Available online at www.sciencedirect.com
ScienceDirect
Research Policy 35 (2006) 1362–1376
research policy
www.elsevier.com/locate/respol

Research quality and efficiency An analysis of assessments and management issues in Dutch economics and business research programs

Tom Groot [a,*], Teresa García-Valderrama [b]

[a] *Vrije Universiteit, Faculty of Economics and Business Administration, De Boelelaan 1105 (Room 2E-72), 1081 HV Amsterdam, The Netherlands*
[b] *University of Cádiz, Spain*

Received 31 May 2005; received in revised form 3 July 2006; accepted 7 July 2006
Available online 30 August 2006

Abstract

Assessments of quality and productivity of academic research programs become more and more important in gaining financial support, in hiring and promoting research staff, and in building academic reputation. Most assessments are based on peer review or on bibliometric information. In this paper we analyze both bibliometric data and peer review assessments of 169 research groups in economics, econometrics and business administration. The evaluations are achieved in two independent rounds in 1995 and in 2001, permitting replication of our study.

The purpose of this study is twofold. In the first part we want to see to what degree bibliometric information relates to peer review judgments. The results convey how evaluators weight different output categories in their final overall judgment of academic quality. The results also have practical meaning, since they indicate what the predictive ability of bibliometric data is for future peer review outcomes. In the second part of this study we aim at explaining differences in research output quality and productivity by organizational factors, like size of the research group, composition of staff, sources of research funding and academic discipline. In this part, a composite indicator is used to represent the review committees' overall assessment. The bibliometric data most strongly related to the peer reviews' overall assessment are used to construct data envelopment analyses' efficiency scores as measure of research productivity.

The main conclusions from our study are that the number of publications in international top journals is the best predictor of peer review assessment results. Changes in the classification of bibliometric information, as introduced in the second evaluation round, do not alter this conclusion. Size of the research group appears to be the only permanent characteristic associated with research quality and productivity. Size is positively related to research quality but negatively related to research productivity. Larger groups appear to have the potential to improve quality, but as groups become larger, they also experience problems in maintaining the research productivity of the research team's members. The remaining organizational characteristics appear to be temporarily related to research quality and productivity. In the first evaluation round, research productivity and quality are associated with the discipline variable: research programs in more quantitative areas and characterized by a higher level of paradigm development like econometrics and operations research achieved higher levels of research quality and productivity than programs in more diverse and less quantitative areas like business administration. This relation however is not permanent, since it becomes insignificant in the second evaluation round. Instead, funding relations become more apparent in the second review round. The relative amount of

* Corresponding author. Tel.: +31 20 598 6040; fax: +31 20 598 9870.
E-mail address: tgroot@feweb.vu.nl (T. Groot).

doi:10.1016/j.respol.2006.07.002

national funding in the research group's funding becomes positively related to academic quality, whereas the portion of income from committed research is negatively related to academic quality of the programs' research output. This may have been caused by the increased importance of alternative sources of research funding in the period of the second review.

Keywords: Determinants of research efficiency; Research quality; Non-parametric analysis; Research in economics and business administration; Longitudinal analysis of micro-units of production

1. Introduction

Researchers and research groups are increasingly being evaluated and financed on the quantity and quality of research output produced. In the United States, tenure and promotion decisions are based on publications in high-quality journals (Ballas and Theoharakis, 2003). In Britain, research groups are financed based on a periodic quality assessment in the "Research Assessment Exercise" (Whittington, 1993, 1997). Similar approaches can be found in other European countries as well, like in The Netherlands, Finland and Spain (Loft et al., 2002). In assessing research output, two approaches are mainly used: *peer review* and *bibliometric methods*. Each method has its assets and liabilities. Only a few studies have tried to compare the outcome of peer reviews with bibliometric results and they are inconclusive in their results.

In the past 10 years, Dutch research programs in economics, econometrics and business administration have been obliged to use both methods simultaneously. This provides the unique opportunity to compare both methods. Our first research question is to what extent research assessments based on bibliometric data and on peer review will lead to similar conclusions. Moreover, the use of both peer review and bibliometric methods allow the construction of reasonably reliable measures of research quality and productivity. These measures can be used to answer our second research question, which is aimed at identifying what the characteristics are of research programs that achieve high quality research output and high productivity. We expect to find factors like size, composition of the research team, sources of research funding and academic discipline. Our study follows a similar research approach as Cherchye and vanden Abeele (2005), but tries to extent their research in two ways. We look at both efficiency and quality indicators for each program. The second extension relates to the data we use: our sample includes two rounds of research assessments, enabling replication in order to verify whether the initial assessments produced stable results.

The empirical data used in our study come from a group of 169 research groups, evaluated in two rounds by three different review committees. The two review rounds provided the opportunity to replicate the study. This replication is only a partial replication, because the way in which the bibliometric data are reported have been changed somewhat between the first and second round. The main conclusions from our study are that the number of publications in international top journals is a good predictor of the peer review assessment results. Changes in the way the bibliometric information is presented do not have a major impact on the peer review results. Size of the research group is positively related to research quality but negatively related to research productivity. Larger groups appear to have the potential to improve quality, but as groups become larger, they may also experience problems in maintaining the research productivity of each the team members.

The following section contains a literature review of the different research assessments methods and of characteristics of successful research programs. Section 3 gives a short overview of the research assessment procedures applied to the sample programs. In the following section we try to identify the correspondence between the peer review method applied and the research programs' bibliometric information (the first research question). In Section 5, we identify the characteristics of the research programs achieving high quality and productivity levels (the second research question). In the final paragraph we present our conclusions and provide suggestions for further research.

2. Literature review

2.1. Peer review and bibliometric data

The evaluation of research output is seen to be the major and growing element in the assessment of academics (Parker et al., 1998; Beattie and Goodacre, 2004). Much effort is invested in designing useful and reliable systems for evaluating research output. In this endeavor, two main approaches can be distinguished: the *peer review method* and the *bibliometric method*. The peer review method is based on perceptions of well-informed experts about different quality dimensions of research production. Peer review opens the possibility to include a wide variety of different quality aspects in the final judgment, taking into account the current position of the researchers, specific problems and opportunities they face and the researchers' current and future potential. The peer review judgment is however inherently subjective: the conclusions from peer reviews may be partly dependent upon individual orientations of the evaluators, based on their research interests, experience and knowledge (Nederhof and van Raan, 1987; Brinn et al., 1996). Review results based on peer review committees are therefore also dependent on the committee's composition. Recently, much effort has been invested to make assessments of research performance more objective by developing bibliometric methods.

