*Willem K.B. Hofstee*

# Kwaliteit van beoordelingen in de context van kunstbeleid

**Wat mooi en lelijk is is niet alleen een kwestie van individuele smaak. Artistieke kwaliteit heeft een intersubjectieve kern, en dat maakt het zinvol door te gaan met beoordelingsprocessen waaraan liefst een behoorlijk aantal gekwalificeerde deskundigen meedoen. Maar dat betekent niet dat het beoordelingsproces zonder moeilijkheden verloopt.**

Beoordeling en beleid zijn zusje en broertje. Hun moeder is *kwaliteit*: 'beoordelen op kwaliteit' en 'kwaliteitsbeleid' zijn pleonasmen. Ze suggereren dat beoordeling en beleid ook uit een andere schoot zouden kunnen zijn geboren. Minder duidelijk is hun vaderschap (mama's baby, daddy's maybe): die komt uit het milieu van de kosten, de afwegingen en de randvoorwaarden. Zusje Beoordeling aardt nog het meest naar moeder Kwaliteit en heeft soms wat weinig oog voor de noemer van de baten/kostenverhouding; broertje Beleid daarentegen deelt met pa de neiging vreemd te gaan in puur politieke richting, en die noemer te verheffen tot wat telt: verdeling van fondsen onder electoraal interessante bevolkingsgroepen. Subsidiëring van low culture, liefst gepaard aan woorden van huwelijkse trouw aan moeder Kwaliteit, is een voorbeeld. Het gezinnetje kent aldus zijn spanningen en problemen.

Deze bijdrage is bewust eenzijdig. Zij is geschreven vanuit het perspectief van wat hier wordt gezien als de zwakste partij, de beoordelaar. Die zwakte is gelegen in het kwaliteitsbegrip zelf, en komt tot uiting in bevindingen omtrent (gebrek aan) overeenstemming tussen beoordelaars. Aan de orde komen vervolgens de beleidsconsequenties die men mede daaruit zou kunnen trekken, de onafhankelijkheid van het beoordelingsproces die nodig is voor dergelijk beleid, en tot slot dingen die de beleidsmaker beter achterwege kan laten, in het belang van die onafhankelijkheid. De bijdrage kan worden gelezen als een onderbouwing van een door de tijd gelouterde praktijk, waarin commissies van professionele beoordelaars aanvragen beoordelen op kwaliteit.

In het navolgende gaat het alleen over het toekennen van subsidies, beurzen, plaatsen en dergelijke, en niet zozeer over prijzen en onderscheidingen. Dat zijn namelijk als het goed is geen beleidsinstrumenten maar feestelijke rituelen. Het beleid bewaart er een gepaste afstand toe. Keuzes die bij particuliere kunstfondsen worden gemaakt, liggen eveneens buiten het bereik van dit betoog, om vergelijkbare redenen.

## (Gebrek aan) overeenstemming

In 1990 voerde de Groningse psychologe Titia Top, met subsidie van het toenmalige ministerie van wvc, een exemplarisch onderzoek uit naar beoordelen in de kunst.[1] Van 149 studenten in de autonome richtingen (schilderen, beeldhouwen etc.)

1. Het onderzoek waaraan in dit artikel wordt gerefereerd, is onderdeel van Titia Tops bredere studie naar loopbanen van vrouwelijke en mannelijke kunstenaars (Top 1993).

aan Nederlandse kunstacademies werd het eindexamenwerk beoordeeld door twaalf professionele beoordelaars: galeriehouders, kunstrecensenten, kunsthistorici en kunstenaars, die veelal zitting hadden in kunstcommissies. Ze reisden de eindexamenexposities van de zeven deelnemende kunstacademies af en beoordeelden onafhankelijk van elkaar het eindexamenwerk op kwaliteit, techniek, originaliteit en artistieke toekomst van de maker.

Deze bijdrage richt zich op een zijaspect van het onderzoek van Top, namelijk de mate van overeenstemming tussen twee beoordelaars – of zo men wil, het gebrek daaraan. Die overeenstemming varieerde van 15 procent (voor artistieke toekomst) tot 21 procent (voor techniek). Voor de onderzoekster was die geringe overeenstemming geen verrassing; juist omdat ze die verwachtte had ze een dozijn beoordelaars ingeschakeld, zodat de gemiddelde beoordeling een redelijke mate van betrouwbaarheid zou hebben. Door beoordelaars zelf, door beoordeelden, en door iedereen die geen systematisch onderzoek naar beoordelingen heeft gedaan, plegen zulke getallen wel als onthutsend te worden ervaren.

