Koninklijke Nederlandse Akademie van Wetenschappen

Kwaliteitszorg in de wetenschap

Van SEP naar KEP: Balans tussen rechtvaardigheid en eenvoud

KNAW-commissie kwaliteitszorg

AMSTERDAM, MAART 2008

Kloveniersburgwal 29, 1011 JV Amsterdam
Postbus 19121, 1000 GC Amsterdam
T 020 551 07 00
F 020 620 49 41
E knaw@bureau.knaw.nl
www.knaw.nl

Voor het bestellen van publicaties: Afdeling Communicatie, 020 - 5510 726

ISBN 978-90-6984-554-8

Het papier van deze uitgave voldoet aan ∞ iso-norm 9706 (1994) voor permanent houdbaar papier.

Inhoud

1. Is de kwaliteitszorg te intensief? 6

2. Recente ontwikkelingen in buiten- en binnenland 9

3. Onderzoeksevaluatie geanalyseerd 12

4. Balans tussen rechtvaardigheid en eenvoud 17

5. Aanbeveling: van SEP naar KEP 19

Literatuur 22
Bijlage 1 Overzicht van de evaluatiepraktijk 24
Bijlage 2 Peer review en bibliometrie, voor- en nadelen 26
Bijlage 3 Samenstelling en opdracht van de commissie 28

1. Is de kwaliteitszorg te intensief?

1.1 Inleiding

Op veel terreinen van overheidsbeleid is het evalueren van prestaties en effecten steeds belangrijker geworden. Het wetenschapssysteem is hiervan niet uitgezonderd, in Nederland noch in andere landen. Dat geldt voor het systeem als geheel, in het bijzonder de instrumenten die worden gebruikt om het systeem te beïnvloeden, als voor de afzonderlijke instituties in het hoger onderwijs en het wetenschappelijk onderzoek. Geen wonder dus dat men bij het nadenken over evaluatie en kwaliteitszorg al snel wordt geconfronteerd met de vraag of het niet allemaal teveel van het goede is, en als er wordt geëvalueerd, of dat wel op de meest efficiënte en effectieve wijze gebeurt. Dat zijn dan ook centrale vragen voor de commissie kwaliteitszorg van de KNAW. Deze commissie heeft in het afgelopen jaar een analyse gemaakt van de beoordeling van kwaliteit van wetenschappelijk onderzoek, deels gebaseerd op literatuuronderzoek, deels op eigen ervaringen en presenteert hier haar bevindingen en conclusies. De commissie heeft geconstateerd dat enerzijds het huidige bestel van evaluaties door velen als te complex en te belastend wordt ervaren (ondanks de bedoeling van het SEP om juist de lasten te verlichten), terwijl anderzijds er te weinig flexibiliteit in evaluaties zit waardoor sommige vakgebieden niet goed kunnen worden beoordeeld.

1.2 Verschillende dimensies van kwaliteit

Het overkoepelende doel van kwaliteitszorg en evaluatie is om het wetenschapssysteem optimaal te laten functioneren in het licht van wetenschappelijke en maatschappelijke doelstellingen. Die komen er kort gezegd op neer dat onderzoek wetenschappelijk gesproken van de hoogste kwaliteit moet zijn, en maatschappelijk gesproken zo nuttig en relevant mogelijk. Maar binnen het wetenschapssysteem functioneren vele actoren die niet allemaal dezelfde behoeften hebben en dezelfde doelen nastreven. Zo is de overheid in het algemeen op zoek naar gegevens die het mogelijk maken op nationaal niveau beslissingen te nemen over financiering, prioritering en eventuele reallocaties. Onderzoekers willen graag weten hoe ze het doen ten opzichte van de (internationale) concurrentie. Lokale bestuurders willen weten of de missie van het instituut wordt waargemaakt en of sommige groepen betere resultaten halen dan andere. Externe sponsoren zijn benieuwd of het doel waarvoor ze geld hebben gegeven wordt behaald. Evaluatie van onderzoek is dus zelden ééndimensionaal, evaluatie betekent vrijwel altijd ‘multi-target-’ en ‘multicriteria-evaluatie’. Het begrip kwaliteit kan dus in meerdere dimensies worden geoperationaliseerd. Het zou eenvoudig zijn als zou blijken dat de verschillende dimensies uiteindelijk samenvallen, maar dat lijkt onwaarschijnlijk.

1.3 Complexiteit van de kwaliteitszorg leidt tot extra belasting van onderzoekers

Het wetenschapssysteem is in de afgelopen decennia steeds complexer geworden en daardoor ook veeleisender voor onderzoekers. Zij merken dit onder meer doordat ze steeds meer externe financieringsbronnen moeten aanboren. Naast NWO en het bedrijfsleven, zijn er veel andere externe bronnen die de laatste decennia in belang zijn toegenomen, zoals de ministeries, de Europese Commissie en de maatschappelijke fondsen. Twintig jaar geleden was gemiddeld hooguit 20 tot 25% van de financiering van wetenschappelijk onderzoek afkomstig van externe bronnen, nu is dat gemiddeld zo'n tien procentpunten meer (A. Versleijen (red) 2007). In sommige onderzoeksgebieden is het lager, maar in andere gebieden is het meer dan de helft, soms wel 80%. De groeiende invloed van buitenaf heeft ook gevolgen voor de evaluatie en kwaliteitszorg. Dat betreft niet alleen de beoordeling van de kwaliteit van individuele onderzoekers, onderzoeksgroepen of projecten, maar ook die van het systeem als geheel. En het gaat niet alleen meer om de intrinsieke kwaliteit maar ook om de maatschappelijke relevantie. Externe financiers, die een steeds groter aandeel in de financiering van onderzoek voor hun rekening nemen, hebben ieder hun eigen verwachtingen ten aanzien van de resultaten. De kwaliteitszorg neemt daardoor tevens de vorm aan van het evalueren van de missie en effectiviteit van instituties en organisaties en van 'bottleneckanalyse' (Arnold 2004; Merkx, van den Besselaar 2008). Die 'bottlenecks' kunnen liggen in verschillende onderdelen van het systeem, zoals de financieringsvormen, de programmering, de organisatie van onderzoek en van loopbanen, de opleiding van onderzoekers en overigens ook in het evaluatiesysteem zelf. Het wetenschapssysteem is dus in toenemende mate 'evaluatie-intensief'. Of het nu gaat om agendering en programmering, allocatie van middelen, selectie van onderzoekers en onderzoeksgroepen, organisatie van onderzoek, beoordeling van de resultaten van onderzoek, overal speelt evaluatie een grote rol. Er wordt veel geëvalueerd, te veel en te vaak volgens velen in de onderzoekswereld. En er wordt op veel verschillende manieren geëvalueerd, en met nogal wat verschillende methoden. Dit alles leidt tot een toenemende belasting van onderzoekers, die ten koste gaat van primaire onderzoekstaken.

Daar staat tegenover dat de kwaliteit van het onderzoek sinds het invoeren van landelijke evaluaties (te beginnen met de Voorwaardelijke Financiering van het midden van jaren tachtig van de vorige eeuw) naar de mening van de commissie vooruit is gegaan, en dat de evaluaties daaraan zeker hebben bijgedragen. Dat betreft dan vooral kwaliteit in de zin van de vergelijking van onderzoek in internationaal perspectief, en van de transparantie van prestaties.

1.4 Tendens naar eenvoud en ontdubbeling

Er bestaat helaas niet één magische methode om al het onderzoek te evalueren en alle vragen te beantwoorden. Wel zien we een tendens naar systematisering van evaluatie op nationaal niveau. In Nederland en Engeland bestaan er al een aantal decennia systematische evaluaties op nationaal niveau, die overigens met enige regelmaat worden aangepast als gevolg van discussies over het functioneren er van. In andere landen is de trend van nationale evaluaties van recentere datum, bv. in Australië het

Research Quality Framework en in Frankrijk het Comité National d'Evaluation de Recherche (CNER).

