwijsverplichtingen minder goed heeft kunnen voorbereiden, wat tot uitdrukking komt in lagere scores. In dat geval meten we in belangrijke mate het overigens triviale gegeven of iemand tegelijkertijd een belangrijke werkgroep moest volgen of niet (om maar een voorbeeld te noemen). Ook is het mogelijk dat (onbedoeld) vragen gesteld worden die eerder behandeld zijn in een onderwijssituatie waar geen opkomst plicht gesteld was. In dat geval wordt een bepaalde groep studenten in het nadeel gesteld, wat tot uitdrukking komt in de scores, die dan gedeeltelijk meten of iemand al dan niet zijn colleges gelopen heeft. Het hoeft nauwelijks betoog dat dat niet de bedoeling van een beoordelingsprocedure kan zijn: is presentie van groot belang, dan moet zij op andere wijze gewaarborgd worden.

Ook hier kan geëxperimenteerd worden. Een mogelijk experiment is twee beoordelingsprocedures op te zetten, die met elkaar vergelijkbaar moeten zijn voor wat betreft hetgeen men wil gaan beoordelen, maar die bijvoorbeeld wat de vorm betreft, of degenen die de beoordeling uitvoeren, verschillen. De groep studenten wordt volgens beide procedures beoordeeld, zodat men vervolgens kan nagaan in welke mate de beide reeksen van scores convergeren. Lopen de scores ver uiteen, dan wordt kennelijk in beide, of één van beide, procedures iets gemeten dat in belangrijke mate kenmerkend is voor die bepaalde wijze van beoordelen, en niet de variabele waarin men geïnteresseerd is: of de studenten zich de in de doelstellingen omschreven vaardigheden hebben eigen gemaakt.

Opgemerkt kan nog worden dat het bovenstaande in belangrijke mate betrekking heeft op evaluatie van het onderwijs (alsook evaluatie van de gehanteerde beoordelingsprocedure)l waarmee weer eens geïllustreerd is welke nauwe relatie er tussen beoordeling en onderwijs bestaat. Dit als contrast met de ongelukkige eenzijdige gerichtheid op het onderscheiden van slechte, en 'goede' studenten die veel beoordelingsprocedures nog kenmerkt.

geldigheid van beoordelingsprocedures voor het nemen van beslissingen over studenten.

[In 1976 komt de COWO-cursus 'Cesurbepaling' uit, de eerste coherente poging om greep te krijgen op dit fenomeen. Ik ben er in 2025 nog steeds mee bezig, zoals deze blog laat zien: <https://benwilbrink.wordpress.com/2025/05/21/waarom-is-zakken-mogelijk-bij-eindexamens-vo/> bw]

De enige beslissing, die werkelijk van belang is bij de summatieve beoordeling waarover we nog steeds spreken, is de beslissing slagen of zakken. Het onderwijs is al achter de rug, dus studieaanwijzingen zijn niet helemaal relevant meer.

Slagen of zakken. Laten we eens nagaan wat er zoal aan die beslissing vast zit.

Om eenvoudig te beginnen: veronderstel dat het de bedoeling is dat iedereen slaagt die aan de minimum doelstellingen heeft beantwoord, zoals moet blijken uit de score die het resultaat is van de beoordelingsprocedure. De vraag naar de geldigheid van deze beslissing is de vraag naar het verband tussen de scores die het resultaat van de beoordeling zijn, en het veronderstelde al of niet voldaan hebben aan de minimum eisen.

Voor beantwoording van die vraag is van belang al hetgeen in het voorgaande gezegd is over doelstellingen. Zijn doelstellingen niet expliciet geformuleerd, dan is in het geheel niet duidelijk wat nu eigenlijk dat minimum niveau is, waar het gelokaliseerd kan worden, hoe nagegaan kan worden of een student het bereikt heeft of niet, etc. Zonder expliciete omschrijving van de minimum eisen waaraan

m.b.t. kennis en vaardigheden voldaan moet kunnen worden, kunnen slechts irrationele beslissingen over zakken en slagen genomen worden. Een eerbiedwaardige traditie is geen enkele waarbor voor rationaliteit van dit soort beslissingen (als de traditie inderdaad eerbiedwaardig is, zou het mogelijk moeten zijn in een handomdraai de verlangde explicitering te geven)

- objectiviteit van de beoordelingsprocedure.

