Jointly published by Akadémiai Kiadó, Budapest and Springer, Dordrecht

Scientometrics, Vol. 77, No. 3 (2008) 415–432
DOI: 10.1007/s11192-007-1950-2

Do editors and referees look for signs of scientific misconduct when reviewing manuscripts? A quantitative content analysis of studies that examined review criteria and reasons for accepting and rejecting manuscripts for publication

LUTZ BORNMANN, IRINA NAST, HANS-DIETER DANIEL

ETH Zurich, Zurich (Switzerland)

The case of Dr. Hwang Woo Suk, the South Korean stem-cell researcher, is arguably the highest profile case in the history of research misconduct. The discovery of Dr. Hwang's fraud led to fierce criticism of the peer review process (at *Science*). To find answers to the question of why the journal peer review system did not detect scientific misconduct (falsification or fabrication of data) not only in the Hwang case but also in many other cases, an overview is needed of the criteria that editors and referees normally consider when reviewing a manuscript. Do they at all look for signs of scientific misconduct when reviewing a manuscript? We conducted a quantitative content analysis of 46 research studies that examined editors' and referees' criteria for the assessment of manuscripts and their grounds for accepting or rejecting manuscripts. The total of 572 criteria and reasons from the 46 studies could be assigned to nine main areas: (1) 'relevance of contribution,' (2) 'writing/presentation,' (3) 'design/conception,' (4) 'method/statistics,' (5) 'discussion of results,' (6) 'reference to the literature and documentation,' (7) 'theory,' (8) 'author's reputation/institutional affiliation,' and (9) 'ethics.' *None* of the criteria or reasons that were assigned to the nine main areas refers to or is related to possible falsification or fabrication of data. In a second step, the study examined what main areas take on high and low significance for editors and referees in manuscript assessment. The main areas that are clearly related to the quality of the research underlying a manuscript emerged in the analysis frequently as important: 'theory,' 'design/conception' and 'discussion of results.'

Received September 20, 2007

Address for correspondence:
LUTZ BORNMANN
ETH Zurich, Zähringerstr. 24, CH-8092 Zurich, Switzerland
E-mail: bornmann@gess.ethz.ch

0138–9130/US $ 20.00

Introduction

In accordance with MERTON'S [1973] norm of organised scepticism, research results in a manuscript have to undergo the peer review process before they count as 'scientific knowledge' in the fullest sense [ZIMAN, 2000]. For journals, there are usually two steps in the peer review process. The editors of a journal first look to see if a manuscript is consistent with the journal's stated purpose and priorities. Some journals publish research in specific areas, such as microbiology; some publish only groundbreaking research of more general interest. If the editors decide that a manuscript is a good fit for the journal, it is then sent on to external expert reviewers, or referees, for in-depth review [SENSE ABOUT SCIENCE, 2005]. The referees "assess the soundness of a manuscript's ideas and results, its methodological and conceptual viewpoint, its quality, its potential impact on the world of science" [CAMPANARIO, 1998, P. 182] and many other aspects to ensure that the quality of the manuscript meets the standards of the scientific community [OFFICE OF MANAGEMENT AND BUDGET, 2004].

According to FLETCHER & FLETCHER [2003] "readers believe that peer review helps them manage information by affirming the scientific validity of published articles" (p. 62). But, can the readers in fact expect this of journal peer review, considering the cases of scientific misconduct that have been disclosed in recent years (for example, Jon Sudbø and Jan Hendrick Schön) (for overviews, see [FOX, 1994; ODLING-SMEE & AL., 2007])? According to a piece in *Nature* [ANON, 2006A], the case of Hwang Woo Suk, the South Korean stem-cell researcher at Seoul National University, represents perhaps the highest profile case in the history of research misconduct. Dr. Hwang and colleagues at several institutions in South Korea and the United States published articles in the *Science* magazine in 2004 and 2005 that were called seminal breakthroughs in stem cell research (see [CYRANOSKI, 2006; NORMILE & AL., 2006]). Some months following publication, the articles were found to be based on fabricated evidence, despite the fact that they had gone through the extensive peer review process of *Science*.