De overeenkomst in deze systemen is gelegen in de wens onderzoeksevaluatie te vereenvoudigen (in de zin van minder administratieve lasten) en meer effectief te maken (doordat de resultaten op enigerlei wijze in nationaal perspectief kunnen worden geplaatst).

Om tot een onderbouwd oordeel over evaluatie en evaluatiedruk te komen zou men zich op zijn minst vragen moeten stellen over:

- Het waarom van de evaluatie: wat wordt ermee beoogd en wat gebeurt er met de uitkomsten?
- Het doel van de evaluatie: wil men een oordeel over de wetenschappelijke kwaliteit (ex post), de belofte voor de toekomst (ex ante), de impact van het onderzoek op de samenleving en de maatschappelijke relevantie?
- Het aggregatieniveau. Evalueren we een individu, groep, faculteit, of instituut / instelling? Of bijvoorbeeld een interuniversitaire onderzoekschool?
- De criteria: zijn deze relevant en geschikt voor het doel?
- De methode: welke (meet)instrumenten en indicatoren worden gebruikt, zijn ze valide en betrouwbaar?
- En belangrijk: Wat zijn de effecten van de evaluaties op het wetenschapssysteem als geheel, of op specifieke vakgebieden? Heeft het evaluatieproces geleid tot de gewenste verbeteringen en worden de resultaten gebruikt voor beleidsbeslissingen?

Deze vragen zijn het afgelopen jaar in verschillende vormen in de commissie Kwaliteitszorg aan de orde geweest.

In de navolgende hoofdstukken wordt een aanzet gegeven tot een systematische benadering van het evaluatievraagstuk. Eerst verbreden we de horizon met een korte blik op het buitenland en de huidige situatie in Nederland met het Standard Evaluation Protocol (SEP). Vervolgens gaan we in op een aantal methodische aspecten van evaluatie en tenslotte behandelen we de meest pregnante problemen en geven we mogelijke oplossingsrichtingen.

2. Recente ontwikkelingen in buiten- en binnenland

2.1 Inleiding

Hoewel de bovengeschetste ontwikkeling naar intensivering van evaluatie overal speelt, zien we dat er verschillen bestaan tussen landen, zowel in aanpak als in implementatie van de uitkomsten van de evaluaties. Overkoepelende evaluatiesystemen op nationaal niveau, zoals in Nederland het SEP, zijn ook in een aantal andere landen te vinden (bv. UK, Australië). In andere landen (Duitsland, Frankrijk, de VS) is er een meer gevarieerd aanbod. En in veel landen is er nog geen sprake van een ver ontwikkeld evaluatiesysteem (een aantal Oost-Europese landen, en bv. Ierland, dat pas in de laatste jaren intensief is gaan investeren in het wetenschappelijk onderzoek).

2.2 Evaluatie buiten Nederland

In de VS is de ranking van universiteiten de dominante evaluatievorm; deze beïnvloedt in hoge mate de werfkracht van universiteiten voor studenten, onderzoekers en onderzoeksmiddelen. Daarnaast evalueren de federal agencies wat er met hun onderzoekfondsen gebeurt, of het specifieke onderzoek een overheidstaak is en of de wetenschappelijke en maatschappelijke doelstellingen van de programma's worden behaald (Michelson 2006). Deze evaluaties, en vooral ook de politieke besluitvorming over onderzoeksbudgetten leiden, tot grote schommelingen in de onderzoeksbudgetten van de verschillende agencies, en dus voor de verschillende onderzoeksthema's.

Het Engelse systeem van de Research Assessment Exercise (RAE) heeft in de loop der tijd verschillende vormen aangenomen en is inmiddels in zijn vijfde fase beland sinds het in 1989 werd ingevoerd (en de zesde fase volgt in de loop van 2008). Volgens Barker (2007) is een belangrijk effect geweest dat onderzoeksgelden zijn herverdeeld naar een klein aantal universiteiten die uitblinken in de traditionele academische disciplines, dit ten koste van multidisciplinair en toegepast onderzoek. Ook heeft de RAE volgens Barker geleid tot een transfermarkt van goede onderzoekers en zijn nogal wat sub-top onderzoeksgroepen verdwenen. In de discussie over de nieuwe (zesde) fase waarin 'research impact' het belangrijkste criterium lijkt te worden, gaan stemmen op die beweren dat de RAE te ver is doorgeschoten, voornamelijk omdat regionale innovatiesystemen verslechteren doordat een bepaald type onderzoek en bepaalde opleidingen zijn verdwenen. Ook is het moeilijk voor nieuwe groepen zonder nog een duidelijke impact, om aan de eisen van het nieuwe RAE te voldoen.

In Australië is onlangs een nieuw Research Quality Framework gelanceerd waarmee al het publiekgefinancierde onderzoek ($5 miljard per jaar) in de komende twee jaar zal worden geëvalueerd. Via dit framework wordt niet alleen de kwaliteit van het

wetenschappelijk onderzoek gemeten, maar ook de 'benefits to the wider community'. Het RQF is buitengewoon uitgebreid en heeft voor een groot aantal vakgebieden specifieke protocollen die beogen aan te sluiten bij wat gewoon is in die gebieden, in termen van output en interactie met de samenleving. Daarnaast spant de overheid zich in om de toegankelijkheid van onderzoek voor gebruikers te verhogen. Een van de manieren is om in de beoordelingspanels niet alleen wetenschappers te laten plaats nemen, maar ook gebruikers.

2.3 Europese ontwikkelingen

In Europa tekent zich op het niveau van de EU ook een beweging af naar een grotere evaluatie-intensiteit. Enerzijds simpelweg omdat er steeds meer geld te verdelen valt (KP7 heeft bijna twee keer zoveel middelen als KP6), anderzijds, zoals blijkt uit het zeer ambitieuze Lissabon (2000) verdrag, omdat er grote politieke druk is om de al dan niet vermeende kennisparadox op te lossen. Europese leiders stelden als doel dat de Europese Unie binnen 10 jaar "the most dynamic and competitive knowledge-based economy in the world" zou zijn. Daartoe zou 3% economische groei nodig zijn en 20 miljoen nieuwe banen. Wetenschap en technologie moeten daarbij een centrale rol spelen. Van wetenschap wordt in die context niet alleen verwacht dat ze van hoge kwaliteit is, maar ook dat ze meer bijdraagt aan de kenniseconomie en aan de oplossing van maatschappelijke vraagstukken (bv. milieu, migratie, water) (Godin 2005). Er zijn (of worden) verschillende evaluatienetwerken actief waarin met regelmaat wordt gesproken over Europese ontwikkelingen op dit gebied, zoals het RTD-evaluation network (uit KP6), de ESF membership fora (over 'evaluation of funding schemes and research programmes' en over 'peer review'), ALLEA Working Group 'Evaluating for Science'. Het eerste netwerk bestaat al een jaar of tien en is vooral bedoeld voor informatie-uitwisseling; de andere drie zijn recent opgericht, waardoor de plannen nog in een voorlopig stadium zijn. De KNAW is in deze netwerken vertegenwoordigd en speelt in het ALLEA-verband een leidende rol. Ook de OESO is actief op het gebied van evaluatie, vooral door middel van conferenties.