Laten we eerst nagaan wat de uitkomst precies betekent. Twee beoordelaars kunnen het volledig met elkaar oneens zijn: de een geeft de toekomstige Rembrandt een 0 en Jansen een 10, de ander respectievelijk een 10 en een 0. De overeenstemming tussen deze twee beoordelaars is hier dus perfect negatief: -100 procent. Tegen die achtergrond is +15 procent tot +21 procent nog niet eens zo laag. Maar drie beoordelaars kunnen het met elkaar al niet meer volledig oneens zijn. Naarmate het aantal beoordelaars groter wordt, nadert de minimale gemiddelde overeenstemming van beneden af tot 0, in plaats van -100 procent. Voor de praktijk: in een meerhoofdige jury is er altijd wel iemand het een beetje met je eens. De beoordelaar, ook maar een mens, ontleent daaraan een gevoel van overeenstemming. Maar dat gevoel is grotendeels illusoir: als een jury zou worden samengesteld uit zuivere dobbelstenen, zou er ook een mate van overeenstemming zijn tussen sommige paren daarvan.

*Als een jury zou worden samengesteld uit zuivere dobbelstenen, zou er ook een mate van overeenstemming zijn*

Als beoordelaars maar wat zouden doen, zou men 0 procent overeenstemming moeten verwachten; als ze het volledig met elkaar eens zouden zijn, 100 procent. Hoe tussenliggende getallen zoals 15 procent en 21 procent precies moeten worden geïnterpreteerd, hangt af van de beoordelingsschaal en van de gekozen index: bij gebruik van schoolcijfers bijvoorbeeld kan niet meer letterlijk van een percentage overeenstemming worden gesproken. Maar ruwweg geldt dat de beoordelaars voor een klein deel gemeenschappelijk oordeelden, en voor het overgrote deel ieder hun eigen weg gingen. Pas bij middeling van beoordelingen wordt het resultaat enigszins acceptabel: de stokpaardjes van de individuele beoordelaars vallen dan tegen elkaar weg. Top berekende dat het gemiddelde oordeel van de twaalf beoordelaars voor 64 procent (artistieke toekomst) tot 76 procent (techniek) zou overeenstemmen met het gemiddelde oordeel van een andere jury van twaalf leden.

Tot troost moge strekken dat beoordelaars op andere terreinen het er niet beter van afbrengen. Men zou kunnen veronderstellen dat het bij beoordelingen van manuscripten of subsidieaanvragen in de exacte wetenschappen ook exacter zou toegaan. Maar die wetenschappen mogen dan exact zijn, de beoordeling ervan is het niet. Bij nader

inzien bestaat er in dit verband natuurlijk ook geen exacte wetenschap: de beoordeling ervan heeft primair betrekking op het creatieve gehalte, net als in de kunst, en daar zijn per definitie geen regels voor.

Betekent dit nu dat er in jury's en commissies net zo goed met dobbelstenen kan worden gegooid? Men hoort dat de 'slachtoffers' vaak zeggen, maar kwaliteit bestaat wel degelijk – al geldt dat met allerlei slagen om de arm. De afgewezen kandidaat heeft weliswaar het volste recht te denken dat de beoordelaar zich schromelijk heeft vergist, want zelfs bij een twaalfkoppige jury is de kans daarop nog steeds aanzienlijk. Maar wie zegt dat kwaliteit niet bestaat, dat beoordelen louter een kwestie is van persoonlijke smaak, en dat mooi en lelijk puur subjectief zijn, heeft ongelijk. Kwaliteit heeft wel degelijk een *inter*subjectieve kern. Die valt nooit helemaal te pakken te krijgen, maar wel te benaderen. Het nuchtere feit dat menselijke beoordelaars meer met elkaar overeenstemmen dan dobbelstenen, geeft aan dat beoordelen meer is dan louter een ritueel. De benadering is beter naarmate het aantal (gekwalificeerde) beoordelaars hoger is.

Men kan vervolgens tegenwerpen dat beoordelaars zich collectief kunnen vergissen, en ook dat komt zeker voor. Maar die vergissing kan alleen worden afgemeten aan het oordeel der geschiedenis, dat wil zeggen het oordeel van latere generaties. Idealiter zou men dus de mensen van alle tijden bij de beoordeling betrekken. In de praktijk moet worden volstaan met de vraag of een kunstwerk toekomst heeft, dat wil zeggen met een poging zich in die toekomst te verplaatsen. Maar in geen geval is er een beter criterium voor kwaliteit voorhanden dan het intersubjectieve.