The discovery of Hwang's fraud resulted in strong criticism of the peer review process of *Science*. MARTIN [2006] wrote, for example, "if the *Science* editorial staff had paid more attention to the science and less to the sensation, and if others had not leapt onto the bandwagon, the impact of this sorry affair might have been much less" (p. 607). And NORMILE & AL. [2006] reported that as a response, *Science* planned to conduct an internal investigation into its handling of the articles by Hwang and had also contacted members of its senior editorial board regarding possible modification of the peer review procedure. But the Hwang case did not just call into question the validity of the peer review process at *Science* but also the peer review system in general: "How could this have happened? Why didn't the peer-review process uncover the fabrication? Do we need to make changes in the way that we conduct and publish research?"

[CHO & AL., 2006, P. 614]. And, according to a piece in *Nature* [ANON, 2006B], "the Hwang fraud saga has already fuelled some misconceptions about how the combination of referees and journal editors actually works" (p. 118).

To answer the question as to why the journal peer review system did not detect scientific misconduct not only in the Hwang case but also in many other cases, an overview is needed of the criteria that referees and editors normally consider when reviewing a manuscript. Do referees and editors even look for indications of scientific misconduct in a manuscript? Or are they far more concerned with other issues when assessing manuscripts? What issues are these? For the present study we carried out a quantitative content analysis of research studies that examined editors' and referees' criteria for manuscript review and their reasons for accepting and rejecting manuscripts for publication. The goal of the analysis was, for one, to produce as complete a catalogue as possible of the different areas examined by peer reviewers in manuscript assessment and, for another, to identify the areas that editors and referees find more and less important when assessing manuscripts. In both evaluations, we were especially interested in the importance that editors and referees place on possible unethical behavior (scientific misbehavior) by authors of manuscripts submitted for publication.

Short reviews of editors and referees' assessment criteria and reasons for accepting or rejecting manuscripts in journal peer review are provided by BYRNE [1998], HIRSCHAUER [2004, PP. 70–71], MEADOWS [1998, PP. 180–183], and WELLER [2002, PP. 49–54, 92–96]. As these reviews describe only one (small) part of the existing literature and did not conduct quantitative content analyses of criteria and reasons in order to determine the underlying dimensions, there was a need to create a more comprehensive overview using quantitative content analysis techniques.

Methods

Research into studies

In a first step of our literature search, we researched published studies on editors and referees' assessment criteria and reasons for accepting or rejecting manuscripts for publication in journal peer review in the reference lists of the short reviews provided by BYRNE [1998, CHAPTER 11], HIRSCHAUER [2004], MEADOWS [1998], and WELLER [2002]. We searched for both publications (journal articles, monographs, collected works, etc.) and grey literature (Internet documents, institutional reports, case reports, etc.). In a second stage, to obtain keywords for the study search in computerized databases, we then prepared a bibliogram [WHITE, 2005] for the studies researched in the first step using RefViz, data visualization and analysis software (Thomson Reuters, Philadelphia, PA, USA). The bibliogram ranks by frequency the words included in the titles and abstracts of the studies researched. We used the words at the top of the

ranking list (such as peer review, manuscript, acceptance, rejection, reason, and criterion) for searches in computerized literature databases (including Web of Science, IngentaConnect, PubMed, Sociological Abstracts, ProQuest Digital Dissertations, PsycINFO, ERIC) and via Internet search engines (for example, Google, Ask Jeeves). In the final step of our literature search, we located all of the citing publications for a series of articles that looked at editors' and referees' assessment criteria or reasons for accepting or rejecting manuscripts in journal peer review and for which there are a fairly large number of citations in Web of Science.

Describing the studies for the quantitative content analysis

For the quantitative content analysis our search strategy yielded a total of 46 studies published between 1967 and 2006. The complete list of studies included is available upon request from the correspondence author. The majority of the studies investigated criteria used in assessing manuscripts and/or reasons for acceptance and rejection of submitted manuscripts in the social and behavioral sciences and in the field of public health. Only a few studies looked at other fields (such as chemistry). The studies analyzed data captured in the period from 1958 to 2002 (16 studies do not provide this information). The data in the studies were collected from referees (or based on their reviews) and/or editors. Two of the studies referred to *scientists* only, so that it was not apparent if these were referees or editors. The assessment criteria and/or reasons for acceptance or rejection of manuscripts were investigated in the studies using (1) surveys of editors or referees, (2) content analyses of referees' comments and editorial rejection letters, and (3) analyses of journal manuscript review forms.

