ELSEVIER

Research Policy 27 (1998) 95–107

research policy

Comparative analysis of a set of bibliometric indicators and central peer review criteria
Evaluation of condensed matter physics in the Netherlands

E.J. Rinia [a,*], Th.N. van Leeuwen [b], H.G. van Vuren [a], A.F.J. van Raan [b]

[a] *Foundation for Fundamental Research on Matter (FOM) P.O. Box 3021, 3502 GA Utrecht, Netherlands*
[b] *Centre for Science and Technology Studies (CWTS), University of Leiden, Wassenaarseweg 52, P.O. Box 9555, 2300 RB Leiden, Netherlands*

Accepted 9 February 1998

Abstract

In this paper first results are presented of a study on the correlation between bibliometric indicators and the outcomes of peer judgements made by expert committees of physics in the Netherlands. As a first step to study these outcomes in more detail, we focus on the results of an evaluation of 56 research programmes in condensed matter physics in the Netherlands, a subfield which accounts for roughly one third of the total of Dutch physics. This set of research programmes is represented by a volume of more than 5000 publications and nearly 50,000 citations. The study shows varying correlations between different bibliometric indicators and the outcomes of a peer evaluation procedure. Also a breakdown of correlations to the level of different peer review criteria has been made. We found that the peer review criterium 'team' shows generally the strongest correlation with bibliometric indicators. Correlations prove to be higher for groups which are involved in basic science than for groups which are more application oriented.

Keywords: Bibliometric indicator; Condensed matter physics; Peer review criteria

1. Introduction

In an increasing amount and variety of studies bibliometric data have been used to assess research performance, particularly in the natural and life (medical, biological) sciences. The assumption underlying these analyses is that bibliometric indicators provide a reliable 'image' of (at least substantial parts of) scientific activity. For instance, number and type of publications are considered to be indications of scientific production. Citation-based indicators are regarded as measures of impact or international visibility of research (Narin, 1976; Garfield, 1979; Martin and Irvine, 1983; Moed et al., 1985).

In this context of research evaluation one distinguishes between the concepts 'impact' and 'quality' of scientific research (e.g., Martin and Irvine, 1983; Martin, 1996; Moed et al., 1985). Quality is perceived of as a broad concept with different aspects. For example, cognitive and methodological quality are distinguished. Impact is assumed to point to

* Corresponding author.

0048-7333/98/$19.00
PII S0048-7333(98)00026-2

another specific quality aspect. Because of the multi-dimensional character of the concept quality, an assessment of quality is in practice always based on the application of a 'mix' of different criteria (van Raan and van Leeuwen, 1995).

In peer review, the main assessment mechanism in science by which programmes, proposals or manuscripts are critically judged by professional colleagues, this is often recognised by the inclusion of several criteria which reviewers are asked to address. The number and the nature of these criteria in the evaluation of research proposals are currently issue of extensive discussions (see for instance NSF, 1996).

It seems to be more generally assumed now that advanced bibliometric methods offer crucial information about research performance that can be seen as complementary to peer opinion (van Raan, 1996; NSTC, 1996). This is particularly important, as peer review also has serious drawbacks (Cole et al., 1978; Horrobin, 1990).

In several earlier bibliometric studies comparisons are made between bibliometric results and the judgement of scholars or experts on the quality of research. (Anderson et al., 1978; Bayer and Fulger, 1966; Chang, 1975; Cole and Cole, 1967; Martin and Irvine, 1983; Nederhof, 1988; Nederhof and van Raan, 1987, Nederhof and van Raan, 1989). These studies revealed a reasonable correspondence between the results of bibliometric analyses on the one hand, and judgements of scientific quality by peers on the other. This is an important basis for the applicability of bibliometric indicators.

The correlations found are significant, but not perfect as can be expected. It is important to know whether more general or systematic patterns can be found for cases where judgements on the basis of bibliometric results do correspond and where—and why—they do not correspond with the opinion of colleagues. For instance a poor correlation was found between citation indicators and the originality of research proposals in application oriented research (van den Beemt and van Raan, 1995).

In order to explore in more detail the correlations between scores on bibliometric indicators and the outcomes of peer judgements, we conducted a specific study as part of a larger bibliometric research project on physics in the Netherlands performed by the Centre for Science and Technology Studies (CWTS, Leiden University) and the Foundation for Fundamental Research on Matter (FOM).

In this paper we discuss the results of a comparative analysis between the outcomes of a peer evaluation recently held (October 1996) of research programmes in condensed matter physics and the results of a bibliometric study on research programmes in academic physics in the Netherlands (van Leeuwen et al., 1996). For this comparison we focus on the bibliometric as well as on the peer review side on several specific elements of the assessments, in order to gain more insight into relevant aspects of the evaluation procedures. We stress that there are always bibliometric elements in each peer evaluation, for example publications in important journals. So peer evaluation and bibliometric assessment will inevitably show some correlation, they are never 'orthogonal dimensions' in evaluation space. The important questions for empirical investigation are therefore *which* particular bibliometric indicators do correlate to *what* extent, and under *what* 'circumstances'.

2. Method

2.1. The FOM condensed matter peer review procedure

As a first element of this analysis we use data of an evaluation held in October 1996 of 62 research programmes in condensed matter physics carried out at universities or (para) academic institutes in the Netherlands. This physics subfield accounts for more than one third of the total output in physics, both in the Netherlands and in the world. (Rinia and de Lange, 1991). The programmes involved have been evaluated within the framework of a periodical peer evaluation by the national working community for condensed matter physics of the Foundation for Fundamental Research on Matter (FOM), the Netherlands Physics Research Council. The results of this assessment are the basis for the funding of the programmes by FOM for the next two years.

