

**KWALITEITSBEOORDELING IN DE
ONTWERPENDE EN CONSTRUERENDE DISCIPLINES**



2010 Koninklijke Nederlandse Akademie van Wetenschappen (KNAW)

© Sommige rechten zijn voorbehouden / Some rights reserved

Voor deze uitgave zijn gebruiksrechten van toepassing zoals vastgelegd in de Creative Commons licentie. [Naamsvermelding 3.0 Nederland]. Voor de volledige tekst van deze licentie zie <http://www.creativecommons.org/licenses/by/3.0/nl/>

Koninklijke Nederlandse Akademie van Wetenschappen

Postbus 19121, 1000 GC Amsterdam

Telefoon + 31 20 551 0700

Fax + 31 20 620 4941

knaw@bureau.knaw.nl

www.knaw.nl

pdf beschikbaar op www.knaw.nl


Basisvormgeving: edenspiekermann, Amsterdam

Opmaak en bureauredactie: Ellen Bouma, Alkmaar

Druk: Bejo druk & print, Alkmaar

Foto cover: Nationale beeldbank/Taco Gooiker

ISBN: 978-90-6984-619-4

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



Dit advies is gemaakt van FSC-papier en gecertificeerd onder nummer CU-COC-804134-N.

KWALITEITSBEOORDELING IN DE ONTWERPENDE EN CONSTRUERENDE DISCIPLINES

EEN SYSTEMATISCH KADER

Koninklijke Nederlandse Akademie van Wetenschappen
Advies Raad voor Technische Wetenschappen, Wiskunde en Informatica,
Natuur- en Sterrenkunde en Scheikunde van de KNAW

VOORWOORD

De grote waarde van de wetenschap is haar natuurlijke diversiteit. Het is een diamant met vele facetten. Deze verscheidenheid uit zich in de keuze van onderwerp, de methode van onderzoek en de uiteindelijke wijze van verslaggeving. Zij is in haar diepste zin een reflectie van de rijkdom aan verschijnselen in de wereld en in onze geest.

De wetenschap verdient het dat haar grote waarde zichtbaar gemaakt wordt op een manier die zo objectief mogelijk is en tegelijkertijd deze diversiteit respecteert, zo niet celebreert. Een optimale kwaliteitsborging is niet alleen nodig om in de buitenwereld draagvlak voor onderzoek te verwerven. Ook de beoefenaars zelf zijn gebaat bij een heldere evaluatie van hun onderzoek. De uitdaging hierbij is de meetmethoden zo dicht mogelijk te laten aansluiten bij de intrinsieke waarde van het onderzoeksveld. Hier valt nog veel werk te verrichten.

Het in kaart brengen van wetenschappelijke publicaties en hun citaties is daarbij een belangrijke eerste stap. Volgens dit criterium blijkt het Nederlands wetenschappelijk onderzoek van wereldklasse te zijn. Het rapport *Wetenschaps- en Technologie-Indicatoren* dat tweejaarlijks in opdracht van het ministerie van OCW wordt opgesteld, heeft dit onlangs nog eens bevestigd. Nederland behoort in termen van citatie-impact tot de top van de wereldranglijst, samen met de Verenigde Staten, Zwitserland en Denemarken. Bovendien zijn wij bijzonder productief.

Natuurlijk is het goed nieuws dat Nederlandse wetenschappers langs deze as gemeten tot de wereldtop behoren. Het is echter belangrijk dat wij ons realiseren dat wetenschappelijke kwaliteit zich niet *alleen* laat meten aan de hand van wetenschappelijke publicaties en citatie-impact. Voor vele wetenschapsdomeinen zijn citaties slechts het halve verhaal, als het niet nog minder is. De standaardmeetmethoden missen daarmee grote delen van belangrijke onderzoeksvelden. Zo zijn de producten van de ontwerpende en construerende wetenschappen niet alleen *peer-reviewed*

tijdschriftartikelen, maar ook *conference proceedings*, ontwerpen, software en constructies. In de geesteswetenschappen verschijnen veel publicaties in boekvorm of in andere talen dan Engels waardoor ze vrijwel onzichtbaar zijn in de citatie-impact. Dergelijke disciplines kunnen daarom minder uit de voeten met de gebruikelijke manier van kwaliteitsbeoordeling. Het is dan ook niet toevallig dat de KNAW verzocht is te adviseren over hanteerbare beoordelingscriteria voor juist deze disciplines.

De KNAW heeft in 2008 in het advies *Kwaliteitszorg in de wetenschap; van SEP naar KEP* al bepleit dat een nieuw protocol voor de kwaliteitsbeoordeling voldoende flexibiliteit moet hebben om de verschillen tussen vakgebieden te accommoderen. Onderzoekers die bijvoorbeeld de maatschappelijke en culturele relevantie van hun werk of de economische waarde naar voren willen brengen, moeten daartoe in de gelegenheid worden gesteld. De KNAW heeft daarom met genoegen deze adviesvragen opgepakt. Het nu voorliggende advies geeft voor de ontwerpende en construerende wetenschappen invulling aan de roep om een flexibeler beoordeling. Ik ben verheugd dat de commissie tot de conclusie komt dat een specifieke set van criteria voor deze disciplines niet nodig is. Wetenschappelijke kwaliteit en maatschappelijke relevantie volstaan. De disciplinespecifieke invulling zit in de indicatoren waarmee deze twee criteria beoordeeld moeten worden. Dit is een belangrijk leidend principe dat tegelijkertijd de universaliteit en de diversiteit van de wetenschap benadrukt.

De commissie geeft aan bewust geen definitieve selectie of onderlinge weging van de indicatoren te maken. Dit is een stap die voor de verschillende beoordelings situaties door de wetenschappers zelf in nauwe samenspraak met de universitaire bestuurders moet worden gemaakt. Ik hoop dat de 3TU-Federatie en de ontwerpende en construerende wetenschappers deze handschoen voortvarend zullen oppakken, zodat de Nederlandse wetenschap voortaan in al haar verscheidenheid kan flonkeren.

Robbert Dijkgraaf
President KNAW

INHOUD

VOORWOORD 5

SAMENVATTING 9

1. INLEIDING 13
 - 1.1 Achtergrond 13
 - 1.2 Samenstelling en opdracht van de commissie 13
 - 1.3 Werkwijze 14
 - 1.4 Relatie met andere initiatieven 15
 - 1.5 Afbakening 16

2. KNELPUNTEN BIJ DE KWALITEITSBEOORDELING 19
 - 2.1 Inleiding 19
 - 2.2 Ervaren knelpunten 21
 - 2.3 Evaluatie ervaren knelpunten en conclusies 23

3. BEOORDELINGSCRITERIA 27
 - 3.1 Inleiding 27
 - 3.2 Beoordelingscriteria 29
 - 3.3 Toelichting op de indicatoren 30
 - 3.4 Het relatieve belang van de criteria en de indicatoren 33
 - 3.5 Profielen 34
 - 3.6 Selectie van *peers* 35
 - 3.7 Conclusies 35

4. INVENTARISATIE VAN BUITENLANDSE ERVARINGEN EN INZICHTEN 37
 - 4.1 Inleiding 37
 - 4.2 Samenvatting ervaringen in het buitenland en conclusies 37

5. CONCLUSIES EN AANBEVELINGEN 41

LITERATUUR 46

BIJLAGEN

1 Geïnterviewden 47

2 Indicatoren ERIC-project 48

3 Kwaliteitsbeoordeling in het buitenland 49

SAMENVATTING

Achtergrond

De 3TU-Federatie heeft aan de KNAW gevraagd te adviseren over de criteria die gebruikt kunnen worden bij de ex ante- en ex postbeoordeling van technisch-wetenschappelijk werk in de ontwerpende en construerende disciplines. Ontwerpende en construerende wetenschappers ervaren namelijk regelmatig problemen bij de beoordeling van de kwaliteit van hun wetenschappelijk werk bij visitaties, benoemingen en onderzoeksaanvragen. De indicatoren voor kwaliteit die daarbij gebruikt worden zijn ontleend aan de meer fundamentele wetenschappen (publicaties in ISI-tijdschriften, impactfactoren, citaties, Hirsch-index) en zijn in de ogen van deze wetenschappers ontoereikend. Ten eerste omdat hun producten niet alleen *peer-reviewed* tijdschriftartikelen zijn, maar ook *conference proceedings*, ontwerpen en constructies. Ten tweede, omdat hun onderzoek in het algemeen meer contextspecifiek en multidisciplinair van aard is dan dat van de meer fundamentele wetenschappen. Tijdschriften in deze gebieden hebben daardoor een lagere impactfactor en worden systematisch lager gewaardeerd.

De KNAW heeft hiertoe een commissie ingesteld onder voorzitterschap van prof. dr. ir. A.W.M. Meijers (TU Eindhoven) met als opdracht een advies op te stellen over de criteria voor de ex ante- en ex postbeoordeling van:

- ontwerpende en construerende activiteiten in de technische disciplines die als wetenschappelijk kunnen worden aangemerkt
- wetenschappelijk onderzoek in de ontwerpende en construerende disciplines

De criteria moeten voldoen aan de volgende vereisten:

- van nut voor organisaties die wetenschappelijke activiteiten financieren (NWO, ed.)
- van nut voor universiteiten bij de beoordeling van wetenschappelijk personeel (onderzoekers / ontwerpers)
- geloofwaardig in internationaal verband

In het voorliggende advies doet de commissie een voorstel voor deze criteria en doet zij verslag van haar bevindingen. Het advies is mede gebaseerd op interviews met 27 wetenschappers uit verschillende disciplines en een inventarisatie van de buitenlandse inzichten en ervaringen.

Algemene criteria met disciplinespecifieke indicatoren

Een belangrijk uitgangspunt voor de commissie is dat de beoordelingsmaatstaven voor een bepaalde discipline niet volledig verschillend kunnen zijn van die van een andere discipline. Dat zou de maatstaven arbitrair maken en opportunistisch. De commissie heeft daarom gestreefd naar een beoordelingskader met twee elementen:

- algemeen geldende kwaliteitscriteria
- disciplinespecifieke indicatoren voor deze criteria. Deze indicatoren geven empirische informatie over de mate waarin een persoon, groep of onderzoeksvorstel voldoet aan de kwaliteitscriteria. Dit kan kwantitatieve maar ook kwalitatieve informatie te zijn

De commissie concludeert dat – in tegenstelling tot wat bij de start van het adviestraject werd verwacht – de kwaliteit van het wetenschappelijke werk in de ontwerpende en construerende disciplines kan worden beoordeeld aan de hand van de twee criteria die ook voor andere wetenschapsgebieden gelden: 1) wetenschappelijke kwaliteit en 2) maatschappelijke relevantie. Er zijn dus geen aanvullende criteria nodig voor ontwerpen en construeren. Uit de internationale inventarisatie blijkt dat deze twee criteria ook internationaal gezien geloofwaardig zijn voor deze disciplines.

Dat er geen andere kwaliteitscriteria nodig zijn, betekent echter niet dat alle wetenschapsgebieden op dezelfde manier kunnen worden beoordeeld. Beoordelen van kwaliteit behoort maatwerk te zijn vanwege de verschillen tussen disciplines (waaronder publicatiecultuur), typen wetenschappelijke activiteiten (ontwerpen, onderzoeken) en beoordelingssituaties (visitatie, benoeming, onderzoeksvorstel). Dit maatwerk komt tot uiting in de keuze en weging van de disciplinespecifieke indicatoren voor de criteria wetenschappelijke kwaliteit en maatschappelijke relevantie. Tabel 3.1 op pagina 29 geeft een overzicht van de indicatoren die de commissie relevant acht voor de ontwerpende en construerende disciplines.

Vervolgstap nodig

De commissie heeft er nadrukkelijk *niet* voor gekozen om een nadere selectie of onderlinge weging van de indicatoren voor de ontwerpende en construerende wetenschappen te maken. Dit is een vervolgstap die door de wetenschappers zelf (*peers*) in nauwe samenspraak met de universitaire bestuurders voor de verschillende beoordelingssituaties moet worden gemaakt. De commissie adviseert het bestuur van de 3TU's om ruimte te scheppen voor een disciplinespecifieke kwaliteitsbeoordeling in de technische wetenschappen en aan de ontwerpende en construerende disciplines te vragen indicatoren en hun relatieve gewicht vast te stellen voor de beoordeling van kwaliteit in die disciplines.

Vergelijking tussen vakgebieden problematisch

Tijdens de interviews in het veld bleek dat de beoordeling van kwaliteit van het wetenschappelijk werk door internationaal gerespecteerde *peers* vrijwel nergens als probleem wordt ervaren. De problemen ontstaan bij vergelijking *tussen* vakgebieden of als de gebieden te breed worden. De commissie vindt dat kwaliteit alleen *binnen* disciplines zinvol kan worden vergeleken. Een *one-size-fits-all*-benadering bij de beoordeling van kwaliteit, hoe gewenst ook vanuit bestuurlijk oogpunt vanwege zijn eenvoud, doet geen recht aan significante disciplinaire verschillen en zal dus altijd bepaalde vakgebieden ten onrechte bevoordelen. De commissie adviseert onderzoeksfinanciers zoals NWO en STW om meer in te zetten op programma's voor disciplines die niet goed passen bij de bestaande gebiedsindeling en de bestaande manier van kwaliteitsbeoordeling. Dat geldt in het bijzonder voor de ontwerpende en construerende wetenschappen. Daarvoor moeten kwaliteitsindicatoren worden gehanteerd die recht doen aan deze disciplines.

Peer review centraal

De commissie is van mening dat *peer review* het uitgangspunt moet zijn voor een goede kwaliteitsbeoordeling en juicht het daarom toe dat de meeste onderzoeksfinanciers gebruik maken van *peer review*. De keuze van deze *peers* is wel cruciaal en heeft invloed op de uitkomst van de beoordeling. Het zou goed zijn om de keuze van deze *peers* meer te objectiveren en transparanter te maken, binnen NWO, visitatiecommissies en benoemingsadviescommissies. De *peers* die betrokken zijn bij de beoordeling van kwaliteit in de ontwerpende en construerende wetenschappen moeten zelf ook worden beoordeeld aan de hand van het brede spectrum van indicatoren en niet alleen op publicatiegedrag.

Belang van publicaties

In dit advies wordt voorgesteld voor kwaliteit ook andere indicatoren dan *peer-reviewed* publicaties te gebruiken. De commissie benadrukt echter dat *peer-reviewed* publicaties ook in de ontwerpende en construerende disciplines belangrijk zijn en blijven. Ze dragen bij aan onderlinge kwaliteitstoetsing, aan verspreiding van kennis en aan verdere verwetenschappelijking. Publiceren is in deze disciplines of onderdelen daarvan niet altijd vanzelfsprekend. De commissie adviseert wetenschappers in de ontwerpende en construerende disciplines waar nodig krachtig een cultuur van *peer-reviewed* publicaties te bevorderen. Het is daarbij van belang te zoeken naar publicatievormen die passen bij de (deel)discipline.

1. INLEIDING

1.1 Achtergrond

Voor ontwerpende en construerende wetenschappers is de beoordeling van de kwaliteit van hun wetenschappelijk werk bij visitaties, benoemingen en onderzoeksaanvragen sinds jaar en dag een probleem. De indicatoren voor kwaliteit die daarbij gebruikt worden, zijn ontleend aan de meer fundamentele wetenschappen (publicaties in ISI-tijdschriften, impactfactoren, citaties, Hirsch-index) en zijn in de ogen van deze wetenschappers ontoereikend. Ten eerste omdat hun producten niet alleen *peer-reviewed* internationale publicaties zijn, maar ook ontwerpen en constructies. Ten tweede omdat hun onderzoek in het algemeen meer contextspecifiek en multidisciplinair van aard is dan dat van de meer fundamentele wetenschappen. Tijdschriften in deze gebieden hebben daardoor een lagere *impactfactor* en worden systematisch lager gewaardeerd. Als een gevolg hiervan ervaren onderzoekers in de ontwerpende en construerende disciplines een gebrek aan wetenschappelijke erkenning en problemen bij het verwerken van externe onderzoeksmiddelen, onderzoeksvisitaties en benoemingen.

1.2 Samenstelling en opdracht van de commissie

Met het oog op de hierboven geschetste problematiek heeft de 3TU-Federatie aan de KNAW gevraagd te adviseren over de criteria die gebruikt kunnen worden bij de ex ante- en ex postbeoordeling van technisch-wetenschappelijk werk in de ontwerpende en construerende disciplines. De KNAW-Raad voor Technische Wetenschappen, Wetenschap, Informatica, Natuur- en Sterrenkunde en Scheikunde (TWINS) heeft daartoe een commissie ingesteld bestaande uit:

- Prof. dr. ir. A. van den Berg, Universiteit Twente
- Prof. dr. ir. R. de Borst, Technische Universiteit Eindhoven

- Prof. dr. P.P.M. Hekkert, Technische Universiteit Delft
- Prof. dr. ir. A.W.M. Meijers (voorzitter), Technische Universiteit Eindhoven
- Prof. dr. R.A. van Santen, Technische Universiteit Eindhoven

Dr. E.E.W. Bruins (STW) was toehoorder. Ir. A. Korbijn (KNAW) trad op als secretaris, dr. J.B. Spaapen (KNAW) als adviseur. Drs. C.S. Tan (KNAW) ondersteunde de commissie bij de interviews en de internationale inventarisatie.

Opdracht

De commissie had als opdracht om een advies op te stellen over criteria voor de ex ante- en ex postbeoordeling van:

- ontwerpende en construerende activiteiten in de technische disciplines die als wetenschappelijk kunnen worden aangemerkt
- wetenschappelijk onderzoek in de ontwerpende en construerende disciplines

De criteria moeten voldoen aan de volgende vereisten:

- van nut voor organisaties die wetenschappelijke activiteiten financieren (NWO e.d.)
- van nut voor universiteiten bij de beoordeling van wetenschappelijk personeel (onderzoekers / ontwerpers)
- geloofwaardig in internationaal verband

In het voorliggende advies doet de commissie een voorstel voor deze criteria en doet zij verslag van haar bevindingen.

1.3 Werkwijze

De commissie heeft allereerst interviews gehouden met 27 representanten uit verschillende disciplines van de technische wetenschappen. Het doel van deze interviews was om na te gaan welke problemen men ervaart met de huidige manier van beoordelen en om na te gaan welke aspecten volgens de belanghebbenden in een nieuwe set beoordelingscriteria meegenomen moeten worden. Een inventarisatie van de ervaren knelpunten is opgenomen in hoofdstuk 2. Een lijst van de geïnterviewden en hun vakgebieden is weergegeven in bijlage 1.

Om een beeld te krijgen van de manier waarop in het buitenland tegen deze problematiek wordt aangekeken en wat we van deze ervaringen kunnen leren, is informatie verzameld over de situatie in het Verenigd Koninkrijk, Australië, Finland, de Verenigde Staten, Noorwegen, Zweden, Frankrijk, Duitsland en Oostenrijk. De resultaten hiervan zijn samengevat in hoofdstuk 4; meer gedetailleerde informatie is weergegeven in bijlage 3.

Op basis van de verzamelde informatie heeft de commissie een analyse gemaakt van de knelpunten bij de beoordeling van kwaliteit in de ontwerpende en

construerende wetenschappen en een voorstel geformuleerd voor geschikte beoordelingscriteria. De bevindingen op hoofdlijnen en een eerste opzet van de criteria zijn in een schriftelijke commentaarrronde aan de geïnterviewden voorgelegd. Het uiteindelijke voorstel voor de criteria is beschreven en toegelicht in hoofdstuk 3.

1.4 Relatie met andere initiatieven

1.4.1 *Evaluating Research in Context*

De 3TU's nemen deel in een drietal *pilots* die in het kader van het ERiC-project (*Evaluating Research in Context*) zijn uitgevoerd. ERiC is voortgekomen uit een project van de Commissie Overleg Sectorraden (COS) over het meten van de maatschappelijke kwaliteit van onderzoek. Tijdens dit project werd een meetmethode ontwikkeld, de sci_Quest-methode. In 2006 werd besloten deze methode verder te ontwikkelen om de praktische toepasbaarheid te onderzoeken en te vergroten. Hiertoe is het platform ERiC opgericht waarin participeren: de HBO-raad, KNAW, NWO, VSNU en de afdeling *Science System Assessment* van het Rathenau Instituut. Het ministerie van Onderwijs, Cultuur en Wetenschap is als waarnemer betrokken. In het kader van dit project zijn drie *pilots* uitgevoerd bij de faculteiten elektrotechniek van de TU Eindhoven, bouwkunde van de TU Delft en werktuigbouwkunde van de Universiteit Twente. De *pilots* hadden tot doel een evaluatiemethode te ontwikkelen om de maatschappelijke relevantie van het onderzoek aan deze faculteiten te kunnen beoordelen. De voorgestelde indicatoren uit de *pilots* bouwkunde en *electrical engineering* zijn weergegeven in bijlage 2. Bij het opstellen van het onderhavige advies is kennisgenomen van de uitkomsten van de ERiC-*pilots*.

Het voorliggende advies is zowel smaller als breder in *scope* dan het ERiC-project. Ten eerste richt dit advies zich op een beperkter aantal disciplines, namelijk de ontwerpende en construerende wetenschappen. Ten tweede gaat het advies niet alleen over de beoordeling van de maatschappelijke relevantie van het wetenschappelijke werk, maar ook de beoordeling van de wetenschappelijke kwaliteit van dat werk.

1.4.2 *Standaard Evaluatie Protocol*

De KNAW, NWO en de VSNU hebben voor de beoordeling van wetenschappelijk onderzoek het *Standard Evaluation Protocol 2009 – 2015* (SEP) vastgesteld [VSNU, KNAW en NWO 2009]. De beoordeling volgens het SEP 2009 – 2015 bestaat uit een zesjaarlijkse externe evaluatie via een zelfevaluatierapport en een *site visit*, en een interne *midterm review*, die tussen twee externe beoordelingen plaatsvindt. Het SEP kent vier beoordelingscriteria en twee niveaus van beoordeling, het instituut als geheel (of faculteit of onderzoekschool) en de onderliggende groepen of programma's. De vier beoordelingscriteria zijn:

1. Kwaliteit (*Quality*)
2. Productiviteit (*Productivity*)
3. Maatschappelijke relevantie (*Societal relevance*)
4. Vitaliteit en haalbaarheid (*Vitality & Feasibility*)

De beoordeling van het wetenschappelijke onderzoek via het SEP dient zowel de verantwoording over de geleverde prestaties als het verbeteren van de kwaliteit in de toekomst.

Een onderscheid tussen dit adviestraject en het SEP is dat het SEP alleen betrekking heeft op complete instituten of de onderliggende groepen of programma's. In dit adviestraject wordt ook gekeken naar de beoordeling van onderzoeksvoorstellen en individuele personen. Gezien het belang van het SEP in Nederland is in het onderhavige advies wel gezocht naar criteria die goed aansluiten bij de SEP criteria.

1.4.3 Beoordelingscriteria geesteswetenschappen

Ook in de geesteswetenschappen levert de huidige manier van kwaliteitsbeoordeling problemen op. In citatie-indexen wegen bijvoorbeeld alleen artikelen en geen boeken mee, en van de artikelen alleen de publicaties in het Engels. Een bijdrage aan het maatschappelijk debat wordt niet in de kwaliteitsbepaling meegenomen. Op verzoek van de commissie Nationaal Plan Toekomst Geesteswetenschappen (commissie Cohen) is de KNAW daarom bezig om in overleg met onderzoekers in het veld en met gebruikers van gegevens over onderzoek in de geesteswetenschappen (bijv. NWO), een eenvoudig en adequaat systeem van kwaliteitsindicatoren te ontwerpen. Het advies over de geesteswetenschappen is op moment van schrijven van dit advies nog niet gereed.

1.5 Afbakening

De opdracht van de commissie was om te adviseren over criteria voor de beoordeling van kwaliteit van (1) ontwerpende en construerende activiteiten die als wetenschappelijk kunnen worden aangemerkt en (2) onderzoek in de ontwerpende en construerende wetenschappen. Voor deze tweeledige opdracht is gekozen omdat er ook binnen meer 'science-achtige' vakgebieden ontwerpende en construerende activiteiten plaatsvinden die met de huidige systematiek niet goed te beoordelen lijken te zijn (bijvoorbeeld het ontwerpen van instrumentatie). Daarnaast vindt er in de ontwerpende en construerende disciplines onderzoek plaats dat multidisciplinair van aard is en dat een eigen signatuur heeft.

In het kader op de volgende pagina wordt globaal geschetst wat verstaan wordt onder ontwerpende en construerende activiteiten. Ontwerpende en construerende disciplines worden door de commissie dan opgevat als die disciplines waarin dit soort activiteiten een belangrijke rol spelen. Voorbeelden zijn industrieel ontwerpen, bouwkunde,

informatica, werktuigbouwkunde, scheikundige technologie, biotechnologie, civiele techniek, maritieme techniek, elektrotechniek en lucht- en ruimtevaarttechniek. De commissie is van mening dat de in hoofdstuk 3 voorgestelde criteria en indicatoren in ieder geval op deze disciplines goed van toepassing zijn.

TECHNOLOGISCH ONTWERPEN

Bij ontwerpen gaat het niet om het verwerven van een beter begrip van de bestaande werkelijkheid (ware of betrouwbare uitspraken daarover) maar om het creëren van iets nieuws. Ontwerpers voegen iets toe aan de werkelijkheid dat eerder niet bestond en dat waarde of functionaliteit heeft voor gebruikers.

Ontwerpen is een generieke activiteit die in veel domeinen voorkomt: van het ontwerpen van een kunstwerk tot het ontwerpen van een nieuwe wet of van een chemische installatie. In de ontwerpende en construerende disciplines gaat het om het ontwerpen van nieuwe technologische artefacten. Dat kunnen producten of processen zijn die fysisch gerealiseerd kunnen worden (gebouwen, telefoons, elektriciteitscentrales, moleculen, chips), maar ook abstracte algoritmes (software). Daarnaast kunnen ook diensten worden ontworpen (internettransacties), levende organismen (genetisch gemodificeerde gewassen), of menselijke ervaringen (*virtual reality*).