Results

Main areas considered in manuscript assessment

In the first step of the analysis, to determine what areas editors and referees examine in journal peer review, all review criteria and reasons for acceptance and rejection found in the 46 studies were captured word-for-word. A total of 682 criteria and reasons were found. Criteria and reasons that were semantically similar and identical in content were combined to form one criterion or reason. This reduced the number of criteria and reasons to 542. In a second step of the analysis, based on the given criteria and reasons, we developed *inductively* a category system with nine main areas considered in manuscript assessment (for example, area 1: 'relevance of contribution,' area 2: 'writing/presentation'). In a third step, all criteria and reasons from the studies were assigned to these nine main areas.

Of the total of 542 criteria and reasons, 485 could be clearly assigned to one of the nine areas for assessment; 42 criteria and reasons could be assigned to two different areas each; and one reason was even assigned to three areas. A total of 13 criteria and reasons were too general to be assigned to a category and were not included in the further analysis (for example, 'recommendation regarding publishing'). The final total, including the multiple assignments but not including the criteria and reasons that could not be categorized, was 572. With this, there were approximately 64 criteria and reasons per area. To gain a better overview, the criteria and reasons within the areas were sorted once again in underlying dimensions within the areas (for example, 'relevance of topic to journal' within the area 'relevance of contribution'). One researcher on our team carried out the assignments of the criteria and reasons to the areas and the underlying dimensions. The assignments were checked by a second researcher of our team, and disagreements were resolved by consensus.

Table 1 shows the nine main areas to which editors und referees' evaluation criteria and reasons for acceptance and rejection of manuscripts could be assigned: (A) 'relevance of contribution,' (B) 'writing/presentation,' (C) 'design/conception,' (D) 'method/statistics,' (E) 'discussion of results,' (F) 'reference to the literature and documentation,' (G) 'theory,' (H) 'author's reputation/institutional affiliation,' and (I) 'ethics' (the nine main areas are highlighted in bold italic in the table). In the table the nine areas are listed in the above (A) to (I) order, which is descending order according to the number of assigned criteria and reasons. The number of criteria and reasons assigned to the nine areas ranged from 148 (A: 'relevance of contribution') to 10 (I: 'ethics'). Depending on the data examined by the individual studies (for example, criteria rated by referees, reasons for rejection, or reasons for acceptance), the criteria and reasons captured are stated in a positive, negative, or neutral form (see the categorizations in Table 1). For example, in the area 'relevance of contribution,' the reason "the topic selected was appropriate" is stated in a positive form, the reason "contains nothing new" is stated in a negative form, and the criterion "appropriateness of topic" is stated in a neutral form.

As Table 1 shows, for each main area considered in assessing a manuscript, up to six different underlying dimensions were used for grouping criteria and reasons that are similar in content. For example, the underlying dimensions in the area 'relevance of contribution' are: 'relevance of topic, in general,' 'relevance of topic to scientific advancement,' 'originality, newness,' 'contribution to practical process,' 'relevance of topic to journal,' and 'relevance of results.' Within the criteria and reasons in a main area stated in a positive, negative, or neutral form, the underlying dimensions are listed in descending order according to the number of criteria and reasons assigned to that dimension. The table shows the underlying dimensions and one to two examples of the criteria and reasons in those dimensions. The complete list of criteria and reasons assigned to the main areas and the underlying dimensions are available upon request from the correspondence author.