The judgement of research by this evaluation procedure is in fact an assessment of the ongoing research programme. The scientific work performed

in the past years is the primary target of evaluation, but also the research planned for the next period is assessed. These plans may be an extension of current research but also new projects can be proposed. So, a mix of both past performance and potential performance is assessed.

The programmes involved are carried out by research groups ('FOM working groups') all of which but two are part of the FOM working community for condensed matter physics. Although the main goal of FOM is the advancement of basic research, this research council recently focused special attention to the application potential of basic research for other sciences, society or industry. Therefore, the evaluation procedure in this working community was split up in two categories. The one category (B) concerns 'curiosity driven' research, the main motivation here is to expand basic knowledge. The other category (A) concerns 'application driven' research which means primarily basic research, but with an application-oriented ('strategic') or technological relevance. Groups were allowed to submit their research programme for evaluation (and, by that, as a request for funding) in both categories. For each category a separate jury of experts was installed. In category A about half of the jury consists of scientists from industrial laboratories.

Each programme was reviewed by on average four and at least three (mostly foreign) referees. Referees were explicitly asked to comment on scientific merit of the programme (relevance, originality, appropriateness of methods), quality and productivity of the group (including past performance), and on the programme as a whole. Programmes submitted for evaluation in the 'application driven' category have been judged also for their strategic/technological relevance. All programmes were also submitted for review to the members of the jury and to the groupleaders of the working community. The paraphrased comments of the referees and, in some cases, of the other peers, were presented to the principal investigators of each programme in order to have their reactions. The complete set of referee and peer comments as well as the reply of the investigators finally was laid down in a protocol. On the basis of this protocol, and the programme involved, the juries rated all programmes according to the principal criteria (quality of the team, goal, method). It should be noted that the description of these three criteria in both categories is to a greater extent similar. However, as discussed above, in category A more emphasis is put on technological relevance and in category B on the advancement of basic knowledge. Besides these three criteria, in category A 'strategic' technological relevance was also considered as a separate criterium. For each criterium a rating was given on a scale of 1 (excellent) to 9 (poor). Finally, an overall judgement was given which served as the basis for grant supply.

Because of the additional criteria concerning strategical and technological relevance in the assessment of programmes in category A, and because of the reviewing by different juries, the overall ratings for the two categories are not standardised. Therefore, a direct comparison of the jury scores between the two categories is not allowed. Thus, correlations between bibliometric indicators and jury ratings are analysed within the context of each category separately.

2.2. The bibliometric research performance assessment

As a second main element of this analysis, we used bibliometric data of the earlier mentioned comprehensive bibliometric study on the research performance of 220 research programmes in physics, which was carried out in the context of an evaluation of academic disciplines at universities in the Netherlands. These programmes relate to the large majority of research groups in academic physics in the Netherlands, among which those in condensed matter physics. This extensive bibliometric study is based on the publication oeuvre in the period 1985–1994 of all senior scientists participating in the selected programmes (van Leeuwen et al., 1996). The process of data-collection and the methodology applied are to a large extent similar to those adopted in previous studies on academic research performed by CWTS, for example, on chemistry in the Netherlands (Moed and Hesselink, 1996). The main data source are the bibliographic data of all papers with a Dutch address published during the period 1985–1994 in journals processed by the Institute for Scientific Information (ISI) for the Science Citation Index (SCI), SSCI and

the A & HCI. A detailed description of the data system is given in Moed et al. (1995) and in a recent overview by van Raan, 1996.

In this 'broad scale' bibliometric study also papers of Dutch physicists with *only* a foreign address are included. As a source of bibliometric data for those publications, CD-ROM versions of the Science Citation Index (SCI) were used. Publication data were carefully verified and missing data were completed by scientists and institutes involved. This latter procedure is a very important part of the CWTS bibliometric studies.

It is of special interest, particularly in relation to the peer review process described above, that in the bibliometric study, scientists are 'linked' to programmes on the basis of the situation as of May 1995. Those who participated in a programme in the past but left before May 1995 (for instance, retirement) were excluded. For senior scientists recently appointed, his or her total publication output and impact generated during the period 1985–1994 has been included in the analysis of the programme concerned. By this choice the bibliometric analysis focuses specifically on the past performance of those who have the task to shape the future of a programme (the 'back to the future' option, instead of analysing 'total' past performance of a group's oeuvre, for instance from the perspective of accountability of research funds). We think that this 'modality' methodologically improves the comparison between peer evaluation and bibliometric assessment.

However, as the peer evaluation procedure took place in 1996 and contains a mix of assessment of both past and potential performance, it might be interesting to further compare in the future the results of this evaluation with bibliometric data of a more overlapping time span, for instance for the period 1994–1998.

The research groups in condensed matter physics and their programmes analysed in the bibliometric study and the programmes submitted to the above discussed peer evaluation procedure are not completely identical. The bibliometric study aimed at an assessment of the total programme involved (based on the publication output of *all* senior researchers), whereas programmes submitted to the peer evaluation procedure in some cases consist of smaller or specific parts of the total programme. A matching of programmes identified in both procedures and of the data concerned was performed on the basis of the *participation of the same senior physicists*. Six programmes in a total of 62 programmes submitted to the peer review procedure could not be matched well with programmes included in the bibliometric analysis. They have been excluded from this study. For the remaining 56 programmes a fairly good match was obtained, though in a few cases a complete overlap between participating senior researchers was not reached. In eight cases the same group submitted their research programme for evaluation in both categories A and B. In these cases the bibliometric results of these programmes were matched twice with the two different jury ratings. The results concerning the correlations between peer review judgements and bibliometric indicators are based on the data of these 56 programmes.