2.4 Nederland

In Nederland zijn onderzoeksevaluatie en kwaliteitszorg over het algemeen ver ontwikkeld, en Nederland wordt dan ook vaak als positief voorbeeld genoemd. De drie belangrijkste organisaties, KNAW, NWO en VSNU trekken al lang gezamenlijk op en namen in 1999 het initiatief tot de oprichting van een commissie die in 2001 een rapport uitbracht over de contouren van een nieuw stelsel van kwaliteitszorg. Dat rapport heeft vervolgens geleid tot het sinds 2003 alom gebruikte Standard Evaluation Protocol (SEP) en de Meta-evaluatiecommissie (MEC), die toezicht houdt op het gebruik van het SEP door de Nederlandse onderzoeksinstellingen. Het SEP kent vier beoordelingscriteria (kwaliteit, productiviteit, relevantie en vitaliteit) en richt zich daarmee op een brede beoordeling (d.w.z. inclusief maatschappelijke relevantie en managementaspecten) van relevante eenheden van onderzoek, waarbij de instellingen vrij zijn in de keuze van de te beoordelen eenheden. Dit heeft geleid tot een waaier aan te beoordelen eenheden, variërend van hele disciplines tot (delen van)

faculteiten, onderzoeksinstituten of onderzoekscholen. Deze eenheden presenteren zich via een zelfevaluatie waarin ze een beschrijving kunnen geven van het eigene van hun onderzoekspraktijk, en eventueel van bijzondere lokale omstandigheden. De besturen van de instellingen benoemen de evaluatiecommissies, die internationaal zijn samengesteld. Het SEP en de zelfevaluatie zijn, meer dan vorige systemen, gericht op een dialoog met de beoordelaar en de instelling (dan op een afstandelijk oordeel) en op verbetering van het onderzoek. In de literatuur wordt in dit verband wel een onderscheid gemaakt tussen een 'jury' model van beoordeling en een 'coach' model.

Hoewel breed geaccepteerd en gebruikt, blijft er ook kritiek op het SEP bestaan. Mede naar aanleiding van het recente rapport van de Meta-evaluatiecommissie (MEC), lijkt de aanvankelijke waardering voor het systeem deels plaats te maken voor onvrede, onder meer omdat de MEC weinig inzicht kan bieden in wat er binnen de instellingen gebeurt met de resultaten van evaluaties (MEC rapport: 12). De meta-commissie is er niet in geslaagd informatie hieromtrent boven tafel te halen. Eerder onderzoek van de Commissie Dynamisering wees uit dat universiteitsbesturen wel rekening houden met de resultaten van evaluaties, maar dat dat niet betekent dat heel goed en excellente groepen worden beloond. Voor zover de commissie heeft kunnen nagaan is het alleen bij de KNAW gebruikelijk om van elke instituutsevaluatie een dossier op de website te publiceren, inhoudende het evaluatierapport zelf en de bestuurlijke commentaren, waaronder een standpunt van het bestuur van de Akademie.

Een tweede reden tot zorg is dat het SEP regelmatig door subsidiegevers wordt omzeild en eigen protocollen worden gehanteerd, die afwijken van het SEP. Dit leidt tot een stapeling van procedures, die elk enigszins van elkaar verschillen en dus tot een extra zware belasting van onderzoekers, die hun tijd daardoor minder kunnen besteden aan het primaire onderzoeksproces.

3. Onderzoeksevaluatie geanalyseerd

3.1 Inleiding

We behandelen hieronder aspecten van kwaliteitszorg die in de huidige discussie over intensivering een rol spelen. Eerst geven we een overzicht van wat er in Nederland wordt geëvalueerd, om zodoende een beeld te krijgen van de 'stapeling' van evaluaties. Vervolgens gaan we in op een aantal methodische aspecten en tenslotte behandelen we verschillen tussen vakgebieden met betrekking tot evaluatie.

3.2 Stapeling van evaluaties

Stapeling van evaluaties is tot op zekere hoogte onvermijdelijk, immers evaluatie en kwaliteitszorg vormen een routinematig onderdeel van de onderzoekspraktijk. Het selecteren van personeel, het bevorderen van medewerkers, het selecteren van projectaanvragen, het reviewen van papers, dat zijn allemaal 'standaard' evaluatieve momenten. Daarnaast zijn er evaluaties die een zwaardere beleidsmatige component kennen, zoals (her)erkenningen van onderzoekscholen en de evaluaties volgens het SEP protocol. Andere voorbeelden zijn: informatieverzameling t.b.v. verkenningen of gebiedsspecifiek beleid, bv. de opzet van Innovatiegerichte Onderzoek Programma´s (IOPs); beleidsgerichte gebiedsevaluaties; evaluatie van BSIK-, FES- en andere grote investeringsvoorstellen (roadmap grote faciliteiten, genomics-zwaartepunten, etc.); de onderwijsevaluaties; de accreditatieprocedures van opleidingen, zoals ook de research masters. Een actieve faculteit of andere eenheid kan meerdere keren per jaar aan een of andere vorm van evaluatie onderworpen zijn.

Bij dergelijke evaluaties wordt vaker de vraag gesteld of zij wel nodig zijn als de functies van het wetenschapssysteem goed functioneren. Bijvoorbeeld, als de projectselectie door onderzoeksponsors (NWO, KNAW, OCW, EZ, andere departementen, de Europese Commissie, collectebusfondsen, etc.) goed werkt, blijkt dan niet 'vanzelf' wat de goede onderzoeksgroepen zijn? Tegelijkertijd is er ook veel kritiek op projectselectie, bv. dat er een systematische bevoordeling plaatsvindt van bepaalde vakgebieden, of binnen vakgebieden van bepaalde theoretische of methodische oriëntaties. Bovendien maken bepaalde vormen van onderzoek (bv. gestandaardiseerd experimenteel onderzoek) het mogelijk veel sneller artikelen te publiceren dan andere, waardoor de kans op hogere cijfers in de beoordeling toeneemt.

Bij evaluaties spelen ook het aggregatieniveau en de tijdsdimensie een rol:

– Van micro (individuele onderzoekers of onderzoeksgroepen) tot macro (universiteiten, de adviesstructuur) niveau.
– Evaluatie ex ante (onderzoeksplannen: project- en programmavoorstellen; de onderzoeksportfolio; verkenningen), of ex post (uitkomsten van onderzoek: projecten en programma's, manuscripten, promoties, visitatie).
– Met korte termijn horizon (financiering van een groep) of met een lange termijn horizon, zoals beslissingen over investeringen in onderzoeksinfrastructuren of

de samenstelling van de Nederlandse onderzoeksportfolio die de agenda voor de komende tien jaar bepalen.

Zoals in het vorige hoofdstuk is aangegeven neemt de intensiteit van evaluaties toe. Dat geldt niet alleen voor het jaarlijks bijhouden van productie en performance indicatoren ('monitoring'), maar ook voor zelfevaluaties, midterm evaluaties etc. De stapeling van evaluaties knelt des te meer omdat het niet mogelijk blijkt de levering van informatie voor de verschillende evaluaties te uniformeren. Het telkens net weer iets anders moeten aanleveren voor verschillende evaluaties zorgt voor een extra belasting van de onderzoekers. Dit alles in ogenschouw genomen lijken de klachten over het stapelen van evaluaties niet ongegrond. In bijlage 1 zetten we alle evaluaties eens op een rijtje, wellicht niet eens uitputtend.

3.3 De meest gebruikte methoden: peer review en bibliometrie

In de praktijk wordt een evaluatie doorgaans uitgevoerd door een commissie bestaande uit peers, experts, stakeholders of een mix van dezen. De meest gebruikte methode om kwaliteit vast te stellen is peer review. In een aantal vakgebieden (natuur- en levenswetenschappen) wordt daarnaast meestal ook gebruik gemaakt van bibliometrisch onderzoek. Men ziet echter de laatste tijd dat ook in de humaniora en de sociale wetenschappen het gebruik van meer kwantitatieve benaderingen toeneemt. Al deze methoden richten zich vooral op de wetenschappelijke kwaliteit, maar er is een groeiende belangstelling voor methoden die de maatschappelijke kwaliteit en impact van onderzoek in beeld brengen en zo een meerdimensionale (zelf)evaluatie kunnen ondersteunen (RMW 2002; SWR/RGW 2005; Spaapen and Dijstelbloem 2007; Laredo & Mustar 2001; Van Ark 2007).