However, bibliometric methods also have their shortcomings. They are restricted to written output, the evaluation results are influenced by the measurement methods applied and the assessment of acceptable publication numbers is partly dependent upon publication and citation "habits" in different fields and specializations (Nederhof and van Raan, 1993). Most studies in research evaluation deal with measurement issues in bibliometric methods. Counting the *number of publications* is the most simple method, but fails to recognize differences in academic quality (Dwyer, 1994; Zivney et al., 1995; Hasselback et al., 2000). Academic quality may be assessed by using the number of *citations* to the original work (Brown and Gardner, 1985; Brown, 1996). Citation analysis also has its weaknesses. The results are heavily dependent on the (number of) journals included in the analysis (Newman and Cooper, 1993; Ballas and Theoharakis, 2003). Not all citations are expressions of acceptance of ideas: articles may be cited in order to critique them (Croom, 1970). Citation behavior can differ from field to field (Moed et al., 1985). Sometimes citation scores are inflated, because of a bias in favor of popular authors and established journals. Also popular topics and fields with many researchers lead to inflated citation scores. Special cases are review articles and methodological papers: they generate more citations on average (Brown and Gardner, 1985). An alternative approach is to weight the number of published articles using the journal's (adjusted) *impact score* (Dyckman and Zeff, 1984; Morris et al., 1990; Newman and Cooper, 1993; van Fleet et al., 2000). Impact scores (the number of times an article in a specific journal is cited by other journals) suffer from similar distortions as citation scores. Journals can be alternatively ranked based on *perceptual ranking* methods, by which individual specialists attach quality scores to journals (Schroeder et al., 1988; Hull and Wright, 1990; Hall and Ross, 1991; Harzing, 2003). Individual perceptions of journal quality may however be influenced by differences in opinion among evaluators about relevant evaluation criteria. These differences may be partly influenced by biases: only academics included in the review exercise express their opinions: in as far as these opinions are not representative for the whole population, the sample evaluators create a biased opinion because of *response bias*. Some evaluators are inclined to rank those journals higher for which they have served as reviewer or contributor (*self-serving bias*) (Brown and Huefner, 1994). Recent studies show that regional differences, affiliations with journals (being editorial board member and/or author in the journal) and research interests impact on accounting journal rankings (Ballas and Theoharakis, 2003).

There seems to be no optimal bibliometric system capable of objectively assessing academic quality without bias and measurement error. It is also not clear to what degree these methods lead to similar outcomes. In two studies, Nederhof and Van Raan conclude that peer review and bibliometric studies are positively related, complementary and mutually supportive (Nederhof and van Raan, 1987, 1993). Another study shows a significant relationship between impact scores and peer review results but no relationship between peer review results and total number of publications (Rinia et al., 1998).

2.2. *Explaining differences in research quality and productivity*

Another stream in the literature tries to explain differences in research productivity and research quality. Some studies find that research productivity is positively related to research programs' *size* and amount of *external funding*. *Commissioned research* for outside parties is negatively related to research efficiency, whereas scientific research funding by a *scientific research council* seems to be positively associated with research efficiency (Johnes, 1988; Cherchye and vanden Abeele, 2005). We therefore expect the proportion of the research team being full-time PhD-students and the proportion of the research team financed by the national research council to be positively related to research quality and productivity. PhD-students are fully committed to producing high-quality research output, since their academic career is largely dependent on their PhD research output. Also the capacity to attract research funds from the national research council can be seen as a confirmation of the research team's proven quality and productivity. A final explanatory variable is the *discipline* in which the research group is active. The field of organizational studies is seen as divers and is characterized by a fairly low level of paradigm development, particularly as compared to economics and (even more) econometrics and operations research (Kuhn, 1970; Pfeffer, 1993). A well-developed paradigm or shared theoretical structure and methodological approaches may provide a necessary, although not sufficient condition for the systematic advancement of knowledge. Fields with a more highly developed paradigm get more funding (Lodahl and Gordon, 1973), experience a greater effect of academic research productivity on pay (Konrad and Pfeffer, 1990), have shorter review times (Beyer, 1978) and are more frequently referenced by other disciplines (Baron and Hannan, 1993; Pfeffer, 1993). A similar argument was made in The Netherlands at the time of our study (Sterken and Leeflang, 2000). We therefore expect the business administration research groups to be less productive than research groups in economics and (even more) in econometrics and operations research.

3. University research funding and evaluation in The Netherlands

3.1. The national research assessment system

Until 1983, government funding of university research in The Netherlands was directly related to the number of students enrolled (Groot, 1988, 1999). In 1983, a system of *conditional financing of research* was introduced, permitting the Minister to allocate funds for research independently of the number of students enrolled. Financing was conditional on the approval of research programs by faculty committees consisting of experts in the relevant fields (Hazeu, 1983; Timmermans, 1984; Groot and van Helden, 1999). The conditional financing system made resources for research and research programs visible, permitting policy makers and university managers to develop research policies and to evaluate research programs. During the first 10 years however, university research had been mainly evaluated by examining the quality of research program *proposals*. In 1993, a new system was introduced to evaluate the *output* of conditionally financed programs. Under authority of the VSNU (the Association of Universities in the Netherlands) a system of periodic, discipline-oriented evaluation of university research from an international perspective was implemented (VSNU, 1990, 1994a). The objectives of this system were threefold: (1) to enhance the quality of research, (2) to inform faculty managers about the achievements of their research programs and (3) to introduce a system of accountability to the public at large of the spending of state funds (VSNU, 1994b; Westerheijden, 1996).

At least once every 5 years, a review committee assesses the research output in each academic discipline. The committee consists of non-Dutch peers with an outstanding reputation in their academic field. The final report of the committee is available to the public (VSNU, 1995). In 1993, the first disciplines evaluated were mechanical engineering, biology, psychology, history and archaeology.

3.2. The first evaluation

In 1995, a committee of six non-Dutch experts representing the areas of general economics, business economics[1] and quantitative economics/econometrics with outstanding international reputation evaluated the output of 92 research programs in eight universities from the 5-year period 1990 to 1994. Each research program sent a "self study" report to the review committee. In the self study, each program presented its objectives, resources and research results in a common format, prescribed by the VSNU (1994b). Resources were reported in total full-time equivalents research input, subdivided according to the funding source, such as university funding, funding by the national research council, and funding from external parties like the European Union or private enterprises. Resources were also reported according to type of researcher: such as number of full-time equivalents (fte) PhD-students and senior staff members. Research results were reported in number of dissertations, number of scientific publications according to quality level of publisher (in case of books; level A is highest, level E is lowest quality) and journal (in case of articles; from levels A to F), and number of professional publications. The quality ranking system used was developed by the Chairmen of the Standing Committees for Research of the faculties of Economics of the VSNU and pre-specifies the quality levels of each of the academic publishers and of 1250 journals (VSNU-VVCW, 1991). Next to this information, each program provided the committee five key journal articles it regards as its best publications.