3. Results and discussion

3.1. Bibliometric indicators used in this study

The bibliometric indicators calculated in the study on Dutch academic physics (see van Leeuwen et al., 1996) are given in Table 1 together with the numerical values for the entire set of groups participating in the FOM (national) working community for condensed matter physics. Output and impact indictors are measured 'cumulatively' during a fixed time period (1985–1994). Indicators are based on all publication and citation data related to this period (Schubert et al., 1989) so called 'total block indicators'. For a detailed description of the methodology used, we refer to Moed et al. (1995) and to van Raan (1996).

The same indicators as given in Table 1 are calculated for each programme separately. The number of papers per programme in the period 1985–1994 varies from 574 to 21. It should be noted that some overlap between programmes exists because of co-authorship. There are four programmes with less than 50 publications.

To give an impression of the scattering of the citation scores within the entire set of 56 programmes: 42 programmes obtain a citation rate above the world-wide field citation rate (CPP/FCSm). Of

Table 1
Indicators of publication output and impact (1985–1994), for the FOM working community for condensed matter physics

Bibliometric indicator[a]		Score 1985–1994
The number of papers (normal articles, letters, notes and reviews) published in journals processed for the CD-ROM version of the Science Citation Index (SCI).	P	5327
The number of citations recorded in SCI journals to all publications involved. Self-citations are included.	C	46,858
The average number of citations per publication, or citation per publication ratio. Self-citations are included.	CPP	8.80
The average number of citations per publication. Self-citations are not included.	CPPex	6.42
Percentage of papers not cited during the time period considered.	%Pnc	28.21
The world-wide average citation rate of all papers published in journals in which a group has published (the group's journal set).	JCSm	7.05
The world-wide average citation rate of all publications in (sub)fields in which the group is active. Subfields are defined by means of SCI journal categories.	FCSm	5.47
The impact of a group's publications, compared to the world-wide average citation rate of the group's journal set (self-citations included).	CPP/JCSm	1.25
The impact of a group's publications, compared to the world-wide citation average in (sub)fields in which the group is active (self-citations included).	CPP/FCSm	1.61
The impact of journals in which a group has published (the group's journal set), compared to the world-wide citation average in (sub)fields covered by these journals.	JCSm/FCSm	1.29
The percentage of self-citations. A self-citation is defined as a citation in which the citing and the cited paper have at least one author in common (either a first author or a co-author).	%SELFCIT	27.05

[a] In this case 'group' is the entire (national) working community.

these, 27 programmes have a score which is significantly above this world average. 14 programmes obtain a citation score below the world-wide field average, of which 5 programmes have a score significantly below average. Concerning the citation rate compared to the world-wide journal average (CPP/JCSm): 34 programmes obtain a citation rate at or above world average. Of these, 16 programmes have a rate significantly above average. 22 programmes have a citation score below the world-wide

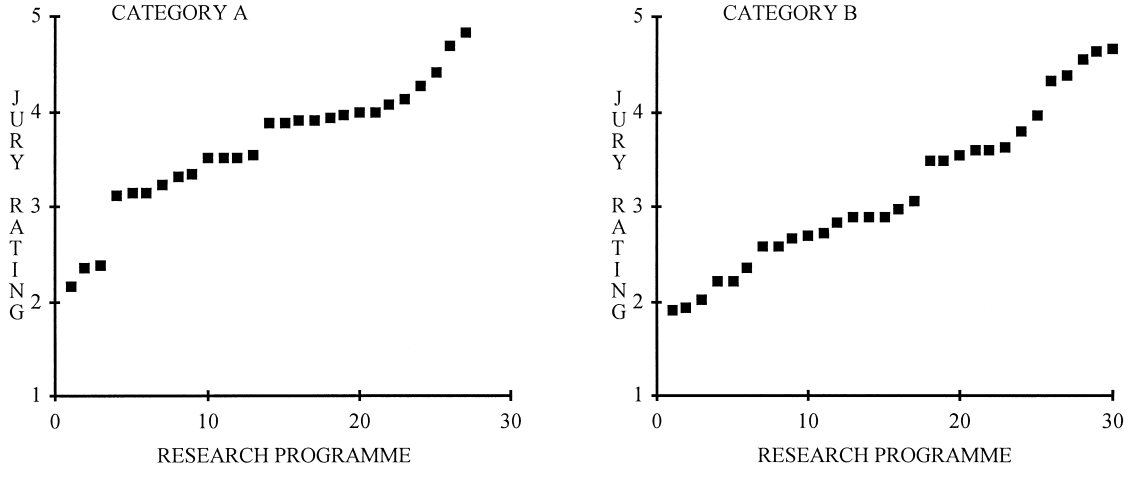

Fig. 1. Jury ratings of the 56 condensed matter physics programmes in two categories (A: application-oriented; B: basic research).