Peer review heeft in de wetenschappelijke wereld een groter draagvlak dan bibliometrisch meten. Er is veel onderzoek gedaan naar het functioneren van peer review procedures. Tijdens een recente OECD conferentie werden de voordelen van peer review als volgt opgesomd: "It is a relatively quick, low cost, fast-to-apply, well known, widely accepted and versatile evaluation method that can be used to answer a variety of evaluation questions throughout the project performance cycle as well as in other applications". Maar peer review is ook onderhevig aan kritiek, onder meer zou peer review niet altijd belangeloos zijn, wordt het gekenmerkt door een zekere mate van conservatisme waardoor innovatief onderzoek niet snel wordt beloond, en zou het moeilijk zijn om multi- en interdisciplinair onderzoek via peer review te waarderen. We sommen de bekendste voor- en nadelen op in bijlage 2.

Bibliometrische methoden hebben het voordeel dat ze 'objectiverend' werken; ze zijn relatief transparant. Met de groei van het wetenschapssysteem wordt het ook steeds moeilijker om een goed overzicht te hebben over meer dan een heel klein deel van de wetenschap en zelfs over de eigen discipline. Daarom lijken sommige op scientometrisch onderzoek gebaseerde indicatoren te prefereren. Deze indicatoren kunnen in beginsel ook een veelheid van aspecten dekken (naast artikelen en citaties ook aantallen PhD's, verworven onderzoekmiddelen). Ze kampen echter met legitimiteitsproblemen, in bepaalde velden meer dan in andere (maar zeker niet simpelweg langs de scheidslijn bèta-levenswetenschappen vs. alfa-gamma wetenschappen). Ook hier is een flink aantal problemen te constateren, die in bijlage 2 worden opgesomd.

Als oplossing wordt vaak gesuggereerd om peer review en bibliometrie te combineren. Dat is niet automatisch een goede oplossing omdat de resultaten van beide evaluaties lang niet altijd hoog correleren (Aksnes & Taxt 2004). In sommige gebieden correleert projectselectie nauwelijks met bibliometrische indicatoren (Van den Besselaar & Leydesdorff 2007).

3.4 Enkele belangrijke problemen bij evaluatie

Ondanks de problemen waarnaar hierboven wordt verwezen (en die worden uitgewerkt in bijlage 1) is peer review vooralsnog een goede en breed geaccepteerde manier van evaluatie. Deze commissie zou daaraan niets willen afdoen. Tegelijkertijd willen wij een open oog houden voor de potentiële zwakheden van peer review en hetzelfde geldt voor de bibliometrie. Bij elke vorm van evaluatie moet ook rekening worden gehouden met een aantal algemene problemen die zich voordoen bij vragen rond de kwaliteit en impact van wetenschappelijk onderzoek. Abstract gesproken komen de meeste evaluatiecriteria op de verschillende aggregatieniveaus overeen: kwaliteit, productiviteit, dynamiek, aantallen PhD's, werfkracht (onderzoekers, studenten, onderzoeksgeld), maatschappelijke opbrengst, etc. Accenten verschillen echter en bovendien: hoe lager het aggregatieniveau, hoe moeilijker de criteria valide zijn te operationaliseren en betrouwbaar te meten. Enkele van de belangrijkste problemen zijn:

– Experts in veel vakgebieden denken vaak intuïtief te weten of een individu (of een artikel) kwaliteit heeft, maar het is moeilijk om dat objectief vast te stellen omdat er geen consensus is over de indicatoren.
– De druk op publiceren en de rol van publicatiecijfers in de beoordeling ('publish or perish') lijkt zo langzamerhand zijn doel voorbij te zijn geschoten. Volgens sommigen leidt dit tot een overproductie aan wetenschappelijke tijdschriften en losse artikelen; gebrek aan inzicht in impliciete aannames bij onderzoek; de onmogelijkheid om zinloze tradities te doorbreken en de overtuiging dat 'theoretisch' gelijk staat aan 'vaag'. Voorts gaat de nadruk op publiceren ten koste van andere taken die onderzoekers hebben, met name onderwijs en maatschappelijke dienstverlening (waaronder bijvoorbeeld belangrijke historische collecties of medische taken vallen). Overigens wordt er in sommige landen nu met een zgn. esteem indicator gewerkt waardoor ook andere taken van een onderzoeker kunnen worden gewaardeerd: prijzen of deelname aan belangrijke commissies, redacties etc. (Barker 2007). Zulke Indicators of Esteem worden ook in Nederland regelmatig gebruikt als een van de criteria voor kwaliteit, maar deze criteria zijn nog weinig concreet.
– De attributie: hoe kun je zinvol aangeven wat de bijdrage van de te evalueren eenheid is ten opzichte van anderen met wie essentiële samenwerking heeft plaatsgevonden (Glaeser et al 2004). Onderzoek laat bv. zien dat papers met internationale co-auteurs zichtbaarder zijn (vaker worden geciteerd), maar tegelijkertijd wordt dan onduidelijker wat citaties eigenlijk meten (de kwaliteit van het onderzoek, of die van het netwerk bijvoorbeeld).

– De maatschappelijke opbrengst van onderzoek is moeilijk vast te stellen op microniveau: innovaties zijn vaak gebaseerd op een veelheid van wetenschappelijke resultaten en de inbreng van andere expertise, en bovendien is de tijdsduur tussen wetenschappelijk onderzoek en toepassing onzeker en gevarieerd. Dat is voor evaluatie op microniveau een veel groter probleem dan op systeemniveau.
– Data zijn beter beschikbaar over het systeem als geheel dan op sommige lagere aggregatieniveaus. Op individueel niveau bijvoorbeeld is de kwaliteit van de data vaak matig (in 10% van de ISI records zitten fouten, zoals verkeerd gespelde namen en instituutsnamen; homogeniseren van namen van personen en instellingen vormt een groot probleem). Inmiddels zijn er ook concurrerende databases (Google, grote Amerikaanse universiteiten) die soms tot opvallend verschillende resultaten leiden, zeker op individueel niveau.
– Een andere ontwikkeling die we hier noemen (maar nog niet uitwerken) is het betrekken van stakeholders bij de beoordeling. In de innovatieliteratuur is veel geschreven over het nut van 'early user involvement', d.w.z. het in een vroegtijdig stadium betrekken van gebruikers bij ontwikkelingen in het wetenschappelijk onderzoek. Meer in het algemeen, zeker waar onderzoek wordt ingezet om maatschappelijke problemen aan te pakken, groeit de aandacht voor de rol van stakeholders in het onderzoek en in de evaluatie van het onderzoek.

3.5 Verschillen tussen vakgebieden

Er zijn belangrijke verschillen tussen disciplines voor wat betreft publicatie- en communicatiegewoonten en ook met betrekking tot interactie met maatschappelijke praktijken. Er is dan ook een redelijk brede consensus dat die verschillen moeten kunnen leiden tot differentiatie in beoordelingspraktijken. Zelfs binnen een vakgebied is differentiatie soms gewenst. Chemische onderzoekers publiceren anders en in andere media dan chemisch technologen. In de commissie zijn allerlei relevante verschillen tussen (en ook binnen) disciplines de revue gepasseerd. Van de belangrijkste verschillen, hieronder opgesomd, zou onderzocht moeten worden wat de eventuele consequenties zijn voor de inrichting van een evaluatiestelsel. Overigens zijn binnen diverse disciplines hierop al activiteiten ontwikkeld, bv. de medische vakgebieden, de sociale wetenschappen, de humaniora, technische disciplines.

– Vakgebieden verschillen in de mate van groepsactiviteit. Sommige gebieden zijn zeer individualistisch, andere werken met grotere of kleinere teams. Dat dit consequenties heeft voor de beoordeling blijkt bijvoorbeeld in de discussie rond het SEP. Dat protocol is grotendeels gericht op groepsniveau en daardoor in de huidige vorm minder geschikt voor individu-gerichte gebieden.
– Vakgebieden verschillen ook met betrekking tot de waarde van criteria en indicatoren. In het ene gebied dat zeer internationaal competitief is zijn publicaties in high impact journals maatgevend, in het andere dat technisch is gericht of lokaal georiënteerd gelden andere maatstaven. In sommige gebieden zijn juist boeken de dominante uitingsvorm, in andere technische artefacten.
– Steeds meer onderzoek is multi-, inter- of transdisciplinair en daardoor moeilijker meetbaar met traditionele indicatoren. Ook is het niet eenvoudig voor die gebieden goede commissies te vinden. Een gevolg is ook dat het minder gemakkelijk is de top van een veld te duiden.

