In: Chris Lorenz (ed)
if you're so smart, why aren't you rich?
Universiteit, Markt & Management
Amsterdam: Boom
2008: 235-242

Evaluitis: de ziekte van de 'gecontroleerde' wetenschap[1]

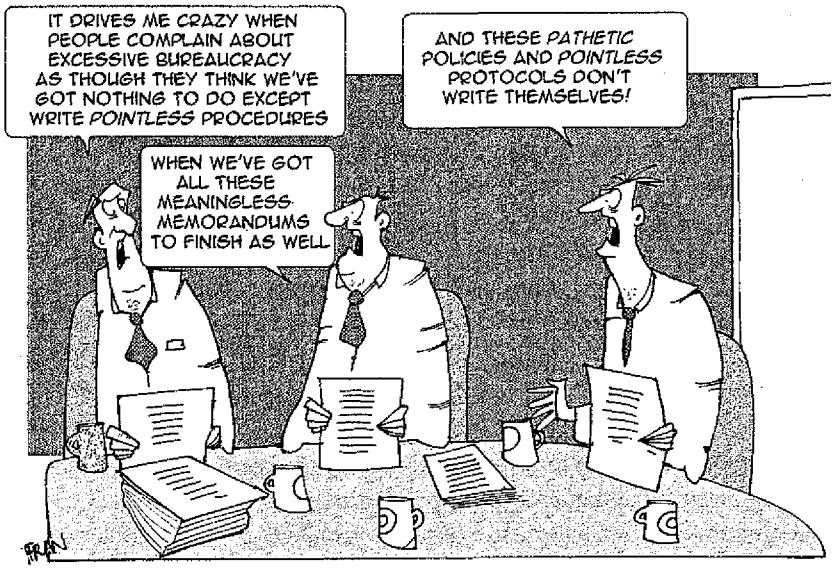
Margit Osterloh en Bruno S. Frey

Een ziekte heeft zich van de wetenschap meester gemaakt: de *evaluitis*. Vandaag de dag worden met steeds kortere tussenpozen hele universiteiten, faculteiten, afdelingen, instituten, onderzoeksgroepen en individuele onderzoekers beoordeeld. Evaluaties en daarvan afgeleide ranglijsten zijn in de wetenschap alomtegenwoordig. Als 'evaluatie' wordt hier het achteraf beoordelen van de prestaties van een organisatie of persoon door externe experts aangemerkt. Evaluaties hebben verborgen en daardoor vaak over het hoofd geziene kosten. Hun nut wordt te hoog gewaardeerd. Voor zover met deze effecten geen rekening wordt gehouden, wordt het nettoresultaat van dit instrument systematisch overschat. In dit geval worden meer evaluaties doorgevoerd dan maatschappelijk zinvol is.

Evaluaties zijn weliswaar in sommige gevallen noodzakelijk, maar ze verbeteren niet altijd het wetenschappelijke systeem. Tegenwoordig gesignaleerde problemen kunnen niet simpelweg door nog zorgvuldiger evaluaties worden opgelost; ze kunnen er zelfs nog erger door worden, want evaluaties vervormen motiverende prikkels. Ze veranderen het gedrag van de betrokken personen op een systematische en onbedoelde wijze, onafhankelijk van de vraag hoe zorgvuldig ze worden doorgevoerd.

Het nut van de beoordeling wordt overschat

Niet alle relevante aspecten van gekwalificeerd werk kunnen van tevoren worden vastgelegd of achteraf worden gemeten. Een evaluatie aan de hand van vooraf vastgelegde criteria geeft de beoordelaars aanleiding om zich overwegend op deze criteria te richten. Vaak nemen de evaluaties het aantal



publicaties als maatstaf. In dat geval zullen onderzoekers nieuwe ideeën of interessante onderzoeksdata als dunne plakjes salami afsnijden en in zo veel mogelijk magere publicaties verwerken. Ondubbelzinnig bewijs voor deze stelling is in Australië te vinden. Midden jaren negentig werden daar het salaris van de wetenschappers en de financiering van de universiteiten aan het aantal publicaties in *peer reviewed*-tijdschriften gekoppeld. Zoals te verwachten was nam het aantal publicaties dramatisch toe – maar de kwaliteit (gemeten aan het aantal citaties) verminderde dienovereenkomstig. Dit zakte zelfs onder het gemiddelde van de OECD-landen.