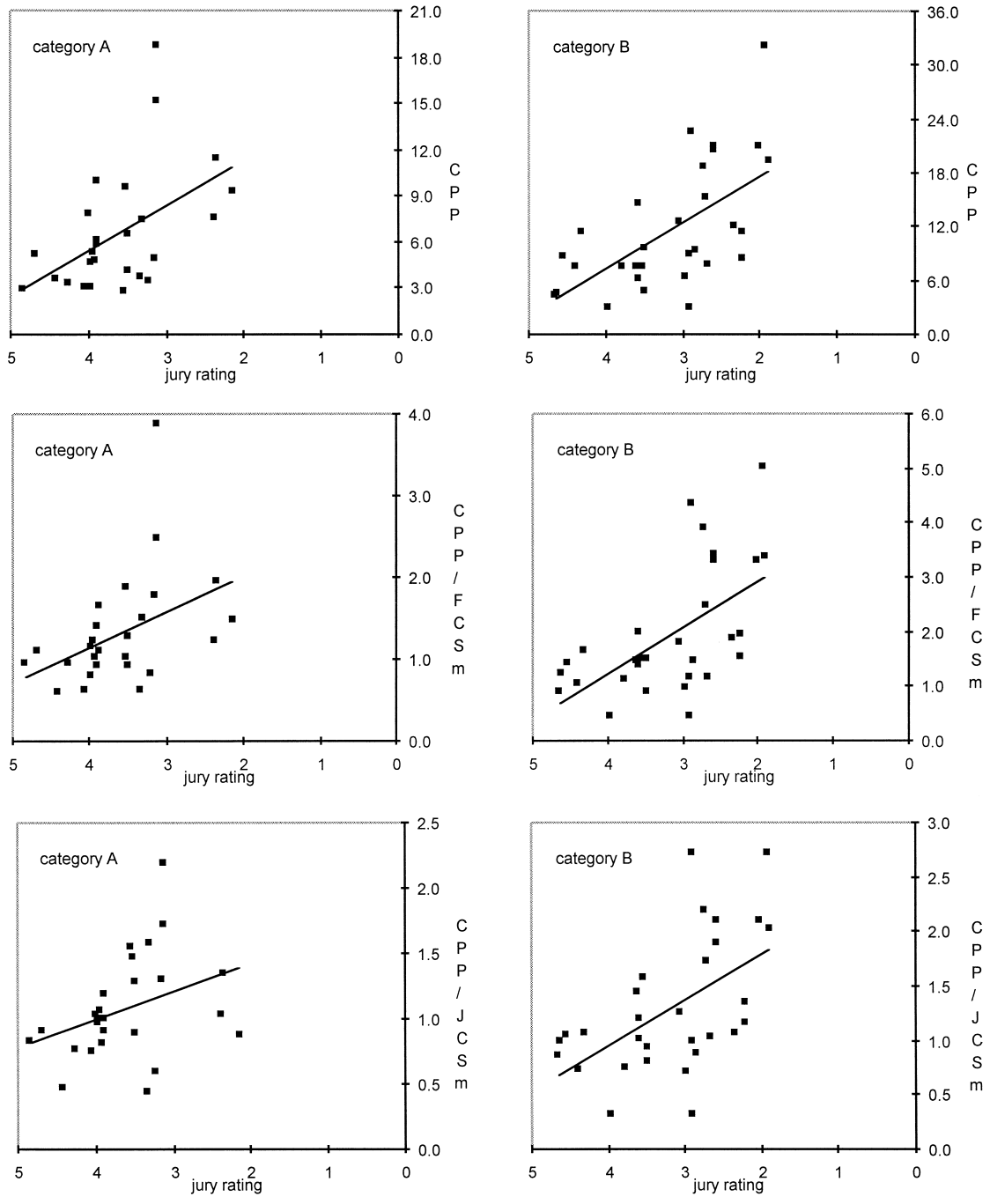

Fig. 2. Graphical display of jury ratings and bibliometric indicators for 56 programmes in two sectors of condensed matter physics (A: application-oriented; B: basic research).

journal average. Of these, 9 programmes have a score which is significantly below average.

3.2. Scattering of jury ratings

For the peer evaluation of programmes in the FOM working community of condensed matter physics, 26 programmes have been reviewed in category A (application-oriented research), and 30 in category B (basic physics).

The jury-ratings for each of the 56 programmes are given in Fig. 1. Whereas individual ratings could be given from 1 (excellent) to 9 (poor), the actual (average) ratings vary between 1.89 and 4.84. Partly this might certainly be explained by the quality of the reviewed programmes, but also by the tendency of the jury to level ratings off at the high and low end of the scale.

The average of the scores amounts to 3.61 in category A, and 3.16 in category B. In Fig. 1 is shown that the differences between these averages are caused by the fact that more programmes receive a (high) rating (between 1 and 3) in the basic physics category (B) than in the application-oriented category A. However, it cannot be concluded that programmes with a main emphasis on extending basic physics knowledge are of higher quality than programmes with a strong strategic component. As explained above, criteria for assessment of programmes and the composition of the jury are not completely identical for the two categories.

3.3. Correlations between bibliometric indicators and overall jury ratings

Linear regression analysis clearly shows different correlations between jury ratings and the various bibliometric indicators. A graphical display of the jury ratings for both categories and the indicators CPP, CPP/FCSm and CPP/JCSm is given in Fig. 2. It is shown that these indicators, though differently, in general correlate positively with peer ratings.

We observe too that all figures reveal a more or less similar pattern: for programmes with relatively high scores on bibliometric indicators, no further differentiation is visible among the jury rates. This finding would confirm our earlier observation that jury assessments level off at the top end of the scale.

For the 56 programmes we calculated for *each* of the eight bibliometric indicators described in Table 1 rank-correlations with the *overall ratings* received in the peer (jury) review procedure. Spearman's rank-correlation coefficients were calculated, as only a small number of ties occurred in the rankings of the scores. The results are given in Table 2.

A first and very interesting finding is that there appears to be *no significant correlation* between jury ratings and the number of journal publications (P) produced by the programmes involved. The other, citation-based bibliometric indicators correlate quite differently with peer review. All bibliometric indicators containing actual citation rates, based on citations obtained by papers of a group analysed, (C, CPP, CPPex, CPP/FCSm and CPP/JCSm) do correlate significantly at the '99% confidence' level.