Er zijn vele manieren om ontwerpprocessen te karakteriseren. Sommige leggen het accent op het rationeel nemen van beslissingen op basis van wetenschappelijke modellen, waarbij de oplossingsruimte steeds verder verkleind wordt; anderen proberen ook niet-rationele, creatieve of *art-like* aspecten van het ontwerpproces te verdisconteren. De Amerikaanse *Accreditation Board for Engineering and Technology* karakteriseert ontwerpen als volgt: *Engineering design is the process of devising a system, component, or process to meet desired needs. It is a decision-making process (often iterative), in which the basic sciences, mathematics, and the engineering sciences are applied to convert resources optimally to meet these stated needs (accreditation criteria for engineering curricula 2010-2011, www.abet.org)*. De opties waaruit gekozen kan worden in dit proces zijn overigens niet voorgegeven maar worden daarin juist ontwikkeld.

Een ontwerpproces resulteert in een product: het ontwerp. Dit is een representatie van het beoogde artefact in een of andere vorm (prototype, computermodel, schaalmodel, tekening). Daaraan kan van alles onderzocht worden (prestatie, gebruikersaspecten, produceerbaarheid, kosten, etc). In veel construerende disciplines wordt een ontwerp daarom gezien als een platform voor onderzoek. Sommigen karakteriseren een ontwerp als het analogon in de construerende wetenschappen van een experiment in de natuurwetenschappen [Hee, van en Van Overveld, 2010].

2. KNELPUNTEN BIJ DE KWALITEITSBEOORDELING

2.1 Inleiding

Wetenschappelijk onderzoek kent een grote verscheidenheid. Deze verscheidenheid komt voort uit de verschillende onderzoekstradities en stijlen in de alfa-, bèta- en gammadisciplines. Binnen elk van deze disciplinaire clusters is er opnieuw sprake van verschillende typen van onderzoek. Deze verschillen worden vaak beschreven in termen van het doel van het onderzoek: het streven naar fundamenteel begrip of het oplossen van praktische problemen. Hiermee hangt samen de indeling van het onderzoek in termen van fundamenteel en toegepast onderzoek. De discussie in de literatuur laat zien dat dit geen elkaar uitsluitende categorieën zijn. Een vaak gebruikte categorisering van wetenschappelijk onderzoek is weergegeven in tabel 2.1.

Tabel 2.1 Categorisering van wetenschappelijk onderzoek, ontleend aan [Stokes, 1997, p 73].

		Considerations of use?	
		No	Yes
Quest for fundamental understanding?	Yes	Pure basic research	Use-inspired basic research
	No		Pure applied research

Binnen de technische universiteiten is het spectrum van wetenschappelijke activiteiten nog groter en complexer omdat ook ontwerpen en construeren onderdeel uitmaken van het wetenschappelijke portfolio. Voor technische universiteiten zijn deze activiteiten essentieel. Het is immers onderdeel van hun missie om bij te dragen aan

de oplossing van maatschappelijke problemen of aan het creëren van economische kansen en met name bij het ontwerpen en construeren wordt de verbinding met de praktijk gemaakt.

De hierboven genoemde verscheidenheid in wetenschappelijke activiteiten roept de vraag op of het mogelijk is om met één meetlat de kwaliteit van al deze activiteiten te beoordelen. Kunnen criteria die bijvoorbeeld ontleend zijn aan het fundamentele (natuurkundige) onderzoek probleemloos worden gebruikt bij de beoordeling van de kwaliteit van het toegepaste onderzoek, of van construerende en ontwerpende activiteiten? Deze vraag wordt door de beoefenaars van de ontwerpende en construerende disciplines meestal ontkennend beantwoord en vormt de kern van de controverse over de beoordeling van kwaliteit in deze disciplines.

De discussie hierover speelt al vele jaren. In 2000 bracht het Discipline-overlegorgaan Construerende Techniek (DCT) het advies uit: *Criteria voor de Beoordeling van het Ontwerpen in de Construerende Disciplines* [DCT, 2000]. In 2009 verscheen het advies *Kwaliteitsbeoordeling Ontwerp- en Constructiedisciplines* [Schouten, 2009]. In deze adviezen lag de nadruk vooral op de beoordeling van ontwerpen als wetenschappelijke activiteit. Het onderhavige KNAW advies richt zich niet alleen op ontwerpende en construerende activiteiten maar ook op het daaraan gerelateerde wetenschappelijke onderzoek.

Toenemende urgentie

Het belang van een goed beoordelingskader voor kwaliteit in de ontwerpende en construerende disciplines is in de loop van de jaren alleen maar groter geworden. Ten eerste worden onderzoeksprojecten en tijdelijk wetenschappelijk personeel in toenemende mate alleen nog extern gefinancierd. De ontwerpende en construerende disciplines zijn daarbij in competitie met andere disciplines en moeten het opnemen tegen natuurwetenschappelijke disciplines die beter passen bij de bestaande kwaliteitscriteria. Ten tweede is kwaliteitszorg aan de universiteiten steeds belangrijker geworden (externe visitaties, interne *midterm reviews*) en daaraan worden consequenties verbonden door bestuurders. De criteria voor wetenschappelijke kwaliteit die daarbij gehanteerd worden zijn dan cruciaal. Dit probleem speelt overigens niet alleen bij de ontwerpende en construerende disciplines. In het algemeen geldt dat toegepast multidisciplinair onderzoek minder goed scoort op traditionele kwaliteitscriteria. Het gaat vaak om kleinere deelgebieden waarin multidisciplinaire tijdschriften met lagere impactfactoren een belangrijke rol spelen, waarbij het citatiegedrag meestal asymmetrisch is: er wordt in de multidisciplinaire tijdschriften veel verwezen naar artikelen in monodisciplinaire tijdschriften, maar omgekeerd veel minder.

In de volgende paragraaf wordt aangegeven welke knelpunten in de praktijk worden ervaren. De informatie is afkomstig uit de interviews die de commissie gehouden heeft met vooraanstaande wetenschappers in het veld (zie bijlage 1).

2.2 Ervaren knelpunten

Vergelijken van kwaliteit tussen vakgebieden problematisch

De meeste geïnterviewden in de technische wetenschappen vinden dat de ontwerpende en construerende wetenschappen niet goed te beoordelen zijn met de traditionele beoordeling op basis van wetenschappelijke publicaties in ISI-tijdschriften met hoge impactfactoren. De mate waarin men daar zelf last van ondervindt, bijvoorbeeld bij het verwerven van onderzoeksmiddelen, hangt af van het gebied waarin men werkt. De beoordeling van kwaliteit binnen een vakgebied door *peers* wordt daarentegen vrijwel nergens als een probleem ervaren. Onderling weet men goed wie de topwetenschappers onder de collega's zijn. De problemen ontstaan bij vergelijking tussen vakgebieden.

Ontbreken van een duidelijk loket voor ontwerpende en construerende disciplines

Het vergelijken van de kwaliteit van wetenschappelijke werk tussen vakgebieden gebeurt onder andere bij het in competitie verwerven van onderzoeksfinanciering. Als vakgebieden teveel van elkaar verschillen is dit problematisch. Onderzoekers in de ontwerpende en construerende wetenschappen ondervinden problemen omdat meer fundamenteel onderzoek meestal als wetenschappelijk hoogstaander wordt gezien. Het ontbreken van een duidelijk loket voor de ontwerpende en construerende wetenschappen wordt door veel geïnterviewden als gemis ervaren. De wetenschappers die aangeven geen problemen te hebben met de huidige manier van kwaliteitsbeoordeling, ontvangen vaak onderzoeksfinanciering uit het bedrijfsleven of uit meer specifieke subsidieprogramma's zoals 'Point-One' of de Innovatiegerichte Onderzoeksprogramma's (IOP).

De recente overheveling van 100 miljoen euro van de eerste naar de tweede geldstroom heeft het ontbreken van een loket bij NWO voor de ontwerpende en construerende wetenschappen extra urgent gemaakt. Op dit moment bestaan er volgens een aantal geïnterviewden geen gelijke kansen voor deze disciplines bij NWO.

Voor de ontwerpende en construerende disciplines is een extra complicatie dat op een aantal gebieden, bijvoorbeeld de weg- en waterbouw of de volkshuisvesting, intermediaire onderzoeksorganisaties zijn opgeheven, minder kennisintensief zijn geworden of geen onderzoek meer financieren. Hierdoor zijn er minder mogelijkheden voor derde geldstroomonderzoek, terwijl de kansen in tweede geldstroomprogramma's waarbij betrokkenheid van maatschappelijke *stakeholders* wordt geëist afnemen. Ook daardoor is de urgentie van een loket bij NWO toegenomen.

Grote verschillen in publicatiecultuur te weinig erkend

Geen van de geïnterviewden betwist het belang van publiceren als methode van kennisverspreiding en toetsing van resultaten door *peer review*. De grote nadruk op *peer reviewed*-artikelen in ISI-tijdschriften doet echter te weinig recht aan de grote verschillen in publicatiecultuur tussen vakgebieden. Binnen de informatica bijvoorbeeld hebben sommige *peer-reviewed conference proceedings* een hogere status dan publicaties in ISI-tijdschriften. Ook het aantal publicaties dat een onderzoeker jaarlijks kan produceren verschilt sterk per discipline. In de construerende disciplines gaat veel tijd zitten in het ontwerp en de constructie voordat daarover gepubliceerd kan worden. Diverse geïnterviewden rapporteren een steeds grotere druk om te publiceren in hoge impacttijdschriften als *Science* en *Nature* terwijl die tijdschriften door vakgenoten niet worden gezien als leidend voor hun vakgebied.

Geïnterviewden uit disciplines als industrieel ontwerpen en bouwkunde geven aan dat publicaties voor hen in toenemende mate van belang zijn als onderdeel van een strategie van verwetenschappelijking van hun vakgebied. Daarbij moet wel rekening worden gehouden met het feit dat de vorm en het medium waarin gepubliceerd wordt passend is voor de betreffende discipline.

Afnemende rol van ontwerpen en construeren in de technische disciplines

Ontwerpen en construeren zijn pijlers van de ingenieurswetenschappen en vormen een brug naar de praktijk. In een ontwerp of constructie wordt kennis uit verschillende disciplines geïntegreerd voor het realiseren van functies van een systeem ten behoeve van *real-life*-situaties. Veel geïnterviewden constateren dat de verwetenschappelijking in sommige technische wetenschappen (bijvoorbeeld civiele techniek) geleid heeft tot een 'monoscience-achtige cultuur' waarin het ontwerpen en construeren van technische systemen minder aandacht krijgen dan nodig is en er vaak een grote kloof met de praktijk is ontstaan. Dat is een ongewenste situatie.

Grotere afstand tussen universitaire bestuur en werkvloer

Veel geïnterviewden ervaren dat kwaliteitsbeoordeling van wetenschappelijk werk aan de universiteiten is verworden tot een bureaucratisch instrument. Vanwege de toegenomen afstand tussen werkvloer en bestuur is het moeilijk om op bestuurlijk niveau goed zicht te hebben op de kwaliteit van het onderzoek in een bepaalde discipline. Er is daarom behoefte aan 'objectieve' en bij voorkeur kwantitatieve informatie over de wetenschappelijke kwaliteit van verschillende groepen. Bibliometrische analyses of andere cijfermatige indicatoren worden dan gebruikt om een vergelijking te maken tussen die groepen. De ontwerp- en construerende wetenschappen zijn hierbij in het nadeel ten opzichte van de disciplines die beter passen bij de bestaande kwaliteitscriteria. Dat nadeel betreft niet alleen de financiering van die vakgebieden, maar zeker ook de wetenschappelijke reputatie en erkenning.

2.3 Evaluatie ervaren knelpunten en conclusies

Verschillen tussen vakgebieden

De commissie onderschrijft de opvatting dat het vergelijken van wetenschappelijke kwaliteit tussen verschillende vakgebieden zinloos is. Zo'n vergelijking veronderstelt dat er kwaliteitsindicatoren bestaan die voor de betreffende disciplines in gelijke mate gelden of zouden moeten gelden. Die aanname is onjuist. De onderzoekstradities, publicatiegewoontes, soorten activiteiten en producten als tijdschriftartikelen, *proceedings*, boeken, artefacten en ontwerpen zijn daarvoor te divers. Uitspraken waarin bijvoorbeeld de wetenschappelijke kwaliteit van personen in verschillende vakgebieden worden vergeleken op basis van ISI-tijdschriftartikelen en impactfactoren hebben daarom geen betekenis. Ook de veelgebruikte Hirsch-index biedt voor een vergelijking geen basis. Hooguit kan een vergelijking worden gemaakt tussen gegevens die voor de disciplines genormeerd zijn. Het is dan mogelijk om te zeggen dat wetenschapper A tot de top 5 procent behoort in discipline X en dat wetenschapper B tot de top 25 procent behoort in discipline Y.

Kwaliteit kan dus alleen binnen disciplines zinvol worden vergeleken. Een *one-size-fits-all*-benadering, hoe gewenst ook vanuit bestuurlijk oogpunt vanwege zijn eenvoud, doet geen recht aan significante disciplinaire verschillen en zal dus altijd bepaalde vakgebieden ten onrechte bevoordelen. Deze opvatting is in lijn met eerdere conclusies van de KNAW commissie kwaliteitszorg [KNAW-commissie kwaliteitszorg, 2008].

Peer review

De beoordeling van wetenschappelijke kwaliteit binnen disciplines gebeurt meestal door *peers*. *Peeroordelen*, hoe zorgvuldig en genuanceerd zij ook tot stand komen, kunnen echter nooit volledig objectief zijn en recht kan doen aan alle verschillen en controverses binnen een discipline. Er is sprake van een niet te vermijden persoonlijke *bias* en er spelen altijd subjectieve elementen mee. Verschillende *peers* zullen daarvoor tot gedeeltelijk andere oordelen kunnen komen. Ook bestaat er een risico dat een klein aantal *peers* een discipline domineren en dat eerdere beoordelingen het latere oordeel vertekenen [KNAW, 2008]. Ondanks deze beperkingen vindt de commissie dat *peer review* de beste manier van het beoordelen van kwaliteit is. Voorwaarde is wel dat een vakgebied voldoende homogeen is en voldoende (internationale) omvang heeft om onafhankelijke *peeroordelen* mogelijk te maken. Als hieraan is voldaan leveren kwaliteitsoordelen door *peers* óók binnen de ontwerpende en construerende wetenschappen betrekkelijk weinig problemen op.

Bibliometrisch onderzoek

De beperkingen van *peer review* hebben mede geleid tot het doen van bibliometrisch onderzoek (naast de hierboven genoemde groter wordende afstand tussen bestuur

en werkvloer). Het is echter een illusie om te denken dat een volledig kwantitatieve/ bibliometrische benadering van wetenschappelijke kwaliteit mogelijk is, of *peer review* zou kunnen vervangen. Niet in de laatste plaats omdat *peers* in het vakgebied nodig zijn om de bibliometrische gegevens over dat vakgebied te kunnen beoordelen en duiden. Bibliometrisch onderzoek kan naar de mening van de commissie hooguit aanvullend zijn op *peer review*. Beide benaderingen hebben hun beperkingen, en het is belangrijk om deze beperkingen voor ogen te houden. Kwaliteitsoordelen moeten daarom altijd met voorzichtigheid gehanteerd worden.

Momentopname

Een extra reden voor voorzichtigheid is dat kwaliteitsoordelen altijd momentopnames zijn en dat de kwaliteit van mensen en hun werk in de loop van de tijd kan variëren. Beoordelingscommissies moeten zich hiervan bewust zijn.

Publicaties ook voor ontwerpende en construerende wetenschappen van belang

De commissie hecht veel waarde het publiceren van wetenschappelijke resultaten in *peer-reviewed* tijdschriften, ook voor vakgebieden als industrieel ontwerpen en bouwkunde. Ten eerste omdat voorafgaande aan een publicatie toetsing door *peers* plaatsvindt van de resultaten van wetenschappelijk werk. Ten tweede omdat door publicaties de kennis die opgebouwd is wordt ontsloten en verspreid. De publicatievormen moeten wel passen bij het vakgebied. Een eenzijdige focus op publicaties in monodisciplinaire tijdschriften met een hoge impactfactor is niet wenselijk, zeker niet voor een technische universiteit waar onderzoek met een multidisciplinair en toepassingsgericht karakter ook belangrijk is. De commissie heeft met instemming kennisgenomen van het toenemende belang dat wetenschappers in vakgebieden als industrieel ontwerpen en bouwkunde toekennen aan *peer-reviewed* publicaties als onderdeel van de wetenschappelijke output.

Criteria sturend bij richting onderzoek

De criteria waarmee de kwaliteit van wetenschappelijk werk wordt beoordeeld spelen zelf weer een rol bij beslissingen over welke typen van activiteiten worden ontplooid aan de universiteiten. Zij werpen hun schaduw vooruit omdat onderzoekers anticiperen op de criteria waarmee hun wetenschappelijk werk beoordeeld zal gaan worden en op basis waarvan ook de financiering van hun werk zal plaatsvinden. Kwaliteitscriteria zijn daarom niet neutraal maar normerend voor een vakgebied. Naast hun rol in het ex postbeoordelen van wetenschappelijk werk spelen ze ook een ex antero rol bij de sturing van het onderzoek. In dit verband onderschrijft de commissie de zorg van een aantal geïnterviewden dat een eenzijdige focus op criteria die passen bij

natuurwetenschappelijk onderzoek in sommige technische disciplines ertoe geleid heeft dat ontwerpen en construeren en het daaraan gerelateerde multidisciplinaire onderzoek onder druk zijn komen te staan. Het is gezien de missie van een technische universiteit van groot belang om een goed evenwicht te bereiken tussen de verschillende soorten onderzoek die eerder onderscheiden zijn (paragraaf 2.1, tabel 2.1) en de ontwerpende en construerende activiteiten. Dat kan alleen maar lukken als de kwaliteitscriteria die gebruikt worden daarop zijn toegesneden.

Onderzoeksfinanciers

Ook voor onderzoeksfinanciers zoals NWO en STW geldt dat een *one-size-fits-all*-benadering geen recht doet aan relevante disciplinaire verschillen en dus altijd bepaalde vakgebieden ten onrechte zal bevoordelen. Verschillende geïnterviewden vinden dat er bij NWO geen gelijke kansen bestaan voor de ontwerpende en construerende disciplines. Meer in het algemeen geldt dat binnen de structuur van NWO multi- of interdisciplinair onderzoek niet vanzelfsprekend past en moet concurreren met disciplinair onderzoek dat veel beter uit de voeten kan met de gehanteerde kwaliteitscriteria. Ook zijn er gevestigde belangen in het geding waartegen nieuwe vakgebieden het moeten opnemen. De commissie is van mening dat NWO het probleem onderkend heeft met het initiëren van gebiedsoverstijgende programma's. Dit is een lijn die verder versterkt moet worden, bijvoorbeeld door ook programma's te starten tussen gebieden van NWO en STW. Wat STW betreft hebben verschillende geïnterviewden aangegeven dat de maatschappelijke relevantie van het onderzoek te beperkt wordt geïdentificeerd met kortetermijninteresse van bedrijven in gebruikerscommissies. Het zou volgens de commissie goed zijn ook hier een meer gedifferentieerde benadering te kiezen en ook onderzoeksprojecten mogelijk te maken die een relevantie op de langere termijn hebben, ook als daarvoor nog geen bedrijven kunnen worden gevonden. Dat is ook voor de ontwerpende en construerende disciplines van belang.

Transparante doelstelling van belang

Ten slotte stelt de commissie vast dat kwaliteitscriteria moeten zijn afgestemd op het doel dat een kwaliteitsoordeel dient. De beoordeling van een onderzoeksvoorstel is een andere beoordeling dan die van de *past performance* van een onderzoeksgroep of van het functioneren van een medewerker bij bevordering of vaste aanstelling.

CONCLUSIE 2.1

Wetenschappelijke kwaliteit kan alleen binnen disciplines zinvol worden vergeleken. Een *one-size-fits-all*-benadering bij de beoordeling van kwaliteit, hoe gewenst ook vanuit bestuurlijk oogpunt vanwege zijn eenvoud, doet geen recht aan significante disciplinaire verschillen en zal dus altijd bepaalde vakgebieden ten onrechte bevoordelen.

CONCLUSIE 2.2

Binnen de ontwerpende en construerende disciplines zijn er betrekkelijk weinig problemen met het vaststellen van de kwaliteit van wetenschappelijk werk of het aanwijken van excellente wetenschappers door *peers*.

CONCLUSIE 2.3

Voor technische universiteiten is het van belang dat de criteria voor wetenschappelijke kwaliteit toegesneden zijn op de verschillende typen onderzoek en activiteiten die daar plaatsvinden. Een eenzijdige focus op criteria die ontleend zijn aan de natuurwetenschappen bemoeilijkt een goed evenwicht tussen fundamenteel onderzoek, toegepast onderzoek en ontwerpende en construerende activiteiten.

CONCLUSIE 2.4

Het publiceren van wetenschappelijke resultaten in *peer-reviewed* tijdschriften, boeken en *proceedings*, is ook voor de ontwerpende en construerende disciplines van groot belang. De publicatievorm moet wel passen bij de aard van het vakgebied.

3. BEOORDELINGSCRITERIA

3.1 Inleiding

In dit hoofdstuk worden criteria en indicatoren geformuleerd die volgens de commissie passend zijn voor de beoordeling van kwaliteit in de ontwerpende en construerende wetenschappen.

Terminologische opmerking

Kwaliteit is een waarderend begrip waarmee wordt aangegeven hoe goed iets is. Het is een generiek begrip dat van toepassing is op objecten, processen en personen. Voor die verschillende situaties gelden verschillende beoordelingscriteria. De kwaliteit van een publicatie bijvoorbeeld wordt anders bepaald dan de kwaliteit van de computer waarop deze is gemaakt. Uitspraken over kwaliteit zeggen niet zoveel als daarbij niet wordt aangegeven op welke criteria deze kwaliteitsoordelen gebaseerd zijn. Hoe goed iets voldoet aan een beoordelingscriterium is meestal niet direct waarneembaar. Daarom wordt gewerkt met indicatoren. Deze geven empirische informatie over de mate waarin een object, proces of persoon voldoet aan een bepaald criterium. Dat hoeft niet per se kwantitatieve informatie te zijn, zoals bijvoorbeeld het aantal publicaties, het kan ook gaan om meer kwalitatieve informatie, zoals bijvoorbeeld het oordeel van collega's over de kwaliteit van het tijdschrift waarin die publicaties staan.

Uitgangspunten bij de beoordeling

Rekening houden met disciplinaire verschillen

Zoals eerder besproken zijn er significante verschillen tussen disciplines. Maar ook binnen het relatief beperkte gebied van de ontwerpende en construerende wetenschappen is de variatie groot. In brede kring wordt gedeeld dat de huidige praktijk van

kwaliteitsbeoordeling deze diversiteit onvoldoende in acht neemt. Met deze diversiteit moet naar mening van de commissie rekening worden gehouden in de beoordelings-systematiek.

Differentiatie noodzakelijk...

Gegeven de verschillen tussen disciplines en gegeven de verschillen tussen beoordelings-situaties moet volgens de commissie differentiatie een uitgangspunt zijn bij het beoordelen van wetenschappelijke kwaliteit. Dit betreft:

- differentiatie naar (deel)vakgebied
- differentiatie naar type activiteit binnen een vakgebied: van onderzoeken tot ontwerpen en construeren
- differentiatie naar het object van beoordeling: een medewerker, een onderzoeksvoorstel, of een onderzoeksgroep

...maar overkoepelend raamwerk nodig

Aan de andere kant meent de commissie dat de beoordelingsmaatstaven voor een bepaalde (deel)discipline niet volledig verschillend kunnen zijn van die van andere disciplines. Dat zou deze maatstaven arbitrair maken en opportunistisch, terwijl ook de indruk zou kunnen ontstaan dat aan de ontwerpende en construerende disciplines 'lichtere' eisen worden gesteld dan aan andere disciplines. Een overkoepelend beoordelingskader is daarom gewenst waarbinnen de genoemde diversiteit een plaats moet krijgen. De commissie heeft daarom gestreefd naar overkoepelende criteria voor de beoordeling van kwaliteit die ook breder van toepassing zijn, met indicatoren die per discipline kunnen verschillen of een ander belang kunnen hebben.

Peer review centraal

In het vorige hoofdstuk is al aan de orde gekomen dat volgens de commissie *peer review* het uitgangspunt moet zijn bij de beoordeling van de kwaliteit van publicaties, onderzoekers, onderzoeksgroepen en onderzoeksvorstellen. Voorwaarde hiervoor is dat een discipline voldoende homogeen is, voldoende (internationale) omvang heeft en dat er voldoende goede *peers* beschikbaar zijn.

Bibliometrie ondersteunend en niet vervangend

Bibliometrische gegevens zijn naar het oordeel van de commissie ondersteunend voor kwaliteitsoordelen van *peers* en niet vervangend. Dit kwam eerder aan de orde. Voor het bepalen van de waarde van bibliometrische indicatoren (citatiescores, impactfactoren, H-index, etc.) is de definitie van wat als een vakgebied wordt beschouwd cruciaal en geen triviale aangelegenheid.

Transparante doelstellingen

Kwaliteitsbeoordelingen zoals onderzoeksvisitaties worden in praktijk vaak gebruikt als *ranking*- en pr-instrument, niet als instrument om mogelijke verbeteringen te

identificeren. Het doel van kwaliteitsbeoordelingen moet voorafgaand aan de beoordeling duidelijk worden gemaakt en de beoordeling moet ook in het licht van deze doelstelling worden gezien.

3.2 Beoordelingscriteria

Van vijf naar slechts twee criteria

In eerste instantie heeft de commissie vijf beoordelingscriteria overwogen, aansluitend bij de benadering die de Royal Academy of Engineering in het Verenigd Koninkrijk heeft voorgesteld voor de beoordeling van kwaliteit in de technische disciplines. Deze criteria waren: 1) publicaties, 2) impact, 3) innovativiteit, 4) betrokkenheid van externe *stakeholders* en 5) reputatie van de betrokken wetenschappers. Voor deze criteria was tijdens de consultatieronde van de commissie met het veld in Nederland veel steun. Bij nadere uitwerking bleken sommige van deze criteria echter indicatoren te zijn die beter bij andere criteria konden worden ondergebracht. Publicaties in gerenommeerde tijdschriften vormen bijvoorbeeld geen zelfstandig criterium maar zijn een indicator voor het criterium wetenschappelijke kwaliteit. Ook de betrokkenheid van externe *stakeholders* bij het onderzoek is geen criterium maar een indicator voor de maatschappelijke relevantie van het betreffende onderzoek. Over innovativiteit als zelfstandig criterium heeft de commissie lang gearzeld. Uiteindelijk is besloten dat innovativiteit twee aspecten combineert en daarom niet zelf een criterium is: wetenschappelijke vernieuwing en maatschappelijke toepasbaarheid. Het eerste aspect komt in wetenschappelijke publicaties tot uitdrukking, het tweede aspect kan door indicatoren voor maatschappelijke relevantie worden weergegeven. De reputatie van wetenschappers ten slotte is naar de mening van de commissie een afgeleide grootte die zowel wetenschappelijke als maatschappelijke aspecten omvat.