Het aantal publicaties kan ook verhoogd worden door wetenschappers die geen bijdrage aan het onderzoek geleverd hebben als coauteur op te voeren, waarbij die wetenschappers dan op een andere manier voor compensatie zorgen. De golf van publicaties zorgt er bovendien voor dat een leger van deskundigen moet worden ingezet om de publicaties te beoordelen. De werkdruk leidt er dan onvermijdelijk toe dat steeds oppervlakkiger beoordelingen worden geschreven of dat de beoordelingen door assistenten worden opgesteld.

Het meten van onderzoeksprestaties door middel van citaties leidt eveneens tot systematische vertekeningen. Er worden citatiekartels gevormd en de stimulans om zich te wijden aan modieuze thema's, waar veel aandacht voor is neemt toe. Het toepassen van wetenschappelijke kennis in de praktijk of op andere vakgebieden wordt daarentegen verwaarloosd, omdat publicaties in algemeen toegankelijke boeken en lezingen voor een lekenpubliek, net als activiteiten op het terrein van advisering en scholing, geen citaten in wetenschappelijke publicaties opleveren.

Wordt tenslotte het aantal begeleide promovendi als 'prestatiemeter' gebruikt, dan worden de eisen verlaagd. Als uitweg voor deze problemen wordt vaak gekeken naar het verwerven van middelen uit de derde geldstroom. Die zeggen echter niets over de zin of de productiviteit van het onderzoek dat met dit geld wordt gefinancierd. Toch is dit criterium populair omdat geldstromen bijzonder makkelijk te meten zijn. Wordt een wetenschappelijke instelling hierop beoordeeld, dan is ze gedwongen om op zoek te gaan naar middelen uit de derde geldstroom en om tegelijkertijd minder goed meetbare onderzoeks- en onderwijsactiviteiten te veronachtzamen. Dit criterium is echter voor vele vakgebieden onzinnig, vooral op het gebied van de geesteswetenschappen. In de tweede plaats ontstaan systematische prikkels om *te veel* onderzoeksgelden aan te vragen en om *in*efficiënt onderzoek te doen zodra de omvang van de derde geldstroom als doorslaggevend criterium van 'prestaties' wordt beschouwd.

Deze voorbeelden laten zich eenvoudig vermenigvuldigen. Het resultaat zijn steeds omslachtigere evaluatieprocessen, die op de beroemde wedstrijd tussen de egel en de haas lijken. Er ontstaat een *ratrace* die de wetenschap niet verbetert, maar die alleen tot hogere kosten leidt. De kosten van de voor evaluaties benodigde bureaucratie en de opkomende evaluatie-industrie zijn nu al enorm en worden door de koppeling van de beloning aan 'prestaties' nogmaals verhoogd. Bovendien hebben de onderzoekers steeds minder tijd voor hun eigenlijke taken. Ze worden gedwongen om permanent óf te evalueren, óf geëvalueerd te worden.

Daar komt nog bij dat juist baanbrekend onderzoek vaak tegen de heersende wetenschappelijke mening ingaat. Dit type onderzoek wordt daarom in eerste instantie slecht beoordeeld en daar komt soms pas na tientallen jaren verandering in. Goed of zelfs revolutionair wetenschappelijk onderzoek onderscheidt zich doordat het nieuwe criteria genereert en tegen het heersende paradigma moet opboksen. Talrijke voorbeelden zijn te vinden in de literatuur die aansluit bij Thomas Kuhns werk over 'wetenschappelijke revoluties' of bij Ludwig Flecks werk over 'wetenschappelijke denkcollectieven'.