Al deze verschillen zouden gevolgen moeten hebben voor de inrichting van evaluatieprocedures. De vraag is echter of het wel zo efficiënt is om voor elk gebied eigen procedures in te richten . Vanuit de overheid en andere bestuurlijke gremia is er een permanente druk om alle onderzoek zoveel mogelijk langs dezelfde meetlat te leggen. En liefst een zo eenvoudig mogelijke meetlat. Het SEP bijvoorbeeld biedt daartoe in de ogen van de overheid onvoldoende mogelijkheden en men wenst dan ook in het nieuwe SEP aanpassingen die bv. een landelijke vergelijking van gebieden mogelijk maakt. Deze tegenstelling is een gevolg van het feit dat evaluaties verschillende doelstellingen kunnen hebben voor verschillende actoren. Bezien vanuit het doel 'middelenallocatie' komt men tot andere eisen aan een evaluatie (en dus criteria en indicatoren) dan bezien vanuit het doel 'bewaking en verbetering van kwaliteit.' Een van de grote uitdagingen in de nabije toekomst is beide gezichtspunten te verzoenen in één stelsel. Het huidige SEP loopt tot 2009. In het najaar van 2007 begon het overleg over een nieuw SEP, waarbij de huidige analyse van de werking van het SEP uiteraard een belangrijke rol speelde. Belangrijke vragen zijn in hoeverre de uitkomsten van al die evaluaties inderdaad leiden tot beter onderzoek en of beleidsorganen en de politiek er voldoende inzicht door krijgen betreffende de kwaliteit en effectiviteit van het wetenschapssysteem. De commissie kwaliteitszorg is bereid om zich in deze discussie te mengen, uiteraard in overleg met de MEC. Een mogelijke oplossing voor de tegenstelling (elk gebied een eigen procedure, of één procedure voor alle gebieden) zou zijn om een 'kern' evaluatieprotocol te hanteren (een 'basis SEP'). Zo'n kern protocol ('KEP') zou aanzienlijk eenvoudiger moeten zijn dan het huidige protocol, in lijn met de wensen van de onderzoekers om de evaluatiebelasting te reduceren (zie ook het volgende hoofdstuk). Maar het zou ook rechtvaardiger moeten zijn, dat wil zeggen dat het voldoende recht moet doen aan de verschillen tussen vakgebieden.

4 Balans tussen rechtvaardigheid en eenvoud

4.1 Inleiding

In Nederland en elders is de laatste jaren gezocht naar een oplossing voor bovengenoemde problemen, in het bijzonder op twee punten: de (te) zware beoordelingslast en de onrechtvaardige effecten van gehanteerde criteria en indicatoren in verschillende vakgebieden. Het SEP was bedoeld om andere beoordelingen overbodig te maken (hetgeen niet is gelukt) en om een zekere flexibiliteit te bieden aan de te evalueren eenheden en vakgebieden om het eigene aan de onderzoekspraktijk in de evaluatie te betrekken. Dit laatste is ook nog niet erg gelukt, maar er zijn wel vorderingen gemaakt op dat gebied. Dit hoofdstuk is de opmaat voor een concreet voorstel van de commissie om het huidige stelsel rechtvaardiger en eenvoudiger te maken. Het voorstel zelf wordt in het volgende hoofdstuk uitgewerkt. Volgens de commissie zou het huidige Standard Evaluation Protocol (SEP) moeten worden vervangen door een Kern Evaluatie Protocol (KEP). Het KEP zou zowel voor de onderzoekers die gegevens moeten aanleveren, als voor beoordelaars een lastenverlichting moeten brengen.

4.2 Naar een rechtvaardiger en eenvoudiger systeem

De Commissie Kwaliteitszorg vindt dat de kwaliteitszorg in Nederland op twee punten sterk moet worden verbeterd:

1. Het moet eenvoudiger: Er zijn te complexe en te weinig gestandaardiseerde evaluaties, er is bovendien een gebrek aan onderlinge afstemming van de diverse evaluaties. Dit wordt ook wel 'stapeling' genoemd; een probleem dat wordt verergerd doordat onderzoekers steeds weer (net) iets anders moet aanleveren. Dit is het meest genoemde probleem in de commissie en wordt ook breder gevoeld. In feite was het al een grote zorg voor de minister van OCW toen het SEP werd ingesteld. Het was juist expliciet de bedoeling van het SEP een einde te maken aan de veelheid en diversiteit van evaluaties. In het protocol wordt ook expliciet vermeld dat gegevens van een evaluatie drie jaar lang gebruikt moeten kunnen worden in andere evaluaties. In de praktijk blijkt dat echter niet te werken. De commissie is van mening dat onderzoek, onderzoekers en onderzoeksgroepen inderdaad nog steeds te vaak – al dan niet overlappend – worden geëvalueerd, waarbij de meerwaarde van de verschillende evaluaties ten opzichte van elkaar op zijn minst onduidelijk is.
2. Het moet rechtvaardiger: Er bestaat een aanzienlijke diversiteit van onderzoeksen dus ook evaluatiepraktijken in de verschillende disciplines en vakgebieden. Deze verschillen zouden op de een of andere manier een plek moeten krijgen in de evaluatiepraktijk. Die wordt echter gedomineerd door criteria en indicatoren die geaccepteerd zijn in de natuur- en levenswetenschappen, maar niet (of minder) in de humaniora en sociale wetenschappen. Daarnaast is er onderzoek

dat zich nadrukkelijk richt op een breder publiek dan de wetenschap alleen. De beoordeling van dat onderzoek vergt eveneens een andere benadering dan het tellen van publicaties en citaties. Ten aanzien van dit laatste zijn de laatste jaren weliswaar vorderingen geboekt, maar deze zijn nog niet breed geaccepteerd.

Oplossing. De commissie stelt voor om het huidige systeem waar mogelijk te vereenvoudigen, maar ook rechtvaardiger te maken. De uitdaging is tussen beide een balans te vinden.

Er zou moeten worden uitgegaan van één evaluatiesysteem, een aanpassing van het huidige SEP, zodanig dat alleen de meest relevante informatie hoeft te worden geleverd: een kern-SEP, of korter een Kern Evaluatie Protocol: KEP. Dat protocol zou tevens voldoende flexibiliteit moeten hebben om de verschillen tussen vakgebieden te accommoderen. Onderzoekers die bijvoorbeeld de maatschappelijke en culturele relevantie van hun werk, of de economische en / of technologische waarde, naar voren willen brengen, moeten daartoe in de gelegenheid zijn. Fondsen en andere geldverdelende instanties zouden vervolgens bereid moeten zijn om zoveel mogelijk hun evaluaties hierop af te stemmen. Het aantal hoofdcriteria in het huidige SEP zou kunnen worden teruggebracht van vier naar drie: kwaliteit, productiviteit, relevantie, en bij elk van die drie criteria zou preciezer – in de vorm van een beperkt aantal indicatoren of indicaties – moeten worden vastgesteld wat daaronder wordt verstaan. Onder relevantie zou als expliciete indicator 'inverdiencapaciteit' kunnen worden toegevoegd, zowel voor wat betreft materiële als voor personele inbreng. Het vierde SEP-criterium 'vitality and feasibility' is weliswaar van belang, maar moeilijk te waarderen in een incidentele site visit. Het gaat bij dit criterium in feite om de vraag of de onderzoeksgroep goed wordt geleid en een goede vitale structuur heeft (bv. een goede balans tussen in- en uitstroom). De beoordeling daarvan is beter op zijn plaats in de regelmatige gesprekken tussen bestuurders en onderzoeksleiders, die op de instellingen zelf plaatsvinden.