From Table 2 we observe that the highest correlations are obtained for the average number of citations per publication, in- and excluding self-citations (CPP, CPPex) and for the (relative) citation indicator using the *field-based world averages as reference standard*, CPP/FCSm. The relative citation indicator using the *journal-based world average as reference standard*, CPP/JCSm, correlates significantly with jury-ratings in both categories. The absolute number of citations (C) also correlates significantly with peer judgements, though less than the before

Table 2
Spearman's rank-correlation coefficients (r_S) of bibliometric indicators (1985–1994) and overall jury ratings in category A: application-oriented and category B: basic research

Period 1985–1994	Category A (r_S)	Category B (r_S)
Indicator		
P	0.37	0.16
C	0.54(+)	0.47(+)
CPP	0.57(+)	0.65(+)
CPPex	0.51(+)	0.68(+)
%Pnc	−0.35	−0.13
FCSm	0.29	0.01
CPP/FCSm	0.57(+)	0.63(+)
JCSm	0.48(+)	0.28
CPP/JCSm	0.46(+)	0.58(+)
JCSm/FCSm	0.47(+)	0.51(+)
%Selfcit	−0.19	−0.63(+)

A '+' sign indicates that correlation is significant at a confidence level of 99%.

Table 3
Spearman's rank-correlation coefficients (r_S) of bibliometric indicators (1990–1994) and overall jury ratings in category A: application-oriented and category B: basic research

Period 1990–1994	Category A (r_S)	Category B (r_S)
Indicator		
P	0.27	0.25
C	0.35	0.51(+)
CPP	0.30	0.64(+)
CPPex	0.27	0.72(+)
%Pnc	−0.22	−0.13
FCSm	0.23	0.31
CPP/FCSm	0.29	0.66(+)
JCSm	0.36	0.50(+)
CPP/JCSm	0.13	0.50(+)
JCSm/FCSm	0.38	0.53(+)
%Selfcit	−0.06	−0.67(+)

A '+' sign indicates that correlation is significant at a confidence level of 99%.

mentioned indicators. For these indicators (except for the absolute number of citations) higher and more significant correlations are found for programmes in basic physics research (category B) than for programmes with a stronger technological component (category A). The lower correlations found for the application-oriented programmes in category A between most indicators based on actual citation rates (CPP, CPPex, CPP/FCSm and CPP/JCSm) and the jury ratings, are in agreement with earlier results in studies on the relation between bibliometrics and fields of technological research (see for instance le Pair, 1988; van Els et al., 1989). In this respect, however, also factors related to the peer review process should be taken into account. In the case of category A, the jury judged programmes belonging to a large number of subfields (materials science, semiconductor physics, applied physics), whereas programmes in the basic physics category (B) were much more coherent. This additional factor in category A of judging programmes belonging to a broad spectrum of subdisciplines, is reflected by the jury ratings which show a larger dispersion in category A than in category B.

As might be expected relatively low correlation coefficients are found for the 'only-journal-based' indicators (JCSm, for the journals used by the research group; and FCSm, for the journals of the field as a whole), especially in category B. It shows that the status of the journals involved, as reflected by the impact of the journals, do not correspond very well with the quality of a research programme as perceived by peers. However, in the application oriented

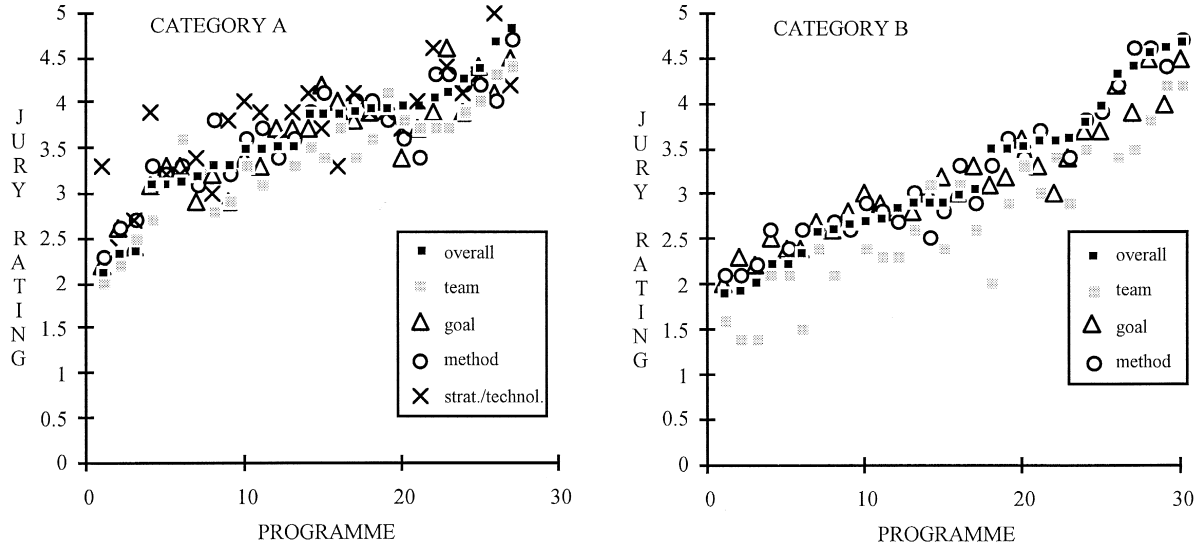

Fig. 3. Jury ratings for overall judgements and the four (category A:application-oriented) and three (category B: basic research) specific criteria of 56 programmes in condensed matter physics.

category A the impact of the journal set (JCSm) correlates slightly with jury ratings.