De beoordelingscriteria kunnen gemanipuleerd worden

Toponderzoek heeft tijd nodig om tot resultaten te komen die beoordeeld kunnen worden en het duurt nog langere tijd totdat haar betekenis binnen de hoofdstroom van de wetenschap wordt erkend. Bij een beoordeling op basis van kortlopende publicatie- en citatiescijfers zouden heel wat baanbrekende onderzoekers weinig kans hebben gehad. Er zijn sowieso heel wat voorbeelden waarin de *scientific community* vernieuwing eerder belemmert dan stimuleert. Een bijzonder dramatisch voorbeeld is Ignaz Semmelweis, de ontdekker van de kraamvrouwenkoorts. Hij moest meer dan dertig jaar wachten op de erkenning en de toepassing van zijn onderzoeksresultaten, die duizenden vrouwen het leven had kunnen redden.

Als een indicator belangrijk wordt voor de eigen positie, stimuleert dat om de indicator in het eigen voordeel te beïnvloeden. De directies van scholen kunnen de beoordeling van hun school verbeteren door de scholieren op specifieke examenvragen voor te bereiden en door slechte leerlingen onder allerlei voorwendsels van de betreffende tests uit te sluiten. Sommige managers beïnvloeden de prestatiecriteria zodra hun inkomen ervan afhangt.

Ze drijven op korte termijn de aandelenkoersen op wanneer ze een deel van hun salaris in de vorm van aandelenopties ontvangen.

Zulke vormen van manipulatie zijn ook in de wetenschap verbreid sinds de onderzoeksprestaties in het kader van evaluaties aan de hand van kwantitatieve criteria worden gemeten. Zo trekken universiteiten graag wetenschappers aan die goed op de betreffende criteria 'scoren' om zo goed bij evaluaties en *rankings* voor de dag te komen. Universiteiten ruziën onderling over de vraag wie Nobelprijswinnaars mogen claimen. Voor het wetenschappelijk klimaat is vooral dit ophemelen van onderzoeksresultaten schadelijk. Dit stimuleert namelijk het uitsluitend publiceren van succesvolle testresultaten en het verzwijgen of zelfs verdoezelen van negatieve resultaten – en dat terwijl het falsificeren van hypotheses tot de kerntaken van de wetenschap behoort. Volgens een onderzoek heeft niet minder dan een derde van de Amerikaanse wetenschappers zich oneerlijk of zelfs frauduleus gedragen door ideeën van anderen over te nemen zonder naar behoren te citeren of door ongewenste onderzoeksresultaten achter te houden.

Nog erger is het vervalsen van onderzoeksresultaten. In experimenten is aangetoond dat personen die zich gecontroleerd voelen in een veel hogere mate bereid zijn om te bedriegen. Recente schandalen laten zien dat dit ook voor de wetenschap van toepassing is.

De met de evaluatie verbonden prestatiebeoordeling beïnvloedt de motivatie om te werken negatief wanneer de betrokkenen het gevoel hebben dat deze evaluatie als controle bedoeld is. Er treedt een verdringingseffect op waardoor de intrinsieke motivatie om te werken afneemt en de extrinsiek bepaalde stimuli belangrijker worden. De totale productie – gemeten aan de vooraf vastgelegde criteria – hoeft niet per se te verminderen en kan zelfs toenemen. Het valt echter te betwijfelen of de gevolgen voor de kwaliteit en de originaliteit van het onderzoek gunstig zijn. Creatief onderzoek kenmerkt zich juist doordat het nieuwe maatstaven creëert, die zich soms maar langzaam door weten te zetten. Creatief onderzoek is daarom in hoge mate aangewezen op de persoonlijke motivatie, die om die reden beter niet ondergraven kan worden. Bovendien geeft de permanente evaluatie een negatief signaal af. Ze toont dat het vertrouwen is opgezegd dat wetenschappers uit eigen beweging goede prestaties leveren in onderzoek en onderwijs. Dit opzeggen van het vertrouwen kan uitsluitend resulteren in een afnemende loyaliteit aan de instelling waar men werkt.

