Sven Hemlin
Henry Montgomery

PEER JUDGMENTS OF SCIENTIFIC QUALITY: A CROSS-DISCIPLINARY DOCUMENT ANALYSIS OF PROFESSORSHIP CANDIDATES

Introduction

On certain occasions in a scientist's life, evaluations of his/her scientific achievements become particularly important, and this is when he/she applies for an academic position. The evaluations are typically made by peers in the scientific world. The procedure of appointing professors and filling other academic positions is essentially the same in many countries, inasmuch as the appointments largely are based on peer evaluations of the applicants' scientific achievements. It goes without saying that these evaluations should be accurate. The resulting academic appointment sometimes represents an investment in a major research project - the scientific achievements of a researcher in the remainder of his/her scientific career.

The present study concerns evaluations of candidates for professorships in a national cross-disciplinary sample of open positions. In particular, we focus on the scientific quality judgments made by the evaluators.

The data in the present study consist of peer evaluation documents concerning the applicants for an open position. These documents end up in recommending a certain candidate for the pertinent position. In Sweden, this recommendation is usually followed by the authorities that are responsible for professorial appointments.

Professorial evaluation documents have a unique value as a basis for studies of quality judgments within different disciplines. In Sweden, these evaluations are made in all research areas. They often give detailed descriptions and evaluations of research efforts in the area related to the open position. For these reasons, professorial evaluation documents may be used for examining similarities and differences in scientific quality judgments across a broad spectrum of academic disciplines.

The present study was part of a research programme on how scientists conceive scientific quality. In our first study (Hemlin & Montgomery, 1990) 22 professors from various disciplines were interviewed about their views of scientific quality. The results of that investigation indicated that judgments of a research effort could be described in terms

Science Studies, Vol. 6 (1993), No. 1, 19—27.

of a number of attributes or scientific values (e.g., originality, stringency, correctness) which may be combined with a number of aspects or phases of the research effort (e.g., problem, method, results). This distinction could also be made of criteria of scientific quality suggested in the literature (e.g., Chase, 1970; Frantz, 1968; Kuhn, 1977). We also identified several factors which interact with the evaluation of the research effort (e.g., the type of quality indicators used, the research policy connected with the evaluations, characteristics of the researcher and his/her environment). In a subsequent study (Hemlin, in press), our conceptual system was used in a questionnaire on how scientists view scientific quality. The questionnaire was answered by a sample of 224 scientists.

Our two previous studies indicated that scientists from different sectors of the academic world use approximately the same conceptual system when they describe what they mean by scientific quality. However, the stress laid on particular components of the system (e.g., specific attributes or aspects) may vary across disciplines.

In the present study we have shifted our attention from how scientists view the concept of scientific quality (theoretical level) to how they actually make judgments of scientific quality (practical level). The natural question then is how the two levels correspond to each other. To what extent will the similarities and differences in how scientists from different disciplines describe their views of scientific quality reoccur when actual quality judgments are examined? Another aim of the present study was to investigate how the decision was made when selecting the candidate for the professor position. The presentation of the findings of the latter purpose will be dealt with in a forthcoming paper (see also, Montgomery & Hemlin, 1991).

Method

Documents. Professorial evaluation documents concerning 31 professorships were collected from the University of Göteborg and Chalmers University of Technology in Göteborg. All documents except one concerned the years 1981-1984. A document from 1975 was included in the sample to increase the representation of appointments in the social sciences, which happened to be under-represented in the period 1981-1984. Original documents were copied at the registry of the two universities in the beginning of 1985. The 31 professorships were distributed as follows across faculties: the faculty of humanities (4 cases), the medical faculty (10 cases), the natural sciences faculty (7 cases), the social sciences faculty (3 cases), and the technical faculty (7 cases).

Procedure. All value statements were marked in each document, i.e., all judgments in which a positive or negative evaluation was made of a research effort or the researcher him/herself. Also, descriptive statements regarding the qualifications of the applicant were marked (e.g., teaching experience, number of supervised doctorate students reaching a Ph.D., number of scientific articles, number of citations). Each delimited judgment or statement was coded into the following four overall categories, (a) the object, that is which kind of object was focused on by the peer's judgment (a single paper/research effort or the research/researcher as a whole), (b) the aspect, i.e., if the problem, the method, the theory, results, reasoning or writing style or no aspect was focused on, (c) the attribute, i.e., the criteria of good science which were associated with the aspect (e.g., stringency, novelty, beauty). This category also included various descriptive statements such as the number of published papers and the number of supervised doctoral students reaching a Ph.D., (d) the value, which could be negative, positive or neutral (e.g., "high quality" was coded positively (+), "to question the chosen method" was coded negatively (-) on the attribute Correctness, and "written a textbook" was coded neutral. The coding procedure was applied in accordance with a manual in which each coding category was defined and exemplified (see Montgomery & Hemlin, 1991). The coding manual comprised 64 categories. In total,

some 8 000 statements were coded into the four main categories listed above (i.e., each statement was coded in four ways). This means that about 30 000 codings were carried out. Codings were performed by three research assistants one of whom was the first author of this paper. The coding reliability was tested for a sample of documents by examining the agreement between different judges who had coded the same data. The average interjudge agreement was 80 %.