The committee evaluates each program with respect to four criteria: *scientific quality* ("what is the quality

[1] Most business economics programs in The Netherlands are equivalent to business administration programs in other countries, like in marketing, finance, accounting and management.

of the outcome? What position has been achieved in the national and international context?"), *scientific productivity* ("what is the amount of scientific output produced, given the size of the group and the quality of the output?"), *scientific and societal/technological relevance* ("what significance does the research performed have for the development of the scientific field, for the discipline of economics (widely construed) and for society?") and its *long-term viability* ("what is the long-term viability of the program and its prospects bearing in mind national and international competition?"). The committee's evaluation according to each of these criteria is expressed in a score on a five point Likert-type scale and a short explanatory report. The scores are fully anchored for three of the four criteria (see for more details Appendix A).

3.3. The second evaluation

In 2001, the VSNU organized a second round of evaluations, covering the years 1995–2000. This is not an exact replication of the first round, because some adjustments have been made. The most important adjustment is the installation of two committees: one for Economics (including business economics) and one for Business Administration and Management (BA&M). The first committee covered economics, econometrics and some business economics programs, whereas the second committee covered business administration programs. The business economics programs were given the option to choose by which committee they wanted to be evaluated. Most accounting and finance programs opted for the economics committee, while most organization, mar-

Table 1
Research input and output categories used in the 1995 and 2001 evaluations

	Inputs	Explanation
Input-categories 1995 and 2001		
Academic staff directly funded by government	PhD students	Trainee research assistant
	Other	All other categories of scientific staff
Academic staff funded by national research councils	PhD students	Trainee research assistant funded by the national research council NWO
	Other	All other categories of scientific staff
Academic staff funded by third parties	PhD students	Trainee research assistant
	Other	All other categories of scientific staff
	Outputs	
Output-categories 1995		
Dissertations	PhD theses (I/I, I/E, E/I—from own institution (I) or external (E)/prepared in own institution (I) or external (E))	
Scientific publications	Journal articles (national and international) in six categories: A–F (A highest, F lowest)	
	Books (national and international) in five categories: A–E	
	Contributions to books (national and international) in five categories: A–E	
Annotations	Bibliographically traceable result of scientific research	
Professional publications	(part of) books, journal articles or technical descriptions aimed at dissemination outside the academic audience	
Output-categories 2001		
Dissertations	PhD theses (I/I, I/E, E/I – from own institution (I) or external (E)/prepared in own institution (I) or external (E))	
International academic publications	Articles in international journals	
	International books	
	Contributions in international books	
	International proceedings	
Dutch academic publications	Articles in Dutch journals	
	Dutch books	
	Contributions in Dutch books	
Professional publications	Research reports	

keting and information systems programs opted for the second committee. In order to secure consistency in judgment, two evaluators were member of both committees. The committee for Economics reviewed 60 programs from eight universities, considerably less than the 90 programs in the 1995 evaluation, due to closure of unsuccessful programs, mergers of some programs into larger units and relocation of others to the business administration evaluation (Westerheijden, 1996). The committee for Business Administration and Management (BA&M) reviewed 19 programs from six universities. Another major difference is in the categorization of publications in the self studies: the VSNU decided to require a less detailed system of output categorization. Next to the three types of dissertations (this was not altered), three main categories of academic publications were distinguished: international journal articles (divided into refereed and non-refereed), other academic publications (books, chapters and proceedings), and Dutch publications (divided into refereed and non-refereed). Finally, a category of professional publications and scientific reports was proposed to report on commissioned research and applied work. Some minor differences exist between the two groups: in economics, an option was given in the international journals category and in the books category to distinguish between 'top' publications and 'other'. In case such distinction was made, the faculty was asked to give a clear definition for the distinction

Table 2
Descriptive statistics

	1995 evaluation ($n = 90$)				2001 evaluation ($n = 79$)			
	Mean	Minimum	Maximum	S.D.	Mean	Minimum	Maximum	S.D.
Input variables								
Total staff	24.4	1.2	126.7	19.4	34.0	2.8	106.5	22.0
PhD students	8.2	0	48	7.8	15.0	0	47.5	12.2
Staff university	18.3	1.2	75.7	21.7	25.8	2.4	79.1	18.6
Staff national funded	2.4	0	18.9	3.8	3.2	0	17.8	4.0
Staff third party funded	3.6	0	67.6	10.0	4.9	0	81.4	11.0
PhD students (% of total staff)	33.7	0	87	18.6	42.0	0	70	15.9
Staff national funded (%)	9.1	0	75	14.3	10.1	0	65.3	12.0
Staff third party funded (%)	10.6	0	83	16.2	12.2	0	93.6	19.4
Output variables: bibliometric data								
Dissertations	4.9	0	22	4.1	7.6	0	22	5.4
Journal articles A			2.9	0	24	4.6		
Journal articles B	11.7	0	47	11.6				
Journal articles C	21.0	0	171	27.3				
Journal articles D	29.7	1	154	28.4				
Journal articles E	21.2	0	137	20.9				
Journal articles F	8.9	0	58	9.9				
Books A	0.5	0	7	0.9				
Books B	1.4	0	12	1.9				
Books C	3.5	0	26	4.2				
Books D	3.8	0	45	6.7				
Books E	13.4	0	228	37.9				
International journal articles					58.1	1	220	41.0
International books					8.8	0	43	9.1
Contributions in int.books					30.6	2	169	26.3
International proceedings					24.6	0	187	35.6
Dutch journal articles					30.8	0	239	39.6
Dutch books					7.2	0	33	8.1
Contributions in Dutch books					17.9	0	95	21.3
Research reports					52.0	1	202	46.8
Committee's qualifications								
Quality	3.1	1	5	0.9	3.6	2	5	0.8
Productivity	3.0	1	5	0.9	3.3	2	5	0.7
Relevance	3.4	1	5	0.8	3.5	2	5	0.6
Viability	3.1	1	5	0.9	3.4	1	5	0.9
Overall judgment	3.2	1.2	5	0.8	3.5	2	5	0.7

used. In the business administration group, an option was given to add working papers, research reports and external reports to the 'other academic publications' category. In order to secure comparability between the research programs, we decided to keep differences in output variables to a minimum. An overview of the input and output variables used in the 1995 and 2001 evaluations are presented in Table 1.

Both committees applied the same peer judgment categories *academic quality*, *productivity*, *relevance* and *long-term viability* as was used in the 1995 evaluation. However, they introduced a slightly different definition of the *relevance* criterion by giving it a restrictive interpretation focusing exclusively on practical contributions to professional or policy ends. By excluding relevance for the scientific community, the committees hoped to make divergent interpretations of this variable less likely. The committees made sure they applied similar methodologies, including the use of the same fully anchored Likert-type five points scales that were taken from the 1995 evaluation. Also the procedures the committees followed were identical to the 1995 procedures, except for the fact that the 2001 committees held meetings with the coordinators of *all* research programs and not with a selection of coordinators. The evaluation committee for Economics presented their report in August 2002 (VSNU, 2002b), the committee for Business Administration and Management followed in December (VSNU, 2002a).

The input variables, output variables and committees' final assessments are summarized in Table 2.

