Evaluatie van onderzoeksprestaties

Universiteiten, overheidsorganen en economische literatuur hechten steeds meer belang aan het evalueren van academici. Zowel het onderwijs als het onderzoek van individuele professoren en onderzoeksgroepen worden veelvuldig vergeleken. Het aggregeren en vergelijken van diverse onderzoekscriteria is door specialisatie en verschillende achtergrondfactoren echter een heikel punt.

De meeste universiteitsgebonden, nationale of internationale studies op het gebied van wetenschapsbeleid en onderzoeksevaluatie hanteren omwille van de eenvoud eendimensionale evaluatiecijfers. Denk bijvoorbeeld aan een waardering van de onderzoeksreputatie via een enquête of een *peer review* (Groot, 2003), het aantal publicaties in erkende wetenschappelijke tijdschriften (Sax *et al.*, 2002) of het aantal citaties van het wetenschappelijk werk (Lee en Bozeman, 2005). Zo ook de ESB Top 40 (Jolink, 2007; 2008) die een relatieve rangschikking geeft van de publicatieprestaties van een selectie van economen aan Nederlandse universiteiten. De beperkte uitzonderingen die meer dan één dimensie opnemen, komen niet veel verder dan twee dimensies, zoals publicaties en citaties bij de h-index (Hirsch, 2005). De betrouwbaarheid en de geldigheid van dergelijke indicatoren is beperkt. Wetenschappelijk onderzoek is immers te complex en te veelzijdig om door eendimensionale evaluatiecijfers beoordeeld te worden. Avital en Collopy (2001) en Hattie en Marsh (2002) argumenteren dan ook dat een betrouwbare en representatieve evaluatie van wetenschappelijk onderzoek vraagt om het samenbrengen van wat academici op de verschillende heterogene onderzoekscriteria hebben gerealiseerd. Als meerdere dimensies gewogen worden, moeten onderzoekers daarenboven ook breder presteren. Beide artikelen benadrukken daarnaast het belang om de prestatiebeoordelingen voor invloedrijke karakteristieken en ook de werkomgeving van de individuele onderzoeker te corrigeren; zulke beoordelingen zijn niet of moeilijk controleerbaar voor de onderzoeker.

Het samenbrengen van onderzoeksprestaties op verschillende outputcriteria in één evaluatiecijfer brengt echter twee belangrijke conceptuele problemen met zich mee. Ten eerste is er het wegen en aggregeren van de prestaties van de onderzoekers op de verschillende criteria in één beoordelingscijfer. De weinige studies (Kyvik, 1990) die meerdere outputcriteria betrekken in onderzoeksevaluaties, passen gewoonlijk gelijke of vaste uniforme gewichten toe bij het aggregeren van de onderzoeksoutput. De vraag is echter of een dergelijke gewichtenkeuze verdedigbaar is. Een keuze voor gelijke gewichten impliceert namelijk een gelijk belang voor alle onderzoekscriteria. Zo zullen onder dergelijke gewichtenschema's de publicaties in internationale en aan een peer review onderworpen vakbladen even zwaar wegen als niet-gepubliceerde onderzoekspapers. Anderzijds, het gebruik van uniforme gewichten impliceert dat alle onderzoekers eenzelfde mening hebben over de impact van de onderzoekscriteria. Gezien de variëteit in de persoonlijkheden en opinies van de onderzoekers lijkt dit een moeilijk verdedigbare keuze. Massy en Wilger (1995) stellen dan ook dat enige variatie in de gewichtenkeuze, zowel met betrekking tot de criteria als tot de onderzoekers, noodzakelijk is voor een correcte onderzoeksbeoordeling.