Bij de bescheiden mate van overeenstemming moet nog worden bedacht dat een deel van de consensus onder de tafel blijft. Dat deel heeft betrekking op aperte wankwaliteit. Die doet zich namelijk zelden voor. Dat komt doordat er voorafgaand aan de beoordeling allerlei voorselectieprocessen hebben plaatsgevonden. De eindexaminandi bijvoorbeeld vormden een selectie uit degenen die ooit met de studie waren begonnen, en die hadden zichzelf van tevoren al geselecteerd. Zo gaat dat in het algemeen: voorselectie- en zelfselectiemechanismen zorgen dat kandidaten niet al te zeer uit de toon vallen. Enerzijds verlicht dat de taak van de beoordelaars, maar anderzijds verkleint het hun onderlinge overeenstemming. Als er een enkele keer toch een onzinnige aanvraag of kandidatuur is binnengeslopen, maken beoordelaars daar vrij unaniem korte metten mee. Als het er veel zouden zijn – wat zelden of nooit voorkomt – dan zouden de beoordelingen een grotere mate van unanimiteit weerspiegelen. Maar die overeenstemming komt dus in de praktijk niet aan de oppervlakte.

### Enkele beleidsconsequenties

In een beleidscontext vallen uit het bovenstaande twee conclusies te trekken. De ene is dat beoordelingsprocedures het voornamelijk van hun vooruitwerkende kracht moeten hebben. Ze slagen er niet erg goed in een betrouwbare rangordening aan te brengen, maar het *feit* dat er beoordeeld wordt, roept voor- en zelfselectieprocessen in het leven die een zekere kwaliteitsgarantie bewerkstelligen. Beoordelingen hebben aldus 'afschrikkingspotentieel' (zij het niet op iedereen). Als er niet zou worden beoordeeld, zou naar verwachting het aantal gegadigden hoger en de gemiddelde kwaliteit lager zijn. Daar staat natuurlijk tegenover dat een enkele briljante gegadigde walgt van elke beoordeling, en zich dus alleen meldt als er *niet* wordt beoordeeld. Maar doorgaans

zal dat aantal niet groot genoeg zijn om het kwaliteitsverlagende effect te compenseren.

De tweede conclusie is van heel andere aard. Ruwweg gezegd: beoordelaars zijn betrekkelijk goed in het scheiden van evident kaf van het koren. Ze blijken nogal slecht in het maken van gradueel onderscheid boven die drempel. Zo gezien is de meest aanvaardbare beoordeling een *marginale*. Die stelling heeft vérstrekkende beleidsconsequenties. Het bijpassende systeem zou namelijk een openeindbestel zijn in plaats van een vast budget of een numerus fixus: iedere als serieus beoordeelde aanmelding voor een plaats, subsidie, beurs en dergelijke zou moeten worden gehonoreerd. Aan dit 'technische' argument kan iets principiëlers worden toegevoegd, namelijk dat iemand die zich kwalificeert, recht kan doen gelden op een voorziening. Dat recht is geen kwestie van meer of minder, maar van al dan niet.

*Beoordelaars zijn betrekkelijk goed in het scheiden van evident kaf van het koren*

Aan de orde is hier het fundamentele onderscheid tussen systemen van vergelijkende selectie en die van honorering. Selectie doet zich bijvoorbeeld voor bij vervulling van een vacature: de best beoordeelde kandidaat krijgt de baan, ongeacht het feit dat ook anderen geschikt waren. Het achterliggende principe ligt in de bedrijfseconomische sfeer: de selecterende instantie maximaliseert via de selectie een bepaald nut. Honorering daarentegen heeft betrekking op verworven rechten van het individu, dat zich heeft gekwalificeerd voor bijvoorbeeld een opleiding. In de praktijk lopen zulke principes natuurlijk vaak door elkaar. Bij de toelating tot de medicijnenstudie bijvoorbeeld kan men redeneren dat de BV Nederland (of de ondernemende universiteit) maximaal rendement uit haar plaatsen moet halen, wat neerkomt op vergelijkende selectie; of men kan vinden dat alle gekwalificeerden gelijke rechten hebben, zodat de eventuele inperking daarvan moet geschieden zonder aanzien des persoons, dus via wachtlijst of loting. De verschillende varianten van gewogen loting vormen praktisch-politieke compromissen tussen de twee principes.

In het verband van kunstbeleid valt te overwegen dat de bedrijfseconomische invalshoek voornamelijk als een tang op een varken slaat. Uiteraard blijft het mogelijk alles op die noemer te brengen: hoe mooier de kunst, des te hoger het bruto nationaal product. Maar zelfs in deze commerciële tijd zal die vorm van beleidsdenken door de meeste mensen als barbarij worden beschouwd. De logische consequentie trouwens, die hier en daar ook wordt getrokken, zou eerder liggen in de afschaffing van subsidies, beurzen en dergelijke. Gegeven echter het voortbestaan daarvan, ligt de andere interpretatie meer voor de hand: de overheid stelt talent in staat zich verder te ontplooien, uit overwegingen van beschaving. Bij die honoreringsgedachte past marginale beoordeling en een flexibel budget.