4. Correspondence between peer review and bibliometric method

4.1. Peer review assessments' dimensions

Although the committee labeled each of the four dimensions differently, attaching different meaning to each of them, the scores on the four dimensions appear to be highly correlated (Pearson correlation coefficients are between 0.628 and 0.796, $p < 0.01$). Common factor analysis (without rotation) identifies only one dimension for each of the two evaluation rounds. We treat the two committees in the 1995 round independently, leading to three independent samples.

In the 1995 evaluation round, the evaluation criteria load on a single common factor, explaining almost 80% of the variance. The 2001 evaluation criteria load on a common factor, explaining nearly 60%. As Table 3 indicates, the decrease in eigenvalue is partly due to a decline in factor loadings for the evaluation criterion "relevance." We expect this to be a result of the altered definition used in the 2001 evaluations, as explained before. In order to work with a measure of research quality that is consistent over time, we decided to proceed with a composite measure of research quality that is calculated as the mean score based on three evaluation criteria: quality, productivity and viability. The resulting overall quality rating variables turn out to be reliable: the Cronbach alphas are 0.90, 0.82 and 0.85 for the 1995 evaluation, the economics 2001 and the business administration 2001 evaluations respectively (Nunnally, 1978; Hair et al., 1998).

4.2. Correspondence between peer review and bibliometric information

In order to assess the correspondence between the committees' overall judgments and the bibliometric data we applied ordinal regressions for each assessment, using the committees' composite quality score as the dependent variable. Since only 19 business administration programs were evaluated in 2001, we decided to group these together with the economics programs evaluated in the same year. Although we cannot be absolutely sure that both committees followed the same evaluation logic, they both applied the same procedures and joint

Table 3
Common factor analysis on committees' judgments

	1995 evaluation	2001 evaluation	
		Economics	Business administration
Common factor			
Eigenvalue	3.161	2.350	2.374
Variance explained	79%	58%	59%
Evaluation criteria (factor loadings)			
Quality	0.921	0.838	0.899
Productivity	0.889	0.825	0.832
Relevance	0.853	0.418	0.282
Viability	0.892	0.891	0.892

Table 4
Ordinal regression of bibliometric data on the committees' overall judgments

Bibliometric variables	1995 evaluation, estimates (Wald)[a]	Bibliometric variables	2001 evaluation, estimates (Wald)[b]
Dissertations	0.072 (0.830)	Dissertations	0.037 (0.679)
Articles A[c]	0.296 (17.779)***	International journal articles	0.027 (19.266)***
Articles B[d]	0.130 (18.636)***	International books	−0.024 (1.187)
Articles C	0.017 (1.988)	International book chapters	−0.005 (0.361)
Articles D	0.009 (0.657)	International proceedings	0.007 (2.428)*
Articles E	0.000 (0.001)	Dutch journal articles	0.006 (2.100)*
Articles F	0.020 (0.439)	Dutch books	−0.029 (0.605)
Books A	0.085 (0.132)	Dutch book chapters	−0.003 (0.019)
Books B	−0.043 (0.110)	Professional publications	−0.011 (8.148)***
Books C	0.064 (0.729)		
Books D	−0.065 (2.359)		
Books E	−0.019 (6.137)**		
Chi-square	86.744 ($p<0.000$)		58.713 ($p<0.000$)
Nagelkerke	0.665		0.587
n[e]	81		68

*$p<0.10$; **$p<0.05$; ***$p<0.01$.
[a] Ordinal regression, using the Logit link function (evenly distributed categories in the dependent variable).
[b] Ordinal regression, using the Probit link function (normally distributed latent variable).
[c] Articles A is the number of top publications in heavily refereed journals. VSNU provided a list of A-journals.
[d] Articles B is the number of excellent publications in – mostly – English and heavily refereed journals, or top publications in heavily refereed collections of English articles. VSNU provided a list of B-journals.
[e] The sample populations used in the analyses is smaller than in Table 2 indicated because of listwise deletion of cases. This is due to missing dependent variables: in some cases the committee decided not to give a score, for instance when a program indicates it will close in the future, a viability score was not issued.

memberships were used to harmonize decision making as much as possible. Table 4 shows to what extent committees' overall judgments correspond with bibliometric data. The high number of different output categories in the 1995 bibliometric system does not lead to a better prediction of the peer review results than the more condensed system used in 2001. In both systems, the number of international journal publications appears to be the best indicator of research quality. Interesting is the significant negative indication for Books E in the 1995 sample and for professional publications in the 2001 evaluations. This is not due to a multicollinearity problem (VIF scores vary from 1.2 to 1.7). Books E in the 1995 sample consists of reports in the Dutch language not published by a publishing company. Most research reports published under a university label fall in this category, making it comparable to the professional publications category in the 2001 sample. Research programs producing a high number of applied research papers receive a lower quality assessment. This may be due to the fact that these research programs allocate less effort to producing contributions to the academic literature. The quality of the research reports will also be missed by the evaluations committees' judgments because most research papers in these categories are written in the Dutch language.

5. Determinants of research programs' quality and productivity

5.1. Data envelopment analysis

In order to make a clear distinction between quality and productivity, we computed productivity scores using data envelopment analysis (Cooper et al., 2000; García-Valderrama and Groot, 2002). Data envelopment analysis (DEA) is a non-parametric linear-programming technique that enables to define the relative efficiency of decision making units (DMUs). DEA is an extreme point method, comparing multiple inputs and outputs of each sample program to the reference optimal sample programs. Since we do not know the functional form of input–output relationships in research programs, application of the DEA methodology allows us to work with optimal combinations of inputs and outputs for each program.[2] The relative technical efficiency of a program is calculated by the ratio of the weighted sum of outputs

[2] Although it is possible to surpress or enlarge the impact of an individual input or output in the efficiency scores (Thanassoulis et al., 1995).

and inputs to be selected as the Pareto optimal efficiency measure of each program subject to the constraints that an efficiency of 1 is the maximum achievable. The efficiency measure is the radial distance from the optimal production frontier. This frontier is based on the levels of each output category that is produced by the most efficient programs (see for a more detailed explanation Appendix B). DEA has also some drawbacks. Like other extreme point methods, DEA is not able to distinguish between random measurement error and inefficiency. Its nonparametric approach allows comparisons with peers/reference groups, but not with "absolute" optimal programs. DEA results are also sensitive to outliers: unique combinations of inputs and outputs will lead to higher efficiency scores than more common combinations (Anderson et al., 2000).

The output variables used in these analyses are selected from the most important variables contributing to research quality as depicted in Table 4. For the 1995 evaluation we use the number of articles A and B, for the 2001 evaluations we select the output variables number of international journals, international proceedings and Dutch journal articles. The number of full-time equivalents PhD-students and research staff are used as input variables. In order not to include economies of scale in the DEA-scores we use the CCR constant returns to scale models (Charnes et al., 1978). This allows us to test in the second stage what the impact of size on productivity is, jointly with the impact of other research programs' characteristics. The resulting DEA efficiency scores correlate best with the review committees' scores on the productivity item (Spearman correlation coefficient of 0.584; $p < 0.000$ for the 1995-assessment and 0.272; $p < 0.02$ for the 2001 assessment), indicating convergent and discriminant validity of the measure.