Een tweede belangrijk probleem betreft het corrigeren van de onderzoeksbeoordelingen voor exogene factoren die op korte termijn niet controleerbaar zijn voor de onderzoeker. De academische literatuur (Hattie en Marsh, 2002; Sax *et al.*, 2002) toont aan dat dergelijke factoren een significante invloed kunnen hebben op de door de onderzoeker geleverde prestaties. Bovendien is er ook de ervaring van de onderzoekers zelf, die leert dat onderzoeksomstandigheden niet alleen een voordelige maar ook een nadelige invloed kunnen hebben op de productiviteit van onderzoekers (Lee en Bozeman, 2005). Traditionele onderzoeksevaluaties houden meestal geen rekening met de invloed van dergelijke factoren. Ze zijn dan ook vaak vertekend in het voordeel van de onderzoekers met de meest gunstige kenmerken of die werken in de beste omstandigheden. Voor hen is het namelijk gemakkelijker om goede evaluaties te realiseren. Het tegenovergestelde geldt dan weer voor de onderzoekers die werken onder minder gunstige karakteristieken en condities. Voor hen is het moeilijker om een goede beoordeling te krijgen.

Een endogene niet-parametrische weging

Om de gewichten van de verschillende onderzoekscriteria te bepalen bij het aggregeren tot één indicator, wordt een niet-parametrische wegingsmethode voorgesteld. Deze techniek bevat vooraf nauwelijks of geen veronderstellingen met betrekking tot hoe de gewichten er moeten uitzien, dit in tegenstelling tot de parametrische methoden. Bijgevolg hangen de uiteindelijke evaluatiescores niet af van gemaakte veronderstellingen, iets wat een voordeel is gezien de onduidelijkheid over de exacte impact van de onderzoekscriteria. De voorgestelde methode is sterk

KRISTOF DE WITTE
Postdoc aan de Universiteit Maastricht en onderzoeker aan de Katholieke Universiteit Leuven

NICKY ROGGE
Onderzoeker aan de Hogeschool-Universiteit Brussel

gelieerd aan het DEA-model (*Data Envelopment Analysis*): een efficiëntiemetingstechniek waarbij voor elke observatie wordt nagegaan in welke mate deze erin slaagt om vaak meerdere inputs of gebruikte middelen om te zetten in meerdere outputs of geproduceerde uitkomsten (Groot, 2003). Hierbij wordt de efficiëntiescore van iedere observatie bepaald ten opzichte van zijn benchmark-observatie. Een observatie wordt pas efficiënt bevonden wanneer zij erin slaagt om met het huidige niveau aan inputs de hoogst mogelijke output te generen. In dat geval zal de geëvalueerde observatie haar eigen referentie-observatie zijn en dus als efficiënt bestempeld worden. Meer bepaald betreft het een variant op het DEA-model, namelijk de BoD-methode (*Benefit of the Doubt*) (Melyn en Moesen, 1991). Deze methode verschilt van de DEA-techniek doordat de focus uitsluitend op de outputzijde ligt, namelijk het wegen en aggregeren van verschillende outputs. De kerngedachte hierbij is dat informatie over de gewichten afgeleid kan worden van de outputdata zelf. De BoD-methode laat in het bijzonder toe dat iedere onderzoeker de onderzoekscriteria waarin hij beter presteert ten opzichte van zijn collega's zwaarder laat doorwegen in zijn beoordeling. Criteria waarin hij relatief zwak presteert, mogen dan weer minder zwaar doorwegen. Om te vermijden dat onderzoekers bepaalde criteria zouden negeren door het toekennen van een nulgewicht of dat ze aan bepaalde criteria een te zwaar onrealistisch gewicht zouden toekennen, kunnen experts de vrijheid in de non-parametrische gewichtenkeuze beperken (Cherchye *et al.*, 2007; De Witte en Rogge, 2009b). De achterliggende intuïtie van de BOD-methode is dat relatief sterke prestaties kunnen worden gezien als indicaties dat de geëvalueerde onderzoeker de specifieke criteria als belangrijk beschouwt. De BoD-methode zoekt dus binnen de gewichtsrestricties voor iedere onderzoeker naar de optimale gewichten die resulteren in een maximale onderzoeksbeoordeling ten opzichte van de collega-onderzoekers. De op BoD gebaseerde beoordelingsscores kunnen variëren tussen 0 en 1, waarbij een hogere score een betere algemene onderzoeksprestatie aangeeft. Gezien de mogelijk significante invloed van externe karakteristieken en werkomstandigheden, zoals van geslacht, leeftijd en percentage onderzoekstijd, is het van belang om de prestatie-evaluaties van onderzoekers te corrigeren voor exogene invloeden. Bovendien worden de onderzoeksbeoordelingen ook robuust gemaakt tegen eventuele onnauwkeurigheden in de data of de impact van extreme observaties. Zo is er een onderzoeker die 27 sterk gelijkaardige rapporten schreef die apart gevaloriseerd worden. Er wordt daarom een variant van de BoD-methode voorgesteld die als basisidee hanteert om een onderzoeker alleen te vergelijken met collega-onderzoekers die over een soortgelijke mix van karakteristieken en werkomstandigheden beschikken (De Witte en Kortelainen, 2008; De Witte en Rogge, 2009a; 2009b).