De commissie heeft uiteindelijk geconcludeerd dat er in de kern van de zaak maar twee beoordelingscriteria zijn voor de kwaliteit van activiteiten in de ontwerpende en construerende wetenschappen:

1. wetenschappelijke kwaliteit
2. maatschappelijke relevantie

Deze conclusie is opmerkelijk omdat de voorgestelde criteria niet afwijken van de reeds bestaande criteria voor andere wetenschapsgebieden. Anders dan in eerste instantie gedacht, wordt er dus géén aparte set van criteria voor kwaliteit in de ontwerpende en construerende wetenschappen voorgesteld. De bovenstaande twee criteria worden breed gedragen in het wetenschappelijke veld en komen goed overeen met wat internationaal gezien gebruikelijk is (zie hoofdstuk 4). Zij lijken een overkoepevend raamwerk te bieden voor de beoordeling van kwaliteit in alle wetenschapsgebieden.

Gegeven dit overkoepelende raamwerk pleit de commissie vervolgens voor differentiatie ten aanzien van de indicatoren die gebruikt worden om te bepalen hoe goed iets aan de twee beoordelingscriteria voldoet:

1. differentiatie naar discipline: de aard en het relatieve belang van de gebruikte indicatoren zal in de verschillende wetenschapsgebieden verschillend zijn; de ontwerpende en construerende wetenschappen kennen bijvoorbeeld ook andere indicatoren voor wetenschappelijke kwaliteit dan aantallen en citaties in *peer-reviewed* tijdschriften.
2. differentiatie naar het type activiteit in een bepaalde discipline: onderzoeken, ontwerpen of construeren.
3. differentiatie naar het object van beoordeling: werk of persoon. Bij de beoordeling van een subsidieaanvraag voor een onderzoeksproject zijn vaak andere indicatoren relevant dan bij de beoordeling van een persoon voor een benoeming of promotie.

In tabel 3.1 zijn de twee beoordelingscriteria en de daarbij horende indicatoren weergegeven die volgens de commissie geschikt zijn om kwaliteit in de ontwerpende en construerende wetenschappen te beoordelen. Zoals gezegd zijn niet alle indicatoren even relevant of hebben ze hetzelfde relatieve gewicht voor elke (deel)discipline, activiteit of beoordelingssituatie. De tabel beoogt de belangrijkste typen van indicatoren weer te geven. Indicatoren voor de kwaliteit van wetenschappelijk onderwijs zijn niet in de tabel opgenomen en vallen buiten de opdracht van dit advies.

3.3 Toelichting op de indicatoren

In deze paragraaf worden de indicatoren in de tabel voor zover nodig nader toegelicht.

3.3.1 Criterium wetenschappelijke kwaliteit

Wetenschappelijke publicaties

Zoals eerder vermeld zijn wetenschappelijke publicaties ook voor de ontwerpende en construerende wetenschappen een belangrijke indicator voor wetenschappelijke kwaliteit. De vorm van deze publicaties kan echter verschillen voor de verschillende vakgebieden (bijv. tijdschriften, *proceedings*, elektronische publicaties).

Overige output: ontworpen artefacten

De commissie vindt dat bij de beoordeling van de kwaliteit van het wetenschappelijke werk in de construerende en ontwerpende wetenschappen ook ontworpen artefacten (producten, processen, software) een rol moeten kunnen spelen. Hierbij moeten wel enkele kanttekeningen worden geplaatst. Het gaat hierbij alleen om ontworpen artefacten die een wetenschappelijke bijdrage leveren aan het vakgebied, niet om artefacten die ontworpen zijn in het kader van het onderwijs. Essentieel voor een wetenschappelijk ontwerp is dat het nieuwe kennis oplevert die generiek toepasbaar is.

Tabel 3.1 kwaliteitsindicatoren voor de technische wetenschappen

	INDICATOREN VOOR WERK	INDICATOREN VOOR PERSOON
WETENSCHAPPELIJKE KWALITEIT	<p>Wetenschappelijke publicaties Artikelen in <i>peer-reviewed</i> tijdschriften (aantal en soort tijdschrift) Artikelen in <i>peer-reviewed conference proceedings</i> (aantal en soort proceedings) Wetenschappelijke boeken bij gerenommeerde uitgevers, of significante bijdragen daaraan (aantal en soort) Citaties van individuele artikelen Impactfactoren van artikelen van tijdschriften waarin is gepubliceerd</p> <p>Overige output Peer-reviewed artefact (ontwerp) + documentatie. Hieronder valt ook het ontwerp van software</p> <p>Wetenschappelijke impact (ex post) Gebruik van wetenschappelijke producten door andere wetenschappers (artefacten, methoden, meetinstrumenten, tools, standaarden en protocollen)</p> <p>Potentiële wetenschappelijke impact Mogelijke bijdrage aan theorie- en modelvorming, methodeontwikkeling, werkingsprincipes of ontwerpconcepten</p>	<p>Ontvangen blijken van wetenschappelijke erkenning Ontvangen wetenschappelijke prijzen Lidmaatschappen van prestigieuze organisaties zoals academies van wetenschappen Prestigieuze subsidies zoals Veni-, Vidi-, Vici-beurzen, of ERC grants Eredocoraten Visiting professorships</p> <p>Redacteurschappen Chief/full editorschips internationaal wetenschappelijk tijdschrift/boek/conference proceedings</p> <p>Mate waarin men door peers wordt benaderd als expert Adviseurschappen in wetenschappelijke kring (NWO, visitaties, etc.) Keynote lezingen op wetenschappelijke congressen Lidmaatschappen programmacommissies Deelname internationale beoordelingscommissies van wetenschappelijke programma's/instituten of wetenschappelijke adviesraden/instituten</p> <p>Wetenschappelijke impact over gehele loopbaan Citatiescore van persoon Bijdrage aan wetenschappelijke schoolvorming</p>
MAATSCHAPPELIJKE RELEVANTIE	<p>Gebruik van resultaten door externe stakeholders (ex postimpact) Bijdrage aan oplossing maatschappelijke problemen Marktintroducties en nieuwe projecten in industrie Inkomsten door gebruik van resultaten Spin-offs met industrie Gebruikte patenten Gebruikte artefacten (ontwerpen, software)</p> <p>Gebruik van resultaten door professe (ex post impact) Gebruik van artefacten, methoden, meetinstrumenten, tools, standaarden en protocollen</p> <p>Betrokkenheid van externe stakeholders bij wetenschappelijk werk (potentiële maatschappelijke impact) Bedrijven of maatschappelijke organisaties betrokken bij de begeleiding van onderzoeksprojecten (bijvoorbeeld in gebruikerscommissies) Derde geldstroomfinanciering door mogelijke gebruikers (bijv. industrie) Publieke financiering gerelateerd aan maatschappelijke kwesties Valorisation grants</p> <p>Bijdrage aan verspreiding van kennis Professionele publicaties en voordrachten, niet-wetenschappelijke publicaties, tentoonstelling etc. over onderzoeksresultaten</p>	<p>Mate waarin men door externe stakeholders wordt gezien als expert Advies- en consultancywerkzaamheden (gericht op gebruikers) Leidende positie in industrieel onderzoek (bijv. directeur R&D-afdeling)</p> <p>Mate waarin men binnen de professie gezien wordt als expert Oeuvreprijzen (bijvoorbeeld architecten) Overzichtstentoonstellingen</p> <p>Bijdrage aan verspreiding van kennis Activiteiten gericht op popularisering wetenschap, onderwijs en bijdragen aan het publieke debat Training van professionals PhD's met eerste baan in relevante praktijk</p>

De documentatie bij het artefact moet hierover inzicht geven. Als een ontwerp nieuwe kennis oplevert dan is deze kennis meestal ook te publiceren in wetenschappelijke tijdschriften, waardoor kwaliteitsborging via *peer review* gegarandeerd wordt. Dat deze kennis in beginsel publiceerbaar is, betekent nog niet dat hij ook altijd gepubliceerd is. Daarvoor kunnen allerlei goede redenen zijn, die bijvoorbeeld te maken kunnen hebben met de bedrijfsmatige omgeving waarin deze kennis tot stand is gekomen en waarin publiceren niet gebruikelijk of zelfs niet toegestaan is.

Artefacten kunnen als indicator voor wetenschappelijke kwaliteit een rol spelen bij de benoeming van hoogleraren die een tijd in het bedrijfsleven hebben gewerkt. Gezien het belang van dit soort kandidaten voor de technische universiteiten moeten deze wetenschappelijke prestaties mee kunnen wegen. Het is echter essentieel dat in zo'n geval een oordeel van *peers* wordt gevraagd over de wetenschappelijke merites van een (beperkt) aantal goed gedocumenteerde ontworpen artefacten.

Wetenschappelijke impact

Onder wetenschappelijke impact wordt verstaan de mate waarin het wetenschappelijke werk gebruikt wordt door andere wetenschappers. Een van de meest voorkomende vormen van gebruik is het citeren van wetenschappelijk werk. Een andere indicator is het gebruik van nieuw ontwikkelde wetenschappelijke producten door collega wetenschappers. Met wetenschappelijk producten wordt hier bedoeld artefacten, methoden, meetinstrumenten, hulpmiddelen, standaarden, protocollen etc. Het gebruik van wetenschappelijke resultaten door niet-wetenschappers valt niet onder dit criterium maar onder het criterium maatschappelijke relevantie.

Potentiële wetenschappelijke impact

Deze indicator speelt een rol bij de beoordeling van de kwaliteit van onderzoeksvoorstellen. Deze worden mede beoordeeld op hun mogelijke bijdrage aan het vakgebied.

Ontvangen blijken van wetenschappelijke erkenning

Blijken van erkenning zijn een belangrijke graadmeter voor de wetenschappelijke kwaliteit van een persoon. Voorwaarde hiervoor is dat het blijken van erkenning zijn die door *peer review* tot stand zijn gekomen. Voorbeelden zijn Spinozapremies, eredoctoraten, lidmaatschappen van prestigieuze organisaties als academies van wetenschappen en prestigieuze subsidies zoals *Vici's* of ERC Grants.

Mate waarin men door peers wordt benaderd als expert

De mate waarin een wetenschapper door collega's wordt gezien als expert is een belangrijke indicator voor kwaliteit omdat binnen een vakgebied meestal goed duidelijk is wie de topwetenschappers zijn. Dit kan bijvoorbeeld blijken uit *keynote* lezingen op wetenschappelijke congressen, lidmaatschappen van programmacommissies en uit adviseurschappen in wetenschappelijke kring.

Wetenschappelijke impact over gehele loopbaan

Hieronder wordt verstaan de wetenschappelijke *impact* die een persoon heeft gehad gedurende zijn of haar loopbaan. Dit kan bijvoorbeeld blijken uit de citatiescore van een persoon. Ook het oordeel van *peers* over de mate waarin een wetenschapper heeft bijgedragen aan wetenschappelijke schoolvorming kan hier worden meegewogen.

3.3.2 Criterium maatschappelijke relevantie

Gebruik van resultaten

Hieronder vallen de indicatoren die een maat zijn voor het gebruik van de wetenschappelijke kennis of producten buiten het wetenschappelijke veld. Het daadwerkelijke gebruik is meestal alleen te beoordelen na afronding van het onderzoek. Dat gebruik kan vele vormen aannemen: van de bijdrage aan een maatschappelijk probleem tot het commerciële gebruik in een bedrijf. Een aparte categorie betreft het gebruik van resultaten door de beoefenaars van een professie in de maatschappij.

Potentiële impact

De maatschappelijke relevantie van onderzoek kan op verschillende tijdschalen beoordeeld worden. Toegepast onderzoek heeft een kortere tijdshorizon dan fundamenteel onderzoek. Voor het langere termijnonderzoek is het echter notoir moeilijk om de maatschappelijke relevantie vast te stellen. Hier kan gewerkt worden met indicatoren die de mate van interesse van maatschappelijke *stakeholders* in het onderzoek uitdrukken. Die *stakeholders* zijn overigens niet alleen bedrijven, maar ook ministeries of internationale organisaties. Voor het kortere termijnonderzoek is de betrokkenheid van *stakeholders* door co-financiering van het onderzoek een indicator. Deze betrokkenheid kan bovendien de kans op daadwerkelijk gebruik van de resultaten sterk vergroten.

Bijdrage aan verspreiding van kennis

Zowel bij de beoordeling van wetenschappelijk werk als bij de beoordeling van een persoon moet de bijdrage aan de verspreiding van kennis worden meegewogen als onderdeel van maatschappelijke relevantie. Kennisverspreiding draagt immers bij aan de oplossing van maatschappelijke problemen of het creëren van economische kansen. Dit is ook een onderdeel van de missie van een technische universiteit.

3.4 Het relatieve belang van de criteria en de indicatoren

Beoordelen is maatwerk en voor elke beoordelingssituatie moet voorafgaand aan de beoordeling worden bepaald welke van de twee criteria en welke van de indicatoren daarvoor het belangrijkste zijn. Voor een voltijds hoogleraar die een wetenschappelijke groep moet leiden, ligt dit anders dan voor een parttime hoogleraar die vooral de brug naar de praktijk moet slaan. Voor een persoonsgebonden Vidi-beurs ligt dit

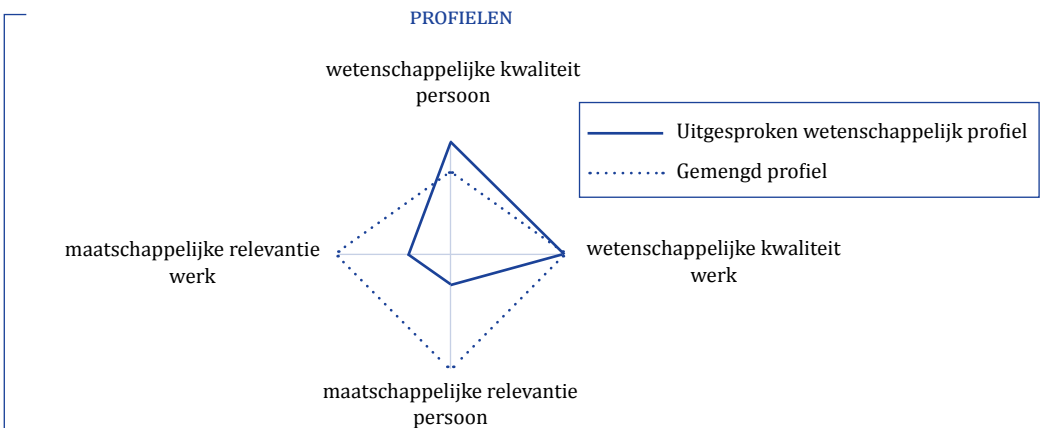
weer anders dan voor een aanvraag in de open competitie van NWO- of voor een STW-aanvraag. Om nog maar niet te spreken van de verschillen per vakgebied. Kortom, er bestaat niet één antwoord op de vraag hoe kwaliteitsbeoordeling moet plaats vinden.

De bijdrage van dit advies is dat er een systematisch raamwerk wordt voorgesteld voor het beoordelen van kwaliteit in de ontwerpende en construerende wetenschappen dat recht doet aan disciplinaire verschillen en ontworpen is voor een veelheid van beoordelingssituaties.

3.5 Profielen

De beoordeling van een persoon of onderzoeksgroep op basis van de bovenstaande tabel mag niet resulteren in een enkel cijfer – dat zou immers ontkennen dat er verschillende kwaliteitscriteria en indicatoren zijn. Het resultaat is een profiel. Zo'n profiel kan grafisch worden weergegeven. Het meest voor de hand liggend is een weergave waarbij een totaal oordeel wordt gegeven voor elk van de kwadranten van de tabel (in plaats van een oordeel voor elk van de indicatoren in dat kwadrant). Dat kan bijvoorbeeld gebeuren op een schaal van laag naar hoog (cijfers suggereren in dit geval een schijn-nauwkeurigheid). Maar een andere weergave kan ook wenselijk zijn.

Het gebruik van profielen maakt een discussie mogelijk over welke profielen wenselijk zijn. Dat kan gebeuren op verschillende aggregatieniveaus (personen, groepen). Die profielen kunnen per discipline en per onderzoeksgroep verschillen. Er kunnen ook verschillen zijn binnen een groep. Een parttime hoogleraar uit de industrie zal een ander profiel hebben dan een *fulltime* hoogleraar die leiding geeft aan de vakgroep. De bedoeling is wel dat een groep als geheel het gewenste profiel heeft. In figuur 3.1 is een voorbeeld opgenomen van twee profielen: een uitgesproken wetenschappelijk profiel en een gemengd profiel.



Figuur 3.1 Voorbeeld van een uitgesproken wetenschappelijk en een gemengd profiel.

3.6 Selectie van *peers*

Een kernprobleem bij *peer review* is het vaststellen wie de geschikte *peers* zijn. Dat is een probleem bij visitaties, maar ook bij de beoordeling van onderzoeksvoorstellen of het aanstellen van wetenschappelijk personeel. Met name in kleine landen en kleine vakgebieden waar iedereen elkaar kent, bestaat het risico dat *peer review* leidt tot vriendjespolitiek. Het is daarom wenselijk om op een meer objectieve manier vast te stellen wie geschikte *peers* zijn in een bepaalde situatie. Hierbij kan de voorgestelde tabel met criteria en indicatoren nuttig zijn. *Peers* kunnen namelijk worden gedefinieerd als die vakgenoten die goed scoren op de indicatoren die kenmerkend zijn voor een bepaalde discipline. Kandidaten voor beoordelingscommissies kunnen dan worden beoordeeld op basis van deze score. Daarnaast kan de onafhankelijkheid van *peers* worden bevorderd door ook internationale *peers* in commissies op te nemen.

3.7 Conclusies

CONCLUSIE 3.1

In de ontwerpde en construerende disciplines kan de kwaliteit van het wetenschappelijke werk worden beoordeeld aan de hand van twee criteria: 1) wetenschappelijke kwaliteit en 2) maatschappelijke relevantie. Een aparte set van criteria is voor deze disciplines niet nodig. Voor elk van de twee criteria bestaan er indicatoren die aangeven hoe goed men op de criteria scoort.

CONCLUSIE 3.2

Het beoordelen van kwaliteit is maatwerk vanwege de verschillen tussen disciplines, typen wetenschappelijke activiteiten (ontwerpen, onderzoeken) en beoordelings-situaties (visitatie, benoeming, onderzoeksvoorstel). Het voorliggende advies biedt een systematisch beoordelingskader dat recht kan doen aan deze verschillen door aan de indicatoren voor wetenschappelijke kwaliteit en maatschappelijke relevantie een bepaald belang of gewicht toe te kennen (tabel 3.1).

4. INVENTARISATIE VAN BUITENLANDSE ERVARINGEN EN INZICHTEN

4.1 Inleiding

Gezien het internationale karakter van wetenschappelijk onderzoek is het van belang dat de in het vorige hoofdstuk voorgestelde criteria in internationaal perspectief geloofwaardig zijn. Bovendien heeft de commissie gebruik willen maken van de ervaringen in het buitenland. Daarom is voor een aantal landen onderzocht hoe men tegen de kwaliteitsbeoordeling van wetenschappelijk onderzoek aankijkt, welke criteria men gebruikt en wat hiermee de ervaringen zijn. Waar mogelijk is specifiek gekeken naar de ontwerpende en construerende wetenschappen. In het Verenigd Koninkrijk (VK) en Australië zijn zeer uitgebreide systemen opgezet voor de landelijke onderzoeksbeoordeling. Daar is veel discussie geweest over geschikte criteria en is ook veel achtergrondonderzoek gedaan. De situatie in deze landen is daarom in detail bekeken. In Finland is gekeken naar de evaluatie van de Aalto Universiteit vanwege de degelijke aanpak en de grote aandacht voor de ontwerpende en construerende wetenschappen in deze evaluatie. Tot slot is van een aantal voor Nederland belangrijke landen kort nagegaan hoe die met kwaliteitsbeoordeling omgaan. De informatie over de individuele landen is te vinden in bijlage 3. In paragraaf 4.2 zijn de belangrijkste bevindingen en conclusies weergegeven.

4.2 Samenvatting ervaringen in het buitenland en conclusies

De meeste buitenlandse beoordelingssystemen hebben betrekking op alle wetenschapsgebieden en gaan meestal slechts zijdelings in op de specifieke problemen bij het beoordelen van de ontwerpende en construerende wetenschappen. Er zijn wel een

paar opvallende observaties die relevant zijn voor dit advies over de construerende en ontwerpende wetenschappen.

Gebruik van bibliometrische gegevens

In de beoordelingsmethodes van het VK, de Finse Aalto Universiteit en Australië wordt veel waarde gehecht aan bibliometrische analyses. Duidelijk is dat er veel discussie is in de betreffende landen over de toepasbaarheid van deze analyses, zeker bij de ontwerpende en construerende wetenschappen. Bij het Britse Research Excellence Framework lijkt men in eerste instantie zwaar op het gebruik van bibliometrische methoden te hebben ingezet maar is het belang ervan uiteindelijk afgezwakt. Op basis van de reacties uit de consultatierondes is men tot de conclusie gekomen dat deze indicatoren niet robuust genoeg zijn en onvoldoende draagvlak in het veld hebben. Ze zijn echter niet uit het beoordelingssysteem verdwenen, de analyses dienen nu als achtergrondinformatie voor *review*commissies. In het Australische systeem is nog wel steeds een prominente plek voor bibliometrie weggelegd. Wel heeft men daar differentiatie toegepast naar vakgebied door gedetailleerd vast te leggen voor welk gebied dit van toepassing is. Voor de ontwerpende en construerende wetenschappen is dit voor alle disciplines behalve voor de bouwkundige disciplines het geval. De *engineering* disciplines worden op vergelijkbare manier als de natuurkunde en chemie beoordeeld. In de Finse evaluatie geeft de evaluatiecommissie aan dat de bibliometrische analyses in het panel *Architecture, design, media en art research* niet bruikbaar waren omdat er te weinig was gepubliceerd in ISI-tijdschriften.

Mogelijkheid om niet-traditionele informatie mee te nemen

Zowel de evaluatie van de Finse Aalto Universiteit als de Australische ERA (*Excellence in Research for Australia*) bieden de mogelijkheid om wat men noemt 'niet traditionele output' in de beoordeling mee te nemen. Hieronder worden bijvoorbeeld verstaan films, websites, tentoonstellingen en creatief 3D-werk (architectuur, design, games, computerprogramma's). Ook hier lijkt deze uitzondering vooralsnog vooral betrekking te hebben op de disciplines architectuur/bouwkunde en industrieel ontwerpen. De beoordeling hiervan vindt plaats door *peer review*commissies die een aantal door de indieners zelf voorgelegde cases beoordelen. In het Australische systeem is daarbij expliciet beschreven dat de indieners zelf moeten aangeven waaruit de wetenschappelijke component van het werk bestaat. De ervaring met de evaluatie van de Aalto University in Finland leert dat hier nog wel een gewenningsproces nodig is. De evaluatiecommissie wees erop dat het bestuur weliswaar deze mogelijkheid heeft gecreëerd, maar er in de praktijk op aangedrongen is vooral wetenschappelijke artikelen aan te leveren.

Maatschappelijk *impact* internationaal gezien in de belangstelling

Naast de wetenschappelijke kwaliteit van de *output* laten alle besproken evaluatiesystemen maatschappelijke *impact* meewegen in de beoordeling. De manier waarop dat moet gebeuren is nog sterk in discussie.

Redelijk vergelijkbare criteria waarop wordt beoordeeld

Er is redelijke overeenstemming over de criteria waarop de kwaliteit van de wetenschappelijke *output* moet worden beoordeeld. Hoewel de exacte omschrijving verschilt, worden in essentie vrijwel altijd de criteria: Wetenschappelijke kwaliteit, Maatschappelijke relevantie en in wat mindere mate Productiviteit gebruikt.

Manier waarop *peer review*commissies worden samengesteld nergens beschreven

*Peer review*commissies spelen in de verschillende beoordelingsprocessen een cruciale rol. Dit betekent dat de kwaliteit van het beoordelingsproces in hoge mate afhangt van de kwaliteit van de *peer review*commissies. Zeker gezien de enorme breedte van het onderzoeksterrein waar sommige panels zich over moeten buigen, is het de vraag of de leden van de *peer review*commissies wel voldoende '*peers*' zijn voor het betreffende onderzoek. Opvallend is dat de meeste van de besproken beoordelingsprocedures niet ingaan op de manier waarop de *peer review*commissies samengesteld worden of de manier waarop de leden worden geselecteerd.

CONCLUSIE 4.1

In de beoordelingssystemen van toonaangevende landen zoals het Verenigd Koninkrijk, Australië, Verenigde Staten en Finland zijn wetenschappelijke kwaliteit en maatschappelijke relevantie de belangrijkste criteria.

CONCLUSIE 4.2

In verschillende buitenlandse beoordelingssystemen is het voor de ontwerpde en construerende wetenschappen mogelijk om resultaten zoals creatief werk, websites, producten in de beoordeling mee te laten wegen. Deze mogelijkheid is vaak wel beperkt tot de meer artistieke richtingen.

CONCLUSIE 4.3

In vrijwel alle beoordelingssystemen wordt veel aandacht besteed aan bibliometrische informatie. De waarde die eraan wordt gehecht verschilt.

CONCLUSIE 4.4

Het in dit advies voorgestelde beoordelingskader is geloofwaardig in internationaal verband.

5. CONCLUSIES EN AANBEVELINGEN

In hoofdstuk 2 is aangegeven welke knelpunten er zijn bij de huidige beoordeling van de kwaliteit van wetenschappelijk werk in de ontwerpde en construerende wetenschappen. In hoofdstuk 3 is een nieuw beoordelingskader geschetst dat tegemoet komt aan deze knelpunten. Dat kader is in hoofdstuk 4 geplaatst in een internationale context. In dit laatste hoofdstuk wordt de reikwijdte van het advies besproken en de gevolgen voor de bestaande praktijk. Tevens wordt aangegeven welke vervolgstappen nodig zijn. Ten slotte worden enkele aanbevelingen geformuleerd. Eerst worden nogmaals de belangrijkste conclusies hieronder weergegeven.