5.2. Research quality and productivity

In a second stage analysis we aim at explaining differences in research quality and productivity among research programs. For quality we expect to find unbiased estimators using OLS regression on the reviewers' composite quality indicator. For productivity scores, using a two-stage approach may lead the first stage to generate biased results because of the missing second stage independent variables (Grosskopf, 1996; Wang and Schmidt, 2002; Coelli et al., 2005) and because of possible serial correlations between de DEA efficiency estimates (Simar and Wilson, 2007). Other studies report unbiased first-stage DEA and second stage estimations using simulations based on different data generating processes (Banker and Natarajan, 2001), different forms of production functions and error distributions (Bardhan et al., 1998). A one-step model in which both the stochastic frontier as well as the relationship with program characteristics are estimated simultaneously implies the dependent variable being summarized in a single dependent variable (Coelli et al., 2005). We used Limdep to verify our two-stage results and found support for the main findings of the 1995 sample but not for the 2001 sample. The reason for the lack of support for the 2001 data may be found in the multidimensionality of the constructed single output variable and possible collinearity problems among the independent variables.

Research programs' quality and productivity of research is expected to be dependent on the size of the research group, composition of the research staff, funding source and discipline in which the research group is operating. Size is measured by the total number of full-time equivalent staff employed in the research program during the 5-year period evaluated. Composition of the research staff is measured by the proportion of PhD-students in the total full-time equivalents research staff. Funding source relates to two variables: the proportion of full-time equivalents research staff funded by the National Research Council and funded by commissioned research projects for third parties. The discipline is identified by two dummy variables, one for economics and one for business administration. The default value (zero) relates to econometrics, mathematical decisions theory and operations research. The relation between research programs' characteristics and the committees' overall quality scores is determined using ordinary least squares regressions. The correspondence between research program's characteristics and research productivity could not be analyzed using OLS regressions, given the skewed and truncated distribution of the DEA scores. We therefore applied truncated Tobit regression, using 0 and 1 as truncation points. The results are presented in Table 5.

5.3. Discipline

In the first evaluation round, discipline appears to be related to both research quality and research productivity. The more analytic fields, like econometrics and operations research, appear to produce more and higher quality research output than research groups in fields like economics and business administration. This finding supports the general impression at the time of the first evaluation round that research groups in more analytic fields had more opportunities to excel, because they were contributing to well-developed and well-organized disciplines that facilitate more academic publications that are more frequently cited. The replication study how-

Table 5
Correspondence between research programs' organizational characteristics, research quality (t-value) and research productivity (z-value)

	1995 evaluation		2001 evaluation	
	OLS regression (quality[a], β[b])	Tobit regression[c] (productivity[a], B[b])	OLS regression (quality[a], β[b])	Tobit regression[c] (productivity[a], B[b])
Constant	–	0.162 (1.969)**	–	0.361 (0.001)***
Staff				
Total fte staff	0.500 (5.052)***	−0.000 (−0.202)	0.403 (3.468)***	−0.002 (−1.989)**
% PhD students	0.186 (1.698)*	0.224 (1.607)	−0.012 (−0.105)	0.285 (1.715)*
Funding sources				
% national funding	−0.120 (−1.239)	−0.008 (−0.044)	0.225 (1.864)*	0.475 (2.216)**
% third party funding	0.011 (0.099)	0.039 (0.247)	−0.199 (−1.671)*	−0.036 (−0.251)
Discipline				
Economics (dummy)	−0.311 (−2.558)**	−0.188 (−3.396)***	−0.086 (−0.502)	−0.073 (−0.991)
Business. administration (dummy)	−0.313 (−2.444)**	−0.239 (−3.439)***	−0.034 (−0.195)	−0.025 (−0.348)
F	7.380***	0.321	3.075***	0.185
Adj. R^2	0.324	0.256	0.146	0.092
n	81	82	74	69

Note. Use of log(total fte staff) as size variable in the models does not change the main findings. $^{*}p<0.10$; $^{**}p<0.05$; $^{***}p<0.01$.
[a] Dependent variables.
[b] Independent variables.
[c] Tobit regression truncated sample (0,1), quadratic hill climbing procedure.

ever does not confirm our original finding: discipline does not turn out to be a structural factor in predicting research quality and productivity, but a more transitory factor. The 1995 evaluation was the first evaluation in its kind in The Netherlands, and lead to a critical review of some research programs, notably in the area of business administration. It looks as if these criticized programs improved their research output considerably between the two evaluation rounds, rendering the discipline variable insignificant. Cherchye and vanden Abeele (2005) find some specialization types to be related to efficiency. Given data restrictions, we were not able to divide our sample into specialization types. The impact of our discipline variable appears to have been diminished over time.

5.4. Size

Size appears to be an important variable: in both evaluation rounds it is positively related to research quality. We additionally included a quadratic term in staff numbers to check for nonlinearity. This term appears to be significant in both samples and of a negative sign, indicating decreasing economies of scale ($t=-3.014$, $p<0.004$ for the 1995 sample; $t=-2.658$, $p<0.010$ for the 2001 sample). We also tested non-linearities in the relationships between size and quality by slicing the size variable into eight portfolios of equal percentiles. We used size as fixed variables and the other variables of Table 5 as covariates in a general linear model (GLM). The decreasing economies of scale relationship between size and quality in both samples was confirmed and appears to be monotonic. We could not find an optimal size of the research group related to research quality: larger groups have a lower marginal productivity gain from a unit increase in size than smaller research groups.

The present analyses do not establish causality: large research groups may provide sufficient funds and researchers to produce a productive environment that is conducive to high-quality research performance. But also reversely: high quality research output may attract funds that permit research groups to grow. At the same time we found a surprising *negative* relation between size and research productivity in the 2001 sample (the negative sign was already present in the 1995 sample, but not significant). In both samples, the partial correlation coefficient between productivity and size, controlling for quality, is negative and significant (for the 1995 sample: −0.356, $p<0.001$; for the 2001 sample: −0.374, $p<0.001$). A quadratic term for total staff turned out to be nonsignificant for both samples, indicating a linear relation between size and productivity. Larger research groups produce higher quality research output but appear to have more difficulties managing the research productivity of each of its members than smaller research groups.

5.5. Staff composition and funding

Staff composition measured in the proportion of PhD-students participating in the program seems to have a positive, but not very convincing and stable associations with research quality and productivity. The positive relation with productivity seems to be more stable across the two samples than the positive association with quality.