In category B the indicator expressing the relative impact of the journals in which a group publishes, comparing the *journal-based world average citation rate with the field-based world average citation rate* (JCSm/FCSm), correlates significantly with jury ratings, as is the case with the indicator expressing the percentage of selfcitations (%Selfcit). A negative correlation found between the percentage of self-citations and jury ratings in category B means that in the case of basic physics programmes in condensed matter physics lower levels of self citation correspond significantly with higher jury ratings.

When bibliometric indicators are calculated for the shorter and more recent period 1990–1994 (Table 3), for programmes in the application oriented category A correlations with jury ratings decrease compared with the correlations found for the period 1985–1994 and significant correlations disappear. In the basic physics category B, most indicators which correlate significantly in the whole period 1985–1994 also correlate significantly in the period 1990–1994.

3.4. Correlations between bibliometric indicators and specific criteria

Next to the *overall* assessment, the juries also gave ratings for four specific criteria in their assessment: *team*, *goal*, *method*, and *strategical and technological relevance*.

The criterium *team* consists of an assessment of the researchers and the research team. The criterium *goal* includes the choice of the problem and the relevance for the advancement of basic physics knowledge and 'urgency' of the programme. The criterium *method* concerns originality of methods

Table 4
Spearman's rank-correlation coefficients (r_S) of bibliometric indicators and jury ratings for the overall judgements and specific criteria. (A: application-oriented; B: basic research)

Indicator	Jury rating[a]				
	Overall (r_S)	Team (r_S)	Goal (r_S)	Method (r_S)	S&T (r_S)
Category A					
P	0.37	0.44	0.34	0.27	0.39
C	0.54(+)	0.55(+)	0.42	0.40	0.50(+)
CPP	0.57(+)	0.52(+)	0.40	0.48(+)	0.44
CPPex	0.51(+)	0.47(+)	0.34	0.43	0.37
%Pnc	−0.35	−0.36	−0.22	−0.37	−0.34
FCSm	0.29	0.44	0.30	0.28	0.06
CPP/FCSm	0.57(+)	0.40	0.34	0.46(+)	0.52(+)
JCSm	0.48(+)	0.51(+)	0.37	0.47(+)	0.25
CPP/JCSm	0.46(+)	0.35	0.28	0.36	0.46(+)
JCSm/FCSm	0.47(+)	0.30	0.33	0.48(+)	0.33
%Selfcit	−0.19	−0.18	−0.11	−0.21	−0.08
Category B					
P	0.16	0.21	0.13	0.13	
C	0.47(+)	0.50(+)	0.41	0.39	
CPP	0.65(+)	0.67(+)	0.60(+)	0.54(+)	
CPPex	0.68(+)	0.68(+)	0.63(+)	0.55(+)	
%Pnc	−0.13	−0.08	−0.09	−0.10	
FCSm	0.01	0.02	0.07	0.01	
CPP/FCSm	0.63(+)	0.63(+)	0.60(+)	0.52(+)	
JCSm	0.28	0.29	0.25	0.26	
CPP/JCSm	0.58(+)	0.58(+)	0.54(+)	0.50(+)	
JCSm/FCSm	0.51(+)	0.48(+)	0.53(+)	0.44(+)	
%Selfcit	−0.63(+)	−0.59(+)	−0.64(+)	−0.53(+)	

[a]A '+' sign indicates that correlation is significant at a confidence level of 99%.

and contribution to extending basic physics knowledge. Finally *strategical and technological relevance* was judged as separate criterium for programmes in category A. Though all criteria contain elements of past performance and future prospects, the criterium *team* is most explicitly concerned with past performance, whereas the criteria *method* and *goal* represent a mix of both elements (past performance and future prospects).

The mutual dependence of these three and four specific criteria, respectively, for each programme is shown in Fig. 3. Linear correlation coefficients between the overall ratings and the criteria *team*, *goal* and *method* range from 0.83 to 0.95 with those for the *team* deviating most. Correlations between the overall ratings and the judgements for the specific criteria are slightly higher in category B than in the application oriented category A. This finding might also be related to the fact that programmes in the latter category concern more heterogeneous subfields than in the basic condensed matter physics category B.

A quite remarkable finding is that the correlation between the overall rating for programmes in the application-oriented category A and ratings for the *strategical and technological relevance* criterium is less ($r^2 = 0.65$) than between the other criteria. This might indicate that for category A, covering programmes with an emphasis on technological relevance, the jury still attaches a relatively great value to the basic scientific merits of a programme.

All criteria can be seen as representing different aspects of the broad concept of quality of scientific research. It may be expected, however, that aspects related to the first three criteria (competence of the team, relevance of the problem and originality of the methods) are more directly related to impact and visibility as measured in the international scientific literature than the criterium of strategical and technological relevance. To investigate this, in Table 4 Spearman's rank-correlation coefficients are calculated for bibliometric indicators and each of the four specific criteria used by the juries. Taking the r_S values for the specific criteria we observe that, in general, the highest correlations are found between bibliometric performance indicators and the criterium *team*. This is the case for both categories, with again higher and more significant correlations in the basic physics category B. As explained above, the criterium *team* is indeed the most explicit criterium for the assessment of 'past performance' of a research group by peer review. It confirms the expectation that citation-based indicators relate well with past achievement. In category A (application-oriented research) the quality of the team as perceived by peers, shows a correlation with not only the C, CPP and CPPex indicator but also with the JCSm indicator, which is based on only the impact of the journals used by the research group.

For the 'basic physics' programmes (category B) lowest correlations are found between bibliometric indicators and the criterium *method*, for the 'applied physics' programmes (category A) lowest correlations are found between bibliometric indicators and the criterium *goal*.