De data

Er zijn prestatiebeoordelingen opgesteld voor de 73 onderzoekers aan het departement Economie en Management aan de Hogeschool-Universiteit Brussel (HUB) voor de periode 2006–2008. In een BoD-model volstaat 73 observaties ruimschoots voor een adequate analyse. De dataset omvat gegevens verzameld via drie bronnen: de officiële onderzoeksgegevens, administratieve data en enquêtedata (Hattie en Marsh, 2002). De officiële prestatiegegevens omvatten de realisaties van de onderzoekers in een selectie van negen outputcriteria. Deze negen outputcriteria en hun maximale impact in de beoordeling werden vastgelegd door de onderzoeksraad van het departement van de HUB. De onderzoeksraad bestaat uit het departementsbestuur en de onderzoekers. De criteria worden weergegeven in tabel 1. De administratieve gegevens en de enquêtedata omvatten informatie over de exogene individuele karakteristieken van de onderzoeker, factoren die werkomstandigheden typeren en de motivaties van de onderzoeker aan de hand van vijf enquêtevragen die zijn beoordeeld op een vijfpunts-Likertschaal die meer punten toekent naarmate de onderzoeker meer instemt met de vraagstelling (tabel 2).

De onderzoeksbeoordelingen

Een eerste belangrijke bevinding op basis van de onderzoeksdata en de gecorrigeerde BoD-evaluatiescores is de grote ongelijkheid in de onderzoeksoutput. Bijvoorbeeld, 48 van de 81 onderzoekers slaagden er niet in om in de periode 2006–2008 een internationale publicatie te verzorgen in een tijdschrift dat op de Thomson Master lijst staat. Bovendien, 60 van de 118 internationale publicaties in tijdschriften op deze lijst zijn het werk van slechts acht onderzoekers, ongeveer tien procent van het aantal onderzoekers aan de HUB. Soortgelijke opmerkingen gelden ook voor de acht andere onderzoekscriteria. Deze bevinding suggereert dat er een aanzienlijke heterogeniteit bestaat tussen de onderzoekers aan de HUB. Merk op dat er tien onderzoekers zijn die er niet in slaagden om ook maar één publicatie te realiseren, ongeacht het criterium. Bijgevolg krijgen zij een gecorrigeerde BoD-prestatiescore gelijk aan 0, de laagst mogelijke score. De grote ongelijkheid in de productiviteit van onderzoekers is typerend voor academisch onderzoek, ongeacht de discipline (Ramsden, 1994). De gemiddelde BoD-evaluatiescore van 0,693 toont aan dat de gemiddelde HUB-onderzoeker zijn aantal publicaties in de negen outputcriteria proportioneel laat toenemen met ongeveer 31 procent, als hij zo goed zou presteren als de *best practice*. Deze score is aannemelijk en reeds lovenswaardig, gezien de specifieke context van het Vlaamse hoger niet-universitair onderwijs in het algemeen en de positie van de HUB in het bijzonder. Als gevolg van het accrediteringsproces in het Vlaams hogeronderwijs moeten namelijk hogescholen die een academisch masterdiploma willen afleveren ook inspanningen leveren op het vlak van wetenschappelijk onderzoek. Sinds de start in 2004 vergt dit een sterke onderzoeks-