Het SEP zou dus concreter en eenvoudiger moeten: een KEP, met drie ipv vier criteria en een compacte verslaglegging voor wat betreft de zelfevaluatie, aangegeven door een maximale lengte per onderdeel. Voor de uitwerking van de beoordelingsprocedure, en de nieuwe elementen daarin, kan worden geleerd van andere evaluatiepraktijken die zich momenteel in Nederland en daarbuiten ontwikkelen.

5. Aanbeveling: van SEP naar KEP

De intensivering van evaluaties in de afgelopen drie decennia heeft naar de mening van de commissie aanvankelijk goed gewerkt: de kwaliteit van het onderzoek is op drie manieren verbeterd: Nederland scoort over het algemeen goed in internationaal perspectief, slecht onderzoek is er uitgelicht en het evaluatiesysteem is transparanter geworden. Het huidige nationale systeem, het SEP, is redelijk tot goed beoordeeld door de Meta-evaluatiecommissie, maar lijkt vooralsnog niet tot minder beoordelingslast te leiden terwijl het toch daarvoor mede was bedoeld. Deels komt dat doordat verschillende financiers verschillende eisen stellen, en deels ook omdat beoordelende organisaties overlap niet altijd kunnen of willen voorkomen. Dit zou als gezegd grotendeels opgelost kunnen worden door een kern-protocol met voldoende flexibiliteit om recht te doen aan de verschillen tussen vakgebieden. Daarnaast zou aan de kant van de financiers de harmonisatie moeten worden bevorderd.

Op weg naar een nieuw kern-SEP (KEP) ziet de commissie de volgende twee hoofdvragen om uit te werken:

A. Wat moet minimaal worden opgenomen in een KEP?
B. Zijn er naast een KEP nog addenda nodig, bv. voor de maatschappelijke of technologische relevantie?

A. Kern-evaluatie protocol: KEP

Het Kern-evaluatieprotocol (KEP) bestaat uit een aangepaste tekst van het huidige SEP met als doel vereenvoudiging en rechtvaardigheid voor alle disciplines. De belangrijkste aanpassingen die wij hier noemen zijn drie i.p.v. vier hoofdcriteria: kwaliteit, productiviteit en relevantie. En daarnaast een sterk vereenvoudigde zelfevaluatie. De zelfevaluatie moet alleen de hoogstnoodzakelijke informatie bevatten. De dikke pakketten die soms aan evaluatiecommissies worden aangeboden bevatten veel informatie die nooit wordt gelezen. Ten aanzien van publicaties is het voor een review commissie bijvoorbeeld interessanter om een aantal van de kern-publicaties te zien, dan de volledige lijst met publicaties van de afgelopen vijf jaar. Van de onderdelen in de zelfevaluatie zou bovendien een maximale lengte moeten worden aangegeven.

Hieronder volgt wat volgens de commissie in het zelfevaluatierapport dient te staan:

PARAGRAFEN VAN DE ZELFEVALUATIE

1. Doel van het onderzoek [maximaal 0,5 pg]
2. Samenstelling van de groep op basis van twee indicaties: totaal aantal medewerkers per functiecategorie (inclusief buitenpromovendi) en een overzicht van de financieringstromen (intern en extern). Geen verdere details, indien gewenst zijn die te vinden in jaarverslagen e.d. [max. 1 pg]
3. Onderzoeksomgeving en inbedding, nationale en internationale positionering, aantal gastonderzoekers (met en zonder eigen fondsen) [maximaal 1 pg]
4. Kwaliteit:
   a. 3-5 sleutelpublicaties per (sub)groep
   b. 3-5 belangrijkste resultaten/hoogtepunten relevant voor het vakgebied, per (sub)groep
   c. aantal artikelen in de top-10% tijdschriften relevant voor het vakgebied; idem in de top-25%
   d. 3-5 belangrijkste boeken of hoofdstukken in boeken, voor zover relevant
5. Output:
   a. aantal artikelen in gerefereerde tijdschriften
   b. aantal boeken, hoofdstukken in boeken
   c. aantal gerealiseerde promoties en aantal promoties 'in pocket'
6. Inverdiencapaciteit bij competitieve fondsen, nationaal en internationaal
7. Academische reputatie per onderzoeksleider (prijzen, uitnodigingen voor voordrachten op grote congressen, organisaties van congressen, editorships, lidmaatschappen academies) [max. 1 pg]
8. Valorisatie in brede zin: sociaal-culturele relevantie en / of technische of economische impact [max. 1 pg] (zie ook B. hieronder)
9. Haalbaarheid van het voorstel of programma, beschikbare infrastructuur en methodologie [max. 1 pg]
10. Toekomstvisie, inclusief kansen en bedreigingen [max. 1 pg]

B. Addenda?

Na ampele discussie beantwoordt de commissie de vraag of er naast het KEP nog gebruik zou moeten worden gemaakt van addenda om specifieke informatie toe te voegen aan de zelfevaluatie negatief. Men zou kunnen denken aan addenda voor bv. de maatschappelijke relevantie of voor technologische impact. De commissie acht deze informatie zeer relevant, maar vindt de gedachte van een beknopte zelfevaluatie te belangrijk om die op te offeren aan extra addenda. De commissie is van mening dat dergelijke extra informatie heel goed kan worden gegeven binnen de bovenstaande lijst met criteria, bv. onder 1, 3, 8 en 10.

Ten aanzien van de bredere relevantie van onderzoek merkt de commissie tot slot nog het volgende op. Om de beoordeling rechtvaardiger te laten verlopen is het nodig om ook andere activiteiten van onderzoekers dan die welke puur op de wetenschappelijke gemeenschap zijn gericht, tot hun recht te laten komen. Veel onderzoek vindt tegenwoordig plaats in de context van maatschappelijke vraagstellingen; er is sprake

van een groei in de publiek-private samenwerking en als gevolg hiervan neemt de druk toe om onderzoek te 'valoriseren'. Deze valorisatie is niet alleen bedoeld in economische zin maar, ook in de zin van een bijdrage aan sociale, culturele, politieke en welzijnsprocessen. Dit punt wordt bij voorkeur aangegeven met indicatoren of indicaties die worden gedragen door het vakgebied. Voor technische vakken kan men bv. denken aan patenten of samenwerking met de industrie; in het (bio)medisch onderzoek bv. aan klinische toepasbaarheid, protocollen; in de geesteswetenschappen aan tentoonstellingen; in de sociale wetenschappen bijdragen aan onderwijsinnovatie. Het zijn maar voorbeelden. KNAW, NWO en VSNU hebben, zoals al eerder vermeld, een gezamenlijk project dat zich richt op de ontwikkeling van indicatoren voor valorisatie in verschillende vakgebieden (www.eric-project.nl). Het project beoogt tevens binnen vakgebieden consensus te bevorderen op het gebied van de evaluatie van maatschappelijke kwaliteit van wetenschap. Op de genoemde site is ook informatie te vinden over de inmiddels brede waaier van valorisatie-methodieken, nationaal en internationaal die worden gebruikt en of getest.

Literatuur

Adviesraad voor het Wetenschaps en Technologiebeleid, Alfa en gamma stralen. AWT: Den Haag, 2007.

Arnold, Erik, Evaluating research and innovation policy: a systems world needs systems evaluations. Research Evaluation 13 (2004) 3-17.

Aksnes, Dag W., and Randi Elisabeth Taxt, Peer review and bibliometric indicators: a comparative study at a Norwegian university. Research Evaluation 13 (2004) 33-41.

Barker, Katherine, The UK Research Assessment Exercise: the evolution of a national research evaluation system, Research Evaluation 16 (1) March 2007 3-12.

Commissie Dynamisering, Investeren in dynamiek, eindrapport deel 1, april 2006.

Dietz, Ton, 'Het CERES-systeem van prestatiemeting', lezing van de directeur van de Interuniversitary Research School for Resource Studies for Development (CERES) op 19 maart 2007, Maagdenhuis Amsterdam, Folia 25 13.