Table 1. Areas (e.g., 'relevance of contribution') and dimensions (e.g., 'relevance of topic, in general') to which editors and referees' assessment criteria and reasons for acceptance and rejection of a manuscript can be assigned (the dimensions of one area are listed in the columns in ascending order by the number of criteria and reasons)

(A) 'relevance of contribution' (148 criteria and reasons out of 45 studies; 36 positively, 47 negatively, and 65 neutrally formulated criteria and reasons)		
positively formulated criteria and reasons	negatively formulated criteria and reasons	neutrally formulated criteria and reasons
relevance of topic, in general (11 criteria and reasons; e.g., the topic selected was appropriate; important, timely, relevant, critical, prevalent problem)	**relevance of topic, in general** (12 criteria and reasons; e.g., low priority; unimportant or irrelevant topic)	**relevance of topic, in general** (24 criteria and reasons; e.g., relevance of subject; appropriateness of topic)
relevance of topic to scientific advancement (9 criteria; e.g., contribution: increment to the current literature (fills gaps in current body of knowledge); an advancement of knowledge)	**relevance of topic to journal** (12 criteria and reasons; e.g., inappropriate subject for journal; unsuited to the readership)	**relevance of topic to scientific advancement** (15 criteria and reasons; e.g., scientific importance of topic; pertinence to current research in the discipline)
originality, newness (8 criteria and reasons; e.g., new/novel treatment of subject; research offers a new perspective on an existing problem)	**originality, newness** (10 criteria and reasons; e.g., contains nothing new; idea not unique)	**originality, newness** (9 criteria and reasons; e.g., originality/novelty; the creativity of ideas in the article)
contribution to practical progress (5 criteria and reasons; e.g., practical, useful implications; informative and useful)	**relevance of topic to scientific advancement** (7 criteria and reasons; e.g., material well outside the mainstream; no significant addition to current body of knowledge)	**relevance of topic to journal** (9 criteria and reasons; e.g., the relevance of the article to the journal's focus; interest to readers)
relevance of topic to journal (2 criteria and reasons; e.g., manuscript content: it is on the same topic as a number of articles recently published in the journal)	**relevance of results** (4 criteria and reasons; e.g., results are inconclusive, incomplete)	**contribution to practical progress** (7 criteria and reasons; e.g., practical implications; contribution to practice)
relevance of results (1 criterion: positive findings)	**contribution to practical progress** (2 reasons; e.g., clinically not applicable)	**relevance of results** (1 reason: results)
(B) 'writing/presentation' (143 criteria and reasons out of 44 studies; 22 positively, 78 negatively, and 43 neutrally formulated criteria and reasons)		
positively formulated criteria and reasons	negatively formulated criteria and reasons	neutrally formulated criteria and reasons
writing style (13 criteria and reasons; e.g., the paper was well written; professional appearance)	**writing style** (32 criteria and reasons; e.g., writing style: incoherent, obscure, jargon, cluttered, bad tone; unprofessional appearance)	**writing style** (17 criteria and reasons; e.g., presentation: quality of writing; thoroughness)
quality of specific parts of manuscript (5 criteria and reasons; e.g., high quality abstract; purpose of research is clearly stated)	**quality of specific parts of manuscript** (28 criteria and reasons; e.g., subjects insufficiently described; tables/figures need clarification)	**quality of specific parts of manuscript** (15 criteria and reasons; e.g., clarity of problem, hypothesis, and assumptions; description of statistical analyses)
correctness (2 criteria: sentences are grammatically correct; sentences are properly punctuated)	**organization/length of manuscript** (10 criteria and reasons; e.g., lacks organization /needs reorganization; too long)	**organization/length of manuscript** (9 criteria and reasons; e.g., organization of manuscript; manuscript length)

Table 1. (cont.)