Ter voorkoming van misverstanden: ook bij marginale beoordeling worden fouten gemaakt, namelijk iets niet honoreren dat wel kwaliteit heeft en iets honoreren dat onvoldoende kwaliteit heeft. Ook daar blijft het nodig met een voldoende aantal beoordelaars te werken om een redelijke mate van intersubjectiviteit te bereiken. Maar marginale beoordeling levert in totaal minder fouten op dan vergelijkende selectie.[2]

Een aantal kanttekeningen hierbij. De eerste is dat marginale beoordeling niet tot een overvloedig beroep op belastinggelden hoeft te leiden: dat hangt af van de onafhankelijkheid van het beoordelingsproces. De juryleden in het onderzoek van Top kenden gemiddeld een 6+ toe aan de eindexaminandi, dus daar moeten nogal wat onvoldoenden tussen hebben gezeten.

2. Zie bijvoorbeeld Hofstee (1999) pp. 36 e.v.; dit leerstuk uit de beoordelingstheorie kent zijn nuances en vertakkingen, maar is robuust genoeg om op te varen.

De tweede kanttekening heeft betrekking op de situatie waarin moet worden gewerkt met een vast budget dat niet toereikend is om alle serieuze claims te honoreren. In zo'n geval is het ongepast zulke gegadigden te berichten dat hun claim is 'afgewezen', of in het algemeen, dezelfde terminologie te gebruiken als bij ongerechtvaardigde claims. De bedoeling van beoordelingen is onder andere dat de beoordeelde er wijzer van wordt, al was het maar omdat de beoordeelde van vandaag de beoordelaar van morgen is; denigrerende feedback is in dat opzicht niet passend. Men zou, naar waarheid, kunnen berichten dat de fondsen niet toereikend waren en dat de commissie, in het besef van haar feilbaarheid, een prioritering heeft moeten aanbrengen in de aanvragen die in beginsel voor honorering in aanmerking kwamen. (Het alternatief van zuivere loting tussen de positief beoordeelde aanvragen is weliswaar in abstracto verdedigbaar, maar is voor weinig betrokkenen aanvaardbaar.)

Mijn betoog levert geen legitimering voor een rangordening van positief beoordeelde claims, en geen aanwijzingen voor het opstellen van zo'n rangorde. In de praktijk levert het middelen van de oordelen van de commissieleden vanzelf een rangorde op, die vervolgens in de discussie kan worden bijgesteld. De nadruk ligt hier op de feilbaarheid, en zelfs de oneigenlijkheid van rangordening.

### Onafhankelijk beoordelen

Hierboven werd gerefereerd aan de onafhankelijkheid van het beoordelingsproces. De vraag dient zich aan hoe ver die onafhankelijkheid binnen het vigerende kunstbeleid reikt: als kunstbeoordelaars zich zouden begeven in een scenario van sociale belangenbehartiging van hun sector, dan laat zich aanzien dat een open budgetregeling binnen de kortste keren zou exploderen. 'Beoordeling' is vast gekoppeld aan 'onafhankelijkheid', maar die onafhankelijkheid moet vooral begrepen worden als een pantser tegen belangen en voorkeuren. En dat zijn er nogal wat.

*'Beoordeling' is vast gekoppeld aan 'onafhankelijkheid', een pantser tegen belangen en voorkeuren*

In de eerste plaats heeft de gegadigde een persoonlijk belang bij een positieve beoordeling. Gegadigden kunnen hun belang bepleiten, al dan niet letterlijk proberen de beoordelaar om te kopen, en dergelijke. In de tweede plaats kunnen derden proberen invloed op de beoordelaar uit te oefenen. Zo kan een overheid een emancipatiebeleid ten gunste van bepaalde groepen gegadigden voeren dat alleszins legitiem kan zijn maar niet met beoordeling moet worden verward. (Het argument is natuurlijk ingewikkelder, want de beoordelaars kunnen zelf tot de conclusie komen dat hun maatstaven aan seksisme of etnocentrisme onderhevig waren.) In de derde plaats dient de beoordelaar zich onafhankelijk ten opzichte van medebeoordelaars op te stellen: men mag zich wel door elkaar laten overtuigen, maar niet laten beïnvloeden. Last but not least dient de beoordelaar te abstraheren van persoonlijke belangen én voorkeuren. Vooral dat laatste is subtiel, want enerzijds kan de beoordelaar moeilijk ergens anders op varen dan op de eigen ontwikkelde smaak, maar anderzijds is beoordelen een 'ambt', hetgeen refereert aan algemene maatstaven. Ervaren beoordelaars hebben er trouwens in de praktijk niet veel moeite mee een hoog kwaliteitsoordeel toe te kennen aan iets wat ze zelf verfoeien.