Disciplineoverleg Orgaan Rechtsgeleerdheid (DRG), Naar prestatie-indicatoren voor rechtswetenschappelijk onderzoek, Maart 2007.

Glaeser, Jochen, et al, Intraorganisational evaluation: are there 'least evaluable units'? Research Evaluation 13 (2004) 19-32.

Godin, Benoit, Measurement of Science and Technology: 1920 to the Present, London: Routledge, 2005.

Gubba, E.G. and Y.S. Lincoln, Fourth Generation Evaluation, Newbury Park, CA: Sage Publications, 1989.

Langfeldt, Liv. Expert panels evaluating research: decision-making and sources of bias. Research Evaluation 13 (2004) 51-62.

Laredo, Ph. & P. Mustar, Laboratory activity profiles: An exploratory apporach, Scientometrics 47/3, 515-539 (2000).

Merkx, Femke, et al, Evaluation of Research in Context; a quick scan of an emerging field. Den Haag: Rathenau Instituut / ERiC, 2007.

Merkx, Femke, and Peter van den Besselaar, Positioning Indicators for cross-disciplinary challenges: the Dutch coastal defense research case, Research Evaluation 20 (2008) (Forthcoming).

Meta Evaluatie Commissie, Trust, but verify. Amsterdam 2007.

Michelson, Approaches to research and development performance assessment in the US: an analysis of recent evaluation trends. Science and Public Policy 33 (2006) 546-560.

Moed, H.F., Citation Analysis in Research Evaluation, Springer, 2005.

Oostveen, Anne-Marie et al, Research evaluation – an overview. Science System Assessment Rapport 0707. Den Haag: Rathenau Instituut (in voorbereiding).

Raad voor de Medische Wetenschappen, The societal impact of applied health research. Amsterdam: KNAW, 2002.

Raad voor de Medische Wetenschappen, Gezondheidsonderzoek: het investeren waard, KNAW, 2007.
Spaapen, Jack, Huub Dijstelbloem and Frank Wamelink, Evaluating research in context. A method for comprehensive assessment, 2nd edition, The Hague: COS 2007.
Sociaal Wetenschappelijke Raad (SWR) en Raad voor de Geesteswetenschappen (RGW), Judging research on its merits, Amsterdam: KNAW, 2005.
Van den Besselaar P., & L. Leydesdorff, Research budget allocation and bibliometric indicators – an assessment. Science System Rapport 0707. Den Haag: Rathenau Instituut (te verschijnen).
Versleijen, A. (red.), Dertig jaar publieke onderzoeksfinanciering in Nederland (1975-2005); historische trends, actuele discussies. Science System Rapport 0703. Den Haag: Rathenau Instituut 2007.
Wenneras & Wold, Nepotism and sexism in peer review, Nature 387 (1997) 341-343.
Werkgroep Kwaliteitszorg Wetenschappelijk Onderzoek en standpuntbepaling KNAW, NWO en VSNU, Kwaliteit verplicht. Naar een nieuw stelsel van kwaliteitszorg voor het wetenschappelijk onderzoek, Amsterdam, 2001.
Working Party on Innovation and Technology Policy, Peer review: its uses, demands and issues, OECD workshop on rethinking evaluation in science and technology, 29-30 October 2007, Paris, DSTI/STP/TIP(2007)13.

Bijlage 1 overzicht van de evaluatiepraktijk

Het onderstaande overzicht beoogt niet volledig te zijn, maar wel een goede indruk te geven van de breedte van de evaluatiepraktijk waarmee onderzoekers worden geconfronteerd. Het geeft aan hoe complex deze praktijk is, of althans is geworden. De cruciale vraag bij elk evaluatiemoment is hoeveel inspanning het kost om aan de gestelde informatievraag te voldoen en vervolgens wat de baten zijn van die inspanningen.

Er zijn in het algemeen drie niveaus van evaluatie te onderscheiden:

A. Evaluaties gericht op individuele onderzoekers:
   1. Bij selectieprocedures, door de toekomstige werkgever.
   2. In (jaarlijkse) functioneringsgesprekken, door de leidinggevende.
   3. In beoordelingen, idem.
   4. Bij aanvragen van projecten, door peers en research councils, door potentiele opdrachtgevers, ook internationaal.
   5. Bij de review van de resultaten van projecten, door financiers.
   6. Bij het indienen van artikelen, door peers en editors.
   7. Bij het afronden van promoties, door peers.

B. Evaluaties gericht op groepen onderzoekers (instituten, programma's):
   8. Zelfevaluatie
   9. Midterm evaluatie
   10. Jaarlijkse monitoring, METIS (verschil faculteit-universiteit-vakgebied)
   11. SEP: onderzoeksinstituut, faculteit
   12. ECOS: (Her)erkenning onderzoeksschool

C. Evaluaties gericht op het (supra)nationale wetenschapssysteem
Deze evaluaties gebeuren in toenemende mate in internationaal (Europees) verband, via incidentele studies, of structureel via bv. de Europese commissie, de OECD, of nationaal via het ministerie van OCW, of via het Rathenau (Science Systems Assessment).

Bij alle evaluaties, op alle verschillende aggregatieniveaus, moeten beslissingen worden genomen over de voornaamste doelen, criteria, indicatoren en methoden. Op elk niveau spelen voor dat niveau specifieke vragen en problemen die van invloed zijn op de wijze waarop de evaluatie wordt georganiseerd. We lopen ze langs:

Systeemniveau

Een aantal – niet uitputtend uiteraard – voorbeelden van vragen / problemen zijn:

– Gaat de selectie van projecten goed? Komt het geld bij de beste onderzoekers?
– Is het loopbaansysteem goed georganiseerd, zodat jonge veelbelovende onderzoekers alle ruimte krijgen om zich te ontwikkelen en om nieuwe onderzoeksrichtingen te exploreren?
– Is de organisatie van onderzoek in facultaire onderzoeksinstituten en in landelijke onderzoekscholen met programma's functioneel in alle disciplines?
– Leidt het academische rangenstelsel tot voldoende uitdaging? Of moeten we af van het idee dat UD en UHD ook eindfuncties kunnen zijn? Is de hiërarchische structuur van het functiegebouw wel functioneel? Wat is de plaats van de Principal Investigator hierin?
– Kunnen sponsors van onderzoek hun vraag goed articuleren? (discussie over de besteding van de onderzoeksmiddelen uit FES).

Instituutsniveau

Instituten doen vaak meer dan alleen wetenschappelijk onderzoek: er wordt onderwijs gegeven; er wordt gewerkt aan wetenschappelijke collecties; er zijn vormen van dienstverlening aan de wetenschap en aan de samenleving. In een evaluatie komen al deze aspecten aan de orde en ook de wijze waarop het instituut wordt geleid. Voor de evaluatie betekent dit dat óf gekozen moet worden om alle aspecten in samenhang te beoordelen, hetgeen effect heeft op de methode en bijvoorbeeld ook op de samenstelling van de externe commissie, óf om verschillende evaluaties te organiseren voor verschillende aspecten (bijvoorbeeld SEP voor het onderzoek, ECOS voor de PhD opleidingen), hetgeen dan kan leiden tot een versterkt gevoel van stapeling van evaluaties.

Groepsniveau

Bij onderzoeksgroepen speelt de vraag wat precies te evalueren. Gaat het om de groep als geheel, of om de individuele onderzoeksleiders? Is een goede onderzoeksgroep een verzameling heel goede onderzoekers, of een groep met een goede leider en een goed en productief programma? Wat zegt dat over vorm en doel van zelfevaluaties van groepen? Is het nuttig om op dit niveau extern te evalueren of is hiervoor juist de leiding van het instituut verantwoordelijk?