organization/length of manuscript (1 reason: deviation in length towards brevity)	**correctness** (5 criteria and reasons; e.g., defective tables or figures; errors in the article)	**publication guidelines** (2 criteria: presentation: conformance to publication guidelines; adherence to journal's stylistic guidelines)
publication guidelines (1 reason: guidelines followed)	**publication guidelines** (3 reasons; e.g., failure to follow guidelines)	
(C) 'design/conception' (92 criteria and reasons out of 43 studies; 19 positively, 40 negatively, and 33 neutrally formulated criteria and reasons)		
positively formulated criteria and reasons	negatively formulated criteria and reasons	neutrally formulated criteria and reasons
conceptual framework: logic and correctness (8 criteria; e.g., research seems unbiased in research design)	**conceptual framework: logic and correctness** (20 criteria and reasons; e.g., conceptual: pre-execution (e.g., conceptual basis for study poor or incomplete))	**quality and appropriateness** (16 criteria and reasons; e.g., adequacy of research design)
quality and appropriateness (6 criteria and reasons; e.g., design of study is adequate; thorough and complete)	**quality and appropriateness** (10 criteria and reasons; e.g. inadequate research design)	**conceptual framework: logic and correctness** (9 criteria; e.g., logical rigor; conceptualisation)
quality of sampling (2 reasons: sample size sufficiently large; good sampling)	**quality of sampling** (6 criteria and reasons; e.g., sample too small or biased; sampling problem)	**quality of sampling** (5 criteria and reasons; e.g., sample and setting: appropriateness)
generalizability (2 criteria; e.g., study has wide generalizability)	**generalizability** (4 criteria and reasons; e.g., pilot study research, with little evidence of generalizability)	**replicability** (2 criteria; e.g., replicability of research techniques)
replicability (1 criterion: replicability of the review (being able to arrive at the same conclusions as the author))		**generalizability** (1 criterion: generalizability and validity of results)
(D) 'method/statistics' (72 criteria and reasons out of 34 studies; 8 positively, 30 negatively, and 34 neutrally formulated criteria and reasons)		
positively formulated criteria and reasons	negatively formulated criteria and reasons	neutrally formulated criteria and reasons
correctness and appropriateness (3 criteria and reasons; e.g., accurate statistical data; methods are adequate to test the research questions)	**correctness and appropriateness** (19 criteria and reasons; e.g., inappropriate procedures/methodology; statistics inadequately handled	**correctness and appropriateness** (14 criteria and reasons; e.g., appropriateness of statistical analysis; data analysis and results: warranted assumptions and appropriate error rates)
method/statistics, in general (2 reasons: good methodology; good analysis)	**quality of operationalization and measurement** (6 reasons; e.g., measurement (e.g., measure is indirect, superficial, not the best); scores insufficiently reliable or unknown reliability)	**method/statistics, in general** (10 criteria and reasons; e.g., statistical analyses)
newness (2 reasons; e.g., novel, unique approach to data analysis)	**method/ statistics, in general** (4 reasons; e.g., poor methodological; poor analysis)	**quality of operationalization and measurement** (10 criteria and reasons; e.g., instrumentation and data collection; operationalization of key constructs)

Table 1. (cont.)

quality of operationalization and measurement (1 criterion: operational definitions are adequate)	**newness** (1 criterion: manuscripts without new data: it discusses a new statistical test or a new data collection technique and contains no new data)	
(E) 'discussion of results' (45 criteria and reasons out of 31 studies; 10 positively, 16 negatively, and 19 neutrally formulated criteria and reasons)		
positively formulated criteria and reasons	negatively formulated criteria and reasons	neutrally formulated criteria and reasons
correctness, adequacy, and objectivity (5 criteria; e.g., objectivity in reporting results; conclusions properly based in the results)	**correctness, adequacy, and objectivity** (14 reasons; e.g., interpretations/conclusions not warranted by data)	**correctness, adequacy, and objectivity** (8 criteria and reasons; e.g., logical rigor; reasonableness of conclusions)
clarity (3 criteria; e.g., existence and clarity of a take-home-message; existence of and persuasiveness in arguing for a well-articulated point of view)	**breadth of interpretation** (2 criteria; e.g., data presented with limited discussion of implications)	**discussion of results, in general** (5 criteria and reasons; e.g., discussion and conclusion)
breadth of interpretation (2 criteriaand reasons; e.g., depth (intensive examination of specific area))		**breadth of interpretation** (4 criteria; e.g., discussion and conclusions: derivation of implications/limitations; depth)
		clarity (2 criteria; e.g., clarity of conclusions/generalizations)
(F) 'reference to the literature and documentation' (27 criteria and reasons out of 26 studies; 3 positively, 10 negatively, and 14 neutrally formulated criteria and reasons)		
positively formulated criteria and reasons	negatively formulated criteria and reasons	neutrally formulated criteria and reasons
coverage of relevant literature (3 criteria and reasons; e.g., thorough, focused, up-to-date review of the literature)	**coverage of relevant literature** (10 criteria and reasons; e.g., ignorance of relevant literature; ignorant of previously published work on the same subject)	**coverage of relevant literature** (10 criteria and reasons; e.g., literature review: linkage to most relevant literature; mastery of relevant literature)
		reference to the literature and documentation, in general (4 criteria and reasons; e.g., literature review; use of bibliography)
(G) 'theory' (24 criteria and reasons out of 22 studies; 5 positively, 11 negatively, and 8 neutrally formulated criteria and reasons)		
positively formulated criteria and reasons	negatively formulated criteria and reasons	neutrally formulated criteria and reasons
newness, interest of theory (5 criteria and reasons; e.g., study proposes a new theory to explain existing research findings; research is of theoretical interest and importance)	**Contribution to/importance of theory** (6 criteria and reasons; e.g., direct replications that added no new dimension to theory)	**theory, in general** (5 criteria and reasons; e.g., theoretical model; theoretical orientation)
	theory, in general (5 reasons; e.g., holes in the theory; theoretical framework is unsound)	**Contribution to/importance of theory** (3 criteria; e.g., importance of topic: theoretical importance; the theoretical relevance of the question investigated)