In de tweede plaats is er de koppeling tussen 'beoordeling' en 'in commissie', misschien wel de belangrijkste garantie voor onafhankelijkheid. Functioneren in een com-

missie belichaamt 'commitment' aan het vinden van intersubjectieve maatstaven. Commissies worden ingesteld in het vertrouwen dat hun oordelen kunnen worden verantwoord, aangezien die tot stand zijn gekomen in 'herrschaftsfreie Dialog': argumenten die daarin houdbaar zijn bevonden, vormen een geschikte basis voor zo'n verantwoording. Uiteraard kan het instrument van de commissie worden misbruikt, en kan een commissie intern slecht functioneren. Maar deelname aan commissiewerk betekent normaliter dat men zich voegt naar het appèl dat daarvan uitgaat. Beoordelen is in veel opzichten een paradoxale en onmogelijke opdracht, maar in de commissie steunt de lamme de blinde.

Voor een goed begrip moet worden aangetekend dat de winst van de onderlinge discussie vrijwel uitsluitend procedureel is; de beoordelingen zelf worden er inhoudelijk niet beter door. Als men van twee jury's de individuele oordelen-zonder-discussie zou middelen, zou men tussen die gemiddelde oordelen meer overeenstemming vinden dan tussen de eindoordelen-na-discussie. Dit komt doordat in de discussie de beoordelingen onderling gecontamineerd raken, zodat de wet van de grote aantallen (waardoor stokpaardjes tegen elkaar worden uitgespeeld) zijn werk niet meer volledig kan doen. De tevredenheid van de commissieleden met een bereikte consensus wordt dus in het algemeen niet gerechtvaardigd door verbetering van het oordeel: die verbetering is illusoir. De discussie heeft een heel andere, meer preventieve functie: ze houdt de beoordelaars bij de les en bij het beoordelingsscript. Het is een gezamenlijke investering in de onafhankelijkheid van de beoordeling.

De derde begripskoppeling bestaat tussen 'beoordelen' en 'op eigen merites': men zou dat als onafhankelijkheid ten opzichte van maatstaven in het verleden kunnen beschouwen. Een beoordeling is geen meting, al willen omringende bureaucraten en autoritaire beoordelaars zelf dat nogal eens zo doen voorkomen. De maatstaven waartegen wordt beoordeeld zijn 'dynamisch': ze staan onder invloed van hetgeen de beoordeelde ter tafel brengt. Dat houdt niet in dat iedere gegadigde een eigen recht zou hebben om bijpassende maatstaven te kiezen (waar dat toe zou leiden, wordt geïllustreerd door de eindexaminandi in het onderzoek van Top, die hun eigen werk gemiddeld zo'n anderhalve punt hoger waardeerden dan de jury). Het houdt wel in dat beoordelen onvermijdelijk een creatief aspect heeft. De beoordeelden hebben de toekomst, althans in beginsel. Ze zullen de grenzen van het vak verleggen, en daarmee de maatstaven. In veel gevallen komt daar weliswaar niets van terecht; net als vorige generaties zullen ze voornamelijk doodlopende wegen inslaan, oude wijn in nieuwe zakken gieten, en meedrijven op de golven van kortstondige modes. Maar een enkeling, in een begenadigd moment, zal tot dan toe onzichtbare barrières doorbreken en iets regelrecht uit de hemel plukken dat het hele ondermaanse circus van scholing en beoordeling de moeite waard maakt. Die mogelijkheid verlangt een open geest van de beoordelaar.

Geen beoordelaar van vlees en bloed kan pretenderen een helder zicht te hebben op zulke toekomstige transformaties. Maar zoals het de kunst van de rij-instructeur is *niet* in te grijpen, bestaat de missie van de beoordelaar uit het openhouden van de toekomst. Beoordelen op eigen merites is aldus meer een attitude dan een techniek. Het is de volgehouden poging zich te verplaatsen in wat de nieuwkomer bezielt, en alleen af te haken als die speurtocht louter leegte oplevert. Zoals Wim Kans militaire medewerker iemand is die militair niet tegenwerkt, jaagt de beoordelaar op een bescheiden resultaat: het ware talent althans geen strobreed in de weg te hebben gelegd.