CONCLUSIE 2.1

Wetenschappelijke kwaliteit kan alleen binnen disciplines zinvol worden vergeleken. Een *one-size-fits-all*-benadering bij de beoordeling van kwaliteit, hoe gewenst ook vanuit bestuurlijk oogpunt vanwege zijn eenvoud, doet geen recht aan significante disciplineverschillen en zal dus altijd bepaalde vakgebieden ten onrechte bevoordelen.

CONCLUSIE 3.1

In de ontwerpde en construerende disciplines kan de kwaliteit van het wetenschappelijke werk worden beoordeeld aan de hand van twee criteria: 1) wetenschappelijke kwaliteit en 2) maatschappelijke relevantie. Een aparte set van criteria is voor deze disciplines niet nodig. Voor elk van de twee criteria bestaan er indicatoren die aangeven hoe goed men op de criteria scoort.

CONCLUSIE 3.2

Het beoordelen van kwaliteit is maatwerk vanwege de verschillen tussen disciplines, typen wetenschappelijke activiteiten (ontwerpen, onderzoeken) en beoordelings-situaties (visitatie, benoeming, onderzoeksvoorstel). Het voorliggende advies biedt een systematisch beoordelingskader dat recht kan doen aan deze verschillen door aan de indicatoren voor wetenschappelijke kwaliteit en maatschappelijke relevantie een bepaald belang of gewicht toe te kennen (tabel 3.1).

CONCLUSIE 4.4

Het in dit advies voorgestelde beoordelingskader is geloofwaardig in internationaal verband.

Reikwijdte advies

Conform de adviesaanvraag van de drie technische universiteiten aan de KNAW is het beoordelingskader in de eerste plaats bedoeld voor de ontwerpende en construerende wetenschappen. De commissie verwacht echter dat het ontwikkelde beoordelingskader ook bruikbaar is voor andere (technische) disciplines. Ten eerste omdat de twee criteria die onderscheiden worden (wetenschappelijke kwaliteit en maatschappelijke relevantie) goed aansluiten bij de bestaande praktijk van kwaliteitsbeoordeling (SEP, beoordeling van onderzoeksvoorstellen, etc.). Ten tweede omdat het voorgestelde kader zodanig flexibel is dat er recht kan worden gedaan aan relevante verschillen tussen disciplines. Die flexibiliteit komt van de indicatoren die gebruikt worden om te bepalen hoe goed wetenschappelijk werk aan de twee kwaliteitscriteria voldoet. Deze indicatoren hoeven niet voor elk wetenschapsgebied identiek of even belangrijk te zijn. Er kunnen zelfs discipline-specifieke indicatoren worden toegevoegd. Het voorgestelde kader is daarom niet alleen bruikbaar voor de ontwerpende en construerende disciplines maar bijvoorbeeld ook voor meer fundamentele (natuur)wetenschappen. De daarvoor veel gebruikte kwaliteitsindicatoren (hoge impacttijdschriftpublicaties, etc.) vormen immers een deelverzameling van de indicatoren in tabel 3.1. Het ontwikkelde kader is dus generiek, maar om operationele waarde te krijgen moet het discipline-specifiek worden gemaakt door de keuze van de kwaliteitsindicatoren en hun relatieve belang. Wetenschappers uit de meer fundamentele disciplines kunnen bijvoorbeeld aangeven dat de kwaliteit van hun werk onder andere met de indicatoren hoge impacttijdschriftpublicaties bepaald kan worden.

Inpasbaarheid in bestaande praktijk

Er zijn verschillende situaties waarbij uitspraken moeten worden gedaan over kwaliteit: de beoordeling van een onderzoeksvoorstel, de beoordeling van een onderzoeksgroep en de beoordeling van een medewerker bij bevordering of aanstelling. De commissie is van mening dat het voorgestelde advies betrekkelijk eenvoudig in te passen

is in de bestaande beoordelingspraktijk. Ook het SEP-protocol dat gebruikt wordt bij visitaties is compatibel met het hier voorgestelde beoordelingskader.

De afwijzing van een *one-size-fits-all*-benadering bij de beoordeling van kwaliteit heeft als consequentie dat in de bestaande praktijk beleidsoordelen waarschijnlijk een explicietere rol gaan spelen bij het maken van keuzes in situaties waar een vergelijking tussen disciplines moet worden gemaakt. Bijvoorbeeld bij het in competitie verdelen van onderzoeksmiddelen door NWO in gebieden die heterogeen zijn, of bij de toekenning van wetenschappelijke prijzen (NWO-Spinozapremie). Deze keuzes kunnen niet louter op basis van een kwaliteitsoordeel gemaakt en gemotiveerd worden. Het vergelijken van kwaliteit tussen verschillende disciplines is immers zinloos omdat er appels met peren worden vergeleken (persoon p_1 of onderzoeksvoorstel o_1 scoort waarde w_1 op indicatorenset i_1 en persoon p_2 of onderzoeksvoorstel o_2 scoort waarde w_2 op indicatorenset i_2). Hooguit kunnen scores die genormeerd zijn voor de betreffende disciplines met elkaar worden vergeleken (persoon p_1 behoort tot de top-5 procent van scorende wetenschappers op indicatorenset i_1 in discipline d_1 terwijl persoon p_2 behoort tot de top-25 procent scorende wetenschappers op indicatorenset i_2 in discipline d_2). Maar zelfs dan is de betekenis hiervan niet evident.

Adviesaanvraag 3TU

De commissie juicht het toe dat de Colleges van Bestuur van de 3TU's de positie van ontwerpende en construerende activiteiten en het daaraan gerelateerde onderzoek aan de orde hebben gesteld. Deze activiteiten staan onder druk door een beoordelingskader voor kwaliteit dat daarop niet voldoende is toegesneden. Hierdoor is óf dreigt in sommige gebieden het evenwicht tussen fundamenteel onderzoek, toegepast onderzoek en ontwerpende en construerende activiteiten verstoord te worden. Dit is gezien de missie van de technische universiteiten een ongewenste situatie. Als de drie Colleges van Bestuur het voorliggende advies overnemen en overbrengen aan hun wetenschappelijke staf dan wordt daarmee ook het signaal afgegeven dat voor technische universiteiten kwalitatief hoogwaardige ontwerpende en construerende activiteiten van belang zijn en een onlosmakelijk deel van de technische wetenschappen uitmaken.

De eerste stap voor de implementatie van het advies is dat de ontwerpende en construerende wetenschappen zelf bepalen welke kwaliteitsindicatoren uit de tabel op hen van toepassing zijn en wat het relatieve belang daarvan is. Deze indicatoren zullen in internationaal verband binnen de discipline geloofwaardig moeten zijn. De bepaling hiervan kan alleen door *peers* in die disciplines gebeuren. Het ligt voor de hand om de bestaande cycli van kwaliteitszorg hiervoor te gebruiken. De besturen van de 3TU's kunnen bijvoorbeeld aan een discipline vragen om – voorafgaande aan een visitatie – kwaliteitsindicatoren vast te stellen die de visitatiecommissie kan gebruiken bij de beoordeling van het onderzoek. Zo'n commissie, die uit internationale *peers* bestaat, kan ook worden gevraagd om over deze indicatoren een oordeel uit te spreken.

AANBEVELING 1

De commissie adviseert de besturen van de 3TU's om ruimte te scheppen voor een disciplinespecifieke kwaliteitsbeoordeling in de technische wetenschappen en aan de ontwerpende en construerende disciplines te vragen indicatoren en hun relatieve gewicht vast te stellen voor de beoordeling van kwaliteit in die disciplines. Deze indicatoren moeten geloofwaardig zijn in internationaal verband.

Onderzoeksfinanciers

De commissie is van mening dat de manier van kwaliteitsbeoordeling niet *a priori* bepaalde vakgebieden mag bevoor- of benadelen. Voor NWO/STW zou dit een kernpunt moeten zijn van de beoordelingssystematiek. Het in dit advies ontwikkelde beoordelingskader kan hulpmiddel zijn bij het vaststellen van kwaliteitsindicatoren die recht doen aan disciplinaire verschillen.

Met de introductie van gebiedsoverstijgende programma's toont NWO zich in ieder geval bewust van het belang van programma's die niet binnen de bestaande kaders passen en waarvoor ook andere kwaliteitsindicatoren kunnen gelden. Deze lijn moet volgens de commissie verder worden versterkt, bijvoorbeeld door ook programma's tussen gebieden van NWO en STW te starten. Dat is voor de ontwerpende en construerende disciplines, die een loket missen bij NWO, van groot belang. De commissie raadt STW aan om een meer gedifferentieerde benadering te kiezen ten aanzien van de maatschappelijke relevantie van onderzoeksprojecten. Ook projecten waarvoor bedrijven vanwege de langere termijn nog geen directe interesse tonen zouden voor financiering in aanmerking moeten kunnen komen. De ontwerpende en construerende disciplines hebben hier belang bij omdat op een aantal gebieden intermediaire onderzoeksorganisaties zijn opgeheven (zie hoofdstuk 2).

De commissie juicht het toe dat de meeste onderzoeksfinanciers gebruik maken van *peer review*. De keuze wie deze *peers* zijn is wel cruciaal en heeft grote invloed op de uitkomst van de beoordeling. De speelruimte van beoordelingscommissies om af te wijken van *refereerapporten* is namelijk klein. Het zou goed zijn om de keuze van *peers* meer te objectiveren en transparanter te maken. Dat komt de kwaliteit van het *peer review* proces ten goede. Tabel 3.1 kan daarbij van nut zijn. De *peers* die betrokken zijn bij de beoordeling van kwaliteit in de ontwerpende en construerende wetenschappen moeten zelf ook worden beoordeeld aan de hand van het brede spectrum van indicatoren en niet alleen op publicatiegedrag.

AANBEVELING 2

De commissie adviseert NWO en STW om meer in te zetten op programma's voor disciplines die niet goed passen bij de bestaande gebiedsindeling en de bestaande manier van kwaliteitsbeoordeling. Dat geldt in het bijzonder de ontwerpende en construerende wetenschappen. Daarvoor moeten kwaliteitsindicatoren worden gehanteerd die recht doen aan deze disciplines. De commissie adviseert STW om ook projecten mogelijk te maken met een maatschappelijke relevantie op de langere termijn waarvoor bedrijven nog geen concrete belangstelling hebben.

Wetenschappers in de ontwerpende en construerende wetenschappen

Een op de ontwerpende en construerende wetenschappen toegesneden wijze van kwaliteitsbeoordeling vraagt om een actieve inzet van de wetenschappers in deze disciplines. Die is nodig voor het vaststellen, wegen en operationaliseren van kwaliteitsindicatoren. Daarnaast zullen zij actief binnen NWO en STW moeten participeren als *peers* bij de beoordeling van onderzoeksvorstellen. Het is de commissie opgevallen dat dit maar op beperkte schaal gebeurt.

Het in dit advies geschetste beoordelingskader maakt het gebruik van andere indicatoren voor wetenschappelijke kwaliteit dan *peer-reviewed* publicaties mogelijk. Dit neemt niet weg dat de commissie en veel geïnterviewden vinden dat *peer-reviewed* publicaties ook in de ontwerpende en construerende disciplines belangrijk zijn en blijven. Ze dragen bij aan onderlinge kwaliteitstoetsing en aan verspreiding van kennis. Publiceren is in deze disciplines of onderdelen daarvan niet vanzelfsprekend. Het is van belang om daar te zoeken naar publicatievormen die passend zijn bij de (deel) discipline. Daarbij kunnen nieuwe media een rol spelen.

AANBEVELING 3

De commissie adviseert wetenschappers in de ontwerpende en construerende disciplines om daar waar nodig krachtig een cultuur te bevorderen van *peer-reviewed* publicaties. Die dragen bij aan de toetsing van resultaten, de verspreiding van inzichten en de verdere verwetenschappelijking van de discipline.

AANBEVELING 4

De commissie adviseert de universiteiten en NWO/STW om de keuze van *peers* transparanter en objectiever te maken. Dat komt de kwaliteit van het *peer review* proces ten goede. Het in dit advies ontwikkelde beoordelingskader met discipline-specifieke kwaliteitsindicatoren kan daarbij een rol spelen. De keuze van de juiste *peers* is van belang voor de samenstelling van visitatiecommissies en beoordelingscommissies, alsook voor het *referee* proces van onderzoeksvorstellen.

LITERATUUR

- Aalto University (2009a). *Research Assessment Exercise 2009, Panel reports.*
- Aalto University (2009b). *Research Assessment Exercise 2009 Terms of reference for research assessment panels.*
- Aalto University (2009c). *Bibliometric report 2003-2007.*
- DCT Disciplineoverlegorgaan Construerende Techniek (2000). *Criteria voor de beoordeling van het ontwerpen in de construerende disciplines.*
- ERiC publicatie 1001 (2010). *Handreiking Evaluatie van maatschappelijke relevantie van wetenschappelijk onderzoek.*
- ERiC publicatie 1002 (2010). *Pilot Study at Faculty of Architecture TU Delft, Final Report (Confidential).*
- ERiC publicatie 1003 (2010) *Evaluating Research in Context Pilot Study at the Faculty of Electrical Engineering – TU/e (Confidential).*
- ERiC publicatie 1004 (2010). *Maatschappelijke relevantie van Werktuigbouwkundig onderzoek. Eindrapport ERiC-project CTW Universiteit Twente (vertrouwelijk).*
- Hee, K. van, Overveld, K. van, (2010). *Criteria voor het beoordelen van een technologisch ontwerp.*
- HEFCE (2009). *Research Excellence Framework, Second consultation on the assessment and funding of research .*
- KNAW-commissie kwaliteitszorg (2008). *Kwaliteitszorg in de wetenschap, Van SEP naar KEP: Balans tussen rechtvaardigheid en eenvoud.*
- Royal Academy of Engineering (2008). *Research Excellence Framework.*
- Royal Academy of Engineering (2009). *Research Excellence Framework, second consultation. Response to the Higher Education Funding Council for England.*
- Schouten, M.J.W. (2009). *Kwaliteitsbeoordeling Ontwerp- en Constructiedisciplines.* Technische Universiteit Eindhoven.
- Stokes, D.E. (1997). *Pasteur's Quadrant; Basic Science and Technological Innovation,* Brookings Institution Press, Washington.
- VSNU, KNAW en NWO (2009), *Standard Evaluation Protocol 2009-2015,* beschikbaar op www.knaw.nl/sep.

BIJLAGE 1

GEÏNTERVIEWDEN

Prof. dr. P.M.G. Apers, Universiteit Twente, informatica
Prof. dr. ir. J.A. Battjes, TU Delft, civiele techniek
Prof. ir. A. Beukers, TU Delft, lucht-en ruimtevaartechiek
Prof. dr. C.J.P.M. de Bont, TU Delft, industrieel ontwerpen
Prof. dr. ir. A.C.J.M. Eekhout, TU Delft, bouwkunde
Prof. dr. K.M. van Hee, TU Eindhoven, wiskunde en informatica
Prof. dr. ir. J.J. Heijnen, TU Delft, bioprocestechnologie
Dr. ir. F.D. van der Hoeven, TU Delft, bouwkunde
Prof. dr. ir. F.J.A.M. van Houten, Universiteit Twente, werktuigbouwkunde /
industrieel ontwerpen
Prof. dr. ir. J.T.F. Keurentjes, TU Eindhoven, scheikundige technologie
Prof. ir. H. Leegwater, TU Eindhoven, scheikundige technologie
Prof. dr. ir. M.C.M. van Loosdrecht, TU Delft, bioprocestechnologie
Prof. dr. ir. J.B.O.S. Martens, TU Eindhoven, industrieel ontwerpen
Prof. dr. ir. B. Nauta, Universiteit Twente, elektrotechniek
Prof. dr. C.J. Overbeeke, TU Eindhoven, industrieel ontwerpen
Prof. dr. ir. W.A. Poelman, Universiteit Twente, industrieel ontwerpen
Prof. ir. J.M. Post, TU Eindhoven, bouwkunde
Prof. dr. A.F.J. van Raan, Universiteit Leiden, kwantitatieve analyse
Prof. dr. ir. M.J.W. Schouten, TU Eindhoven, industrieel ontwerpen
Prof. dr. ir. M. Steinbuch, TU Eindhoven, werktuigbouwkunde
Dr. ir. S. Silvester, TU Delft, industrieel ontwerpen
Dr. P.E. Vermaas, TU Delft, filosofie
Dr. J. Visschers, NIKHEF, instrumentatieontwerp
Prof. dr. ir. H.J. de Vriend, Deltares, civiele techniek
Prof. dr. ir. P.J.V.V. van Wesemael, TU Eindhoven, bouwkunde
Prof. dr. ir. J.J. van Wijk, TU Eindhoven, informatica
Prof. dr. H.W. Zandbergen, TU Delft, technische natuurkunde

BIJLAGE 2

INDICATOREN ERiC-PROJECT

In het kader van het ERiC-project zijn een drietal *pilots* gehouden bij de faculteiten *electrical engineering* van de TU Eindhoven, bouwkunde van de TU Delft en werktuigbouwkunde van de Universiteit Twente. Deze *pilots* hadden tot doel een evaluatiemethode te ontwikkelen die gebruikt kan worden om de maatschappelijke relevantie van het onderzoek aan deze faculteiten te beoordelen. De voorgestelde indicatoren uit de *pilots* bouwkunde en *electrical engineering* zijn hieronder weergegeven.

Tabel BL.2.1 Indicatoren voor maatschappelijke relevantie van bouwkundig onderzoek [ERiC, 2010]

Aspect van maatschappelijke relevantie	Indicatoren
Verspreiding van kennis	Professionele publicaties, niet-wetenschappelijke publicaties, tentoonstellingen etc. Verspreiding van technologie, artefacten, standaarden Advies- en consultancywerkzaamheden Popularisering, onderwijs en bijdragen aan het publieke debat Training van professionals, mobiliteit van opgeleiden Masterscripties en afstudeeropdrachten die vragen uit de praktijk beantwoorden
Interesse van stakeholders	Aantal onderzoekers met relevante praktijkervaring in de sector(en) waar het onderzoeksprogramma zich op richt Publieke financiering gerelateerd aan maatschappelijke kwesties Derde geldstroomfinanciering door mogelijke gebruikers Samenwerking met maatschappelijke stakeholders in onderzoek, tests en evaluaties Consortia met niet-academische organisaties
Impact en gebruik van resultaten	Inkomsten door gebruik van resultaten Zichtbaarheid in publiek debat / publieke <i>mediarankings</i>

Tabel BL 2.2 Indicatoren voor maatschappelijke relevantie voor electrical engineering [ERiC, 2010]

Aspect van maatschappelijke relevantie	Indicatoren
Verspreiding van kennis	PhD's in industrie MA's in industrie <i>Proofs of concept</i> Presentaties op gespecialiseerde conferenties
Interesse van <i>stakeholders</i>	Gezamenlijke <i>roadmaps</i> Presentaties op uitnodiging Valorisatiegrants Industriële financiering Uitwisseling van staf (uitwisselingsprogramma's zoals Casimir, kennisregeling) Parttime professors vanuit / in industrie Consortia met industrie
Impact en gebruik van resultaten	Marktintroductie en nieuwe projecten in industrie <i>Spin offs</i> met industriecontacten Patenten en octrooien

BIJLAGE 3

KWALITEITSBEOORDELING IN HET BUITENLAND

BL3.1 Verenigd Koninkrijk, *Research Excellence Framework*

De discussie in het Verenigd Koninkrijk (VK) over de kwaliteitsbeoordeling van wetenschappelijk onderzoek staat momenteel vooral in het teken van het *Research Excellence Framework* (REF). Het REF is een nieuw proces voor de beoordeling van wetenschappelijk onderzoek in het VK en is bedoeld als opvolger van de *Research Assessment Exercise* (RAE). De laatste RAE is gehouden in 2008. Een belangrijk bezwaar van de laatste RAE is dat het een veel te omvangrijke en arbeidsintensieve operatie is geworden. Het doel van de REF is om excellent onderzoek te identificeren en de *impact* daarvan te beoordelen. De beoordeling vindt plaats op het niveau van samenhangende groepen onderzoekers. Er worden geen individuele onderzoekers of complete instituten beoordeeld. De eerste beoordelingen volgens het REF zijn gepland in 2013 en zullen dan de verdeling van de financiering in 2014 bepalen. De manier waarop de REF exact zal worden uitgevoerd is nog onderwerp van discussie. In deze paragraaf worden de hoofdlijnen van de REF beschreven zoals die eind 2009 bekend waren [HEFCE, 2009]. Tevens wordt ingegaan op de belangrijkste discussiepunten waarbij gezien de focus van dit advies de nadruk wordt gelegd op de kritiek vanuit ingenieurswetenschappen.

Manier van beoordelen

In het voorstel voor de REF staat *peer review* centraal. Al het onderzoek dat ter beoordeling wordt voorgedragen wordt door *expert panels* beoordeeld op drie criteria:

1. *Output quality*. Het belangrijkste doel van de REF is om onderzoek te identificeren dat internationaal gezien als excellent kan worden betiteld. Hierbij mogen allerlei vormen van *output* ter beoordeling worden aangeboden.
2. *Impact*. Met dit criterium wordt beoordeeld in hoeverre wetenschappers aantoonbare bijdragen leveren aan de economie, maatschappij, beleidsontwikkeling, cultuur en kwaliteit van leven. Dit wil men gaan beoordelen door middel van *case-study's*.
3. *Environment*. Dit criterium is bedoeld om de kwaliteit van de onderzoeksomgeving mee te wegen. Hieronder wordt o.a. de *researchstrategie*, de ontwikkeling van de staf en training van onderzoekers verstaan.

Niet alle criteria wegen even zwaar. Voor het totaaloordeel denkt met aan de weging: *output quality*: 60 procent, *impact*: 25 procent en *environment*: 15 procent. Het totaaloordeel bestaat uit een sterrensysteem oplopend van *unclassified* (minder dan goed of niet te beoordelen) tot *four stars (exceptional)*.

De indeling van het onderzoek in verschillende *review panels* is nog onderwerp van discussie. Men voorziet een tweetrapsstructuur met vier hoofdpanels met daar- onder dertig subpanels. Het grootste deel van de technische wetenschappen zal in dit voorstel worden beoordeeld in één groot *engineeringpanel* waarin alle onderzoek zit op het terrein van elektrotechniek, chemische technologie, werktuigbouw en lucht- en ruimtevaart, mijnbouw, civiele techniek, materiaalkunde. Onder hetzelfde hoofdpanel maar onder een ander subpanel vallen dan: *computer science and informatics, mathematics, physics, chemistry en earth systems*. Opvallend is dat architectuur in het panel *Architecture, the built environment, town and countryplanning* is ondergebracht dat zelfs onder een ander hoofdpanel valt.

Discussiepunten

Gebruik bibliometrie

Er is veel discussie geweest over de bruikbaarheid van citatie-indicatoren en andere kwantitatieve indicatoren. Gekeken is of in de meer *science* georiënteerde gebieden deze indicatoren de *expert review* zouden kunnen vervangen. Uiteindelijk is men tot de conclusie gekomen dat deze indicatoren niet robuust genoeg zijn en ook onvoldoende draagvlak hebben in het wetenschappelijke veld. Desalniettemin wil men deze indicatoren een grotere rol laten spelen dan voorheen in de RAE. De bedoeling is nu dat de indicatoren worden gebruikt door de panels als hulpmiddel om hun oordeel op te baseren. Dit zou in ieder geval moeten gebeuren voor de panels in de gebieden medisch, biologie, de meer fysisch georiënteerde natuurwetenschappen, psychologie, engineering en informatica. Hoe dit precies moet gaan gebeuren wordt vooralsnog aan de panels overgelaten.

Een groot deel van de discussie spitst zich nog toe op het criterium *impact*, omdat erg onduidelijk is welke *impact* er precies meegenomen moet worden en wie nu de waardering krijgt voor bepaalde ontwikkelingen. Dit criterium is ook nieuw t.o.v. de RAE. Bij de beoordeling moeten de te beoordelen *units* één of meerdere *cases* indienen aan de hand waarvan ze hun *impact* beschrijven. *Impact* wordt daarbij gezien als invloed op economische, sociale, *public policy*, culturele of kwaliteit-van-leven-aspecten. Als indicatoren voorziet men indicatoren die:

- aangeven hoeveel *research*inkomsten er zijn gegenereerd uit gebruikers zoals bedrijven, overheid, en liefdadigheidsinstellingen.
- een maat geven voor de mate van samenwerking met gebruikers.
- specifiek zijn voor een bepaald onderzoeksgebied. Deze zouden geselecteerd moeten worden uit een soort menu. Een voorstel voor dit type indicatoren is te vinden in [HEFCE (2009)].

Kritiek vanuit de ingenieurswetenschappen

In een grootschalige consultatieronde zijn verschillende belanghebbenden gevraagd commentaar te geven op de beoogde opzet. Voor dit advies is het commentaar van de Royal Academy of Engineering het meest interessant [Royal Academy of Engineering, 2009]. Hun belangrijkste bedenkingen zijn:

- Men onderschrijft dat het gebruik van citatie-analyse niet robuust genoeg is om gebruikt te worden bij de beoordeling van onderzoeksresultaten en betwijfelt of dit ooit het geval zal worden gezien de breedte van het vakgebied en de verschillende typen van *output*. Bovendien wijst men erop dat de verschillen in publicatiegewoontes tussen de verschillende subdisciplines bijzonder groot zijn.
- Het gebruik van citatie-informatie als informatie voor de panels wordt toegejuicht maar men wijst er wel op dat beschikbare informatie uit citatiedatabases vaak ontoereikend is voor de *engineering*disciplines waardoor er grote verschillen kunnen ontstaan in de aangeleverde informatie in de verschillende subdisciplines.
- Men onderschrijft dat de *impact* van onderzoek meegenomen moet worden in de beoordeling. Dit is echter gezien de breedte van de verschillende soorten *engineering output* een lastig proces. Het vergt tijd om hier ervaring mee op te doen. De voorgestelde weefactor van 25 procent voor dit criterium wordt daarom in dit stadium te hoog geacht.
- In de *Research Assessment Exercise* waren industriële leden volwaardig lid van de *expert panels*. In het huidige voorstel van de REF zijn zij alleen betrokken bij de beoordeling van de *impact*. De Royal Academy of Engineers pleit ervoor dat industriële leden bij het *engineering panel* betrokken zijn bij de beoordeling van alle criteria.
- De Royal Academy of Engineering heeft sterke bedenkingen bij het onderbrengen van alle *engineering*onderzoek in één panel. Ten eerste omdat hierdoor een enorme hoeveelheid onderzoek (omvang ongeveer 4.400 fte onderzoekers) door één panel moet worden beoordeeld. Ten tweede omdat men verwacht dat hierdoor de concurrentie tussen de *engineering*disciplines onderling toeneemt waardoor er uiteindelijk minder geld voor de *engineering*wetenschappen komt.