In de mate waarin zij in objectiviteit tekort schiet, kan minder vertrouwen gesteld worden in de scores op grond waarvan de beslissingen genomen worden, Hetzelfde geldt t.a.v. de betrouwbaarheid. Hoe groter de onbetrouwbaarheid in de scores, hoe groter ook het aantal foute beslissingen zal (en dat er foute beslissingen genomen worden is onvermijdelijk. Deze fouten zijn van twee soorten: studenten zakken terwijl ze dat 'eigenlijk' niet verdienen; en studenten slagen die het 'eigenlijk' niet verdienen. Hogere onbetrouwbaarheid verhoogt het aantal fouten, dat is altijd het geval. De enige manipulatie mogelijkheid bestaat in het verschuiven van aantal fouten van de ene soort naar de andere, M.a.w. wanneer het belangrijker is om te vermijden dat studenten onterecht zakken dan dat studenten onterecht slagen, dan kan de zakslaag grens z6 verschoven worden dat minder fouten van de eerste soort gemaakt worden, en dus tegelijkertijd meer fouten in onterecht laten slagen.)

- inhoudelijke geldigheid.

Het laten zakken is gerefereerd aan de inhoudelijke geldigheid: is de inhoudelijke geldigheid niet gewaarborgd, dan kan ook voor afwijzen geen overtuigende argumentatie gegeven worden.

- begripsmatige geldigheid (behandeld op blz 8,9).

Ook deze is vereist, anders zouden we studenten afwijzen op gronden die niet relevant zijn.

Uit het bovenstaande zou de indruk kunnen ontstaan dat laten zakken van studenten vaak erg moeilijk te rechtvaardigen is. Die indruk is juist. Het is natuurlijk wel mogelijk om afspraken te maken

over de grens die tussen en zakken getrokken zal worden, afspraken die erg duidelijk moeten zijn. Anderzijds is van belang de ernst van een afwijzende beslissing:

Hoe ernstiger de gevolgen van afwijzing voor de student, des te meer gewicht komt toe aan de op de vorige bladzijde genoemde overwegingen. Hetzelfde geldt overigens ook voor de onderwijsinstelling: grote aantallen afwijzingen kunnen een extra belasting van toch al schaarse faciliteiten betekenen.

Maatregelen die de omvang van het fenomeen 'zakken' kunnen beperken, en volkomen onnodige vervelende consequenties kunnen wegnemen:

Zorg er voor dat het onderwijs kwalitatief op zo'n hoog niveau komt, dat slechts bij wijze van uitzondering nog studenten op grond van de summatieve beoordeling afgewezen hoeven worden. Om dat te kunnen bereiken is niet alleen een beter gestructureerd onderwijs nodig, maar vooral tijdige informatie aan de student, bijvoorbeeld d.m.v. formatieve toetsing, waarop later nog uitvoeriger zal worden ingegaan.

Het laatste was voor de nabije toekomst. Wat direct al gedaan kan worden is het instellen van een zeer snelle herkansing. Er is niets op tegen om herkansingsgelegenheid te geven binnen een week na de uitslag. 'We kunnen daarbij bedenken dat de student vaak niet meer hoeft te doen dan nog eens wat extra aandacht schenken aan onderwerpen waarop zijn prestaties zwak waren. Voor degenen die horen tot de categorie 'onterecht afgewezen' geldt zelfs dat zij de eerste keer gewoon 'pech' gehad hebben, en zij zouden in principe zonder extra voorbereiding voor de herkansing op kunnen komen (natuurlijk weet niemand of hij behoort tot de categorie 'terecht' of 'onterecht' afgewezen, maar dat is ook niet van belang voor de strekking van ons argument). Een snelle herkansing dus.

Ij een snelle herkansing hoort natuurlijk ook een snelle uitslag van de eerste beoordeling. Bij mondelinge procedures kan vrijwel onmiddellijk de uitslag gegeven worden, Bij meerkeuze toetsing is het mogelijk gebruik te maken van ponskaarten met uitdruk

mogelijkheid, die met eenvoudige apparatuur onmiddellijk bij inlevering gescoord kunnen worden ~computer verwerking voor het verkrijgen van gegevens over de kwaliteit van de toets kan altijd later nog gebeuren). Bij schriftelijk werk zou er naar gestreefd moeten worden dit op de snelst mogelijke termijn nagekeken te hebben.