It is shown for most indicators that the numerical values of the correlations for the specific criteria are to a certain degree interrelated. This may be explained by the mutual dependence of the separate specific criteria as shown by the jury ratings mentioned before (see also Fig. 3).

4. Conclusions

Undoubtedly, a lot has been published on the comparison between quantitative indicators and peer review. Nevertheless, few studies offer hard empirical with sufficiently broad differentiation in type of indicators and type of research. Part of these latter studies consider the correlation between indicators and peer review in the publishing process (e.g., Daniel, 1993, Korevaar and Moed, 1996). Other studies compare bibliometric indicators and peer judgements based on questionnaires among scholars or experts (e.g., Anderson et al., 1978; Martin and Irvine, 1981).

In this study detailed evidence based on a sample of 56 research programmes in condensed matter physics in the Netherlands has been analysed in which expert assessment was the basis for funding decisions concerning these programmes.

Our conclusions are summarized as follows.

(a) Positive and significant but no perfect correlations are found between a number of bibliometric indicators and peer judgements of research pro-

grammes in condensed matter physics in the Netherlands.

(b) At the level of overall peer judgements, we find the highest correlations between peer judgement and the average number of citations per publication (CPP, CPPex) and the citation averages normalised to world average (CPP/JCSm and particularly CPP/FCSm).

The latter two indicators compare the average number of citations per publication with the world standards of the corresponding journals and fields. These indicators containing normalised citation rates are included because citation characteristics may vary considerably among journals and fields, which may be a main cause for differences between citation averages. It is shown that both indicators correlate with peer judgements at almost the same degree as the average number of citations. For both categories of programmes it is shown that the mean citation rate normalised to the world-wide field average gives a slightly higher correlation with peer ratings than the mean citation rate normalised to the world-wide journal average.

(c) The impact of journals in which is published by a programme, as reflected by the mean journal citation rates, alone does not correlate well with the quality of these programmes as perceived by peers. From a bibliometric point of view this might be expected, as the mean citation rate of a journal or a journal category is only partly related to its (relative) impact. For another part it also reflects differences between citation patterns of the subfield(s) concerned. Moreover, the impact of research published within one journal may differ largely. The low correlations found in this study between peer ratings and the average citation rate of the journals used by the research group (JCSm) and of the journals of the fields involved (FCSm), support conclusions that journal impact factors alone do not provide a sufficient basis for assessing research performance.

(d) Correlations between bibliometric indicators and expert judgements are higher in the case of 'curiosity driven' basic research than in the case of 'application driven' research. On the one hand this might be explained by a diminished bibliometric visibility of fields of applied science, on the other hand the circumstance of peer reviewing programmes belonging to a large number of subfields, as was the case in the 'applied physics' category (A), may have played a role.

(e) At the level of the four specific criteria used by juries, the highest correlation is found between ratings for bibliometric indicators and the criterium 'team'. A.M. Weinberg mentioned this criterium—the assessment of the competency of researchers and the research team—as one of the two 'internal criteria' for scientific choice which are most often applied when a panel decides on a research grant (Weinberg, 1963). The results presented in this study confirm that judgement of the team, as an important element in peer review closely related to the assessment of past performance, correlates significantly to a number of citation-based indicators. From the analysis of the jury ratings on each criterium, however, it appeared too that a high overall rating is not uniquely determined by having a good research team. Programmes are judged on other aspects as well.

(f) When bibliometric indicators are calculated for a shorter period, correlations with peer judgements tend to decrease, especially in the case of application oriented research.

(g) A negative correlation is found between the percentage of self-citations (%Selfcit) and jury ratings for basic physics programmes in category B, The level of selfcitation can be interpreted negatively as giving an indication of the relative isolation of a group. However, more positively it can be explained as an indication of the uniqueness of the research carried out by the programme. Lower jury ratings given to programmes with a high level of selfcitation, as is found in this study for programmes in basic condensed matter physics, may indicate that selfcitation here is more related to isolation than to uniqueness.

With respect to the level of agreement between different bibliometric indicators and peer judgements, it should be noted that at the basis of this analysis are data from two different evaluation studies. As explained before programmes and the researchers belonging to them identified in the peer evaluation procedure and in the bibliometric analysis were more or less identical, but did not match perfectly. A more perfect match between programmes analysed might be obtained in future studies, for instance by tuning in more precise on scientists, research outcomes and publications which are evalu-

ated by the different methods. In such cases correlations between bibliometric indicators and peer opinion might even prove to be stronger.

The numerical values of the correlations between bibliometric indicators and expert assessments, although in a number of cases clearly significant, also indicate that the two assessment methods are certainly not completely dependent. The empirical findings clearly indicate that both assessment methods offer supplementary information about the performance of research groups or programmes. Though in our view a final quality judgement on a research group can be given by peers only, on the basis of their expert insight into the content and nature of the work conducted by that group, we also conclude that bibliometric indicators may provide important additional information to peers evaluating research performance. Thus, bibliometric indicators act as a *support* to peer review, for instance in cases of incorrect or biased views of peers on a group's scientific quality.

The supplementary information offered by bibliometric indicators may also be relevant for other science policy purposes. Incidental, exceptional differences between results obtained by both methods may yield interesting cases which can function as an eye-opener. Often such cases are interesting topics for further research. It may point at limitations of the bibliometric methodology, but also to biases in the peer review procedure, for instance because the research to be evaluated is not identical with the expertise of the members of a peer college (for a more extensive discussion on the relation between peer review and bibliometric indicators, see van Raan, 1996).