BL3.2 Australië, Excellence in Research for Australia

Op initiatief van de Australische minister voor innovatie, industrie en wetenschap, senator Kim Carr, wordt sinds 2008 een ambitieuze methode ontwikkeld voor de evaluatie van wetenschappelijk onderzoek. Dit zogenaamde *Excellence in Research for Australia* (ERA)-initiatief, wordt ontwikkeld door de Australian Research Council (ARC) samen met het Department of Innovation, Industry, Science and Research en is gebaseerd op een combinatie van *metrics* en *expert review*. In 2009 is ervaring opgedaan met een beperkte proef in twee clusters: *physical, chemical and earth sciences* en *humanities and creative arts*. In 2010 gaat de beoordeling voor alle vakgebieden volgens deze methode van start.

Met deze exercitie wil men informatie verzamelen over de kwaliteit van het onderzoek, nationale sterktes en zwaktes identificeren, excellentie, opkomende gebieden en kansen opsporen en positionering in de internationale onderzoeksarena bepalen.

Manier van beoordelen

Voor deze evaluatie wordt het onderzoek in acht clusters met min of meer vergelijkbare disciplines ingedeeld:

1. *Physical, chemical and earth sciences*
2. *Humanities and creative arts*
3. *Engineering and environmental sciences*
4. *Social, behavioural and economic sciences*
5. *Mathematical information and computing sciences*
6. *Biological and biotechnological sciences*
7. *Biomedical and clinical health sciences*
8. *Public and allied health sciences*

Deze disciplines zijn weer onderverdeeld in een groot aantal subdisciplines. De discipline *engineering* bestaat bijvoorbeeld uit de subdisciplines *aerospace engineering*, *automotive engineering*, *biomedical engineering*, etc. De betrokken instellingen worden in principe beoordeeld op het niveau van deze subdisciplines. Alleen als zo'n subdiscipline te weinig *output* heeft om een goede beoordeling te maken, vindt beoordeling op niveau van de overkoepelende discipline plaats. De beoordeling heeft dus geen betrekking op complete faculteiten of op individuele onderzoekers. ERA gebruikt een reeks van indicatoren om de kwaliteit van het onderzoek te meten (zie tabel BL 3.1).

De eenheden worden geëvalueerd door *Research Evaluation Committees* (RECs), bestaande uit internationaal erkende experts op het gebied van onderzoeksevaluaties en de betreffende disciplines. Er zijn geen richtlijnen gegeven voor de manier waarop de REC-leden worden geselecteerd. De voorzitter van de REC vraagt de overige leden ieder individueel de eenheid te evalueren op grond van de aangeleverde data. In een later stadium komt de REC samen om de resultaten te bespreken. Alleen in relevante (sub)disciplines voeren de REC-leden, zonodig aangevuld met externen, een *peer review* uit voor bepaalde vormen van *research output*. In een zeer uitgebreide disciplinematrix staat per discipline aangegeven welke *output* hiervoor in aanmerking komt.

De instellingen worden verzocht op (sub)disciplineniveau gegevens aan te leveren. Deze gegevens moet men invoeren in het *System to Evaluate the Excellence of Research* (SEER) waarna ze elektronisch worden verwerkt. Op grond van de aangeleverde gegevens wordt het indicatorenpatroon en de hoeveelheid *output* per indicator vastgesteld. De indicatoren worden alleen aan de REC doorgegeven als een bepaalde disciplinespecifieke drempelwaarde is overschreden.

Tabel BL3.1. Beoordelingscriteria in Excellence in Research for Australia

Categorie	Indicatoren
<i>Ranked outlets</i>	Publicaties in wetenschappelijke tijdschriften Publicaties in <i>reviewed-conference proceedings</i>
<i>Citation analyses</i>	Relatieve citatie-impact (tegen wereldgemiddelde en Australisch gemiddelde voor het betreffende veld) Verdeling van publicaties in vergelijking met de rest van de wereld en in vergelijking met het betreffende veld
<i>Research volume and activity</i>	Totaal aan wetenschappelijk <i>output</i> voor verschillende resultaten, aantal fte van betrokken onderzoekers
<i>HERDC Research income</i>	Aantal subsidies Inkomsten per subsidie Totale inkomsten voor onderzoek Inkomsten voor onderzoek per fte Verhouding <i>research</i> inkomsten per fte in vergelijking met discipline specifieke <i>benchmark</i>
<i>Esteem</i>	Redacteur bij A- en A*-tijdschriften Bijdrage aan een prestigieus <i>work of reference</i> <i>Curatorial role</i> Gekozen <i>fellowship</i> van <i>learned academies</i> Nationale competitieve <i>research-fellowships</i> Prestigieuze prijzen en erkenningen
<i>Applied</i>	Toegekende patenten Kwekersrechten Geregistreerde ontwerpen Inkomen via commercialisatie wetenschap NHMRC-(gezondheidszorg)handleidingen

Output niet alleen wetenschappelijke publicaties

Voor bepaalde onderzoeksdisciplines is het mogelijk om wat men noemt niet-traditionele *output* ter beoordeling op te voeren. Voor de technische wetenschappen is deze mogelijkheid echter beperkt tot de disciplines *built environment* en *design* (w.o. architectuur). In geen van de overige *engineering*disciplines is dit toegestaan. Onder niet-traditionele *output* wordt dan bijvoorbeeld verstaan architectonische ontwerpen, creatief 3D-werk, design, games, computerprogramma's, tentoonstellingen.

Indien een werk wordt opgevoerd ter beoordeling als niet-traditionele *output* moeten de onderzoekers nadrukkelijk aangeven waaruit de wetenschappelijke component bestaat. Dit gebeurt in een verklaring waarin de volgende aspecten worden geadresseerd:

Onderzoeksachtergrond:

- veld
- context
- onderzoeksvraag

Bijdrage aan wetenschap:

- innovatieve component
- genereren nieuwe kennis

Significantie wetenschap:

- bewijs van excellentie

BL3.3 Finland, evaluatie Aalto University

Eind 2009 is het universitaire systeem in Finland grondig hervormd. De financiële en administratieve autonomie van de universiteiten werd vergroot. Ze werden opgedeeld in instellingen vallend onder *public law* of stichtingen vallend onder *private law*. Vanaf eind 2009 vielen de stichtingen niet meer onder het *state budgetary system*. Momenteel wordt een nieuw vorm van fondsverstrekking uitgewerkt die meer nadruk legt op strategische aspecten dan op indicatoren afgeleid van opleidings-/ onderwijstaken (in het bijzonder aantal studenten, promoties etc.). In het kader hiervan is men nu op zoek naar nieuwe kwaliteitsindicatoren.

Om een beeld te krijgen van de manier waarop in Finland over de kwaliteitsbeoordeling van wetenschappelijk onderzoek wordt gedacht, is gekeken naar de recent uitgevoerde evaluatie van de Aalto University in Helsinki. Hoewel deze evaluatie slechts betrekking heeft op één universiteit wordt deze hier toch besproken, omdat het een recente en grootschalige evaluatie is van een universiteit met een belangrijk aandeel faculteiten uit de ontwerpende en construerende wetenschappen.

De Aalto universiteit is in 2009 ontstaan uit een fusie tussen de Helsinki School of Economics, de Helsinki University of Technology en de University of Art en Design Helsinki. In datzelfde jaar heeft een *peer review*commissie (PRC), bestaande uit 62 leden afkomstig uit twintig landen, de universiteit geëvalueerd over de periode 2003-2008. De evaluatie had als doel om:

- de wetenschappelijke kwaliteit en de maatschappelijke *impact* van de nieuwe universiteit te bepalen;
- de groepen te identificeren die potentie hebben om internationaal te excelleren;
- het evaluatieproces zelf te analyseren.

De Aalto universiteit wil de aanbevelingen gebruiken om de onderzoeksstrategie te definiëren en *research practices* te ontwikkelen, inclusief het strategisch toewijzen van onderzoeksgelden.

Manier van beoordelen

De evaluatie is uitgevoerd door een PRC, verdeeld over 9 panels

1. *Chemical technology and materials*
2. *Electronics and electrical engineering*
3. *Mathematics and physics*
4. *Computer science en information technology*
5. *Mechanical engineering*
6. *Civil engineering en urban en regional studies*
7. *Business technology, economics en finance*
8. *Marketing, management and applied business research*
9. *Architecture, design, media en art research*

Het zijn negen losse *peer review*commissies met ieder vijf tot zeven leden. Ieder panel heeft de opdracht een discipline te evalueren aan de hand van een zelfevaluatie, een bibliometrische analyse en een bezoek ter plaatse. Voorafgaand aan de visitatie hebben de disciplines zelf uitgebreide informatie aangeleverd.

Alleen actieve, aan de universiteit verbonden onderzoekers mochten resultaten aanleveren. Daarnaast is een bibliometrische analyse uitgevoerd. De PRC werd verzocht speciale aandacht te geven aan de internationale onderzoekspositie. Het internationale onderzoeksveld heeft dan ook steeds als ijkpunt gefungeerd.

Aan de hand van een bezoek ter plaatse en de aangeleverde informatie (zelfevaluatie en bibliometrische analyse) heeft de PRC per discipline cijfers toegekend aan de criteria:

1. Wetenschappelijk kwaliteit
2. Wetenschappelijke invloed (*science impact*)
3. Maatschappelijke invloed (*societal impact*)
4. Onderzoeksomgeving (*research environment*)
5. Toekomstig wetenschappelijke potentie (*future research potential*)

Deze cijfers varieerden van 1 tot 5 waarbij 5 staat voor uitmuntende internationale zichtbaarheid en 1 voor internationaal opkomend.

De PRC heeft de uitkomsten van de bibliometrische analyse, met name de NCSf (citatiescore genormaliseerd naar onderzoeksveld) expliciet meegenomen in het oordeel van de verschillende disciplines. Mede hiermee is zichtbaarheid in de (inter)nationale onderzoeksarena bepaald. De discipline *architecture, design, media* en *art research* vormt hierop een uitzondering; hier zijn de bibliometrische analyses niet meegenomen omdat er te weinig ISI-publicaties zijn.

Output niet alleen wetenschappelijke publicaties

De PRC geeft aan dat de disciplines 1 t/m 8 goed te evalueren zijn met de huidige, op *science* gerichte criteria. De PRC geeft aan dat de Aalto universiteit op papier de ruimte schept om alle vormen van wetenschappelijke *output* (vooral naast publicaties ook artefacten) mee te nemen, maar dat er in praktijk vanuit het bestuur sterk aangestuurd is op het leveren van *science* gerelateerde *output*. Het gevolg was dat vooral de eenheden uit de discipline *architecture, design, media* en *art research* veel geringere hoeveelheden *output* (alleen hun artikelen) hebben aangeleverd dan ze bij de bezoeken van de PRC bleken te hebben. Het betreffende panel was van mening dat het bestuur hierdoor een ernstige onderwaardering van de waarde van het wetenschappelijke onderzoek en de maatschappelijke relevantie heeft gecreëerd.

Meer informatie over deze evaluatie is te vinden in [Aalto University, 2009a, b, c].

BL3.4 Verenigde Staten, National Science Foundation

In de Verenigde Staten is gekeken naar de werkwijze van de National Science Foundation (NSF). De NSF is een onafhankelijk federaal agentschap met als doel: *'to promote the progress of science; to advance the national health, prosperity, and welfare; to secure the national defense...'* Met een jaarlijks budget van 6,9 miljard dollar subsidieert NSF ongeveer 20 procent van het federaal gefinancierde fundamentele onderzoek aan de Amerikaanse universiteiten en colleges. In tegenstelling tot de in de vorige paragraaf beschreven beoordelingssystemen richt NSF zich vooral op ex ante-beoordeling van onderzoeksvoorstellen. Naast de NSF zijn in de VS nog veel private, missiegeoriënteerde en industriële fondsverstrekkers actief. Deze leunen meestal minder op duidelijk vastgelegde criteria.

Centraal in de beoordeling van onderzoeksvoorstellen bij de NSF is *peer review*. Daarbij worden voorstellen voornamelijk beoordeeld op twee criteria:

1. Wat is de wetenschappelijke merites van de voorgestelde onderzoeksactiviteit?
2. Wat is de bredere impact van de voorgestelde onderzoeksactiviteit?

Daarnaast kunnen er afhankelijke van het programma nog aanvullende criteria worden opgelegd.

Manier van beoordelen

Het voorstel komt binnen bij een *program officer* die het doorsluis naar een *review panel*. Dit panel, bestaande uit drie tot tien leden, beoordeelt het voorstel op bovenstaande criteria. Op basis van de oordelen van de panelleden stelt de *program officer* een analyse en een aanbeveling op. De *division director* van NSF neemt op basis van deze aanbeveling een definitief besluit over toekenning.

De indieners van een onderzoeksvoorstel wordt gevraagd om suggesties te doen voor geschikte referenten en men mag ook aangeven wie geen geschikte referenten zijn.

BL3.5 Noorwegen

Wetenschappelijk onderzoek in Noorwegen wordt vooral uitgevoerd door (algemene) universiteiten, gespecialiseerde universiteiten, *university colleges* en onderzoeksinstituten. Het ministerie van onderwijs en onderzoek is hoofdcoördinator en de betreffende vakministeries zijn belast met en verantwoordelijk voor het promoten en financieren van onderzoek dat op hun terrein valt. Een deel van het budget wordt direct door het ministerie verstrekt. Dit geschiedde voorheen met name op grond van studenten-aantallen. Sinds 2003 is financiering ingedeeld in a) basisfinanciering, b) onderwijs en c) onderzoek. De onderzoeksfinanciering is gebaseerd op prestaties en strategische overwegingen. Om prestaties te meten kijkt men naar:

- aantal afgeronde bachelors
- masters
- PhD's
- externe samenwerking
- samenstelling personeel (aantal professoren)
- aantal EU- en RCN-beurzen
- patenten en *output* in de vorm van wetenschappelijke publicaties

Hiervoor kent men aan de publicaties weegfactoren toe afhankelijk van de publicatievorm (artikel, boek) en het -kanaal (tijdschrift, website, uitgever, in geval van boek).

Universiteiten moeten alle gegevens invoeren in een landelijk documentatiesysteem, omdat de financiering mede wordt bepaald op grond van de documentatie in dit systeem.

Naast de hierboven geschetste basisfinanciering kent het Research Council Norway (RCN) competitieve beurzen toe. Ingediende voorstellen worden beoordeeld op grond van een groot aantal criteria aan de hand van *peer review*.

Ook voert het RCN landelijke evaluaties uit bij door RCN gesubsidieerde projecten, op instellingsniveau en op disciplineniveau. Doel van de evaluaties is om de kwaliteit, de relevantie en de efficiëntie van het onderzoek te verhogen en sterktes en zwaktes te identificeren op nationaal en instellingsniveau. Bovendien gebruikt het RCN de uitkomsten bij het opstellen van een nationale onderzoeksstrategie. Daarnaast zijn de uitkomsten een hulpmiddel bij het toewijzen van onderzoeksgelden. De RCN werkt momenteel aan een *follow-up*-model voor deze evaluaties.

BL3.6 Zweden

In Zweden wordt wetenschappelijk onderzoek uitgevoerd aan veertien universiteiten en 25 instellingen voor hogeronderwijs. Het meeste van het met publieke gelden uitgevoerde onderzoek vindt plaats bij de instellingen voor hogeronderwijs. Een deel van het onderzoek wordt direct gefinancierd door de overheid. Bovendien financiert de overheid de *research councils* en de sectorale *research agencies*.

De grootste subsidieverstrekker is het Swedish Research Council (SRC) dat een jaarlijks budget van vier miljard Zweedse kronen ter beschikking stelt. Het SRC kent beurzen toe aan onderzoek verdeeld over zes gebieden waaronder *natural and engineering sciences*. Onderzoeksvorstellen worden voorgelegd aan veertig panels waar ze aan de hand van *peer review* geëvalueerd worden.

De op twee na grootste fondsverstrekker is de Swedish Agency for Innovation Systems (VINNOVA) met een budget van 2,15 miljard Zweedse kronen. Dit agentschap verstrekt fondsen aan zes onderzoeksgebieden waaronder *natural sciences and technology*. Zweden kent geen landelijk ex post-onderzoeksevaluatiesysteem.

BL3.7 Frankrijk

Het Franse hogeronderwijsstelsel kent onderzoeksinstituten, universiteiten en *grandes écoles*. Wetenschappelijk onderzoek wordt hoofdzakelijk aan de instituten uitgevoerd. Hierbij bepalen, financieren en evalueren de instituten zelf hun onderzoek. Momenteel wil de overheid (zie *Stratégie Nationale de Recherche et d'Innovation* (SNRI); de nationale onderzoeks- en innovatiestrategie) de universiteiten meer autonomie geven en de onderzoeksinstituten omvormen tot agentschappen die de onderzoeksmiddelen projectmatig verdelen. Onderzoek zou volgens deze strategienota verplaatst moeten worden naar de universiteiten. Daarnaast kunnen instituten subsidie ontvangen via het Agence Nationale de la Recherche (ANR).

Frankrijk kent geen landelijk ex post-evaluatiesysteem voor onderzoek uitgevoerd aan universiteiten, *grandes écoles* en instituten

BL3.8 Duitsland

In Duitsland zijn de publieke onderzoeksinstituten in vier netwerken georganiseerd:

- Max Planck Gesellschaft (MPG)
- Fraunhofer Gesellschaft (FhG)
- Helmholtz-Gemeinschaft Deutscher Forschungszentren (HGF)
- Wissenschaftsgemeinschaft Wilhelm-Gottfried-Leibniz (WGL)

De universiteiten worden in eerste instantie gefinancierd door de betreffende zestien deelstaten. De grootste federale fondsverstrekker is de Deutsche Forschungsgemeinschaft (DFG) met een jaarlijks budget van 1,3 miljard euro. De DFG verstrekt fondsen aan alle disciplines van humaniora tot *engineering* en krijgt geld van zowel de federale als de deelstaatsoverheden volgens een vaste verdeelsleutel. DFG valt onder auspiciën van het Bundesministerium für Bildung und Forschung (BMBF).

Andere grote fondsverstrekkers zijn de private Alexander von Humboldt Foundation en de Deutscher Akademischer Austausch Dienst (DAAD), die internationale (uitwisselings)programma's en projecten financiert.

Er vindt geen onderzoeksevaluatie op nationaal niveau plaats maar de DFG produceert wel een classificatielijst (*Förder ranking*), gebaseerd op de volgende criteria:

- aantal positieve subsidietoewijzingen DFG
- samenstelling onderzoeksgroep
- fondsverwerving uit derde geldstroom
- positie in door DFG gesubsidieerde programma's
- aantal DFG-*reviewers*
- aantal Alexander von Humboldt *visiting* onderzoekers
- aantal DAAD-wetenschappers en -studenten
- deelname aan EU-*framework*programma's
- publicaties in internationale tijdschriften

Daarnaast hebben de deelstaten Nordrhein-Westfalen (gebaseerd op externe financiering en het aantal PhD's) en Niedersachsen eigen evaluatie systemen ontwikkeld. Niedersachsen heeft bovendien een eigen academisch adviescommissie. In opdracht van het ministerie van Economische zaken zijn in Baden-Württemberg in 2008 de toegepaste onderzoeksinstituten geëvalueerd. Geen van deze evaluaties is direct aan subsidiering gekoppeld. Internationaal gezien lopen deze evaluatiesystemen niet voorop.

Ook het zgn. *Exzellenzinitiative*, een subsidieprogramma van de DFG, heeft als doel heeft om *toplevel* onderzoek te identificeren en de kwaliteit van het hogeronderwijs en wetenschap in de brede zin te bevorderen en te verbeteren. Met dit initiatief wil Duitsland zijn internationale concurrentiepositie verbeteren. De onderzoeksvoorstellen, ingestuurd door de hele Duitse wetenschappelijke gemeenschap, worden aan de hand van *peer review* beoordeeld. De uitgekozen voorstellen worden in een staalkaart gepresenteerd.

BL3.9 Oostenrijk

Wetenschappelijk onderzoek in Oostenrijk wordt gesubsidieerd door:

- Austrian Research Promotion Agency (FFG)
- Austrian Science Fund (FWF)
- Austria Wirtschaftsservice (AWS)

Deze fondsverstrekkers vallen onder het Ministry of Science and Research (BMWF), het Ministry of Transport, Innovation and Technology (BMVIT) en het Ministry of Economy, Family and Youth (BMWFJ) maar strategische verantwoordelijkheden tussen ministeries en fondsverstrekkers zijn niet geheel duidelijk en de verantwoordelijkheden van de verschillende fondsverstrekkers overlappen elkaar.

In 2003 is de National Foundation for Research, Technology and Development opgericht om een constante financiële buffer voor de grote fondsverstekkers te garanderen en de continuïteit van langdurige, grote projecten te waarborgen. Het geld is afkomstig van de Austria National Bank (OeNB) en het European Recovery Fund (ERP).

Oostenrijk kent geen landelijk ex post-evaluatiesysteem voor onderzoek. Wel streeft het platform Research & Technology Evaluation (Fteval) naar het opzetten van een transparant evaluatiesysteem in het kader van optimale strategische planning van RTD-beleid. Daarnaast wil het Fteval een evaluatiecultuur in Oostenrijk bevorderen in samenwerking met de beleidsmakers op het gebied van technology- en research-beleid.

QUALITY ASSESSMENT IN THE DESIGN AND ENGINEERING DISCIPLINES



© Royal Netherlands Academy of Arts and Sciences

Some rights reserved.

Usage and distribution of this work is defined in the Creative Commons License, Attribution 3.0 Netherlands. To view a copy of this licence, visit: <http://www.creative-commons.org/licenses/by/3.0/nl/>

Royal Netherlands Academy of Arts and Sciences

PO Box 19121, NL-1000 GC Amsterdam

T +31 (0)20 551 0700

F +31 (0)20 620 4941

knaw@bureau.knaw.nl

www.knaw.nl

pdf available on www.knaw.nl


Basic design edenspiekermann, Amsterdam

Typesetting: Ellen Bouma, Alkmaar

Printing: Bejo druk & print, Alkmaar

Foto cover: Nationale beeldbank/Taco Gooiker

ISBN: 978-90-6984-620-0

The paper for this publication complies with the  iso 9706 standard (1994) for permanent paper



Mixed Sources

Productgroep uit goed beheerde bossen
en andere gecontroleerde bronnen.

www.fsc.org Cert no. CU-COC-804134-N
© 1996 Forest Stewardship Council

This publication is made from FSC-certified paper with number CU-COC-804134-N.

QUALITY ASSESSMENT IN THE DESIGN AND ENGINEERING DISCIPLINES

A SYSTEMATIC FRAMEWORK

Royal Netherlands Academy of Arts and Sciences
Advisory report KNAW TWINS Council

FOREWORD

One of the great merits of science is its natural diversity. It is a diamond with many different facets. Its variety can be seen in the subjects studied, the research methods used, and the way results are ultimately reported. At its most profound, it reflects the rich variety of phenomena that fill our world and the depth and breadth of our minds.

Science deserves to showcase its merits in a way that is as objective as possible but also respects, if not celebrates, its variety. The highest standards of quality assurance are needed not only to build support for research beyond the scientific community, but also because transparent assessment of research is beneficial for scientists themselves. The challenge is to ensure that the method used to measure quality is compatible with the intrinsic value of the relevant field of research. There is much work ahead in that respect.

One crucial first step is to survey scientific publications and their citations. According to this criterion, the Netherlands has world-class researchers, an assertion recently confirmed by the report *Science and Technology Indicators*, published every other year at the behest of the Dutch Ministry of Education, Culture and Science. In terms of citation impact, the Netherlands is at the top of the world rankings, along with the United States, Switzerland and Denmark. Our research community is also exceptionally productive.

It is of course gratifying to hear that Dutch scientists are among the best in the world according to these measures, but it is important to realise that research quality cannot be measured *solely* on the basis of scientific publications and citation impact. In many fields, citations tell only half the story or less, and the standard assessment methods miss out on large areas of research as a result. For example, the products of the design and engineering disciplines consist not only of peer-reviewed journal articles, but also conference proceedings, designs, software and structures. In the humanities, many publications are in book form, or in a language other than English,

making them virtually invisible in terms of citation impact. These disciplines come in for a harder time when the customary quality assessment measures are used. It is no accident, then, that the Academy has been asked to advise on practical assessment criteria precisely for these disciplines.

In 2008, the Academy argued in its advisory report *Kwaliteitszorg in de wetenschap; van SEP naar KEP* [Quality Assurance in Research: From SEP to CEP] that any new quality evaluation protocol must offer enough flexibility to accommodate differences between disciplines. Researchers who wish to emphasise the societal and cultural relevance of their work or its economic value should be able to do so, for example. The Academy was therefore delighted to advise on this matter. The present advisory report responds to the call for a more flexible assessment method for the design and engineering sciences. I am pleased that the committee has concluded that a separate set of criteria is not necessary for these disciplines. Research quality and societal relevance are sufficient. The discipline-specific factors are reflected in the indicators to be used to assess these two criteria. This is an important guiding principle that emphasises both the universality and the diversity of research at one and the same time.

The committee indicates that it has very deliberately not selected or weighted indicators. That is something that scientists themselves must do in the various assessment situations, in close consultation with university administrators. I hope that 3TU Federatie and design and engineering researchers will waste no time in accepting this challenge so that from now on, Dutch science can sparkle in all its diversity.