Tabel 1 **Criteria voor de evaluatie van onderzoeksprestaties.**

Onderzoekscriteria	Maximum gewicht
1. Internationale publicatie (Thomson Master List)	15
2. Internationaal boek, op basis van eigen wetenschappelijk werk, als auteur	15
3. Nationaal boek, op basis van eigen wetenschappelijk werk, als auteur	10
4. Voltooid onderzoeksrapport indien extern gefinancierd door opdrachtgever, onderzoek met voldoende looptijd	10
5. Nationale wetenschappelijke tijdschriften, hoofdstuk in internationaal wetenschappelijk boek, volledig artikel in internationale proceedings met peer review	7
6. Promotor van een extern gefinancierd project	5
7. Volledig artikel in nationale proceedings met peer review	7
8. Hoofdstuk in wetenschappelijk nationaal boek	7
9. Research of discussion paper in HUB- of andere reeks	5

Tabel 2

Factoren die onderzoeksprestaties beïnvloeden.

	Model 1		Model 2	
	Impact	p-waarde	Impact	p-waarde
Karakteristieken onderzoeker				
Geslacht (vrouw = 1)	Gunstig*	0,018	Gunstig**	0,000
Doctoraat (= 1)	Gunstig*	0,044	Gunstig*	0,032
Geaffilieerd aan andere universiteit	Gunstig**	0,000	Gunstig**	0,002
Leeftijd	Gunstig	0,906	Gunstig	0,,888
Werkomstandigheden				
Aanstellingspercentage	Gunstig**	0,001	Gunstig	0,992
Voorbehouden onderzoekstijd	Gunstig*	0,035	Ongunstig	0,380
Voorbehouden tijd lesgeven	Ongunstig	0,366	-	-
Motivatiefactoren				
Vraag 1: Onderzoek geeft me voldoening			Gunstig**	0,004
Vraag 2: Tijd is een belangrijke beperking voor mijn onderzoek			Gunstig*	0,038
Vraag 3: Een loonstijging zou me aanmoedigen tot meer onderzoek			Gunstig	0,944
Vraag 4: Boordeling van eigen onderzoekscapaciteiten			Gunstig**	0,000
Vraag 5: Meeste samenwerking is binnen mijn departement			Gunstig	0,422

* Significant op vijfprocent-niveau; ** significant op éénprocent-niveau.

inspanning van de docenten. Het is daarom niet verwonderlijk dat, gemiddeld genomen, er nog een sterke efficiëntiewinst mogelijk is in het BoD-model.

De invloed van de potentieel invloedrijke, externe factoren wordt in twee modellen achterhaald (tabel 2). Model 1 corrigeert de prestatiebeoordelingen van de 73 onderzoekers voor externe, individuele karakteristieken van de onderzoekers. Alle individuele karakteristieken vertonen een significante en positieve invloed op de onderzoeksprestaties. Met andere woorden: gemiddeld gezien presteren vrouwelijke onderzoekers, die een doctoraatsdiploma hebben en geaffilieerd zijn aan andere onderzoeksdepartmenten substantieel beter. Daarenboven suggereert Model 1 dat onderzoekers die voltijds actief zijn aan de HUB en die over meer onderzoekstijd beschikken gemiddeld gezien een significant hogere onderzoeksproductiviteit tonen.

De significante invloed van deze exogene werkomstandigheden verdwijnt echter wanneer de selectie van externe factoren wordt uitgebreid met een aantal karakteristieken die peilen naar de motivaties van de onderzoekers (model 2). De individuele karakteristieken van de onderzoeker blijken, gemiddeld gezien, een positieve en significante invloed te hebben op de productiviteit van de onderzoekers aan de HUB. De resultaten van de vijf motivaties van de onderzoekers tonen aan dat de onderzoekers die meer voldoening halen uit academisch onderzoek ook beter presteren, al werd de richting van de causaliteit niet nader bekeken. De onderzoekers die tijd als een belangrijke beperking zien in hun onderzoek, hebben gemiddeld gezien een hogere onderzoeksproductiviteit. De beoordeling van de onderzoeker over de eigen onderzoekscapaciteiten blijkt positief en significant gerelateerd te zijn aan de uiteindelijke prestatiebeoordeling op basis van het gecorrigeerde BoD-model. Het belang dat de onderzoeker hecht aan een koppeling tussen het loon en de onderzoeksprestaties, en het feit dat de onderzoeker het meest samenwerkt met mensen aan het eigen onderzoeksdepartement, is niet-significant gerelateerd aan de productiviteit van de gemiddelde onderzoeker. Niet alle motivatiefactoren blijken dus een grote impact te hebben.