The coded statements were divided into two groups based on which part of the evaluation document the judgments belonged to. The first and larger part normally consisted of a description of each of the applicants' qualifications and research activities (single judgments). The single judgments section is often preceded by an introductory section, which was not analysed in the present study. In the second and final part of the document the applicants were compared with each other to achieve a decision of the "best" applicant (comparative judgments). Very often the three "best" applicants were ranked in this second part of the peer evaluation. In a few cases, peers left out the single judgments part and immediately started to compare the candidates' qualifications.

Results

Results will be presented in three main sections. Firstly, some quantitative data (e.g., number of pages, number of applicants) of the evaluation documents are described. Secondly, findings across research areas are presented. Thirdly, results concerning differences between research areas are shown.

Quantitative data on the documents

The volume of peer evaluation documents varies considerably between different research areas. The longest documents were written in the humanities and the social sciences. In these areas a peer writes on the average approximately seven times more about every applicant than a peer does in the technical sciences and about three and two times as much, respectively, as is true in medicine and the natural sciences. The average number of pages for peer documents across all professorships were 34.5 pages and 6.4 pages per applicant (the mean number of applicants per case was 5.4). In 12 cases peers jointly evaluated the merits of the applicants, but the majority of cases consisted of separate evaluations of the candidates. The former variant was frequent in the technical sciences, but also occurred in other sciences, except for the social sciences. The number of items in single judgments was 8 267 and in comparative judgments 2 401.

Overall pattern

Objects. In the first part of the document (single judgments) evaluations of individual scientific works were as common as evaluations of the total production (53.4% and 46.6%, respectively). However, in the second part of the document in which candidates are compared, the total research production or the candidate him/herself is focussed on (84.4%).

Aspects. The majority of judgments were non-specific with respect to aspects of the research effort (63.1% in single judgments and 79.1% in comparative judgments). More aspects were specified in single judgments than in comparative judgments. Methods, Problems and Results were the overall three most often mentioned aspects. In single judgments Methods were noticed in 12.3% of the judgments made. For comparative judgments, the most often mentioned aspect was the Problem (6.4%). The Theory aspect was rarely mentioned in either case. Methods, Reasoning and Writing Style drew less attention when evaluators compared candidates than when researchers were evaluated individually.

Attributes. The three most common attributes were in order Stringency (13.5% in single judgments and 3.5% in comparative

Table 1. Percentages of aspects in judgments of single researchers in specific research areas.

Aspect	Research Area				
	Hum	Med	Nat	Soc	Tech
No aspect specified	53.7	71.3	67.9	53.3	74.5
Problem	7.7	3.0	6.7	6.7	1.9
Method	8.5	10.3	10.7	10.5	8.0
Theory	4.4	0.7	2.2	7.7	4.3
Results	6.0	10.5	7.9	5.7	3.9
Reasoning	10.1	1.9	2.3	8.7	3.0
Writing Style	9.6	2.3	2.3	7.4	4.4
Sum	100.0	100.0	100.0	100.0	100.0

Note. Hum = Humanities, Med = Medical sciences, Nat = Natural sciences, Soc = Social sciences, Tech = Technical sciences.

judgments), Novelty/Originality (9.6% and 6.5%, respectively) and Productivity (6.6% and 8.2%, respectively). It may be noted that the three attributes all concerned the applicants' scientific production. Educational experience followed next in frequency. Another interesting finding was that Breadth (3.2% and 6.0%, respectively) of research was more frequently mentioned than Depth (1.1% and 1.7%, respectively).

Values. Across research areas 75% of all judgments of single researchers were positive and 25%, were negative. The corresponding figures for comparative judgments were 81.7% and 15.3%, respectively. It may be noted that this finding is not perfectly equivalent to the so-called golden section hypothesis of relations between positive and negative judgments being 62:38 (see Benjafield & Adams-Webber, 1976).

Combinations of aspects and attributes

A large number of combinations of attributes and aspects occurred in the evaluation documents (in all 44 unique combinations excluding combinations with General Evaluative Statement). Stringent Writing Style (10.3%), Stringent Methods (7.5%), and New/Original Results (2.9%) were the three most common combinations of aspects and attributes.