Robbert Dijkgraaf

President of the Royal Netherlands Academy of Arts and Sciences

CONTENTS

FOREWORD 5

SUMMARY 9

1. **INTRODUCTION** 13
 - 1.1 Background 13
 - 1.2 Committee members and committee's task 13
 - 1.3 Procedure 14
 - 1.4 Relationship with other initiatives 15

2. **PROBLEMS ASSESSING QUALITY** 19
 - 2.1 Introduction 19
 - 2.2 Problems encountered 21
 - 2.3 Evaluation of perceived problems and conclusions 23

3. **ASSESSMENT CRITERIA** 27
 - 3.1 Introduction 27
 - 3.2 Assessment criteria 29
 - 3.3 Explanation of indicators 30
 - 3.4 Relative importance of the criteria and the indicators 33
 - 3.5 Profiles 33
 - 3.6 Selecting peers 34
 - 3.7 Conclusions 35

4. **SURVEY OF LESSONS LEARNED ABROAD** 37
 - 4.1 Introduction 37
 - 4.2 Summary of lessons learned abroad and conclusions 37

5. **CONCLUSIONS AND RECOMMENDATIONS** 41

BIBLIOGRAPHY 46

APPENDICES

- 1 List of interviewees 47
- 2 ERiC project indicators 48
- 3 Quality assessment abroad 49

SUMMARY

Background

3TU Federatie – an alliance between the three Dutch universities of technology – has asked the Royal Netherlands Academy of Arts and Sciences to advise on the criteria used for ex-ante and ex-post assessments of research output in the design and engineering disciplines. Scientists in design and engineering regularly encounter problems in the assessment of the quality of their research output, whether that assessment takes place within the context of an external evaluation, an academic appointment or promotion, or an application for funding. The quality indicators used in such situations are borrowed from the more basic sciences (publications in ISI journals, impact factors, citations, the Hirsch Index) and are, in the eyes of these scientists, inadequate. First of all, they argue, their output consists not only of peer-reviewed international publications, but also of conference proceedings, designs and works of engineering. Secondly, they point to the fact that their research is generally more context-specific and multidisciplinary than that carried out in the more basic sciences. Design and engineering journals therefore have a lower impact factor and are consistently given a lower rating.

The Royal Academy installed a committee to look into this question, chaired by Prof. A.W.M. Meijers (Eindhoven University of Technology). The committee's assignment was to draft an advisory report on the criteria to be used for the ex-ante and ex-post assessment of:

- design and engineering activities in technical disciplines that can be considered scientific in nature;
- research in the design and engineering disciplines.

The criteria had to satisfy the following requirements:

- they had to be useful to organisations that fund research activities (NWO, etc.);
- they had to be useful to universities in assessing academic staff (researchers/designers);
- they had to be credible within an international context.

In the present advisory report, the committee proposes a set of pertinent criteria and reports on its findings. The report is based in part on interviews conducted with 27 scientists working in various disciplines and on a survey of the systems used and lessons learned abroad.

General criteria with discipline-specific indicators

One important point for the committee is that the assessment standards for a particular discipline should not differ completely from those used in other disciplines. That would make such standards arbitrary and opportunistic. The committee has therefore attempted to develop an assessment framework consisting of two elements:

1. generally applicable quality criteria;
2. discipline-specific indicators for those criteria. These indicators provide empirical information on the extent to which a person, group or research proposal satisfies the quality criteria. Such information may be either quantitative or qualitative in nature.

Contrary to what it had expected at the start of its advisory processes, the committee has concluded that assessments of research output quality in the design and engineering disciplines can be based on the two criteria used to assess output in other scientific disciplines: 1) research quality and 2) societal relevance. In other words, supplementary criteria are unnecessary for design and engineering. The committee's international survey has shown that these two criteria are also considered credible for these disciplines in other countries.

Although no additional or different quality criteria are necessary, that does not mean that every scientific discipline can be assessed in the same way. Assessing quality should be a question of fine-tuning owing to the differences between disciplines (including their publication cultures), categories of scientific activity (designing, research) and assessment situations (external evaluation, appointment, research proposal). Such fine-tuning is expressed in the discipline-specific indicators selected under the criteria 'research quality' and 'societal relevance', and the relative importance assigned to each indicator. Table 3.1 summarises the indicators that the committee regards as important for the design and engineering disciplines.

Following stage

The committee has explicitly *not* chosen to narrow down the choice of indicators for the design and engineering sciences, or to determine their relative importance. That is the following stage, and the necessary decisions must be taken by the scientists themselves (peers) in each of the various assessment situations, in close consultation with university administrators. The committee advises the board of 3TU to create sufficient scope for discipline-specific quality assessment in the technical sciences, and to ask the design and engineering disciplines to determine the indicators and their relative importance for assessing quality in those disciplines.

Difficult to compare different fields

It became clear during the interviews that virtually no one has encountered problems when quality assessments of research output are conducted by internationally respected peers. The problems arise when comparisons are made *between* disciplines, or when the disciplines themselves are too broad. The committee believes that quality benchmarks can only be meaningful when they are made *within* disciplines. As appealingly simple as administrators may find the 'one size fits all' approach to assessing quality, it does not do justice to significant differences between disciplines and will therefore always give certain ones an unfair advantage. The committee advises research funding bodies such as the NWO and the STW to devote more attention to programmes focusing on disciplines that do not fit easily into existing categories or the present quality assessment method. This applies in particular to the design and engineering sciences; the quality indicators used for these disciplines must do them justice.

Peer review crucial

The committee believes that proper quality assessment should be based on peer review, and it applauds the fact that most research funding bodies make use of the peer review system. It is vital, however, to select the right peers, as this will influence the outcome of the assessment. The peer selection processes used by the NWO, external evaluation committees and appointment advisory committees should be more objective and transparent. Peers involved in assessing quality in the design and engineering sciences must themselves be assessed according to a broad range of indicators, and not only on their publication behaviour.

Importance of publications

This advisory report proposes using indicators other than peer-reviewed publications to assess quality. The committee emphasises, however, that peer-reviewed publications are also important for the design and engineering disciplines, and will remain so. Such publications encourage mutual quality checks, help disseminate knowledge, and contribute to the 'scientification' of disciplines. Scientists working in these disciplines or in sub-disciplines do not always publish as a matter of course, however. The committee advises them to lend their full support to a culture of peer-reviewed publications. It is important to identify publication methods that are appropriate for the discipline or sub-discipline concerned.

1. INTRODUCTION

1.1 Background

Scientists in the design and engineering disciplines have long found assessment of the quality of their research output problematical, whether that assessment takes place within the context of an external evaluation, an academic appointment or promotion, or an application for funding. The quality indicators used in such situations are borrowed from the more basic sciences (publications in ISI journals, impact factors, citations, the Hirsch Index) and are, in the eyes of these scientists, inadequate. First of all, they argue, their output consists not only of peer-reviewed international publications, but also of actual designs and works of engineering. Secondly, they point to the fact that their research is generally more context-specific and multidisciplinary than that carried out in the more basic sciences. Design and engineering journals therefore have a lower impact factor and are consistently given a lower rating. As a result, experienced researchers in the design and engineering disciplines meet with little recognition in the scientific community, and often have trouble arranging external research funding or external evaluations, or gaining academic appointments.

1.2 Committee members and committee's task

In view of the foregoing problem, 3TU Federatie – an alliance between the three Dutch universities of technology – asked the Royal Netherlands Academy of Arts and Sciences to advise on the criteria used for ex-ante and ex-post assessments of research output in the technical disciplines of design and engineering. The Academy's Council for Technical Sciences, Mathematical Sciences and Informatics, Physics and Astronomy and Chemistry (TWINS Council) assembled a committee consisting of the following members:

- Prof. A. van den Berg, University of Twente
- Prof. R. de Borst, Eindhoven University of Technology
- Prof. P.P.M. Hekker, Delft University of Technology
- Prof. A.W.M. Meijers (chairperson), Eindhoven University of Technology
- Prof. R.A. van Santen, Eindhoven University of Technology

Dr E.E.W. Bruins (STW Technology Foundation) served as an observer. A. Korbijn (Royal Academy) acted as the committee's secretary, and Dr J.B. Spaapen (Royal Academy) as an adviser. C.S. Tan (Royal Academy) assisted the committee during its interviews and international survey.

Assignment

The committee's assignment was to draft an advisory report on the criteria to be used for the ex-ante and ex-post assessment of:

- design and engineering activities in technical disciplines that can be considered scientific in nature;
- scientific research in the design and engineering disciplines.

The criteria had to satisfy the following requirements:

- they had to be useful to organisations that fund research activities (NWO, etc.);
- they had to be useful to universities in assessing academic staff (researchers/designers);
- they had to be credible within an international context.

In the present advisory report, the committee proposes a set of pertinent criteria and reports on its findings.

1.3 Procedure

The committee began by interviewing 27 representatives of various technical disciplines. The purpose was to investigate problems that the interviewees had encountered with the present system of assessment and to explore which aspects the stakeholders felt should be included in any new set of assessment criteria. A list of the problems they noted is given in Section 2. The interviewees and their specialisations are listed in Appendix 1.

To explore how this problem is tackled abroad and what lessons the Netherlands can learn from other countries, the committee also collected relevant information on the United Kingdom, Australia, Finland, the United States, Norway, Sweden, France, Germany and Austria. The results are summarised in Section 4; more detailed information is presented in Appendix 3.

Based on the information it had gathered, the committee analysed the problems that arise in assessing quality in the design and engineering disciplines and formulated a proposal for suitable assessment criteria. The main findings and an initial version of these criteria were passed on to the interviewees with a request for their written comments. The final proposal for the criteria is described and explained in Section 3.

1.4 Relationship with other initiatives

1.4.1 Evaluating Research in Context (ERiC)

The three universities of technology are taking part in three pilots being carried out within the context of the ERiC project (*Evaluating Research in Context*). ERiC is a follow-up to a previous project run by the Consultative Committee of Sector Councils for Research and Development (COS), which focused on how to measure the value of research for society. Part of that project involved developing a measuring method, referred to as the sci_Quest method. In 2006, a decision was taken to continue working on this method so as to explore and improve its usefulness in practice. The ERiC platform was set up for this purpose; its members are the Netherlands Association of Universities of Applied Sciences (HBO-raad), the Royal Netherlands Academy of Arts and Sciences (KNAW), the Netherlands Association for Scientific Research (NWO), the Association of Universities in the Netherlands (VSNU), and the Science System Assessment department at the Rathenau Institute. The Dutch Ministry of Education, Culture and Science acted as an observer. The project involved implementing three pilots within the Faculty of Electrical Engineering at Eindhoven University of Technology, the Faculty of Architecture at Delft University of Technology, and the Faculty of Mechanical Engineering at The University of Twente. The purpose of the pilots was to develop a method for assessing the relevance to society of the research conducted by these faculties. The indicators that emerged from the Architecture and Electrical Engineering pilots are given in Appendix 2. The authors took the results of the ERiC pilots into account when composing the present advisory report.

The present report is both narrower and broader in scope than the ERiC project. To begin with, this report focuses on a smaller number of disciplines, i.e. the design and engineering disciplines. Secondly, it not only considers how to assess the societal relevance of the output concerned, but also how to assess scientific the quality of that output.

1.4.2 Standard Evaluation Protocol

The Academy, the NWO and the VSNU have adopted the *Standard Evaluation Protocol* 2009-2015 (SEP) for the evaluation of scientific research [VSNU, KNAW and NWO 2009]. According to SEP 2009-2015, an assessment consists of an external evaluation conducted once every six years and involving a self-evaluation report and a site visit, and an internal midterm review midway between two external reviews. The SEP uses

four assessment criteria and two levels of assessment: the institute (or faculty or research school) as a whole and the underlying groups or programmes. The four assessment criteria are:

1. Quality
2. Productivity
3. Societal relevance
4. Vitality & Feasibility

The purpose of assessing scientific research by applying the SEP is to account for the institute's past performance and to improve its quality in future. One difference between the present advisory report and the SEP is that SEP only concerns entire institutes or the underlying groups or programmes. The present advisory report, on the other hand, also looks at the assessment of research proposals and persons. In view of the SEP's importance in the Netherlands, the authors of the present advisory report have taken pains to identify criteria that complement those of the SEP.

1.4.3 Assessment criteria in the humanities

The present method of quality assessment is also problematical in the humanities. For example, citation indices only count articles, and not books, and of those articles only the English-language publications. Contributions to public debate are also not taken into account. At the request of the Cohen Committee on the National Plan for the Future of the Humanities, the Royal Academy will join with researchers in the field and users of research data in the humanities (such as the NWO) to design a simple and effective system of quality indicators. At the time of writing, the advisory report on the humanities had not yet been completed.

1.4.4 Disciplinary boundaries

The committee's assignment was to recommend criteria for assessing the quality of (1) design and engineering activities that can be categorised as scientific and (2) research in the design and engineering disciplines. This two-pronged assignment came about because the present system also appears to be inadequate for assessing design and engineering activities (for example instrumentation design) within more 'science-like' disciplines. In addition, some of the research conducted within the design and engineering disciplines is multidisciplinary and unique in nature.

The text box below explains what is meant by design and engineering activities. The committee defines the design and engineering disciplines as those disciplines in which design and engineering activities play an important role. Examples are industrial design, architecture, information science, mechanical engineering, chemical engineering, biotechnology, civil engineering, marine engineering, electrical engineering and

aerospace engineering. The committee believes that the criteria and indicators proposed in Section 3 are in any event appropriate for these disciplines.

TECHNOLOGICAL DESIGN

Design does not involve acquiring a better understanding of reality (true or reliable statements about that reality); rather, it is about creating something new. Designers add something to reality that did not exist before, and that is valuable or functional for users.

Design is a generic activity that takes place in many different domains, from art (a new painting or sculpture), to law (a new piece of legislation) to industry (a new chemicals plant). The design and engineering disciplines are concerned with designing new technological artefacts. These may be products or processes that can take physical form (buildings, telephones, power plants, molecules, chips) or abstract algorithms (software). It is also possible to design services (Internet transactions), living organisms (genetically modified crops), or environments that shape human perception (virtual reality).

There are many different ways to describe design processes. Some descriptions emphasise rational decision-making based on scientific models, with the range of solutions being narrowed at each pass; others try to include the non-rational, creative or art-like aspects of the design process. The American Accreditation Board for Engineering and Technology describes design as follows: '*Engineering design is the process of devising a system, component, or process to meet desired needs. It is a decision-making process (often iterative), in which the basic sciences, mathematics, and the engineering sciences are applied to convert resources optimally to meet these stated needs*' (Accreditation criteria for engineering curricula 2010-2011, www.abet.org).

The options to choose from in this process are not predetermined; they are developed within the process itself.

A design process results in a product: the design itself. This is a representation of the ultimate artefact in a form (prototype, computer model, scale model, drawing) that can be used to conduct all sorts of tests (performance, user aspects, producibility, costs, etc). Many engineering disciplines therefore view a design as a platform for research. Some describe an engineering design as analogous to an experiment in the natural sciences [Hee, van and Van Overveld, 2010].

2. PROBLEMS ASSESSING QUALITY

2.1 Introduction

Research comes in many different forms. We can attribute such variety to the various research traditions and approaches taken in the humanities, science and the social sciences. Each of these disciplinary clusters in its turn conducts many different types of research. The differences are often described in terms of the research aim, either a quest for fundamental understanding or a bid to solve practical problems related to use. This is why we divide research into basic research and applied research. Discussion in the literature shows that these are not mutually exclusive categories. Table 2.1 shows a customary classification of scientific research.

Table 2.1 Classification of scientific research, borrowed from [Stokes, 1997, p. 73].

		Considerations of use?	
		No	Yes
Quest for fundamental understanding?	Yes	Pure basic research	Use-inspired basic research
	No		Pure applied research

Dutch universities of technology undertake a broader and more complex spectrum of research activities because design and engineering are part of their academic portfolio. They are essential activities for these universities; it is, after all, part of their mission to help solve societal problems or generate economic opportunities, and design and engineering in particular serve to build the necessary bridge to the real world.

With scientific research taking on so many different forms, as we saw above, we may well question whether a single quality yardstick can be used to assess them all. Can the same criteria that we use to assess basic research (in the natural sciences) also serve to evaluate the quality of applied research, or of design and engineering activities? Practitioners in the design and engineering disciplines usually answer this question with a resounding no, and it indeed touches on the heart of the controversy concerning the assessment of quality in these disciplines.

The discussion has been brewing for many years. In 2000, the Consultative Body for the Engineering Disciplines (DCT) published a report in which it recommended criteria for assessing engineering design [*Criteria voor de Beoordeling van het Ontwerpen in de Construerende Disciplines*] [DCT, 2000]. Another advisory report on quality assessment in the design and engineering disciplines was published in 2009 [*Kwaliteitsbeoordeling Ontwerp- en Constructie Disciplines*] [Schouten, 2009]. Both reports emphasise that engineering design should be assessed as a research activity. The present advisory report focuses not only on design and engineering activities but also on the research related to those activities.

Growing urgency

The importance of a good quality assessment framework in the design and engineering disciplines has grown more pressing over time. To begin with, it is increasingly the case that research projects and academic staff on temporary appointment are financed solely by external sources. Design and engineering compete with other disciplines in that respect, forcing them to battle disciplines in the natural sciences that are a better fit with the existing quality criteria. Secondly, quality assurance has grown increasingly important at universities (external evaluations, internal midterm reviews), with administrators attaching consequences to the outcome. In that case, the criteria used to assess research quality are crucial. This problem is not exclusive to the design and engineering disciplines. Applied multidisciplinary research generally gets lower marks on traditional quality criteria. The disciplines concerned are often more restricted fields in which multidisciplinary journals with lower impact factors play an important role, and in which citation patterns are usually asymmetrical: frequent references in multidisciplinary journals to articles in monodisciplinary journals, but much less so the other way around.

In the following section, we describe the problems that leading scientists in the field have encountered in their work, based on their interviews with the committee (see Appendix 1).

2.2 Problems encountered

Difficult to compare different fields on quality

Most of the interviewees who work in the technical sciences find that it is difficult to assess design and engineering using traditional methods based on publications in ISI journals with high impact factors. The extent to which the interviewees themselves were troubled by this – for example when applying for research funding – depends on the field in which they work. Almost none of the interviewees, on the other hand, had encountered problems when the quality assessment was conducted by peers working in the same field. Those working in the discipline are well aware of who the top scientists are. The problems arise when disciplines are compared with one another.

There is no single, clearly designated funding channel for the design and engineering disciplines

One of the situations in which the quality of research output in different disciplines is compared is in the competition for research funding. This can be problematical when disciplines differ too much from one another. Researchers in the design and engineering sciences encounter problems because more basic research is generally accorded a higher status. Many of the interviewees see the lack of a clearly designated funding channel for the design and engineering disciplines as a shortcoming. Researchers who say that they do not find the current system of quality assessment problematical often obtain funding from business and industry or from more specific funding programmes such as the Point-One or Innovation-Oriented Research Programmes (IOP).

With EUR 100 million having recently been transferred from the direct to the indirect funding mechanism, the lack of a single funding channel at the NWO for the design and engineering disciplines has become even more urgent. At the moment, say a number of the interviewees, these disciplines simply do not have a fair shot at obtaining funding from the NWO.

An additional complication for design and engineering is that in a number of specific areas – for example civil and hydraulic engineering – the intermediary research organisations have been abandoned, have become less knowledge-intensive, or no longer fund research. This reduces the number of opportunities for contract research; at the same time, indirect funding programmes are offering a diminishing number of opportunities that require the involvement of stakeholders in civil society. That too makes it more urgent to have a single funding channel within the NWO.

Little acknowledgment that publication cultures differ considerably

None of the interviewees dispute the importance of publishing for knowledge dissemination and for the evaluation of output by means of peer review. The emphasis on

peer-reviewed articles in ISI journals, however, pays little regard to the fact that disciplines differ considerably when it comes to their publication cultures. In information science, for example, some peer-reviewed conference proceedings are more prestigious than publication in an ISI journal. The number of publications that a researcher can produce each year also varies considerably from one discipline to the next. In engineering, researchers need to spend a considerable amount of time on designing and engineering before they can even consider publishing their results. Various interviewees reported increasing pressure to publish in high-impact journals such as *Science* and *Nature*, even though fellow researchers in their discipline do not consider these journals important.

Interviewees who work in such disciplines as industrial design and architecture indicate that for them, publications are becoming more important as part of a strategy of 'scientification'. It is important, however, to bear in mind that the form and medium in which articles are published must be appropriate for the relevant discipline.

Declining role of design and engineering in technical disciplines

Design and engineering are the foundations of the engineering sciences and represent a bridge to actual practice. A design or work of engineering integrates knowledge from various disciplines in order to create the functions of a system that will operate in a real-life situation. Many of interviewees observed that the 'scientification' of some technical sciences (for example civil engineering) has led to a 'mono-science-like culture' in which the design and engineering of technical systems receive less attention than is actually necessary, and that there is often a huge disparity with real-life situations. That is an undesirable situation.

Growing gap between university administrators and practitioners in the field

Many of those interviewed feel that the quality assessment of research output at universities is becoming a bureaucratic tool. The growing gap between practitioners in the field and university administrators makes it difficult for the latter to have a clear view of the quality of research in a particular discipline. That is why there is a need for 'objective' and, preferably, quantitative information about the quality of different research units. Comparisons between such units are then based on bibliometric analyses or other numerical indicators. Design and engineering are at a disadvantage in that respect, compared to other disciplines that are a better fit with the existing quality criteria. That disadvantage affects not only the way the disciplines are funded, but also their scientific reputation and the recognition they are given.

2.3 Evaluation of perceived problems and conclusions

Differences between disciplines

The committee agrees that it is pointless to compare the research quality of different disciplines. Such a comparison assumes that there are quality indicators that apply or that should apply equally for the relevant disciplines. That is a false assumption, however. The research traditions, publication habits, and activity and output categories (journal articles, proceedings, books, artefacts, designs) are too diverse. Value judgments – for example about the comparative quality of researchers working in various disciplines based the number of ISI journal articles and impact factors – are therefore not meaningful. Neither does the popular Hirsch Index offer a basis for comparison. At most, it is possible to compare data that has been standardised for a particular discipline. We can then say that scientist A falls into the top 5 percent in discipline X, or that scientist B belongs to the top 25 percent in discipline Y.

Quality benchmarks are therefore only meaningful when they are made *within* disciplines. As appealingly simple as administrators may find the ‘one size fits all’ approach, it does not do justice to significant differences between disciplines and will therefore always give certain disciplines an unfair advantage. This opinion corresponds with previous conclusions by the Academy’s KNAW Quality Assurance Committee [KNAW-commissie kwaliteitszorg, 2008].

Peer review

Within disciplines, the assessment of research quality is usually by means of peer review. No matter how scrupulous and differentiating such reviews are, however, they can never be entirely objective or do justice to all the differences and controversies within a discipline. Personal bias cannot be avoided, and subjective elements will always play a role. Different peers will therefore reach different conclusions, at least to some extent. There is also the danger of a small number of peers dominating a discipline, and of previous assessments influencing later ones [KNAW, 2008]. Despite these limitations, the committee believes that the peer review system is in fact the best quality assessment method. The discipline must, however, be sufficiently homogenous in nature and international enough in scope to make independent peer review possible. If these conditions are satisfied, then a peer review system should also be relatively trouble-free in the design and engineering sciences.

Bibliometric research

Along with the growing gap between administrators and practitioners, the limitations of peer review have prompted the trend towards bibliometric research. It is an illusion to think that an entirely quantitative/bibliometric approach to research quality

is possible, however, or that it can replace the peer review method, not in the least because ultimately, peers are needed to assess and interpret the bibliometric data. At most, the committee believes, bibliometric research can supplement peer review. It is important to realise that both methods have their limitations. Quality assessments should always be conducted with a great deal of care for that reason.

Random indication

An additional reason to proceed with caution is that quality assessments are always mere random indications; as time passes, people and their work can vary in quality. That is something that assessment committees must be aware of.

Publications are also important in design and engineering

The committee attaches great value to the publication of research results in peer-reviewed journals, even in such disciplines as industrial design and architecture. That is because, first of all, assessment by peers precedes publication of such results. Secondly, publications make knowledge available to others and ensure that it is disseminated. The publication method must be appropriate for the discipline concerned, however. A unilateral focus on publications in monodisciplinary journals that have a high impact factor is undesirable, especially for a university of technology where multidisciplinary and application-driven research is also important. The committee has been pleased to note the growing importance that scientists in such disciplines as industrial design and architecture now attach to peer-reviewed publications, as part of their research output.

Criteria steer the direction of research

The criteria used to assess the quality of research output also influence decision-making on the type of activities universities choose to develop. They cast a long shadow because researchers tend to anticipate the criteria that will be used to evaluate their output and to determine whether they will receive funding, and how much. Quality criteria are therefore far from neutral; rather, they set the standard in a discipline. Besides their role in the ex-post evaluation of research output, they also have an ex-ante influence in steering the direction of research. In that respect, the committee shares the concern of a number of interviewees that a unilateral focus on criteria appropriate for the natural sciences may have led in some technical disciplines to undue pressure being placed on design and engineering and the related multidisciplinary research. In view of the mission of a university of technology, it is very important to strike the right balance between the different categories of research described previously (Section 2.1, Table 2.1) and design and engineering activities. That will only happen, however, if the quality criteria used are tailored to that purpose.

Research funding bodies

Research funding bodies such as the NWO and the STW must be aware that the 'one size fits all' approach does not do justice to the differences between disciplines and will therefore always give certain disciplines an unfair advantage. A number of the interviewees believe that the NWO does not offer the design and engineering disciplines equal opportunities. More generally, they find that multidisciplinary or interdisciplinary research does not fit effortlessly into the NWO's structure, and that it must compete with disciplinary research that more readily meets the quality criteria set. There are also vested interests that new disciplines are forced to compete with. The committee believes that the NWO has acknowledged the problem by initiating cross-disciplinary programmes. This initiative merits further development, for example by launching programmes that bring together NWO and STW disciplines. With respect to the STW, the various interviewees indicated that the yardstick for assessing the societal relevance of research is too severely limited to the short-term interests of the businesses on the user committees. According to the committee, it would be a good idea to take a more differentiated approach in this case as well and to also support research projects of long-term relevance, even if no businesses can be found for that purpose as yet. That is also important for the design and engineering disciplines.

Importance of a transparent objective

Finally, the committee considers that the quality criteria should be fine-tuned to suit the objective of the relevant quality assessment. It is one thing to assess a research proposal and quite another to review the past performance of a research group or to evaluate the performance of a member of staff awaiting promotion or a permanent appointment.