In 2001, the funding conditions become significant for both research quality and productivity scores. Quality is positively related to the percentage of research team members financed by the National Research Council and negatively associated with the percentage of researchers funded by third parties in commissioned research contracts. Only national funded research appears to be positively associated with research productivity. The relation between productivity and third party funding appears to be insignificant and of a negative sign. These results corroborates earlier findings that national funding is positively related to research efficiency (Cherchye and vanden Abeele, 2005). Cherchye and Vanden Abeele also found a negative relation between third party funding and research productivity. In the 2001 sample we found a similar negative sign, but it turns out to be non-significant. The positive relationship between national funding and research quality may be explained by the fact that nationally funded research programs are evaluated on quality of the research proposal and track record of the research team before becoming eligible for funding. The positive relation to research productivity may be explained by the strict procedures that are followed for evaluating the progress made during project execution. The mean size of national funding has increased considerably between 1995 and 2001 (refer to Table 2), which may explain why this factor has become significant in both 2001 models. The negative association between third party funding and research quality may be explained by the fact that results generated by most commissioned research may not be readily transferable from the applied field into the academic literature. The mean size of commissioned research staff increased as well between 1995 and 2001. The higher portion of the human resources programs dedicated to applied research, the less opportunities they may have to excel in academic publications.

6. Conclusions

In this paper we tried to answer two questions: firstly, to what extend do research assessments based on bibliometric data and on peer review lead to similar conclusions, and secondly, what are the characteristics of successful research programs. We looked at two rounds of evaluations of research programs in economics, econometrics and business administration in The Netherlands. These evaluations provided the opportunity to confront bibliometric data with peer review assessments, and to replicate our analyses. In both rounds, the three evaluation committees' scores on quality, productivity and viability relate to one underlying dimension explaining between 58% and 79% of the variance. More than half of the variance of the overall quality judgments of the committees can be predicted by a handful bibliometric variables, notably number of publications in top-class and in excellent international refereed journals, number of international proceedings and number of Dutch journal articles.

Looking at the second research question, it turns out that larger research groups attain higher quality research output. Yet at the same time, larger research groups also appear to experience problems in managing productivity. Replication of the 1995 study in 2001 enabled us to see which organizational characteristics are permanent predictors of research output and which ones are only temporarily related to research quality and productivity. Research group size seems to be the only permanent factor related to research quality and productivity in both evaluation rounds. Other characteristics, like the composition of the research team and the academic discipline in which the team operates, appear to be only temporarily related to research output, depending on changes in for instance institutional or financial arrangements.

This study also has its limitations. The results are restricted to research groups in economics, business administration and econometrics only and may not be valid for research groups in other disciplines. The strong part of this study, the replication, also has its weaknesses. The 2001 evaluation is not a precise replication of the 1995 assessment, because the bibliometric data were structured differently and the 2001 evaluation was done by two, partly different, review committees in stead of one. This may have introduced differences in the way research programs have been evaluated, despite all the efforts taken by the evaluators to avoid differences in procedures and criteria used. Another limitation is caused by the cross-sectional data used, since they do not permit to establish causality between research output and organizational characteristics. Longitudinal data may help in analyzing causality among the variables. Lastly, this research shows that institutional and financial factors may play a role in attaining research quality and productivity. However, this influence appears not to be stable over time. More research is needed to find out what combinations of institutional and budgetary arrangements may be most conducive to improve research quality and productivity.

Appendix A. The evaluation committee's measurement scales

Scientific quality

5 = excellent	The program makes important and innovative contributions to its field. It is among the world's leading programs in its field
4 = good	The program makes worthwhile contributions to its field and may contain elements of excellence
3 = average	The program's contributions to its field are of interest, but do not attract international attention
2 = unsatisfactory	The program contributes marginally to its field. It needs improvement in order to contribute significantly
1 = poor	The program's contributions to its field are insignificant. The program should be reoriented or discontinued

Scientific productivity

5 = excellent	The program has a high output of scientific publications of high impact; it produces considerable numbers of PhD theses
4 = good	The program has a regular output of scientific publications with a considerable impact; it successfully communicates its results to a relevant audience, as identified in its mission statement
3 = average	The level of productivity is average
2 = unsatisfactory	The program has produced some scientific output, but both number and quality of publications are below standard
1 = poor	The program has hardly produced scientific output and the output is of marginal interest

Relevance

In assessing relevance, the 1995-committee considered three different aspects: (a) the *scientific relevance*, (b) the *relevance of the field for the discipline of economics* (widely construed), and (c) the *social relevance* of the group's research. Only the general qualification of the grades are used: 5 = excellent, 4 = good, 3 = average, 2 = unsatisfactory, 1 = poor.

Long-term viability

5 = excellent	The program has a well established, leading position in its field and has clear and scientifically promising plans for the future. Availability of highly qualified staff as well as future funding are secured
4 = good	The program's approach has been fruitful and future plans and perspectives seem healthy. There are no major worries concerning availability of competent staff or future funding
3 = average	There are some hesitations about one or more of the aspects mentioned above, but there is a fair expectation that the group can maintain or obtain an adequate position within its field
2 = unsatisfactory	The committee has serious doubts about the program's future position in the field. Radical measures are necessary to secure an adequate contribution in the future
1 = poor	For a number of reasons the program should be discontinued in its present form

Appendix B. Data envelopment analysis

Data envelopment analysis (DEA) is a non-parametric technique that establishes the relative efficiency of firms or individual decision making units (DMUs). The relative technical efficiency of a firm is calculated by forming the ratio of a weighted sum of outputs and inputs to be selected in a manner that calculates the Pareto efficiency measure of each firm subject to the constraints that an efficiency of 1 is the maximum achievable. The efficiency measure is the radial distance from the optimal production frontier. This frontier is based on the levels of each output category that is produced by the most efficient firms.

Two models are most frequently associated with the DEA methodology: the CCR ("Charnes–Cooper–Rhodes")-model measures efficiency under constant returns of scale (Charnes et al., 1978) and produces an "overall efficiency" rating, while the BCC ("Banker–Charnes–Cooper")-model measures efficiency on the assumption of variable returns to scale leading to a "pure technical efficiency" rating (Banker et al., 1984). In this paper we used the CCR model (constant returns to scale), producing an overall efficiency rating. The formulation of a DEA model in fractional programming form is as follows (Cooper et al., 2000; García-Valderrama and Groot, 2002):

$$(FP_0) \quad \text{Max}\, \theta = \frac{\sum_{r=1}^{s} u_r Y_{r0}}{\sum_{i=1}^{m} v_i X_{i0}} \tag{1}$$

subject to

$$\frac{\sum_{r=1}^{s} u_r Y_{rj}}{\sum_{i=1}^{m} v_i X_{ij}} \leq 1 \quad (j = 1, \ldots, n)$$

$$u_r \geq 0 \text{ and } v_i \geq 0$$

In which Y_r denotes amount of output r, and X_i amounts of input i. The weights u_r for each output r and v_i for each input i are supposed to be non-negative. In Eq. (1) j different DMUs are evaluated against each other. The DMU to be evaluated is designated DMU_0 where 0 ranges over 1, 2, ..., n.