Beoordelen gebeurt idealiter onafhankelijk, in commissie en op eigen merites. Een samenvattend begrip is 'professionaliteit'. Die term herbergt accenten op streven naar objectiviteit, collegiale verantwoording, en een oriëntatie op het unieke en individuele. Hij sluit belangenbehartiging, eigengereidheid en bureaucratie uit. Bovendien heeft een professional een cliënt; de professional is er zelfs primair voor de cliënt. Misschien is dat wel een behartigenswaardige implicatie: de beoordelaar is er voor de beoordeelde (al voelt die dat niet altijd zo aan).

### Beleid en beoordeling: structurele spanning

Zijn er manieren om beoordelaars in hun schier onmogelijke taak te ondersteunen, anders dan via de koninklijke weg, namelijk het opereren in commissieverband? Het antwoord is zoiets als ja, maar niet dan na lange aarzeling. Aan de orde zijn beoordelingsformulieren, objectieve indicatoren voor kwaliteit, en weging van beoordelaars. In alle drie gevallen moet men ermee oppassen.

Beoordelingsondersteuning is in de praktijk een eufemisme voor bemoeienis van de beleidsmaker waaraan de beoordelingscommissie rapporteert en die de uiteindelijke beslissing neemt. Tussen beleid en beoordelaar bestaat structureel een spanningsverhouding. De beleidsmaker heeft geen verstand van kwaliteit, en is dus voor de beoordeling daarvan aangewezen op de professionele beoordelaar. Diens beoordelingen hebben de status van advies. Het beleid kan die adviezen naast zich neerleggen uit overwegingen van bestuurlijke of politieke aard; men kan bijvoorbeeld een contingentering hanteren die de kwaliteitsrangorde doorbreekt. Maar zulke ingrepen zijn riskant, zowel uit het oogpunt van externe legitimering van de uitkomst als met het oog op handhaving van goede betrekkingen met de beoordelaars. De beleidsmaker zal er dus de voorkeur aan geven het beoordelingsproces zelf te reguleren, zodat aan haar randvoorwaarden tegemoet wordt gekomen maar de uitkomst niettemin in termen van kwaliteit kan worden gelegitimeerd.

Er kunnen ook andere motieven spelen. Beleidsmakers zijn gewend hun handelen in bureaucratische termen te verantwoorden. Beoordelingen geven hun in dat opzicht weinig houvast; vanuit hun denkwijze hebben ze geen weerwoord als gegadigden of derden zich beklagen over de ondoorzichtigheid van het beoordelingsproces. Dat lukt beter wanneer ze erin slagen dat proces zo veel mogelijk te protocolleren. Zelfs wanneer die protocollering van louter rituele aard is en geen enkele invloed op de beoordeling (of in andere contexten, de behandeling) uitoefent, kan de beleidsmaker er een zaak bij de bestuursrechter mee winnen. Daar moeten beoordelaars en behandelaars dan maar het uitvoeren van wat danspassen, meestal bestaande uit het invullen van lijvige formulieren, voor over hebben. Het ritueel kost tijd en leidt tot lichte gevoelens van vervreemding, maar het is in zoverre onschuldig dat de beoordeling of behandeling er niet in de kern door wordt aangetast.

Beoordelings*formulieren* bestaan uit opsommingen van onderscheiden aspecten van kwaliteit. In het onderzoek van Top waren dat er maar twee, namelijk originaliteit en techniek (naast kwaliteit in het algemeen); uit haar resultaten blijkt dat de beoordelaars daar ook inderdaad een onderscheid tussen maakten.[3] Vaak bestaat de neiging de formulieren veel verder te specificeren. Dat heeft als voordeel dat beoordelaars worden

3. Voor de liefhebbers: de correlatie tussen Originaliteit en Techniek was .54. Met Kwaliteit correleerde Originaliteit .74, en Techniek .85.

geconfronteerd met mogelijke aspecten van kwaliteit die ze anders misschien over het hoofd zouden hebben gezien. De risico's zijn gelegen in de suggestie die van protocollering uitgaat: in de eerste plaats dat 'this is all there is', hetgeen strijdt met het beginsel van beoordeling op eigen merites; in de tweede plaats dat alle aspecten bij elke beoordeelde van belang zouden zijn; in de derde plaats dat getrouwe invulling van het formulier tot een objectief oordeel zou leiden. Doorgewinterde beoordelaars plegen korte metten met formulieren te maken: ze nemen er kennis van, gebruiken ze als geheugensteun, en geven vervolgens een globaal en ongespecificeerd kwaliteitsoordeel.