CONCLUSION 2.1

Quality comparisons are only meaningful when they are made within disciplines. As appealingly simple as administrators may find the 'one size fits all' approach to assessing quality, it does not do justice to significant differences between disciplines and will therefore always give certain ones an unfair advantage.

CONCLUSION 2.2

Those working in the design and engineering disciplines encounter relatively few problems when the peer review system is used to assess the quality of research output or to identify outstanding scientists.

CONCLUSION 2.3

It is important to universities of technology that the criteria for research quality should be fine-tuned to reflect the different types of research and activities that take place there. A unilateral focus on criteria derived from the natural sciences makes it difficult to strike the right balance between basic research, applied research, and design and engineering activities.

CONCLUSION 2.4

The publication of research results in peer-reviewed journals, books and proceedings is of huge importance in the design and engineering disciplines. The publication method must be appropriate for the discipline concerned, however.

3. ASSESSMENT CRITERIA

3.1 Introduction

This section identifies criteria and indicators that the committee believes are suitable for assessing quality in the design and engineering sciences.

Comment concerning terminology

The term 'quality' is an evaluative concept indicating how good something is. It can be applied to objects, processes and persons. Different situations call for different quality assessment criteria. For example, we assess the quality of an article for publication differently than the quality of the computer on which we write it. Quality assessments are relatively meaningless if we do not indicate the criteria on which we have based our evaluation. We usually cannot observe directly how well something meets a certain assessment criterion. That is why we work with indicators, which provide empirical information on the extent to which an object, process or person satisfies a particular criterion. That does not necessarily have to be quantitative information (for example the number of articles someone has published); it can also refer to qualitative information (for example what peers think of the journal in which someone has published an article).

Basic principles for assessment

Take disciplinary differences into account

As discussed previously, disciplines differ significantly from one another. Even within the relatively limited field of the design and engineering sciences, however, there is considerable variety. There is broad agreement that the present system of quality assessment does not take this variety sufficiently into account. In the committee's opinion, however, that is something that an assessment system must certainly do.

Differentiation required...

In view of the differences between disciplines and between assessment situations, the committee believes that differentiation is required in assessments of research quality. This means:

- differentiation between disciplines/sub-disciplines;
- differentiation between types of activity within a particular discipline (research or designing and engineering);
- differentiation between the objects of assessment (a member of staff, a research proposal, or a research group).

...but within an overall framework

On the other hand, the committee also believes that the assessment standards for a particular discipline/sub-discipline should not differ completely from those used in other disciplines. That would make them arbitrary and opportunistic, and perhaps also create the impression that the design and engineering disciplines are subject to 'milder' requirements than other disciplines. There should be an overall assessment framework that allows for the variety described above. The committee has therefore attempted to identify overall quality assessment criteria that also apply in a broader sense, along with indicators that may differ or vary in importance from one discipline to the next.

Peer review crucial

As explained in the previous section, the committee regards peer review as crucial to assessing the quality of publications, researchers, research groups and research proposals. The discipline must, however, be sufficiently homogenous in nature, broad (and international) enough in scope, and have a sufficient number of suitable peers available.

Bibliometric data should supplement, not replace, peer review

The committee believes that bibliometric data should be used to supplement peer review, and not replace it, as discussed earlier in this report. When determining the value of bibliometric indicators (citation scores, impact factors, the Hirsch Index, etc.), it is by no means a trivial matter to define precisely what is regarded as a discipline – indeed, this is a crucial point.

Transparent objectives

Quality assessments, for example external evaluations, are often used as a ranking and PR tool, and not to identify potential improvements. The purpose of the quality assessment should be made clear before the assessment is conducted, and the assessment itself should be viewed in the light of that specific purpose.

3.2 Assessment criteria

Five criteria narrowed down to two

In the first instance, the committee considered five different assessment criteria, in line with a proposal put forward by the Royal Academy of Engineering (United Kingdom) for evaluating quality in the technical sciences. These criteria were: 1) publications, 2) impact, 3) innovativeness, 4) involvement of external stakeholders, and 5) reputation of scientists concerned. There was considerable support for these criteria during the consultations between the committee and practitioners in the Netherlands. When considered in greater detail, however, some of these criteria actually turned out to be underlying indicators for other criteria. For example, publication in leading journals is not an independent criterion but an indicator for the criterion 'research quality'. The involvement of external stakeholders in research is also not a criterion in itself, but rather an indicator for the criterion 'societal relevance'. The committee gave considerable thought to whether innovativeness should be categorised as an independent criterion. Ultimately, it decided that innovativeness combines two aspects, scientific originality and applicability for society, and is therefore not a criterion in and of itself. The first aspect is expressed in scientific publications and the second can be shown by indicators measuring societal relevance. Finally, the committee believes that 'reputation' is a derived variable encompassing both scientific and societal aspects.

In the end, the committee concluded that essentially, there are only two criteria suitable for assessing the quality of activities carried out in the design and engineering sciences:

1. research quality
2. societal relevance

This conclusion is remarkable in that the proposed criteria do not differ from criteria currently applied in other areas of science and scholarship. Contrary to the committee's original assumption, then, it will not propose a separate set of criteria for assessing quality in the design and engineering sciences. The foregoing two criteria have broad support in the scientific community and are similar to what is customary at international level (see Section 4). They appear to offer an overall framework for assessing quality in every scientific discipline.

Within this overall framework, however, the committee would wish to see differentiation with respect to the indicators used to determine how well something satisfies the two assessment criteria. Specifically, that means:

1. differentiation between disciplines: the nature and relative importance of the indicators used will differ from one discipline to the next; in the design and engineering sciences, for example, there are other indicators of research quality than numbers and citations in peer-reviewed journals.

2. differentiation between types of activity within a particular discipline: research, designing, or engineering.
3. differentiation between the objects of assessment: output or person. The indicators used to assess a funding application for a research project often differ from those used to assess a candidate for an appointment or promotion.

Table 3.1 shows the two assessment criteria and related indicators that the committee believes are suitable for assessing quality in the design and engineering sciences. As mentioned earlier, not all indicators are equally relevant or have the same relative significance for each discipline/sub-discipline, activity or assessment situation. The table attempts to show the most important types of indicators. Indicators of the quality of education and teaching are not included, and are beyond the remit of this advisory report.

3.3 Explanation of indicators

In this section, we provide a more detailed explanation of the indicators in the table where necessary.

3.3.1 Research quality

Scientific publications

As mentioned earlier, scientific publications are also an important indicator of research quality in the design and engineering sciences. The publication method may differ from one discipline to the next, however (e.g. journals, proceedings, electronic publications).

Other output: design artefacts

The committee believes that design artefacts (products, processes, software) should also be considered in quality assessments of research output in the design and engineering sciences. It wishes to make a few comments in this connection, however. The artefacts it has in mind are those that make a scientific contribution to the discipline, and not artefacts designed within an educational context. One of the essential characteristics of a scientific design is that it generates new knowledge that can be applied generically. The documentation that accompanies the artefact should be enlightening in that respect. If a design generates knowledge, then that knowledge can usually also be published in peer-reviewed journals, a guarantee of quality assurance. The fact that this knowledge *can* be published does not necessarily mean that it has been published. There may be many good reasons for not doing so, for example because publishing is not customary, or may even be prohibited, in the business environment in which the knowledge was generated.

As an indicator for research quality, artefacts can play a role in the appointment of professors who have worked in industry. Because such candidates are important to

Table 3.1 Quality indicators for the Technical Sciences

	INDICATORS FOR OUTPUT	INDICATORS FOR PERSON
SCIENTIFIC QUALITY	<p>Scientific publications Articles in peer-reviewed journals (no. and type of journal) Articles in peer-reviewed conference proceedings (no. and type of proceedings) Scientific books published by leading publishers or significant contributions to such books (no. and type) Citations of individual articles Impact factors of journals in which articles are published</p> <p>Other output Peer-reviewed artefact (design) + documentation. This also includes software design</p> <p>Research impact (ex-post) Use of scientific products by other researchers (artefacts, methods, measuring instruments, tools, standards and protocols)</p> <p>Potential research impact Possible contribution to development of theories and models, methods, operational principles or design concepts</p>	<p>Recognition by scientific community Membership of prominent organisations such as academies of sciences Prestigious grants (VENI, VIDI, VICI, or ERC Grants) Honorary doctorates Visiting professorships</p> <p>Editorships Chief/full editorship of international scientific journal/book/conference proceedings</p> <p>Considered expert by peers Advisory capacity in scientific circles (NWO, external inspections, etc.) Keynote lectures at science conferences Membership of programme committees Participation in international assessment committees for scientific programmes/institutes or scientific advisory councils/institutes</p> <p>Research impact across the course of career Person's citation score Contribution to developing a 'school of thought'</p>
SOCIAL RELEVANCE	<p>Use of results by external stakeholders (ex-post impact) Contribution to solving societal problems Market introductions and new projects in industry Income generated by use of results Spin-offs with industry Patents used Artefacts used (designs, software)</p> <p>Use of results by profession (ex-post impact) Use of artefacts, methods, measuring instruments, tools, standards and protocols</p> <p>Involvement of external stakeholders in scientific output (potential societal relevance) Businesses or civil-society organisations involved in guiding research projects (e.g. in user committees) Contract financing by potential users (e.g. industry) Public financing related to societal questions Valorisation grants</p> <p>Contribution to knowledge dissemination Professional publications and papers, non-scientific publications, exhibitions and other events related to research results</p>	<p>Considered expert by external stakeholders Advisory and consultancy work (focused on users) Leading position in industrial research (e.g. managing director of R&D department)</p> <p>Considered expert by profession Oeuvre prizes (e.g. architects) Retrospective exhibitions</p> <p>Contribution to knowledge dissemination Activities focusing on popularisation of science, education and contribution to public debate Training of professionals PhDs with their first job in relevant practice</p>

universities of technology, such scientific achievements must also be considered. It is essential, however, to ask peers to assess the scientific merits of a small number of well-documented design artefacts.

Research impact

Research impact means the extent to which research output is used by other scientists. One of the most common methods of use is the citation. Another indicator is how much new research products are used by fellow researchers. By research products, we are referring here to artefacts, methods, measuring instruments, tools, standards, protocols, etc. This criterion does not include the use of scientific results by non-scientists; that type of use falls under societal relevance.

Potential research impact

This indicator plays a role in assessing the quality of research proposals. They are assessed in part for their potential contribution to the discipline.

Recognition by the scientific community

Expressions of recognition are an important gauge of the research quality of an individual. Such recognition must be gained within a peer-review process, however. Examples include the award of science prizes (Spinoza) and other distinctions, honorary doctorates, membership of prominent organisations such as academies of sciences, and prestigious grants and other funding (VICI, ERC Grants).

Extent to which peers regard an individual as an expert

The extent to which a researcher is regarded by fellow scientists as an expert is an important indicator for quality because practitioners generally know precisely who are the leading scientists in their field. Such scientists are invited to give keynote lectures at science conferences, become members of programme committees, and to act in an advisory capacity in scientific circles.

Research impact across the course of a career

This means the research impact of a person throughout his or her career. That impact may become evident from that individual's citation score. It can also be seen in how much peers believe that person has helped develop a particular school of thought.

3.3.2 Societal relevance

Use of results

This covers the indicators that measure the use of scientific knowledge or products beyond the field of science. It is usually only possible to assess actual use after the research has been completed. Such use can take many different forms: it may involve making a contribution to solving a societal problem, or it may mean commercial use by

a business. The use of results by the relevant profession constitutes a separate category.

Potential impact

The societal relevance of research can be assessed on a number of different timescales. The time horizon for applied research is shorter than that for basic research. It is notoriously difficult, however, to identify the societal relevance of long-term research. In such cases, it is possible to work with indicators that express the degree of interest that stakeholders in civil society have in the research. The stakeholders consist not only of businesses, but also of government ministries or international non-profit organisations. In the case of short-term research, co-financing can be taken as an indicator of stakeholder involvement. That involvement also considerably increases the chance that the results will in fact be used.

Contribution to knowledge dissemination

Whether assessing research output or evaluating an individual, the contribution made to knowledge dissemination must also be included as an indicator of societal relevance. Knowledge dissemination helps solve societal problems or create economic opportunities, after all. It is also part of the mission of a university of technology.

3.4 Relative importance of the criteria and the indicators

Assessment is a question of fine-tuning. Before an assessment is carried out, a decision must be taken as to which of the two criteria and which of the indicators are most important in that given situation. The criteria and indicators relevant for a full-time professor who heads a research group will differ from those used for a part-time professor whose job involves bridging the gap between academia and real-life situations. The criteria and indicators for an individual VIDI grant will differ from those for a funding application submitted in an open NWO competition, or from an STW funding application. And we have not even mentioned differences between disciplines. In short, there is no one method for conducting a quality assessment.

This advisory report proposes a systematic framework for assessing quality in the design and engineering sciences, one that takes disciplinary differences into account and that has been developed to serve in a variety of assessment situations.

3.5 Profiles

The assessment of a person or research group according to the above table must not result in a single score – that would, after all, be denying the existence of different quality criteria and indicators. The result should be a profile, which can be shown in diagram form. The most obvious diagram would be one that offers an overall score for

each of the quadrants of the table (instead of a score for each of the indicators in that quadrant). One example would be to use a scale running from low to high (numbers would suggest a false accuracy in this case). Some other form of representation may also be appropriate, however.

The use of profiles raises the issue of which profiles are desirable, and at what level of aggregation (persons, groups). The profiles can vary per discipline and per research group. There can also be differences within a group. A part-time professor working in industry will have a different profile than a full-time professor who heads a faculty. The purpose, however, is to see that the group as a whole is assigned the desired profile. Figure 3.1 gives an example of two profiles, one that is explicitly scientific in nature and one that is mixed.

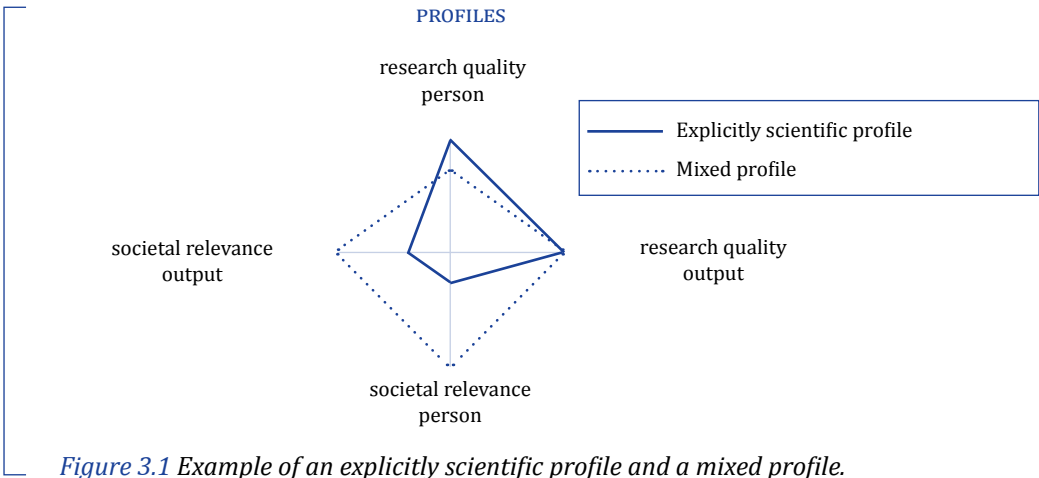


Figure 3.1 Example of an explicitly scientific profile and a mixed profile.

3.6 Selecting peers

One significant issue in the peer review system is how to select the most suitable peers. The problem arises in connection with external evaluations, but also when assessing research proposals or appointing academic staff. In small countries and small disciplines where everyone knows everyone else, there is the risk that peer review will result in an 'old-boy network'. It would be better, then, to use a more objective method for determining which peers are suitable in a given situation. The table of criteria and indicators can be useful in that respect. Specifically, peers can be defined as fellow scientists who get high marks on indicators that typify a particular discipline. Candidates for assessment committees can then be judged on the basis of that score. In addition, peer autonomy can be encouraged by including international peers on committees.

3.7 Conclusions

CONCLUSION 3.1

The quality of research output in the design and engineering disciplines can be assessed on the basis of two criteria: 1) research quality and 2) societal relevance. It is not necessary to identify a separate set of criteria for these disciplines. Each of these two criteria is related to a set of indicators showing how well a person or output scores on the relevant criterion.

CONCLUSION 3.2

Assessing quality is a question of fine-tuning owing to the differences between disciplines, categories of scientific activity (designing, research) and assessment situations (external evaluations, appointments, research proposals). The present advisory report offers a systematic assessment framework that will do justice to these differences by attributing a certain significance or weight to indicators of research quality and societal relevance (Table 3.1).

4. SURVEY OF LESSONS LEARNED ABROAD

4.1 Introduction

Scientific research is an international affair, and in that respect it is important that the criteria proposed in the previous section are also credible from an international perspective. The committee additionally wanted to make use of lessons learned abroad. It therefore looked at how a number of other countries tackle the quality assessment of scientific research, what criteria they use, and what they have learned in the process. Where possible, the committee looked specifically at the design and engineering sciences. The United Kingdom (UK) and Australia have extensive national research evaluation systems. In both countries, the issue of suitable criteria generated considerable discussion, and there was a great deal of background research. The committee therefore investigated the situation in both these countries in detail. In Finland, the committee looked at how Aalto University assesses research quality; this university took a very thorough approach and its evaluation system devotes considerable attention to the design and engineering sciences. Finally, the committee looked briefly at how a number of countries relevant to the Netherlands deal with quality assessment. Information on the individual countries can be found in Appendix 3. Section 4.2 reviews the most important findings and conclusions.

4.2 Summary of lessons learned abroad and conclusions

Most foreign assessment systems cover the entire spectrum of science and scholarship and only tackle specific problems associated with design and engineering indirectly. There are nevertheless a number of relevant observations that can be made within the context of the present advisory report.

Use of bibliometric data

The assessment methods used in the UK, at Finland's Aalto University, and in Australia attach considerable value to bibliometric analyses. The issue of whether these analyses are in fact applicable has clearly generated considerable discussion in these countries, particularly with respect to the design and engineering sciences. Although the UK's Research Excellence Framework at first emphasised the use of bibliometric methods, that emphasis was ultimately toned down. The consultation rounds made clear that such indicators were not robust enough and had too little support from scientists in the field. They have not disappeared entirely from the assessment system, however; they now serve as background information for review committees. In the Australian system, on the other hand, bibliometrics still plays a prominent role, although such analyses are differentiated by discipline and only apply in specifically defined fields. In the design and engineering sciences, they apply in every discipline except the architectural disciplines. The engineering sciences are evaluated in a similar manner to physics and chemistry. The Finnish evaluation committee indicates that the Architecture, Design, Media and Art Research Panel could not actually make use of bibliometric analyses in its reviews owing to the small number of articles published in ISI journals.

Possibility of including non-traditional information

The evaluations produced by both Aalto University and in accordance with Australia's ERA (Excellence in Research for Australia) method allow 'non-traditional output' to be included in the evaluation. This covers films, websites, exhibitions and creative 3D work (architecture, design, games, software). Here again, this exception appears to apply mainly to the disciplines of architecture and industrial design. The evaluations are carried out by peer review committees that assess case studies submitted by the units under review. The Australian system stipulates specifically that the submitting units must themselves identify the research component of the output. The experience gained at Aalto University in Finland shows that this takes some adjustment. The evaluation committee pointed out that while the university administrators had indeed made this possible, when it came right down to it the units were urged to submit mainly scientific articles.

Focus on societal impact from an international perspective

In addition to the research quality of the output, all of the foreign evaluation systems discussed here also consider its societal impact. There is still considerable discussion of the way in which that is supposed to happen.

Reasonably comparable criteria used for assessment

There is a reasonable level of consensus concerning the criteria to be used to assess the quality of research output. Although the precise description differs from one system to the next, the following criteria are virtually always used: research quality, societal relevance and (to a lesser extent) productivity.

Method used to assemble peer review committees not specified

Peer review committees play a crucial role in the various evaluation processes. This means that the quality of the assessment process depends largely on the quality of these committees. Given the enormous variety of fields that some panels are obliged to consider, the question is whether the members of the peer review committees are in fact actually 'peers' to the extent required by particular research projects. What is notable is that most of the assessment procedures discussed here do not stipulate the way in which the peer review committees are assembled or how the members are selected.

CONCLUSION 4.1

The key criteria in the assessment systems of trendsetting countries such as the United Kingdom, Australia, the United States and Finland are research quality and societal relevance.

CONCLUSION 4.2

Various foreign assessment systems allow results such as creative output, websites, and other products to be included in the assessment of quality in the design and engineering sciences. That possibility is often limited to the more 'artistic' fields, however.

CONCLUSION 4.3

Virtually all of the assessment systems considered devoted considerable effort to bibliometric information. The significance of that information differed from one system to the next.

CONCLUSION 4.4

The assessment framework proposed in this advisory report is credible within an international context.

5. CONCLUSIONS AND RECOMMENDATIONS

Section 2 described problems in the present system of assessing the quality of research output in the design and engineering disciplines. Section 3 identified a new assessment framework that avoids these problems. Section 4 considered that framework within an international context. This final section discusses the scope of this advisory report and its implications for existing practice. It also identifies a number of follow-up steps. The section ends with a set of recommendations. First, however, we repeat the key conclusions below.

CONCLUSION 2.1

Quality comparisons are only meaningful when they are made within disciplines. As appealingly simple as administrators may find the ‘one size fits all’ approach to assessing quality, it does not do justice to significant differences between disciplines and will therefore always give certain ones an unfair advantage.

CONCLUSION 3.1

The quality of research output in the design and engineering disciplines can be assessed on the basis of two criteria: 1) research quality and 2) societal relevance. It is not necessary to identify a separate set of criteria for these disciplines. Each of these two criteria is related to a set of indicators showing how well a particular output, unit or person scores on the relevant criterion.

CONCLUSION 3.2

Assessing quality is a question of fine-tuning owing to the differences between disciplines, categories of scientific activity (designing, research) and assessment situations (external evaluations, appointments, research proposal). The present advisory report offers a systematic assessment framework that will do justice to these differences by attributing a certain significance or weight to indicators of research quality and societal relevance (Table 3.1).

CONCLUSION 4.4

The assessment framework proposed in this advisory report is credible within an international context.

Scope of this advisory report

In accordance with the request submitted to the Academy by the three Dutch universities of technology, the assessment framework is intended mainly for the design and engineering sciences. However, the committee anticipates that the framework will also be useful for other technical disciplines. That is first of all because the two criteria identified (research quality and societal relevance) correspond closely to the existing practice of quality assessment (SEP, assessment of research proposals, etc.). Secondly, it is because the proposed framework is flexible enough to do justice to relevant differences between disciplines. That flexibility is the result of the indicators used to determine how well research output satisfies the two quality criteria. These indicators need not be identical or even equally significant in every area of science. Discipline-specific indicators can even be added. The proposed framework is therefore not only useful for the design and engineering disciplines, but also, for example, for the more basic sciences (i.e. the natural sciences. After all, the quality indicators customarily used there (publication in high-impact journals, etc.) are included in the indicators in Table 3.1. The framework is therefore generic in nature, but to become operationally valuable, it must be made discipline-specific by selecting pertinent quality indicators and assigning them a relative importance. Scientists working in the more basic disciplines can, for example, indicate that the quality of their output can be assessed by considering the indicator 'publication in high-impact journals'.

Compatibility with existing practice

There are various situations that require quality assessments to be made: the evaluation of a research proposal; a performance review of a research group; or the evaluation of a staff member being considered for a promotion or permanent appointment. The committee believes that the advice given here will be relatively easy to apply in the existing assessment practice. The SEP used for external evaluations is also compatible with the proposed assessment framework.

Rejection of the 'one size fits all' approach to assessing quality means that policy-making will probably come to play a more important role in decision-making situations involving comparisons between disciplines, for example in the NWO's competitions for research funding in heterogeneous fields or in the award of scientific prizes (Spinoza). Such decisions cannot be made and justified merely on the basis of a quality assessment. Comparing quality between different disciplines is like comparing apples and oranges (person p_1 or research proposal r_1 has a value score of v_1 on indicator

set i_1 and person p_2 or research proposal r_2 has a value score of v_2 on indicator set i_2). At best, it is possible to compare scores that have been standardised for the relevant disciplines (person p_1 scored among the top 5 percent of scientists on indicator set i_1 in discipline d_1 whereas person p_2 scored among the top 25 percent of scientists on indicator set i_2 in discipline d_2). Even here, however, the meaning is not entirely obvious.

Request for advice by 3TU

The committee applauds the fact that the Governing Bodies of the three Dutch universities of technology have raised the issue of the status of design and engineering activities and the related research. These activities are under pressure because the quality assessment framework is not sufficiently compatible with them. As a result, the balance between basic research, applied research, and design and engineering activities has been or is at risk of being upset. This is an undesirable situation, given the mission of the universities of technology. If the three Governing Bodies adopt the advice given here and communicate it to their academic staff, they will have made clear that outstanding design and engineering activities are important to universities of technology and a vital component of the technical sciences.

The first step in the process of implementing the advice is for the design and engineering sciences to determine which quality indicators described in Table 3.1 apply to them and their relevant importance. These indicators must be credible within the discipline in an international context. Only peers working in that discipline can determine which indicators should apply. Most obvious would be to utilise existing cycles of quality assurance for this purpose. The board of 3TU could, for example, ask a discipline to define the quality indicators prior to an external evaluation, so that the evaluation committee is able to use these indicators in its research assessment. A committee consisting of international peers could also be asked to evaluate these indicators.

RECOMMENDATION 1

The committee advises the board of 3TU to create sufficient scope for discipline-specific quality assessment in the technical sciences, and to ask the design and engineering disciplines to identify the indicators and their relative importance for assessing quality in those disciplines. These indicators must be credible in an international context.

Research funding bodies

The committee believes that the method used for quality assessment should not in itself be advantageous or disadvantageous for particular disciplines. This should be a key factor in the assessment system utilised by the NWO/STW. The assessment framework set out in this advisory report can help determine quality indicators that take disciplinary differences into account.

By introducing cross-disciplinary programmes, the NWO is demonstrating that it is in any event aware of the importance of programmes that do not fit within existing frameworks and for which other quality indicators may apply. The committee believes that this approach should be extended, for example by starting up programmes that combine NWO and STW areas. That is very important for the design and engineering disciplines, which do not have a dedicated funding channel at the NWO. The committee advises the STW to take a more differentiated approach to assessing the societal relevance of research projects. Even those projects that do not immediately interest businesses because of their long-term nature should be eligible for funding. That would be of benefit to the design and engineering disciplines, given the fact that various intermediary research organisations in a number of areas have been shut down (see Section 2).