To resolve model (1) it is converted into an equivalent linear form, setting the denominator equal to a constant (unity) and maximizing the numerator as follows:

$$(LP_0) \quad \text{Max}\, \theta = \sum_{r=1}^{s} u_r Y_{r0} \qquad (2)$$

subject to

$$\sum_{r=1}^{s} u_r Y_{rj} - \sum_{i=1}^{m} v_i X_{ij} \le 0 \quad (j = 1, \ldots, n)$$

$$\sum_{i=1}^{m} v_i X_{i0} = 1$$

$$u_r \ge 0 \text{ and } v_i \ge 0$$

In our study inputs are full-time equivalents of PhD-students and other staff, while the outputs are selected from the most important publication categories as depicted in Table 4. From these output categories, articles A are considered of higher quality than articles B, and a similar ranking applies to international journal articles and Dutch journal articles. We used different weight restrictions on our output data in order to model for these quality differences. The restrictions lead to lower mean efficiency scores and standard deviations, but did not change the general ordering of the research groups. A possible explanation is that DEA already assigns optimal weights (multipliers) in solving the dual linear model (Cooper et al., 2000, p. 52). In our sample, the ratio of output weights attached to articles A (u_A) in relation to articles B (u_B) in the final LP solution is 2.78 (u_A^*/u_B^*), indicating that output category Articles A is weighted 2.78 times more than output category Articles B.

Since 1978, DEA is applied in an almost unlimited number of different settings, represented by the list of a bibliography of 472 published papers and dissertations between 1978 and 1992 by Seinford in Cooper et al. (2000). Examples of applications are efficiency analysis of non-profit firms in health care receiving financial support from governments (Hollingworth, 2003), the impact of non-controllable inputs and external conditions on efficiency of schools (Ray, 1991), the implications of extended functions for health care organizations' operating efficiency (Kooreman, 1995), scale and scope efficiencies of housing brokerage firms (Zumpano and Elder, 1994; Anderson et al., 1998), the impact of technological change on trends in efficiency of the textile industry in China (Wei et al., 1995), and efficiency gains of making use of economies of scale in software maintenance (Banker and Slaughter, 1997).

References

Anderson, R.I., Fok, R., Zumpano, L.V., Elder, H.W., 1998. Measuring the efficiency of residential real estate brokerage firms. The Journal of Real Estate Research 16 (2), 139–158.

Anderson, R.I., Lewis, D., Springer, T.M., 2000. Operating efficiencies in real estate: a critical review of the literature. Journal of Real Estate Literature 8 (1), 3–18.

Ballas, A., Theoharakis, V., 2003. Exploring diversity in accounting through faculty journal perceptions. Contemporary Accounting Research 20 (4), 619–644.

Banker, R.D., Charnes, A., Cooper, W.W., 1984. Some models for estimating technical and scale inefficiencies in data envelopment analysis. Management Science 30, 1078–1092.

Banker, R.D., Natarajan, R., 2001. Evaluating contextual variables affecting productivity using data envelopment analysis. School of Management, The University of Texas, Dallas.

Banker, R.D., Slaughter, S.A., 1997. A field study of scale economies in software maintenance. Management Science 43, 1709–1725.

Bardhan, I.R., Cooper, W.W., Kumbhakar, S.C., 1998. A simulation study of joint uses of data envelopment analysis and statistical regressions for production functino estimation and efficiency evaluation. Journal of Productivity Analysis 9, 249–278.

Baron, J.N., Hannan, M.T., 1993. The impact of economics on contemporary sociology. Journal of Economic Literature.

Beattie, V., Goodacre, A., 2004. Publishing Patterns Within the UK Accounting and Finance Academic Community. University of Stirling, Stirling.

Beyer, J.M., 1978. Editorial policies and practices among leading journals in four scientific fields. Sociological Quarterly 19, 68–88.

Brinn, A., Jones, M.J., Pendlebury, M., 1996. UK accountants' perceptions of research journal quality. Accounting and Business Research, 265–278.

Brown, L.D., 1996. Influential accounting articles, individuals, Ph.D. granting institutions and faculties: a citational analysis. Accounting, Organizations and Society 21 (7–8), 723–754.

Brown, L.D., Gardner, J.C., 1985. Using citation analysis to assess the impact of journals and articles on contemporary accounting research (CAR). Journal of Accounting Research 23 (1), 84–109.

Brown, L.D., Huefner, R.J., 1994. The familiarity with and perceived quality of accounting journals: views of senior accounting faculty in leading U.S. MBA programs. Contemporary Accounting Research 11, 223–250.

Charnes, A., Cooper, W.W., Rhodes, E., 1978. Measuring the efficiency of decision making units. European Journal of Operational Research 2, 429–444.

Cherchye, L., vanden Abeele, P., 2005. On research efficiency, a micro-analysis of Dutch university research in economics and business management. Research Policy 34, 495–516.

Coelli, T.J., Rao, D.S.P., O'Donell, C.J., Battese, G.E., 2005. An Introduction to Efficiency and Productiviy Analysis. Springer, New York.

Cooper, W.W., Seiford, L.M., Tone, K., 2000. Data Envelopment Analysis. Kluwer Academic Publishers, Norwell, MA.

Croom, D.L., 1970. Dangers in the use of citation analysis. Nature, 1173.

Dwyer, P.D., 1994. Gender differences in the scholarly actvities of accounting academics: an empirical investigation. Issues in Accounting Education 9 (2), 231–246.

Dyckman, T.R., Zeff, S.A., 1984. Two decades of the *Journal of Accounting Research*: a review article. Journal of Accounting Research 22, 225–297.

García-Valderrama, T., Groot, T.L.C.M., 2002. Controlling the efficiency of university research in the Netherlands. In: Thore, S.A. (Ed.), Technology Commercialization: DEA and Related Analytical Methods for Evaluating the Use and Implementation of Technical Innovation. Kluwer Academic Publishers, Bosten.

Groot, T.L.C.M., 1988. Management van Universiteiten—een onderzoek naar de mogelijkheden voor doelmatig en doeltreffend universitair bestuur. Wolters-Noordhoff, Groningen.

Groot, T.L.C.M., 1999. Budgetary reforms in the non-profit sector: a comparative analysis of experiences in health care and higher education in the Netherlands. Financial Accountability and Management 15 (3/4), 355–378.

Groot, T.L.C.M., van Helden, G.J., 1999. Financieel management van non-profitorganisaties. Wolters-Noordhoff, Groningen.

Grosskopf, S., 1996. Statistical inference and nonparametric efficiency: a selective survey. Journal of Productivity Analysis 7, 161–176.

Hair Jr., J.F., Anderson, R.E., Tatham, R.L., Black, W.C., 1998. Multivariate Data Analysis. Prentice-Hall, Upper Saddle River, NJ.