Er zijn tot op heden nauwelijks empirische onderzoeksresultaten voorhanden over de samenhang van prestatiecontrole en prestaties in de wetenschap omdat het meten van deze prestaties erg lastig is. Uit elders gewonnen

inzichten kan echter de conclusie worden getrokken dat onder bepaalde omstandigheden een variabele financiële beloning tot een geringere bereidheid leidt om te presteren dan een als eerlijk beschouwde vaste beloning, terwijl niet-financiële beloningen (bijvoorbeeld minder onderwijslast) die motivatie niet of nauwelijks verminderen. Niet-verwachte financiële beloning (zoals schenkingen) lijken de intrinsieke motivatie niet te verdringen, symbolische beloningen (bijvoorbeeld onderscheidingen) vergroten deze.

Wanneer instituties of personen aan een evaluatie worden onderworpen, kunnen deze zich daar niet tegen verzetten, ook niet wanneer ze ervan overtuigd zijn dat zo'n evaluatie niet geschikt is voor hun situatie. Doorgaans wordt hen dan voor de voeten geworpen dat ze bang zijn voor de uitslag. Omdat de evaluatie meestal hand in hand gaat met de verdeling van middelen moeten ze tegen beter weten in aan de evaluatie meewerken. Ze doen er zelfs goed aan om *enthousiast* mee te werken. Op die manier wordt een instemming gesuggereerd, die in werkelijkheid helemaal niet bestaat.

Wanneer ze vervolgens bij de evaluatie positief beoordeeld worden, zijn ze dienovereenkomstig verheugd en hopen ze op de bij de positieve evaluatie horende ruimere toewijzing van middelen. De verliezers zullen daarentegen meer moeite doen om zich tegen de gevolgen van de evaluatie teweer te stellen. Daar zijn altijd argumenten voor te vinden: overbelasting door onderwijsverplichtingen en bestuurlijke taken, te weinig middelen of gewoon pech. Achteraf wordt dan geprobeerd om de criteria in het eigen voordeel anders af te wegen.

Evaluaties brengen niet alleen verborgen kosten met zich mee, die meestal over het hoofd worden gezien, maar hun nut wordt doorgaans schromelijk overschat. De door evaluatie verkregen informatie draagt weinig bij aan de verbetering van beslissingen over het toewijzen van middelen voor wetenschappelijk onderzoek. Juist in de *scientific community* is meestal al bekend welke instituten en personen volgens de gangbare criteria bijzonder goed of slecht onderzoek doen. Daar komt nog bij dat de gehanteerde methodes alleen in het hoogste en in het laagste prestatiesegment eensluidende en betrouwbare resultaten opleveren. Voor het middensegment – waarover informatie het hardst nodig is – differentiëren ze op een onbetrouwbare manier.

Evaluaties proberen in de regel het bestaande prestatieniveau in beeld te brengen. Voor politieke beslissingen is die informatie echter meestal niet zo belangrijk. Welke conclusies eruit getrokken kunnen worden is maar de vraag. Moeten slecht beoordeelde instituten en onderzoekers minder middelen tot hun beschikking krijgen? Of moeten ze juist aanvullende middelen

toegekend krijgen zodat ze hun kwaliteit kunnen verbeteren? Moeten omgekeerd de positief beoordeelde instituten en onderzoekers minder middelen tot hun beschikking krijgen omdat ze toch al succesvol zijn? Leiden aanvullende middelen bij hen nauwelijks tot verdere verbetering?

Een zinvolle evaluatie zou de marginale effecten van een verandering in de beschikbare middelen in kaart moeten brengen. Wat zou er gebeuren wanneer een instelling of onderzoeker over meer, dan wel minder middelen zou beschikken? Deze vraag is erg moeilijk te beantwoorden, want het antwoord hangt van een groot aantal factoren af. Bovendien blijft onzeker op welke manier zulke resultaten in het politieke debat gebruikt worden. Dat bewijzen ook de reacties op de zogenaamde 'Exzellenzinitiative' van de Bondsrepubliek Duitsland en haar deelstaten, een breed opgezet programma dat de universiteiten tot 'topprestaties' moet stimuleren.

Ondanks alle twijfelachtige aspecten van evaluaties zou men kunnen argumenteren dat er geen alternatieven zijn. Maar alternatieven zijn er zeker.