*Indicatoren* zijn veel gevaarlijker dan formulieren. Het zijn kwantiteiten met betrekking tot de aanvrager of diens werk; van die kwantiteiten is aangetoond (of wordt beredeneerd) dat ze in het algemene geval voorspellende waarde hebben voor toekomstige kwaliteit. Zoals er in de wetenschap een – weliswaar geenszins perfect, maar wel positief – verband bestaat tussen het aantal geschreven pagina's en de kwaliteit ervan, zal datzelfde gelden voor het aantal doeken of composities van een scheppend kunstenaar, aangezien de mechanismen die aan dat verband ten grondslag liggen vergelijkbaar zijn. De waarde die een indicator aanneemt in een individueel geval kan objectief worden vastgesteld. Door combinatie van een aantal indicatoren kan men achteraf vaak een – nog steeds geenszins perfecte, maar wel redelijk goede – voorspelling leveren van de later gebleken kwaliteit. Vervolgens heeft de toepassing van zulke formules bij nieuwe gevallen twee verleidelijke voordelen: ze geven de beoordelaars een zeker houvast bij hun hachelijke taak en ze vergroten de transparantie van het oordeel tegenover de gegadigden, en daarmee hun rechtszekerheid. Dat laatste is de reden waarom beleidsmakers houden van indicatoren.

Maar intussen zijn de gevaren van indicatoren levensgroot. Het eerste gevaar is dat de beoordelaar erdoor in slaap wordt gesust. Indicatoren zijn per definitie deficiënt: ze zijn niet dekkend en ze zijn niet toegesneden op individuele merites. Ze meten geen kwaliteit, maar iets wat er naast heeft gelegen. Beoordelen is een actief zoekproces dat grote alertheid vraagt; overgave aan indicatorwaarden maakt de beoordeling zelf deficiënt. Nu is dit nog de minste van de twee gevaren, want professionele beoordelaars plegen indicatoren met gepaste achterdocht tegemoet te treden. Men hoort hen wel zeggen dat ze er alleen aandacht aan schenken voorzover ze hun oordeel bevestigen, en beredeneerd kan worden dat dit precies de juiste attitude is (Hofstee 1999, 166ff). Maar dat alles neemt niet weg dat de afbreukrisico's voor de oordeelsvorming levensgroot zijn. Men kan indicatoren eventueel hanteren bij wijze van randconditie, bijvoorbeeld: het bezit van een relevant diploma als voorwaarde voor ontvankelijkheid van een beursaanvraag. Dat is weliswaar arbitrair, maar bestuurlijk eventueel verdedigbaar, en de indicator wordt aldus buiten de lijnen van het beoordelingsproces geparkeerd.

***Indicatoren meten geen kwaliteit, maar iets wat er naast heeft gelegen***

Nog bedreigender voor de kwaliteit is de '*incompatibiliteit*' van bepaalde indicatoren, namelijk die waarvan de gegadigden de waarde zonder veel moeite naar hun hand kunnen zetten. Incompatibiliteit betekent in dit verband onverenigbaarheid van de doelstelling van de beoordeling (stimuleren van kwaliteit) en de uitnodigende werking die de indicator heeft voor het strategische gedrag van de gegadigde. Aansluitend bij het eerder gegeven voorbeeld: als retrospectief zou zijn geconstateerd dat de aantallen com-

posities indicatief zijn voor de kwaliteit ervan, en vervolgens die indicator in het beoordelingssysteem wordt opgenomen, kunnen beginnende componisten hun kansen op honorering van een aanvraag vergroten door, al dan niet met behulp van een computerprogramma, hun 'productie' op te voeren. Niet alleen wordt aldus de beoordeling gecorrumpeerd, maar het creatieve proces zelf raakt aan pervertering onderhevig. Niet zelden vindt een dergelijke verloedering plaats tot innige tevredenheid van de beleidsmaker in kwestie; immers, die constateert dat in reactie op het 'beleid' de 'productie' is toegenomen, en 'dus' de kwaliteit. Bij degenen die aan zo'n regiem onderhevig zijn, is het voorspelbare effect een houding van cynisme. In sectoren zoals hoger onderwijs en gezondheidszorg kunnen de daarmee gepaard gaande burn-outeffecten, veroorzaakt door toepassing van prestatie-indicatoren en benchmarking, ook inderdaad ruimschoots worden aangetroffen.