Differences and similarities among research areas in stress on components of scientific quality

In this section we report data from specific research areas on how much different quality components were stressed. Only single judgments were analysed since the volume of data for comparative judgments in specific areas were judged to be insufficient to allow reliable conclusions.

Objects. Within the humanities and the social sciences single judgments of applicants typically concerned individual papers or research efforts (70.8% and 60.9%, respectively) rather than the total production of the applicants. Judgments within the other disciplinary areas by and large were equally distributed between these two categories although in the natural sciences judgments somewhat more often relied on the total research production or the researcher him/herself (medical sciences 51.6%, natural sciences 37.2%, and technology 53.4%).

Aspects. In most statements no specific aspect of the research effort was focused on (see Table 1). However, in the humanities and the social sciences almost half of the judgments were linked to specific aspects. Reasoning and Writing Style were frequently used aspects in the humanistic and social sciences. Evaluators in the medical and natural sciences, instead, focused on

Table 2. Percentages of attributes in judgments of single researchers in specific research areas.

Attribute	Research Area				
	Hum	Med	Nat	Soc	Tech
Correctness	9.5	2.7	3.4	5.2	3.6
Importance	4.2	6.7	5.3	5.3	2.8
Novelty/Originality	8.9	9.6	11.2	9.6	5.2
Stringency	19.6	5.6	7.2	15.3	8.2
Intrascientific Relevance	1.5	3.0	4.2	2.3	0.9
Extrascientific Relevance	0.5	4.2	1.1	1.2	4.3
International Position	2.7	3.5	4.4	1.2	3.5
Relevance of Subject	3.2	0.8	2.4	2.7	5.6
Breadth	1.8	2.7	3.4	1.9	3.8
Depth	0.7	1.1	1.0	0.3	1.8
Activity/Productivity	5.3	6.1	4.6	3.2	5.3
Knowledge of Subject	2.8	2.9	2.2	4.6	3.3
Tutoring	1.4	4.2	1.1	1.6	3.6
Leadership of Research Proj.	1.2	2.1	1.2	1.7	0.9
Educational Experience	5.5	4.1	6.1	4.6	10.6
Practical/Administrative Exp. 3.5	3.5	8.6	5.3	4.7	8.0
Qualifications for Prof.ship	0.8	2.2	3.0	1.2	0.5
General Evaluative Stat.	17.7	11.7	11.4	17.8	13.5
Various	8.4	15.3	14.4	13.9	12.2
Sum	100.0	100.0	100.0	100.0	100.0

Note. Hum = Humanities, Med = Medical sciences, Nat = Natural sciences, Soc = Social sciences, Tech = Technical sciences.

Results and de-emphasized Theory. The technical sciences exhibited the lowest number of judgments regarding specific aspects.

Attributes. The distribution of attributes across the five research areas was similar across areas (see Table 2). Within the humanities Stringency was the most common attribute and very often linked to the aspects of Reasoning and Writing Style. Intrascientific Relevance was mentioned by evaluators in the humanities, social and natural sciences before Extrascientific Relevance, which was more frequently mentioned by medical and technical scientists. In the same vein, evaluators in the two latter areas were more inclined to use extrascientific qualifications in their judgments than were evaluators in the three remaining subject fields. Merits in education were frequently mentioned by peers in the technical area.

Comparison with previous data

The present results concerning general and subject area specific emphases on particular aspects and attributes were compared to three previous data sets (see Tables 3 and 4). The first data set (Hemlin & Montgomery, 1990) was based on interviews conducted with 22 Swedish professors from different research areas covering the humanities (including theology), the natural sciences (including medicine, the dental faculty and technology), the social sciences (including law) and interdisciplinary research. Among aspects Method, Problem and Results were mentioned more than others, and among attributes Novelty, Correctness and Stringency were emphasized.

The second study (Hemlin, in press) reported two data sets, free answers and ratings, obtained by means of a questionnaire mailed to a random sample of 400 Swedish researchers (response rate 56%) from the five research areas, i.e., the humanities, medicine, the natural sciences, the social sciences and technology.

Table 3 shows that researchers were united in mentioning Methods, Problems and Results in connection with research quality in all studies. Only in data from ratings was

Table 3. Emphasis on specific aspects and attributes in four data sets.