Zinvolle alternatieven bestaan zeer zeker

Met een veranderde institutionele vorm van de wetenschappelijk wereld zouden de permanente evaluaties teruggedrongen en ten dele zelfs vervangen kunnen worden. Wanneer universiteiten stevig met elkaar concurreren is een evaluatie van de kant van de staat overbodig. De studenten kiezen dan die universiteit die volgens hen de beste prestaties levert. De universiteiten hebben de vrijheid om díe studenten uit te kiezen die het beste aan hun ctiteria voldoen en die de reputatie van de universiteit ten goede komen.

Het kan zijn dat studenten hun keuze graag op evaluaties en ranglijsten baseren. Er is een veelvoud van ranglijsten op de markt, die allemaal in meer of mindere mate bepaalde vragen beantwoorden, maar die de onderzoeksprestaties slecht of helemaal niet meten. Ook hier is concurrentie tussen de verschillende ranglijsten beter dan een poging een door de politiek gewenste superranglijst op te stellen, zoals de Duitse Raad voor Wetenschap (Deutscher Wissenschaftsrat) momenteel doet. Ook een zorgvuldig opgestelde 'superranglijst' kan niet verhinderen dat deze de beschreven perverse verandering van motiverende prikkels tot gevolg heeft.

De tegenwoordig gebruikelijke evaluatie achteraf van wetenschappelijke instellingen kan goeddeels vermeden worden wanneer onderzoekend en onderwijzend personeel zorgvuldig geselecteerd wordt. Deze strategie gebruikt de beschikbare middelen met het oog op de toekomst doordat deze

het gewicht op de selectie legt. Daarbij moeten zonder meer de gebruikelijke criteria, zoals het aantal publicaties en de kwaliteit van deze publicaties, worden gehanteerd. Zij garanderen dat aan de wetenschappelijke normen is voldaan en ze geven een indicatie voor de mogelijkheden van de kandidaten. Is iemand eenmaal op basis van strenge criteria tot professor benoemd voor een bepaald terrein van de wetenschap, dan moet die persoon vertrouwen krijgen. Daarom zijn benoemingsprocedures veruit de belangrijkste evaluerende activiteit binnen een wetenschappelijke instelling. Op basis van zorgvuldige selectie kan men erop rekenen dat de benoemde personen de verwachte prestaties leveren, ook zonder de voortdurende dreiging van evaluaties. Sommige van de geselecteerden zullen minder gaan presteren, maar anderen zullen juist door de geboden ruimte gemotiveerd raken om topprestaties te leveren. In de wetenschap moet dat laatste het zwaarst wegen.

Individuele onwilligen en mislukkingen moeten als noodzakelijk kwaad worden gezien, zodat het wetenschappelijke systeem *als geheel* topprestaties kan leveren. Daarentegen kunnen voortdurende evaluaties, in het bijzonder de op resultaat georiënteerde evaluaties, slechts middelmaat garanderen. De als voortdurende controle ervaren beoordelingen bevoordelen slechts een 'normale' wetenschap zonder topprestaties. Onderzoekers als Albert Einstein of Max Planck in de bètawetenschappen en John Maynard Keynes of John Hicks in de economie zouden in het huidige systeem van permanente evaluatie waarschijnlijk niet erg succesvol zijn geweest.

Maar evaluaties van onderzoeksinstellingen kunnen niet helemaal worden vermeden, omdat er anders geen criteria zijn voor de verdeling van de middelen. Deze moet echter vooral proces- en niet resultaatgeoriënteerd plaatsvinden, zoals de Duitse Raad voor de Wetenschap dat tot nu toe deed. De belangrijkste criteria zijn daarbij of zorgvuldige aanstellingsprocedures gegarandeerd zijn en of een hoge mate van autonomie in het onderzoeksproces is gegarandeerd. Op die manier wordt niet alleen rekening gehouden met de bijzondere problemen bij het beoordelen van prestaties in de wetenschap, maar worden ook meteen de belangrijkste stimulansen voor onderzoekers – autonomie en een inspirerende wetenschappelijke omgeving – gecreëerd.

Een volgens deze criteria vormgegeven systeem heeft de Duitstalige wetenschap in het verleden wereldberoemd gemaakt. Het bestaat nog altijd in academische bolwerken als Harvard University, die men in andere gevallen zo graag als voorbeeld neemt.