Piet vergeten mag worden dat de student die zich moet herkansen recht heeft op informatie over de aard en omvang van de tekorten in zijn prestaties. Vragen, antwoorden, beoordelingen en argumentaties voor de beoordelingen zouden hem ter beschikking kunnen, moeten, worden gesteld (de geslaagden overigens ook).

En pas op voor de verleiding om, als er toch een snelle herkansing is, dan maar wat makkelijker tot het nemen van zakbeslissingen over te gaan.

Het kan geen kwaad hier er nog eens op te wijzen! dat de beoordelingsprocedure volkomen doorzichtig moet zijn voor de student: hij moet precies op de hoogte zijn van de aard van de vragen die hij kan verwachten, de onderwerpen waarover gevraagd kan worden, de wijze waarop zijn antwoorden beoordeeld zullen worden, waar de grens tussen slagen en zakken zal liggen, en hoe groot het risico is dat hij zal zakken wanneer in feit e zijn prestaties op het niveau van laten we zeggen een 7 zouden liggen (het laatste heeft betrekking op de betrouwbaarheid van de beoordelingen) (het begrip doorzichtigheid of transparantie is ook door De Groot geïntroduceerd) [De Groot 1970, transcriptie: <https://benwilbrink.nl/publicaties/70degroot.htm>]

Formatieve beoordeling.

Formatieve beoordeling (eveneens een aan Benjamin Bloom ontleende term) heeft betrekking op procedures die de student informatie geven over zijn vorderingen tijdens het onderwijs, zodat hij nog alle gelegenheid heeft om fouten, vergissingen, te weinig inspanning, e.d. te herstellen. De summatieve beoordeling, aan het eind van het onderwijs, heeft geen of weinig waarde in het kader van dat onderwijs zelf: het is een laatste oordeel, vaak ook een Laatste Oordeel helaas. De enige belangrijke functie van een summatieve

beoordeling zou behoren te zijn een evaluatie van het onderwijs, en een brevet van bekwaamheid aan bijna alle studenten. Om dat te kunnen bereiken zijn een aantal maatregelen nodig, onder andere alle maatregelen die al in dit stuk genoemd zijn, gericht op verbetering, grotere doorzichtigheid van de beoordelingsprocedure, formulering van doelstellingen, etc. Daarnaast kan gezocht worden naar methoden om de studenten al onderweg van de nodige feedback, informatie over hun vorderingen, te voorzien, die het voor hen mogelijk maakt met grote zekerheid af te stevenen op een goede kwalificatie op de summatieve beoordeling*

In essentie is het bovenstaande een weergave van het Learning for Mastery model, gepropageerd door Bloom c.s. (zie met name Bloom, Haatinge and Madaus: Handbook on formative and summative evaluation 1971) 9Zie ook deze blog over het 2-sigma experiment <https://benwilbrink.wordpress.com/2018/09/28/benjamin-s-bloom-human-characteristics-and-school-learning/> bw] Men hoeft het echter niet met alle veronderstellingen van dit model eens te zijn, om in te zien dat een inrichting van het onderwijs zoals boven beschreven, onderwijskundig zeer gewenst is.

maatregelen.

De student met een stapel boeken naar zijn kamer sturen, hem na drie maanden een tentamen afnemen, en bij onvoldoende resultaat over drie maanden terug laten komen, is een rampzalige wijze van verzorgen van onderwijs.

Structureer de onderwijs situatie (dat hoeft niet te betekenen dat de student aan het handje door de universiteit geleid wordt), en zorg dat de student zoveel feedback over zijn studie inspanningen krijgt als gewenst is om zonder grote moeilijkheden het normale onderwijs programma te kunnen volgen.

Voor die feedback zijn vele en uiteenlopende methoden mogelijk. Voor de hand ligt om gebruik te maken van een aantal tussentijdse beoordelingen, die de student in staat kunnen stellen eventuele onvruchtbare studiegewoonten bijtijds te herstellen. Andere mogelijkheden zijn het studiemateriaal zo te structureren dat de student zijn eigen feedback kan verzorgen (geprogrammeerde

instructie is daar een extreem voorbeeld van), regelmatige bespreking in werkgroepen of studiegroepjes (die niet noodzakelijk door docenten begeleid hoeven te worden), en andere methoden die de vindingrijke docent samen met zijn studenten kan bedenken.