Emphasized aspects or attributes	Hemlin & Montgomery, 1990 (interview)	Hemlin, (1993) (free answers)	Hemlin, (1993) (ratings)	Present study (single judgm.)
Three most emphasized aspects	Method Problem Results	Method Problem Results	Reasoning Results Method	Method Results Problem
Three most emphasized attributes	Novelty Correctness Stringency	Novelty Stringency Correctness	Correctness Stringency Depth	Stringency Novelty Activity/ Productivity
Emphasis on Breadth vs. Depth	Breadth Depth	Breadth Depth	Depth Breadth	Breadth Depth
Emphasis on intrasc. vs. extrasc. relevance	Intrasc. Extrasc.	Extrasc. Intrasc.	Intrasc. Extrasc.	Intrasc. Extrasc.
Three most emphasized combinations of aspects and attributes	Corr. Method New Results String. Problem	String. Method Corr. Method New Problem	Corr. Method Corr. Results Corr. Reasoning	String. Method String. Writ. St. New Results

Note. Including data from comparative judgments.

there an exception, in that Reasoning was emphasized before Results. The most favoured attributes were Novelty, Stringency, and Correctness in previous studies. Also, Depth was rated high in the questionnaire study (Hemlin, in press). In the present study, Activity/Productivity of the researcher occured more frequently than Correctness. Breadth was generally stressed before Depth, except for the rating data. Intrascientific Relevance and Extrascientific Relevance were equally stressed in the present study. However, the former attribute was more emphasized than the latter in the interview and rating data. In the free answers, the results were reversed. In combinations of aspects and attributes Stringent or Correct Methods were the most frequently mentioned in all four data sets. New Results together with Stringent Writing Style in the present study were the next most frequent combinations. Thirdly, New or Stringent Problem was also stressed in interviews and the free answers to the questionnaire on how scientists view scientific quality.

The three data sets on differences between "soft" and "hard" sciences presented in Table 4, consistently show that Theory, Reasoning, Writing Style and to some extent Problems were stressed by researchers in the "soft" sciences when making judgments on scientific quality. The most emphasized attribute by scientists in these research areas was Stringency. Researchers in the "hard" areas were inclined to stress international relations as an indicator of research quality.

Discussion

Components in scientific value judgments

In general, the results support the usefulness of the fourfold distinction between objects, aspects, attributes, and values as components in judgments of scientific quality. Across research areas the following findings were obtained with respect to each of these components. Firstly, peers shift attention

Table 4. Differences between "soft" and "hard" sciences in emphasis on specific aspects and attributes in three data sets.

Type of difference	Hemlin, (1993) (free answers)	Hemlin, (1993) (ratings)	Present study (single judgm.)
Aspects more emphasized in "soft" sciences	Theory Reasoning	Writing Style Problem Reasoning	Reasoning Theory Writing Style
Attributes more emphasized in "soft" sciences	—	Stringency	Stringency
Attributes more emphasized in "hard" sciences	International relations	International relations	International position

Note. Only statistically significant differences ($p<.05$) listed in Hemlin (1993).

from single papers to the full production when they evaluate candidates and compare candidates, respectively. Secondly, more than two thirds of the judgments do not refer to any specific aspect of the research effort in general. The single most frequently mentioned aspect is Method. Thirdly, three attributes are common, viz., Novelty/Originality, Stringency, and Productivity. Also, Breadth is more frequent than Depth. Fourthly, peers generally make positive judgments. Only about one fourth of all judgments are negative.

It might be concluded that scientists from different areas to a large extent use the same criteria when they evaluate scientific achievements. To the extent that this conclusion is correct, it would be a fruitful task for future research to search for general principles for the conceptual structures underlying judgments of scientific quality. Such research might be helpful for creating cross-disciplinary standards for science evaluations.

Differences between two scientific traditions

Also, there were interesting differences among research areas in the stress laid on particular subcategories of objects, aspects, and attributes. Generally, these differences were consistent with the distinction between "soft" sciences and "hard" sciences. Firstly, "soft" science peers write twice as much about applicants. Secondly, peers in the "soft" sciences make more judgments about individual papers than about the total production, while their colleagues in the "hard" sciences share their attention equally between these two categories. It is not easy to explain why evaluators in the "soft" sciences write more about each candidate than their colleagues in the "hard" sciences. Perhaps, research in the "hard" sciences is easier to evaluate due to the availability of established theories and more exact results. In sciences with competing theories and results less easy to interpret, it might be more difficult to make research evaluations. However, the focus on individual papers in the "soft" sciences may be explained by the fact that scientific work repeatedly is published as books in the humanities and in some social sciences (Garfield, 1979; Line, 1981). This means that the evaluator in the "soft" sciences can concentrate on fewer works, while his/her colleague in the "hard" sciences has more works (e.g., scientific articles)

to examine, leading to less detailed evaluations of each article. The reliance on the amount of international contacts of the candidates as a sign of scientific quality from the "hard" science evaluators, might be interpreted as a support for this conclusion.