Individueel niveau

Bij individuen zijn er in het algemeen drie evaluatiemomenten: tijdens de sollicitatie, tijdens geregelde interne procedures (functioneringsgesprekken e.d.) en wanneer externe aanvragen worden gedaan. Het accent ligt in het laatste geval meestal op de past performance, die veel zegt over de kwaliteit van de onderzoeker en een indicatie geeft ten aanzien van de toekomstverwachtingen. In de eerste twee gevallen speelt dat ook wel een rol, maar wordt er ook naar andere zaken gekeken, bijvoorbeeld leiderschapskwaliteiten, vermogen tot samenwerken, innovatief vermogen, werfkracht voor geld en goede medewerkers.

Bijlage 2 Peer review en bibliometrie, voor- en nadelen

Peer review

Peer review is een van de meest gebruikelijke methoden om wetenschappelijk werk te beoordelen. Van peers wordt gebruik gemaakt op alle niveaus die in de evaluatiepraktijk zijn te onderscheiden (zie bijlage 1). Het heeft veel voordelen: het is relatief snel en gemakkelijk toe te passen, niet duur, breed geaccepteerd en breed inzetbaar, d.w.z. het kan een veelheid van vragen aan. Daartegenover staat wel een aantal nadelen, die deels te maken hebben met het gegeven dat peers mensen zijn (waardoor subjectieve factoren de overhand kunnen krijgen) en deels met het feit dat er zo'n groot beroep wordt gedaan op een beperkt aantal mensen dat de zorgvuldigheid in het geding kan komen (en ook referee fatigue op kan treden). Daarnaast is het steeds meer de vraag of peers wel in staat zijn de vragen te beantwoorden die tegenwoordig vaak in evaluaties aan de orde zijn, vragen met een veel bredere impact dan die welke in de traditionele vakgebiedgerichte evaluaties aan de orde waren.

Omdat het zo'n belangrijk mechanisme is, wordt er van oudsher veel onderzoek gedaan naar het functioneren van peer review. Recent zijn er enkele belangrijke conferenties georganiseerd door de OECD (oktober 2007) en de ESF (2007 en 2008), waarop peer review breed werd geanalyseerd en naar oplossingen werd gezocht. Opgemerkt werd daar onder meer dat het peer review proces steeds meer internationaliseert, hetgeen weer nieuwe problemen met zich meebrengt. Hieronder volgen, bij wijze van illustratie, enkele van de kwesties die in de discussie rond peer review spelen.

- Door de grote druk op veel publiceren lezen reviewers steeds minder van de output. Het is steeds gebruikelijker om alleen enkele 'key-publicaties' te lezen.
- 'Halo effect': als iets eenmaal een goede reputatie heeft, is de kans op een 'automatische' goede review groter.
- Bias vanwege disciplinaire verschillen (Van den Besselaar & Leydesdorff 2007), of paradigmatische verschillen.
- Multi-, inter- en transdisciplinair onderzoek blijken moeilijk te beoordelen met traditioneel peer review: Disciplines disciplineren! 'Peer' review dat aan deze kritiek probeert tegemoet te komen wordt wel 'expert' review genoemd. Dat is echter ook niet onproblematisch. Bij commissies die uit leden met verschillende competenties bestaan, of in ieder geval met verschillende disciplinaire achtergrond, gebeurt het vaak dat elk afzonderlijk lid 'peer' is op hoogstens een deel van het te reviewen onderwerp. Dat leidt vaak tot stilzwijgende en onduidelijke compromissen (Langfeld 2004). Dit is ook een risico bij verkenningen op een breed gebied.

– Peers kunnen belangen hebben. Meer specifiek laat onderzoek zien dat er sprake is van 'Nepotism and sexism in peer review' (Wenneras & World 1997): vrouwen moesten aanzienlijk meer toppublicaties hebben dan mannen om even goed beoordeeld te worden. Daarnaast blijken goede relaties met leden van de commissies sterk te helpen.
– Lage betrouwbaarheid van de ratings. Dit blijkt ook bij de evaluatie van papers en proposals, en zeker niet alleen in de alfa en gammawetenschappen. (Cicchetti 1991 en Rothwell & Martyn 2000 over bijvoorbeeld life sciences).

Bibliometrie

De kritiek op peer review heeft in de vorige eeuw er mede toe geleid dat andere, meer objectieve methoden opkwamen, met name bibliometrische en scientometrische methoden. Vooral in kringen van het wetenschapsbeleid groeide de belangstelling hiervoor, omdat het leek alsof hiermee op een relatief eenvoudige wijze beslissingen waren te nemen over de allocatie van middelen voor het onderzoek. Ook hier bleek echter al snel dat er naast voordelen ook grote nadelen aan bibliometrische methoden kleven.

– Hoe kleiner de te evalueren eenheid, hoe minder valide de methoden blijken te zijn (Glaeser et al 2004).
– De kwaliteit van de indicatoren is onduidelijk: wat meet wat? Hierover is nog veel discussie en er is behoefte aan beter onderzoek om de goede indicatoren te vinden. Een voorbeeld is de aan populariteit winnende H-index, die ook al weer tot veel discussie leidt.
– Bibliometrie is blind voor nieuwe ontwikkelingen, omdat die nog niet in de (gearriveerde ISI geïndexeerde) tijdschriften staan: je moet minstens een aantal jaren terugkijken voor een goede citatieanalyse, past performance zegt weinig over de actualiteit van onderzoek.
– Kwantitatieve criteria lokken perverse effecten uit (vraag je om meer artikelen dan krijg je meer artikelen, vraag je om meer patenten dan krijg je meer patenten; het weglaten uit beoordelingen van laaggeïndexeerde tijdschriften) en dan meet je strategisch gedrag in plaats van kwaliteit. Gebruik van meerdere indicatoren naast elkaar maakt dit natuurlijk moeilijker.
– Hoewel de ISI database de meest gebruikte is om bibliometrische indicatoren aan te ontlenen, ontstaan er ook andere databases, meestal gericht op bepaalde vakgebieden, die tot wezenlijk andere resultaten kunnen leiden, bijvoorbeeld omdat ze (deels) andere media gebruiken of andere output meenemen.

Bijlage 3 Samenstelling en opdracht van de commissie Kwaliteitszorg KNAW

Opdracht

De Commissie Kwaliteitzorg is in 2006 ingesteld om het bestuur van de KNAW gevraagd en ongevraagd te adviseren op het brede vlak van kwaliteitszorg in de wetenschap. De commissie staat onder voorzitterschap van prof. dr. Peter van der Vliet en heeft de volgende leden:

Samenstelling

Prof. dr. Peter van der Vliet, voorzitter, emeritus hoogleraar Fysiologische Chemie, Universiteit Utrecht
Prof. dr. Peter van den Besselaar, hoofd Science Systems Assessment Rathenau Instituut, hoogleraar sociaal-wetenschappelijke informatica UvA
Prof. dr. Trudy Dehue, hoogleraar Wetenschapstheorie en -geschiedenis van de Psychologie, Rijksuniversiteit Groningen
Prof. dr. Els Goulmy, hoogleraar Transplantatiebiologie Histocompabiliteitsantigenen, Universiteit Leiden
Prof. dr. Bas ter Haar Romeny, hoogleraar Oude Testament in de oosters-christelijke traditie, Universiteit Leiden
Prof. dr. Richard Grol, hoogleraar Kwaliteit van Zorg, Radboud Universiteit en Universiteit Maastricht
Prof. dr. Ad Lagendijk, universiteitshoogleraar UvA, onderzoeker: FOM-Instituut voor Atoom- en Molecuulfysica - AMOLF
Prof. dr. Emmo Meijer, Raad van Bestuur, Unilever R&D en hoogleraar Macromoleculaire en organische Chemie, Technische Universiteit Eindhoven
Prof. dr. Theo Mulder, directeur instituten KNAW
Prof. dr. Frits van Oostrom, president KNAW (tot 19 mei 2008)
Dr. Jack Spaapen, coördinator Kwaliteitszorg en onderzoeksevaluatie, KNAW
Prof. dr. Wiecher Zwanenburg, emeritus hoogleraar Franse taal- en letterkunde, Universiteit Utrecht
Drs. Jacco van den Heuvel, beleidsmedewerker KNAW, secretaris commissie kwaliteitszorg