Het streven is naar een eindbeoordeling waar vrijwel iedereen in eerste instantie voor slaagt. Onvermijdelijk moet dat ook een eindbeoordeling zijn waarvoor heel precies is omschreven wat de eisen zijn om te slagen (zoals in de voorgaande bladzijden beschreven), zodat er geen sprake kan zijn van enigerlei verwatering van de 'kwaliteit' van het onderwijs. Al was het alleen maar omdat kwaliteit van het onderwijs niet afgemeten kan en mag worden aan de percentages afgewezen studenten.

In de mate waarin het niet lukt om het gros van de studenten in één keer te laten slagen, moet gezocht worden naar verbetering van het onderwijs materiaal en de gehanteerde methoden van onderwijs en feedback over studieprestaties. (Natuurlijk is niet uit te sluiten dat de eisen voor de summatieve toets echt te hoog zijn gesteld; daar moet discussie en overleg over mogelijk zijn)

Overleg over beoordelingsprocedures.

Eén van de verdiensten van het omstreden voorstel over selectie van de hand van De Groot is dat het sterk de nadruk legt op het overleg tussen alle betrokken partijen over de vorm en inhoud van de selectie procedure. En wat voor selectie geldt, kan in dit geval ook voor beoordelingsprocedures gelden. De Groot (biz 69 van selectie voor en in het hoger onderwijs):

'Weel minder goed functioneert in de huidige situatie het algemene overleg over examen en selectie methoden. onze indruk is althans, dat er in het hoger onderwijs weinig (protest en) overleg over de technische kwaliteiten van examens en toetsen plaats vindt terwijl uit deze stilzwijgende acceptatie bepaald niet mag worden afgeleid dat die examens en toetsen aan de validiteits- en betrouwbaarheidseisen voldoen. Goede procedures voor dit overleg zijn nog niet gevonden; een moeilijkheid is, dat studenten hiertoe weinig gemotiveerd zijn: wie geslaagd is gelooft het wel en gaat

verder, wie gezakt is komt hoogstens individueel, voor zichzelf pleiten."

Wel, het lijkt er veel op dat studenten met dit stuk in de hand heel wat lastige vragen aan hun docent kunnen stellen* Als de laatste dan maar zorgt dat hij ook op de hoogte is van de inhoud van dit stuk, dan ligt het in de lijn der verwachtingen dat er enig vruchtbaar overleg te plegen valt, overleg gericht op verbetering van de gehanteerde beoordelings methoden, waaraan trouwens ook de studenten wel een steentje zouden kunnen (en willen) bijdragen.

maatregelen.

Neem bij kritiek van studentenzijde op gehanteerde beoordelingsprocedures geen verdedigende houding aan, maar zoek samen met hen (en eventueel met inschakeling van hulp van derden, zoals het COWO) naar betere procedures.

Vraag zoveel mogelijk commentaar van studenten over de kwaliteit van de beoordelingsprocedure, van afzonderlijke vragen, over de wijze waarop antwoorden beoordeeld worden, etc. etc. Maak van die commentaar overzichtelijke notities, zodat ze over een jaar nog verwerkt kan worden in de nieuwe beoordelingsprocedures

Nodig studenten eens uit om zelf een volledige beoordelingsprocedure op te zetten, of althans daar uitgewerkte voorstellen voor te doen.

Laat studenten hun eigen vragen, toetsitems e.d. formuleren.

Geef studenten zoveel mogelijk informatie over de beoordeling die ze te wachten staat, maar ook over de vroeger gehanteerde beoordelingsprocedures, inclusief gestelde vragen en de goede antwoorden daarop.

Geef zo mogelijk alle items vrij die in studietoetsen gebruikt kunnen worden. Wanneer er te weinig items beschikbaar zijn om dat redelijkerwijs te kunnen doen, maak er dan zo snel mogelijk een grote serie bij (een goede doelstellingenformulering voor het onderwijs maakt het genereren van items relatief simpel)

Ideaal zou zijn om (sub)faculteitsgewijs deze informatie in een databank te verzamelen.

11 november 1973