Hall, T.W., Ross, W.R., 1991. Contextual effect in measuring accounting faculty perceptions of accounting journals: an empricial test and updated journal rankings. Advances in Accounting, 161–182.

Harzing, A.-W., 2003. Journal Quality List. University of Melbourne, Parkville, Australia.

Hasselback, J.R., Reinstein, A., Schwan, E.S., 2000. Benchmarks for Evaluating the Research Productivity of Accounting Faculty. College of Business, Florida State University, Tallahassee, FL.

Hazeu, C.A., 1983. Voorwaardelijke financiering van onderzoek. Economisch Statistische Berichten 68 (3393), 163–165.

Hollingworth, B., 2003. Non-parametric and parametric applications measuring efficiency in health care. Health Care Management Science 6, 203–218.

Hull, R.P., Wright, G.B., 1990. Faculty perceptions of journal quality: an update. Accounting Horizons, 77–98.

Johnes, G., 1988. Determinants of Research Output in Economics Departments in British Universities. Research Policy 17, 171–178.

Konrad, A.M., Pfeffer, J., 1990. Do you get what you deserve? Factors affecting the relationship between productivity and pay. Administrative Science Quarterly 35, 258–285.

Kooreman, P., 1995. Doelmatigheid meten in de gezondheidszorg. Economisch Statistische Berichten 80 (4011), 493.

Kuhn, T.S., 1970. The Structure of Scientific Revolution. University of Chicage Press, Chicago.

Lodahl, J.B., Gordon, G., 1973. Differences between physical and social sciences in university graduate departments. Research in Higher Education 1, 191–213.

Loft, A., Jorissen, A., Walton, P., 2002. From newsletter to academic journal: creating the European accounting review. European Accounting Review 11 (1), 43–75.

Moed, H.F., Burger, W.J.M., Frankfort, J.G., van Raan, A.F.J., 1985. The use of bibliometric data for the measurement of university research performance. Research Policy 14, 131–149.

Morris, J.L., Cudd, R.M., Crain, J.L., 1990. A study of the potential bias in accounting journal ratings: implications for promotion and tenure decisions. Accounting Educators' Journal, 46–55.

Nederhof, A.J., van Raan, A.F.J., 1987. Peer review and bibliometric indicators of scientific performance: a comparison of cum laude doctorates with ordinary doctorates in physics. Scientometrics 11, 330–350.

Nederhof, A.J., van Raan, A.F.J., 1993. A bibliometric analysis of six economics research groups: a comparison with peer review. Research Policy 22, 353–368.

Newman, J.M., Cooper, E., 1993. Determinants of academic recognition: the case of the Journal of Applied Psychology. Journal of Applied Psychology 78 (3), 518–526.

Nunnally, J.C., 1978. Psychometric Theory. McGraw-Hill, New York.

Parker, L., Guthrie, J., Gray, R., 1998. Accounting and management research: passwords from the gatekeepers, Accounting. Auditing and Accountability Journal 11 (4), 371–402.

Pfeffer, J., 1993. Barriers to the advance of organizatinoal science: paradigm development as a dependent variable. Academy of Management Review 18 (4), 599–620.

Ray, S.C., 1991. Resource-use efficiency in public schools. Management Science 37, 1620–1629.

Rinia, E.J., van Leeuwen, T.N., van Vuren, H.G., van Raan, A.F.J., 1998. Comparitive analysis of a set of bibliometric indicators and central peer review criteria, Evaluation of condensed matter physics in the Netherlands. Research Policy 27, 95–107.

Schroeder, R.G., Payne, D.D., Harris, D.G., 1988. Perceptions of accounting publication outlets. Accounting Educators' Journal, 1–17.

Simar, L., Wilson, P.W., 2007. Estimation and inference in two-stage, semi-parametric models of production processes. Journal of Econometrics 136, 31–64.

Sterken, E., Leeflang, P.S.H., 2000. Een expertlijst van tijdschriften in de bedrijfswetenschappen. Maandblad voor Accountancy en Bedrijfseconomie 74 (4), 73–85.

Thanassoulis, E., Boussofiane, A., Dyson, R.G., 1995. Exploring output quality targets in the provision of perinatal care in England using data envelopment analysis. European Journal of Operational Research 30, 588–607.

Timmermans, W.A., 1984. Rechtspositie van wetenschappelijk (en niet-wetenschappelijk) personeel. Universiteit en Hogeschool 30 (3), 109–127.

van Fleet, D.D., McWilliams, A., Siegel, D.S., 2000. A theoretical and emprical analysis of journal rankings: the case of formal lists. Journal of Management 26 (5), 839–861.

VSNU, 1990. Nota Onderzoeksbeleid. Vereniging van Samenwerkende Nederlandse Universiteiten (VSNU), Utrecht, The Netherlands.

VSNU, 1994a. Gids voor de ondewijsvisitaties in de 2e ronde 1994–1999. Vereniging voor Samenwerkende Nederlandse Universiteiten, Utrecht.

VSNU, 1994b. Quality Assessment of Research, Protocol 1994. Vereniging van Samenwerkende Nederlandse Universiteiten, Utrecht.

VSNU, 1995. Quality assessment of research in economics. Vereniging van Samenwerkende Nederlandse Universiteiten (VSNU), Utrecht, The Netherlands.

VSNU, 2002a. Assessment of research quality in business administration and management. Association of the Universities in The Netherlands (VSNU), Utrecht.

VSNU, 2002b. Assessment of research quality in economics. Association of the Universities in The Netherlands (VSNU), Utrecht.

VSNU-VVCW, 1991. Performance-meting van onderzoek in economische faculteiten. Tweede ad hoc Commissie Voorzitters Vaste Commissie Wetenschapsbeoefening (VCW), Utrecht.

Wang, H.-J., Schmidt, P., 2002. One-step and two-step estimation of the effects of exogenous variables on technical efficiency levels. Journal of Productivity Analysis 18, 129–144.

Wei, Q.L., Sun, B., Xiao, Z.J., 1995. Measuring technical progress with data envelopment analysis. European Journal of Operational Research 80, 691–702.

Westerheijden, D.F., 1996. Een solide basis voor beslissingen, Effecten van onderzoekevaluaties door de VSNU. Vereniging voor Samenwerkende Nederlandse Universiteiten, Utrecht.

Whittington, G., 1993. The 1992 research assessment exercise. The British Accounting Review 25 (4), 383–395.

Whittington, G., 1997. The 1996 research assessment exercise. The British Accounting Review 29, 181–197.

Zivney, T.L., Bertin, W.J., Gavin, T.A., 1995. A comprehensive examination of accounting faculty publishing. Issues in Accounting Education 10 (1), 1–25.

Zumpano, L.V., Elder, H.W., 1994. Economies of scope and density in the market for real estate brokerage services. The Journal of the American Real Estate and Urban Economics Association 22, 497–513.