Thirdly, "soft" scientists specified aspects of the research effort, mostly Writing Style and Reasoning, in half of the judgments, while "hard" scientists made judgments without specifying aspects. The latter peer group de-emphasized the Theory aspect. Again, we find support for the notion (see Hemlin & Montgomery, 1990) that theories are established to a great extent in the "hard" sciences. Hence, theory aspects of research are not focussed on. This result also support an interpretation that most disciplines of the "hard" science group belong to Kuhn's "normal" or paradigmatic sciences. Reversely, scientists in the "soft" and pre-paradigmatic sciences stressed the Theory aspect in evaluations, since well founded and comprehensive theories are lacking in this developmental stage (Kuhn, 1970). In the same vein, the focus on Reasoning and Writing Style can be explained. In absence of definite Results and established Theories, scientific development proceeds, to a great extent, through discourse in the "soft" sciences. Therefore, Reasoning and Writing Style is emphasized. Also, the distinction made by Whitley (1978) between "restricted" and "configurational" sciences accords fairly well with our results. The "restricted" sciences are characterized by sharing common theoretical ideals and basic conceptual assumptions, besides being task specific and using mathematical formalisms. This description is in line with our finding that the "hard" sciences de-emphasized the Theory aspect. The "configurational" sciences match to some degree our "soft" sciences in that objects studied in these disciplines are approached from competing theoretical perspectives. This characteristic goes along with the emphasis on Theory in "soft" sciences. Another feature of "configurational" sciences, described by Whitley (1978), is the great varieties of definitions and analyses of objects which occur in this type of sciences. Our finding that Reasoning and Writing Style are stressed in "soft" sciences may be interpreted as showing the importance of verbal descriptions and delineations of the objects studied in these disciplines.

Fourthly, Intrascientific Relevance was frequently mentioned by "soft" and natural scientists, while Extrascientific Relevance occured more often with medical and technical peers. The last mentioned finding is not surprising, since the medical and technical sciences include several applied scientific subject fields. These findings corroborate largely the conclusions drawn by Hemlin (in press) in the sense that the "soft" and "hard" sciences differ with respect to how they view scientific quality.

Validity of findings

By and large, the present results agree with the findings in our previous studies of how scientists view scientific quality. Across and within research areas the same aspects and attributes tend to be stressed as in the previous studies. The agreement with previous research findings simultaneously support the validity of our previous and present findings. That is, the agreement implies that scientists appear to be proficient in verbalizing the criteria (previous studies) they actually use when evaluating scientific achievements (present study).

REFERENCES

Benjafield, J., & Adams-Webber, J.
1976 "The golden section hypothesis." British Journal of Psychology, 67: 11–15.

Chase, J.T.
1970 "Normative criteria for scientific publication." The American Sociologist, 5: 262–265.

Frantz, T.T.
1968 "Criteria for publishable manuscripts. Personnel and Guidance Journal, 47: 384–386.

Garfield, E.
1979 Citation indexing: Its theory and application in science, technology and humanities. New York: John Wiley and Sons.

Hagstrom, W.
1965 The scientific community. New York: Basic Books.

Hemlin, S.
(in press) "Scientific quality in the eyes of the scientist. A questionnaire study" Scientometrics.

Hemlin, S., & Montgomery, H.
1990 "Scientists' conceptions of scientific quality." Science Studies, 3: 73–81.

Hemlin, S., Montgomery, H., & Johansson, U-S.
1985 "Föreställningar om vetenskaplig kvalitet". (Conceptions of Scientific Quality). (Rapport nr. 4). Göteborgs universitet, Psykologiska institutionen.

Kuhn, T.S.
1970 The structure of scientific revolutions (2nd. ed.). Chicago: The University of Chicago Press. 1977 The essential tension. Chicago: The University of Chicago Press.

Line, M.
1981 "The structure of social science literature as shown by a large-scale citation analysis." Social Science Information Studies, 1: 67–87.

Montgomery, H., & Hemlin, S.
1991 "Judging scientific quality. A cross- disciplinary investigation of professorial evaluation documents." Göteborg Psychological Reports (No. 4), 21.

Whitley, R.
1978 "Types of science, organizational strategies and patterns of work in research laboratories in different scientific fields". Social Science Information, 17: 427–447.

Sven Hemlin
Department of Psychology
University of Göteborg
P.O. Box 14158 S-40020 GÖTEBORG

Henry Montgomery
Department of Psychology
University of Stockholm
S-10691 Stockholm