Besluit

De vooropgestelde BoD-methode biedt een interessant alternatief om de onderzoeksprestaties van academici op verschillende onderzoekscriteria samen te vatten in één evaluatiecijfer, waarbij er een correctie plaatsvindt voor individuele karakteristieken, motivaties en werkomstandigheden. Uit het voorbeeld van de prestatiebeoordelingen opgesteld voor de 73 onderzoekers aan het departement Economie en Management aan de HUB blijkt dat een dergelijke correctie zeker geen overbodigheid is. Zo blijken tal van individuele karakteristieken en motivaties van de onderzoeker een significante impact te hebben op hun prestaties. De methode levert verschillende leerrijke bevindingen op voor de bevoegde instanties. Enerzijds zijn er de cijfermatige en objectieve onderzoeksbeoordelingen die instanties toelaten aanbevelingen te kunnen doen over de toekomstige ontwikkeling van de onderzoekers. Anderzijds zijn er de bevindingen met betrekking tot de invloed van de individuele karakteristieken, motivaties en werkomstandigheden op onderzoeksprestaties, die kunnen helpen bij het opstellen van een beleid dat de onderzoeksproductiviteit ten goede komt.

LITERATUUR

Avital, M. en F. Collopy (2001) Assessing research performance: Implications for selection and motivation. *Sprouts: Working Papers on Information Systems*, 1(14).

Cherchye, L., W. Moesen, N. Rogge, T. van Puyenbroeck (2007) An introduction to 'benefit of the doubt' composite indicators. *Social Indicators Research*, 82(1), 111–145.

Groot, T. (2003) Het beoordelen van onderzoek. *ESB*, 88(4422), 628–631.

Hattie, J. en H. Marsh (2002) The relationship between research productivity and teaching effectiveness: complementarity, antagonistic, or independent constructs? *The Journal of Higher Education*, 73(5), 603–641.

Hirsch, J. (2005) An index to quantify an individual's scientific research output. *PNAS*, 102(46), 16569–16572.

Jolink, A. (2007) Rotterdam regeert in ESB Top 40. *ESB*, 92(4524), 744–745.

Jolink, A. (2008) In de spiegel van de ESB Top 40. *ESB*, 93(4549), 746–748.

Kyvik, S. (1990) Age and scientific productivity. Differences between fields of learning. *Higher Education*, 19(1), 37–55.

Lee, S. en B. Bozeman (2005) The impact of research collaboration on scientific productivity. *Social Studies of Science*, 35(5), 673–702.

Massy, W. en A. Wilger (1995) Improving productivity. *Change*, 27(4), 10–20.

Melyn, W. en W. Moesen (1991) Towards a synthetic indicator of macroeconomic performance: unequal weighting when limited information is available. *Public Economics Research Papers*, 91(17). 1-24.

Ramsden, P. (1994) Describing and explaining research productivity. *Higher Education*, 28(2), 207–226.

Sax, L., L. Hagedorn, M. Arredondo en F. Dicrisi (2002) Faculty research productivity: exploring the role of gender and family-related factors. *Research in Higher Education*, 43(4), 423–446.

Witte, K. de, en N. Rogge (2009a) Accounting for exogenous influences in a benevolent performance evaluation of teachers. *Center for Economic Studies Discussion Papers*, 09(13).

Witte, K. de, en N. Rogge (2009b) To publish or not to publish? On the aggregation and drivers of research performance. *Hogeschool-Universiteit Brussel Onderzoeksartikel*, 2009/15.

Witte, K. de, en M. Kortelainen (2008) Blaming the exogenous environment? Conditional efficiency estimation with continuous and discrete environmental variables. *Center for Economic Studies Discussion Papers*, 08(33).