A final remark is related to the time span concerned in both evaluation procedures analysed in this study. As described above, in the FOM Condensed Matter peer evaluation procedure not only past performance but also future prospects of the programmes have been assessed. The latter aspect might even be more important in the judgement of programmes in the application-oriented category because of the stronger emphasis put on strategical–technological relevance. Bibliometric indicators used in this study, though they may form a part of predictive models (Kostoff, 1995), are by necessity based on an assessment of past performance. Therefore, it might be interesting to include in further studies several time spans in order to test more extensively the (prospective and retrospective) reliability of bibliometric evaluations and peer assessments.

Acknowledgements

We gratefully acknowledge the Netherlands Organisation for Scientific Research (NWO) for financing the major part of the datasystem used in this study. The authors acknowledge Dr. H.F. Moed (CWTS) for his contributions to the bibliometric study of Dutch academic physics, especially the analysis of the 'non-Netherlands' publications

References

Anderson, R.C., Narin, F., McAllister, P., 1978. Publication ratings versus peer ratings of universities. J. Am. Soc. Info. Sci. 29, 91–103.

Bayer, A.E., Fulger, F.J., 1966. Some correlates of a citation measure of productivity in science. Sociol. Educ. 339, 381–390.

van den Beemt, F.C.H.D., van Raan, A.F.J., 1995. Evaluating research proposals. Nature 375, 272.

Chang, K.H., 1975, Evaluation and survey of a subfield of physics: magnetic resonance and relaxation studies in The Netherlands, FOM-Report No. 37175, Utrecht.

Cole, S., Rubin, L., Cole, J.R., 1978, Peer Review in the National Science Foundation, National Academy of Sciences, Washington, DC.

Daniel, H.-D. (1993), Guardians of science: fairness and reliability of peer review, VCH, Weinheim.

Garfield, E., 1979, Citation Indexing—Its Theory and Applications in Science, Technology and Humanities, Wiley, New York.

Cole, S., Cole, J.R., 1967. Scientific output and recognition: A study in the operation of the reward system in science. Am. Sociol. Rev. 32, 377–390.

van Els, W.P., Jansz, C.N.M., le Pair, C., 1989. The citation gap between printed and instrumental output of technological research: the case of the electron microscope. Scientometrics 17, 415–425.

Horrobin, D.F., 1990. The philosophical basis of peer review and the suppression of innovation. J. Am. Med. Ass. (JAMA) 263, 1438–1441.

Kostoff, R.N., 1995. Research requirements for research impact assessment. Res. Pol. 24, 869–882.

Korevaar, J.C., Moed, H.F., 1996. Validation of bibliometric indicators in the field of mathematics. Scientometrics 37, 117–130.

van Leeuwen, Th.N., Rinia, J., van Raan, A.F.J., 1996, Bibliomet-

ric Profiles of Academic Physics Research in The Netherlands, Report CWTS 96-09, Centre for Science and Technology Studies, Leiden.

Martin, B.R., Irvine, J., 1981, Internal Criteria for Scientific Choice: An Evaluation of Research in High-Energy Physics Using Electron Accelerators, Minerva XIX, 408–432.

Martin, B.R., Irvine, J., 1983. Assessing basic research. Some partial indicators of scientific progress in radio astronomy. Res. Pol. 12, 61–90.

Martin, B.R., 1996. The use of multiple indicators in the assessment of basic research. Scientometrics 36, 343–362.

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. Res. Pol. 14, 131–149.

Moed, H.F., de Bruin, R.E., van Leeuwen, Th.N., 1995. New bibliometric tools for the assessment of national research performance: Database description overview of indicators and first applications. Scientometrics 33, 381–422.

Moed, H.F., Hesselink, F.Th., 1996. The publication output and impact of academic chemistry research in the Netherlands during the 1980's: bibliometric analyses and policy implications. Res. Pol. 25, 819–836.

Narin, F., (1976), Evaluative Bibliometrics: The Use of Publication and Citation Analysis in the Evaluation of Scientific Activity, National Science Foundation, Washington, DC.

NSF (1996), National Science Board and National Science Foundation Staff Task Force on Merit Review Discussion Report, NSB/MR-96-15, National Science Foundation, Washington, DC.

NSTC (1996), Assessing Fundamental Science: A Report from the Subcommittee on Research, National Science and Technology Council Committee on Fundamental Science, Washington, DC.

Nederhof, A.J., (1988), The Validity and Reliability of Evaluation of Scholarly Performance. In: van Raan, A.F.J. (Ed.), Handbook of Quantitative Studies of Science and Technology, Elsevier/North-Holland, Amsterdam.

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, 333–350.

Nederhof, A.J., van Raan, A.F.J., 1989. A validation study of bibliometric indicators: The comparative performance of cum laude doctorates in chemistry. Scientometrics 17, 427–435.

le Pair, C., (1988), The citation gap of applicable science, In: van Raan, A.F.J. (Ed.), Handbook of Quantitative Studies of Science and Technology, Elsevier/North-Holland, Amsterdam, ISBN: 0-444-70537-6.

van Raan, A.F.J., van Leeuwen, T.N. (1995). A Decade of Astronomy Research in The Netherlands, Report CWTS 95-01, Center for Science and Technology Studies, Leiden.

van Raan, A.F.J., 1996. Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. Scientometrics 36, 397–420.

Rinia, E.J., de Lange, C., 1991. The Dutch publication output in physics 1979–1988, FOM-Report No. 68726, Utrecht.

Schubert, A., Glänzel, W., Braun, T., 1989. Scientometric datafiles. A comprehensive set of indicators on 2649 journals and 96 countries in all major science fields and subfields, 1981–1985. Scientometrics 16, 3–478.

Weinberg, A.M. (1963), Criteria for Scientific Choice, Minerva I, 159–171.