The committee applauds the fact that most research funding bodies make use of the peer review system. It is vital, however, to select the right peers, as this will have a major influence on the outcome of the assessment. Specifically, assessment committees have very little leeway to deviate from referee reports. It would be a good idea to make the peer selection criteria more objective and transparent. That would improve the quality of the peer review process. Table 3.1 may be useful in that context. Peers involved in assessing quality in the design and engineering sciences must also be assessed according to a broad range of indicators, and not only on their publication behaviour.

RECOMMENDATION 2

The committee advises the NWO and the STW to devote more attention to programmes for disciplines that do not fit easily into existing disciplinary categories and the present quality assessment method. This applies in particular for the design and engineering sciences. The quality indicators used must do justice to these disciplines. The committee advises the STW to support projects that will be relevant to society in the longer term and in which businesses have not yet expressed a tangible interest.

Scientists in the design and engineering sciences

A quality assessment system that is fine-tuned to the design and engineering sciences requires the scientists working in these disciplines to make an active contribution, for example in order to identify the quality indicators, decide on their relative significance, and operationalise them. They will also need to participate as peers in NWO and STW research proposal assessment procedures. The committee has noted that this now only happens on a limited scale.

The assessment framework described in this advisory report makes it possible to use other indicators of research quality than peer-reviewed publications. Nevertheless, the committee and many of the interviewees believe that peer-reviewed publications

are also important for the design and engineering disciplines, and will remain so. Such publications help foster mutual quality checks and the dissemination of knowledge. Scientists working in these disciplines or in specific areas of these disciplines do not always publish as a matter of course, however. It is therefore important to look for methods of publication that are appropriate for the discipline or sub-discipline concerned. New media can play a role in that respect.

RECOMMENDATION 3

The committee advises scientists in the design and engineering disciplines to do their utmost to promote a culture of peer-reviewed publications wherever necessary. Such publications can serve to verify results, disseminate knowledge and contribute to the 'scientification' of the discipline.

RECOMMENDATION 4

The committee advises the universities and the NWO/STW to make the peer selection process more transparent and objective. That will improve the quality of peer review process. The assessment framework and discipline-specific quality indicators described in this advisory report can play a role in this context. It is important to select the right peers when convening external evaluation and review committees and when selecting referees to evaluate research proposals.

BIBLIOGRAPHY

- Aalto University (2009a). *Research Assessment Exercise 2009, Panel reports*
- Aalto University (2009b). *Research Assessment Exercise 2009, Terms of reference for research assessment panels*
- Aalto University (2009c). *Bibliometric report 2003-2007*
- DCT Disciplineoverlegorgaan Construerende Techniek (2000). *Criteria voor de beoordeling van het ontwerpen in de construerende disciplines*
- ERiC publicatie 1001 (2010). *Handreiking Evaluatie van maatschappelijke relevantie van wetenschappelijk onderzoek*
- ERiC publicatie 1002 (2010). *Pilot Study at Faculty of Architecture TU Delft, Final Report (Confidential)*
- ERiC publicatie 1003 (2010) *Evaluating Research in Context Pilot Study at the Faculty of Electrical Engineering – TU/e (Confidential)*
- ERiC publicatie 1004 (2010). *Maatschappelijke relevantie van Werktuigbouwkundig onderzoek. Eindrapport ERiC project CTW Universiteit Twente (vertrouwelijk)*
- Hee, Van K, Overveld, K van, (2010). *Criteria voor het beoordelen van een technologisch ontwerp*
- HEFCE (2009). *Research Excellence Framework, Second consultation on the assessment and funding of research*
- KNAW-commissie kwaliteitszorg (2008). *Kwaliteitszorg in de wetenschap, Van SEP naar KEP: Balans tussen rechtvaardigheid en eenvoud*
- Royal Academy of Engineering (2008). *Research Excellence Framework*
- Royal Academy of Engineering (2009). *Research Excellence Framework, second consultation. Response to the Higher Education Funding Council for England*
- Schouten, M.J.W. (2009). *Kwaliteitsbeoordeling Ontwerp- en Constructiedisciplines*. Technische Universiteit Eindhoven
- Stokes, D.E. (1997). *Pasteur's Quadrant; Basic Science and Technological Innovation*, Brookings Institution Press, Washington
- VSNU, KNAW and NWO (2009), *Standard Evaluation Protocol 2009-2015*, only available at www.knaw.nl/sep

APPENDIX 1

LIST OF INTERVIEWEES

Prof. P.M.G. Apers, University of Twente, Information science
Prof. J.A. Battjes, Delft University, Civil engineering
Prof. A. Beukers, Delft University, Aerospace engineering
Prof. C.J.P.M. de Bont, Delft University, Industrial design
Prof. A.C.J.M. Eekhout, Delft University, Architecture
Prof. K.M. van Hee, Eindhoven University, Mathematics and information science
Prof. J.J. Heijnen, Delft University, Bioprocess engineering
Dr F.D. van der Hoeven, Delft University, Architecture
Prof. F.J.A.M. van Houten, University of Twente, Mechanical engineering/
Industrial design
Prof. J.T.F. Keurentjes, Eindhoven University, Chemical engineering
Prof. H. Leegwater, Eindhoven University, Chemical engineering
Prof. M.C.M. van Loosdrecht, Delft University, Bioprocess engineering
Prof. J.B.O.S. Martens, Eindhoven University, Industrial design
Prof. B. Nauta, University of Twente, Electrical engineering
Prof. C.J. Overbeeke, Eindhoven University, Industrial design
Prof. W.A. Poelman, University of Twente, Industrial design
Prof. J.M. Post, Eindhoven University, Architecture
Prof. A.F.J. van Raan, Leiden University, Quantitative analysis
Prof. M.J.W. Schouten, Eindhoven University, Industrial design
Prof. M. Steinbuch, Eindhoven University, Mechanical engineering
Dr S. Silvester, Delft University, Industrial design
Dr P.E. Vermaas, Delft University, Philosophy
Dr J. Visschers, NIKHEF, Instrumentation design
Prof. H.J. de Vriend, Deltares, Civil engineering
Prof. P.J.V.V. van Wesemael, Eindhoven University, Architecture
Prof. J.J. van Wijk, Eindhoven University, Information science
Prof. H.W. Zandbergen, Delft University, Applied physics

APPENDIX 2

ERIC PROJECT INDICATORS

The ERiC project involved three pilots at the Faculty of Electrical Engineering at Eindhoven University of Technology, the Faculty of Architecture at Delft University of Technology, and the Faculty of Mechanical Engineering at the University of Twente. The purpose of the pilots was to develop a method for assessing the societal relevance of the research conducted within these faculties. The indicators that were proposed after the Architecture and Electrical Engineering pilots are listed below.

Table A2.1 Indicators of societal relevance, Architectural Research [ERiC, 2010]

Aspect of societal relevance	Indicators
Knowledge dissemination	Professional publications, non-scientific publications, exhibitions, etc. Dissemination of technology, artefacts, standards Advisory and consultancy work Popularisation, education and contribution to public debate Training of professionals, mobility of graduates Master's theses and graduation projects that tackle issues arising in actual practice
Stakeholder interest	Number of researchers with relevant practical experience in the sector/sectors to which the research programme relates Public funding related to societal issues Contract financing by potential users Collaboration with stakeholders in research, testing and evaluation Consortiums with non-academic organisations
Impact and use of results	Income generated by using results Visibility in public debate/public media rankings

Table A2.2 Indicators of societal relevance, Electrical Engineering [ERiC, 2010]

Aspect van maatschappelijke relevantie	Indicatoren
Knowledge dissemination	PhDs in industry MAs in industry Proof of concept Presentations at specialist conferences
Stakeholder interest	Joint roadmaps Presentations by invitation Valorisation grants Industrial financing Staff exchanges (exchange programmes such as Casimir, knowledge transfer schemes) Part-time professors from/working in industry Consortiums with industry
Impact and use of results	Market introduction and new projects in industry Spin-offs with industrial contacts Patents

APPENDIX 3

QUALITY ASSESSMENT

ABROAD

A3.1 United Kingdom: Research Excellence Framework

In the United Kingdom (UK), current discussion of research assessment methods focuses mainly on the Research Excellence Framework (REF), a new process for assessing the quality of scientific research and meant to succeed the Research Assessment Exercise (RAE). The most recent RAE was conducted in 2008. One important objection to this latest round is that the RAE had become an unwieldy and labour-intensive operation. The purpose of the REF is to identify excellence in research and to evaluate its impact. The assessment will focus on coherent research units, and not assess individuals or whole institutions. The first REF assessments are scheduled to take place in 2013 and will determine the distribution of funding in 2014. The precise way in which the REF will be conducted is still subject to debate. This section describes the broad outlines of the REF as announced in late 2009 [HEFCE, 2009]. It also looks at the key points of debate, with an emphasis on the criticism levelled by the engineering sciences.

Assessment method

The REF proposal accords a key role to peer review. All research offered for assessment is to be judged by expert panels on three criteria:

1. *Output quality*: the main aim of the REF is to identify research that can be considered excellent within an international context. This covers many different categories of output.
2. *Impact*: this criterion assesses the extent to which scientists have made a demonstrable contribution to the economy, society, public policy, culture and the quality of life. Assessment of impact will be by means of case studies.
3. *Environment*: this criterion will take the quality of the research environment into account. This refers to the research strategy, staff development and the training of researchers.

Not all of the criteria are equally important. The proposal is to weight them as follows for the overall assessment: output quality 60 percent, impact 25 percent and environment 15 percent. The overall assessment will consist of a system of stars, starting

with 'unclassified' (below standard quality or work does not meet the definition of research) and rising to four stars ('exceptional').

Research is to be divided into various review panels, but the details are still subject to debate. A two-tier structure is envisaged, with four main panels and thirty sub-panels. In this proposal, most of the technical sciences will be grouped and assessed in a single large engineering panel covering all research in electrical engineering, chemical engineering, mechanical engineering, aerospace engineering, mining, civil engineering and materials science. Coming under same main panel but a different sub-panel will be computer science and informatics, mathematics, physics, chemistry and earth systems. Architecture has been categorised under the panel for architecture, the built environment, town and country planning, which itself resides under another main panel.

Topics of discussion

Use of bibliometrics

The usefulness of citation and other quantitative indicators has generated a great deal of discussion. The question is whether such indicators could replace the expert review in the more 'science-based' disciplines. Ultimately, the conclusion was that such indicators were not robust enough and would not be acceptable to the sector. Nevertheless, the aim is to let these indicators play a greater role than they do in the RAE. The current idea is for the panels to use the indicators to inform their reviews, in any event in the medical, health, biological and physical sciences, psychology, engineering and computer science. For now, it is up to the panels to decide how they are going to use these indicators.

Much of the discussion until now has focused on the impact criterion because it is very unclear precisely what is being assessed and to whom certain impacts should be attributed. This is also a new criterion compared to the RAE. The units that are being assessed must submit one or more case studies describing their impact. Impact is regarded as the impact on the economy, society, public policy, culture or the quality of life. The following indicators have been proposed:

- indicators of research income generated from users (businesses, government, research charities);
- indicators of collaboration with users;
- indicators specific to a particular area of research. The indicators would be selected from a kind of menu. A proposal for this type of indicator can be found in HEFCE, 2009.

Criticism from the engineering sciences

In a major consultation round, various stakeholders were asked to comment on the new framework. The commentary of the Royal Academy of Engineering is the most pertinent within the context of this advisory report [Royal Academy of Engineering, 2009]. The Academy's key objections are the following.

- The Academy agrees that citation analysis is not robust enough to be used to assess research results and doubts whether it ever will be, given the breadth of the discipline and the various types of output it produces. It also points out that the publication traditions vary considerably from one sub discipline to the next.
- The Academy applauds the use of citation information to inform the panels, but points out that available information from the citation databases is often insufficient for the engineering disciplines, leading to major differences in the information submitted by the various sub disciplines.
- The Academy agrees that the impact of research should be considered in the assessment. Given the many different categories of engineering output, however, this will be a difficult process and it will take time to build up the necessary experience using it. At this stage, the proposed weight factor of 25 percent is therefore considered too much for this criterion.
- In the *Research Assessment Exercise*, industrial members were full members of the expert panels. In the current REF proposal, they are only involved in assessing impact. The Academy is in favour of also involving members from industry in the engineering panel when assessing all the criteria.
- The Academy seriously questions the proposal to have all engineering research reside under a single panel. That is first of all because a single panel will be obliged to review an enormous amount of research (produced by approximately 4400 FTE researchers). Secondly, it expects that competition between the engineering disciplines will increase as a result, ultimately resulting in less money for the engineering sciences.

A3.2 Australia: Excellence in Research for Australia

The Australian Minister for Innovation, Industry, Science and Research, Senator Kim Carr, launched a new initiative in 2008 to develop an ambitious method for evaluating scientific research. The Excellence in Research for Australia (ERA) initiative is being developed by the Australian Research Council (ARC) together with the Department of Innovation, Industry, Science and Research and is based on a combination of metrics and expert review. A trial ERA was run in 2009 in two clusters, physical, chemical and earth sciences and humanities and creative arts. In 2010, the ERA assessment method will be introduced for all disciplines.

The exercise is intended to collect information on the quality of research, identify national strengths and weaknesses, excellence and emerging areas where there are opportunities for development, and to position Australia within the international research arena.

Assessment method

For evaluation purposes, research is divided into clusters of more or less similar disciplines:

1. Physical, chemical and earth sciences
2. Humanities and creative arts
3. Engineering and environmental sciences
4. Social, behavioural and economic sciences
5. Mathematical information and computing sciences
6. Biological and biotechnological sciences
7. Biomedical and clinical health sciences
8. Public and allied health sciences

These disciplines are themselves divided into a large number of subdisciplines. Engineering, for example, consists of the subdisciplines aerospace engineering, automotive engineering, biomedical engineering, etc. The institutions involved are basically evaluated on the basis of these subdisciplines. The higher-level discipline is only assessed if the relevant subdiscipline's output is too meagre for a proper evaluation. Assessment therefore does not cover whole faculties or individual researchers. The ERA method uses a range of indicators to measure research quality (see Table A3.1).

Table A3.1 Assessment criteria in Excellence in Research for Australia

Category	Indicators
Ranked outlets	Publication in peer-reviewed journals Publication in reviewed conference proceedings
Citation analyses	Relative citation impact (against world and Australian HEP benchmarks) Distribution of papers by world centile thresholds, profiled against Australian HEP average
Research volume and activity	Distribution of research output by outlet types and distribution of FTE and headcounts by HEDSC levels
HERDC Research income	Number of grants Average dollar per grant Total income for research Total income per FTE Ratio of total income per FTE compared with discipline-specific benchmark
Esteem	Editorship at A and A+ journals Contribution to a prestigious work of reference Curatorial role Fellowship of a Learned Academy Recipient of nationally competitive research fellowship Recipient of prestigious prizes and distinctions
Applied	Patents Plant breeder's rights Registered designs Research commercialisation income NHMRC-endorsed guidelines (health care)

The units of evaluation are assessed and rated by Research Evaluation Committees (RECs), made up of internationally recognised experts with expertise in research

evaluation and broad discipline expertise. There are no guidelines for selecting the REC members. The REC chair asks the other members to evaluate the unit based on the data provided. Later, the REC meets to discuss the results. Only in the relevant disciplines/subdisciplines do the REC members (assisted by external experts where necessary) conduct a peer review for certain categories of research output. A very detailed discipline matrix shows what output is eligible for this in each discipline.

The institutions are asked to deliver data on at discipline/subdiscipline level. The data must be entered into the System to Evaluate the Excellence of Research (SEER); they can then be manipulated electronically. The indicator pattern and quantity of output per indicator are then determined, based on the data provided. The indicators are only passed on to the REC if a particular discipline-specific threshold value has been exceeded.

Output not limited to scientific publications

It is possible to include 'non-traditional output types' in submissions in certain research disciplines. This option is limited in the technical sciences to the disciplines built environment and design (including architecture). It is not permitted in any of the other engineering disciplines. Non-traditional output includes architectural designs, creative 3D work, design, games, computer software, and exhibitions.

If an output is submitted for evaluation as non-traditional output, the researchers must explicitly identify the research component. This can be done by submitting the Research Statement for ERA Peer Review of Non-Traditional Research Outputs, which addresses the following categories:

Research background:

- field
- context
- research question

Research contribution:

- innovation
- new knowledge

Research significance:

- evidence of excellence

A3.3 Finland: Aalto University evaluation

The university system in Finland was completely overhauled in late 2009. The universities gained more financial and administrative autonomy. They were categorised either as public institutions or private foundations. Starting in late 2009, the foundations were no longer part of the state budgetary system. At the moment, a new approach to funding is being developed that places more emphasis on strategic aspects than on indicators of educational/teaching tasks (no. of students, promotions, etc.). This process also involves a search for new quality indicators.

In order to understand how the Finnish are tackling the quality assessment of scientific research, we looked at a recent evaluation of Aalto University in Helsinki. Although the evaluation only concerned a single university, we decided to include it in this report because it was a recent and large-scale evaluation of a university at which the design and engineering sciences are well represented.

Aalto University was founded in 2009 when the Helsinki School of Economics merged with the Helsinki University of Technology and the University of Art and Design Helsinki. That same year, a peer review committee (PRC) consisting of 62 members from 20 countries evaluated the university over the 2003-2008 period. The purpose of the evaluation was to:

- determine the research quality and societal relevance of the new university;
- identify the groups that had the potential to excel internationally;
- analyse the evaluation process itself.

Aalto University wishes to use the recommendations to define and develop research practices, including the strategic allocation of research funding.

Assessment method

The evaluation was conducted by a PRC divided into nine panels:

1. Chemical technology and materials
2. Electronics and electrical engineering
3. Mathematics and physics
4. Computer science and information technology
5. Mechanical engineering
6. Civil engineering and urban and regional studies
7. Business technology, economics and finance
8. Marketing, management and applied business research
9. Architecture , design, media and art research

This meant nine separate PRCs of five to seven members each. Each panel is meant to evaluate a discipline by means of a self-evaluation, a bibliometric analysis and a site visit. Prior to the site visit, the disciplines themselves provided detailed information.

Only active researchers associated with the university were permitted to submit results. There was also a bibliometric analysis. The PRCs were asked to pay particular attention to the university's international research position. International research hence was consistently used as a benchmark.

Based on the site visit and the information provided (self-evaluation and bibliometric analysis), the PRC awarded marks to each discipline on the following criteria:

1. Research quality
2. Research impact
3. Societal impact
4. Research environment

5. Future research potential

The marks ranged from 1 to 5, with 5 standing for excellent international visibility and 1 for internationally emergent.

The PRC expressly included the results of the bibliometric analysis – specifically the NCSf (citation score standardised by research field) in its assessment of the various disciplines. This helped determine the university's visibility in the national and international research arena. The discipline of architecture, design, media and art research is an exception to this; the bibliometric analyses were not considered in this case because there were too few ISI publications.

Output not limited to scientific publications

The PRC indicates that it was easy enough to evaluate disciplines 1 to 8 using the current science-based criteria. It also indicates that, on paper, Aalto University leaves enough leeway to include all categories of research output (i.e. artefacts in addition to publications), but that the university administrators in fact strongly suggested that science-related output be provided. As a result, the units operating in the discipline of architecture, design, media and art research submitted much smaller quantities of output (i.e. only their articles) than the site visit actually turned up. The relevant panel believes that the administrators had seriously undervalued these units' research and its societal relevance.

For more information about this evaluation, see [Aalto University, 2009a, b, c].

A3.4 United States, National Science Foundation

The committee looked at the National Science Foundation (NSF) in the United States. The NSF is an independent federal agency set up: 'to promote the progress of science; to advance the national health, prosperity, and welfare; to secure the national defense...'. With an annual budget of USD 6.9 billion, the NSF subsidises approximately 20 percent of federally funded basic research at American universities and colleges. Unlike the assessment system described in the previous section, the NSF concentrates mainly on the ex-ante evaluation of research proposals. Besides the NSF, there are many private, mission-based and industrial funding bodies active in the United States. They usually do not rely as heavily on clearly defined criteria.

Peer review is crucial to the NSF's evaluation of research proposals. Such proposals are evaluated mainly according to two criteria:

1. What is the scientific merit of the proposed research activity?
2. What is the broader impact of the proposed research activity?

Additional criteria may also be imposed, depending on the programme.

Assessment method

The proposal is submitted to a program officer, who passes it on to a review panel. The panel, consisting of three to ten members, evaluates the proposal on the above criteria. The program officer then draws up an analysis and a recommendation based on the panel's views. The NSF division director uses this recommendation to inform his or her final funding decision.

Those submitting research proposals are asked to suggest suitable referees. They may also indicate who would not be a suitable referee.

A3.5 Norway

In Norway, research is usually the province of a general or specialist university, university colleges and research institutes. The Ministry of Education is the main coordinating body and the other relevant ministries are responsible for promoting and financing research in their fields. Part of the budget is provided directly by the Ministry of Education and Research. Before, that was usually on the basis of student numbers. Since 2003, funding is divided into a) basic funding, b) education and c) research. Research funding is based on performance and strategic considerations. Performance is measured by looking at:

- number of Bachelors completed;
- number of Masters completed;
- number of PhDs completed;
- external collaboration;
- staff composition (number of professors);
- number of EU and RCN grants;
- patents and output in the form of scientific and scholarly publications.

The publications are weighted depending on the publication method (article, book) and channel (journal, website, publisher if book).

Universities must enter all the data in a national documentation system because funding depends in part on the documentation in this system.

In addition to the basic funding described here, the Research Council Norway (RCN) also awards competitive grants. Proposals are evaluated by means of peer review based on a huge number of criteria.

The RCN also conducts national evaluations:

- of RCN-funded projects
- at institutional level
- at discipline level.

The purpose of the evaluations is to improve research quality, relevance and efficiency and to identify strengths and weaknesses at national and institutional levels. The RCN

also uses the outcomes when drawing up the national research strategy. The outcomes are also helpful when allocating research funding. The RCN is working on a follow-up model for these evaluations.

A3.6 Sweden

In Sweden, research is conducted at 14 universities and 25 institutions of higher education. Most publicly funded research takes place within the latter. Some of that research is funded directly by government. The public authorities also fund the Research Councils and the Sectoral Research Agencies.

The largest funding body is the Swedish Research Council (SRC), which has an annual budget of SEK 4 billion. The SRC awards grants to research projects in six different areas, including the natural and engineering sciences. Research proposals are submitted to forty panels where they are subjected to peer reviews.

The second largest funding body is the Swedish Agency for Innovation Systems (VINNOVA), with a budget of SEK 2.15 billion. This agency makes funding available in six research areas, including natural sciences and technology.

Sweden does not have a national ex-post research evaluation system.

A3.7 France

The French system of higher education consists of research institutes, universities and *grandes écoles*. The institutes are the main centres of research. They decide on and fund their own research and evaluate it themselves. At the moment, the public authorities would like to give the universities more autonomy and turn the research institutes into agencies that distribute research funding on a project-driven basis; see the National Research and Innovation Strategy, SNRI. According to this document, research should be shifted over to the universities. Institutes can also apply for funding through the National Research Agency (ANR).

France does not have a national ex-post evaluation system for research at universities, *grandes écoles* and institutes.

A3.8 Germany

Public research institutions in Germany are divided into four networks:

- the Max Planck Society (MPG)
- the Fraunhofer Society (FhG)
- the Helmholtz Association of German Research Centres (HGF)
- the Wilhelm-Gottfried-Leibniz Scientific Community (WGL)

The universities receive their basic funding from the relevant federal states. The largest federal funding body is the German Research Foundation (DFG), which has an annual budget of EUR 1.3 billion. The DFG funds all the disciplines, from the humanities to engineering, and gets its money from both the federal and the state authorities according to a fixed distribution key. The DFG comes under the auspices of the Federal Ministry of Education and Research (BMBWF). Other major funding bodies include the private Alexander von Humboldt Foundation and the German Academic Exchange Service (DAAD), which finances international exchange and other programmes and projects.

Although research is never evaluated at national level, the DFG does produce a ranking (the Förder ranking), based on the following criteria:

- number of positive funding awards by the DFG;
- composition of research group;
- funding acquired via contract research;
- position in DFG-subsidised programmes;
- number of DFG reviewers;
- number of Alexander von Humboldt visiting researchers;
- number of DAAD researchers and students;
- participation in EU framework programmes;
- articles published in international journals.

In addition, the federal states of North Rhine-Westphalia (based on external funding and number of PhDs) and Lower Saxony (has its own Academic Advisory Council) have developed their own evaluation systems. (The applied research institutes in Baden-Württemberg were evaluated in 2008 at the request of the Ministry of Economic Affairs.) None of these evaluations is linked directly to funding. These evaluation systems are not trendsetting internationally.

The aim of the Excellence Initiative [*Exzellenzinitiative*], a DFG funding programme, is to identify research excellence and to promote and improve the quality of higher education and research in the broader sense. Germany hopes that this initiative will improve its international competitiveness. Research proposals submitted by the entire Germany scientific community are subject to peer review. The proposals that are ultimately selected are then showcased.

A3.9 Austria

Research in Austria is funded by:

- the Austrian Research Promotion Agency (FFG)
- the Austrian Science Fund (FWF)
- the Austria Wirtschaftsservice (AWS)

These funding bodies are part of the Ministry of Science and Research (BMWf), the Ministry of Transport, Innovation and Technology (BMVIT) and the Ministry of Economy, Family and Youth (BMWfJ), but the division of strategic responsibilities between the ministries and the funding bodies is not entirely clear and the responsibilities of the various funding bodies overlap.

In 2003, the National Foundation for Research, Technology and Development was set up in order to guarantee the major funding bodies an ongoing financial buffer and to ensure the continuity of large-scale, long-term projects. The money comes from the Austria National Bank (OeNB) and the European Recovery Fund (ERP).

Austria does not have a national ex-post evaluation system for research. The Research & Technology Evaluation platform (Fteval) is attempting to set up a transparent evaluation system, part of a push to optimise strategic research, technology and development (RTD) policy planning. Fteval also aims to promote an evaluation culture in Austria in cooperation with Technology and Research policy-makers.