*Weging* van beoordelaars, ten slotte, is eerder een hobby van methodologen zoals schrijver dezes (Hofstee 1999, 73ff). In rudimentaire vorm vindt weging plaats bij bijvoorbeeld atletiekjury's, door het schrappen van de twee meest extreme ratings. Dat systeem kan zodanig worden verfijnd dat weging de gemeenschappelijke component in de oordelen, en daarmee de betrouwbaarheid maximaliseert. Het algemene resultaat is dat beoordelaars meer gewicht krijgen naarmate ze meer met de anderen overeenstemmen. Het voor de hand liggende risico is dat de beoordelaar in conformerende zin gaat anticiperen op andermans beoordelingen. In hoeverre dat effect optreedt, hangt overigens af van de overeenstemmingsmaat die men kiest, en van andere procedurele details. Het mechanisme is ook niet bij voorbaat verwerpelijk, aangezien de beoordelaars best mogen worden uitgenodigd zich op een intersubjectief standpunt te stellen. In hoeverre de kwaliteit van de beoordeling erdoor verbetert (of verslechtert) is echter nog onvoldoende onderzocht.

## Besluit

Beoordelaars zijn feilbaar, vaak meer dan ze zelf denken. Professionele beoordelaars zijn niettemin onmisbaar: pogingen om de beoordelaar te vervangen of zelfs maar te ondersteunen door objectieve procedures zullen doorgaans averechtse effecten sorteren. De meest prudente vorm van inschakeling van beoordelaars is marginale beoordeling, gericht op het weren van wankwaliteit. De bijpassende vorm van beleid bestaat uit openeindregelingen. Een vooronderstelling daarbij is dat het beoordelingsproces aan eisen van onafhankelijkheid voldoet. Garanties daarvoor, en voor een mate van beoordelingsbetrouwbaarheid, moeten worden gezocht in het beoordelen in commissie.

## Naschrift

Impliciet in dit betoog is de stelling dat beoordelen van kunst en van wetenschap niet wezenlijk verschillen (al heb ik niet de behoefte om die twee in algemene zin over een kam te scheren). Op die basis zou men zich kunnen verbazen over de samenstelling van panels in de wereld van de kunsten. In de zuivere wetenschap ontlenen beoordelaars hun gezag aan het feit dat ze zelf gevestigde en erkende wetenschapsbeoefenaars zijn; wetenschapshistorici, -filosofen of -journalisten zijn in zulke panels zelden welkom (tenzij het natuurlijk hun eigen discipline betreft). De vraag dringt zich op of

kunsthistorici en -critici, galeriehouders en anderen die zelf geen kunstenaar, schrijver, componist en dergelijke zijn, geroepen zijn scheppende kunstenaars te beoordelen. Uitgaande van het beginsel dat de beoordelaar zich dient te verplaatsen in de gegadigde, op basis van een verwantschap die berust op eigen ervaring, en daaraan het gezag ontleent waarop het vertrouwen van die gegadigde moet berusten, lijkt het antwoord ontkennend te moeten luiden.

**Auteur**

***Willem Hofstee*** is emeritus hoogleraar psychologie aan de Rijksuniversiteit Groningen. Hij heeft zich gespecialiseerd in beoordelingsprocessen.

Dit artikel is een bewerking van de lezing, gehouden op 11 juni 2002 in de Voordrachtzaal van de Boekmanstichting, gewijd aan de beoordeling van kunst. Het integrale verslag van die bijeenkomst wordt samengesteld door de Wiardi Beckmanstichting en is daar te bestellen (020-5512155).

**Literatuur**

Hofstee, W.K.B. (1999) *Principes van beoordeling: methodiek en ethiek van selectie, examinering en evaluatie*. Lisse: Swets.

Top, T.J. (1993) *Art and Gender: Creative Achievement in the Visual Arts*. Dissertatie, Rijksuniversiteit Groningen.

***Abstract***

*The quality of assessments within the context of cultural policy*

Assessors who work with the funding institutions play an important role in current art policy. This justifies the extra attention being paid to how such evaluations are arrived at. Previous studies (Top 1993) have shown that there is little consensus on quality between the experts who are declared as being authorised to assess. However, wherever there is a lack of consensus, this usually involves minor differences, which people try to classify.

Generally, assessment committees strongly agree about quality and non-quality. Consequently, their primary task is to separate the wheat from the chaff. Furthermore, their presence has a purely preventive effect: assessment committees initiate the processes of pre-selection and self-selection even before the first opinion is given.

The fact that assessors need to work within bureaucratic policy frameworks may give rise to tension. Policy officials try to assist the assessors in any way they can: with forms, indicators, and by assessing the assessors. Such assistance does not always have a favourable outcome.

It can be concluded that assessors are the most effective way of avoiding poor quality. This means that they should do no more than the marginal testing of quality and professionalism. This, however, would imply the introduction of an open-ended scheme, i.e. a budget that is not limited beforehand.