Table 1. (cont.)

(H) 'author's reputation/institutional affiliation' (11 criteria and reasons out of 11 studies; 3 positively, 2 negatively, and 6 neutrally formulated criteria and reasons)		
positively formulated criteria and reasons	negatively formulated criteria and reasons	neutrally formulated criteria and reasons
reputation, affiliation (3 criteria and reasons; e.g., you know who the author is and believe that he or she has a justifiable strong reputation in the area he or she writes about)	**reputation** (2 criteria and reasons; e.g., author appears to have weak or inappropriate credentials for the subject matter)	**reputation, affiliation** (6 criteria; e.g., the background and reputation of the author; the scholarship demonstrated in the article)
(I) 'ethics' (10 criteria and reasons out of 13 studies (because of multiple assignments); 0 positively, 5 negatively, and 5 neutrally formulated criteria and reasons)		
positively formulated criteria and reasons	negatively formulated criteria and reasons	neutrally formulated criteria and reasons
	multiple publication (4 criteria and reasons; e.g., previously published elsewhere)	**disciplinary ethics** (5 criteria and reasons; e.g., the ethical sense demonstrated by the author; compatibility with disciplinary ethics)
	secondary analysis (1 reason: only secondary analysis of data presented by others)	
not assignable (13 criteria and reasons, out of 15 studies)		
e.g., recommendation regarding publishing, overall evaluation of manuscript		

Note. The criteria and reasons are shown as they were stated (in a positive, negative, or neutral form). One to two examples of the criteria or reasons assigned to a dimension are given. The areas themselves and the various dimensions within the areas are shown in descending order according to the number of criteria and reasons assigned to the area and dimension.

As Table 1 shows, the greatest number of criteria and reasons (n=148) could be assigned to main area (A), 'relevance of contribution.' Of the 46 studies in total, 45 mentioned an average of three criteria and reasons in this area. 'Relevance of contribution' groups criteria and reasons that refer to the future 'gain' that could result from publication of a manuscript. The possible 'gain' relates to (1) scientific advancement, (2) relevance to journal readers, (3) practical usefulness of the findings (see dimensions in Table 1 (A)). These aspects have to do mainly with the importance, newness, and originality of a study reported on in the manuscript.

With a total of 143 criteria and reasons, 'writing/presentation' (main area (B)) is almost as large in scale as 'relevance of contribution' (main area (A)) (see Table 1). Of a total of 46 studies, 44 named an average of three criteria and reasons that could be assigned to 'writing/presentation.' Noticeable here is the comparatively high number (n=78) of criteria and reasons that are stated in a negative form. Apparently, criticisms of a manuscript frequently refer to this area. 'Writing/presentation' groups together mainly criteria and reasons that refer to the formal quality of a manuscript, such as writing style, written expression, spelling, grammar, and professional appearance of the manuscript. Also in this area are criteria and reasons regarding following the journal's

publication guidelines and regarding appropriate length of the manuscript. Thoroughness of the author also belongs here: does the manuscript contain all of the necessary information in the different sections of the paper, written completely and comprehensibly?

Compared to 'relevance of contribution' and 'writing/presentation,' clearly fewer criteria and reasons (n=92) could be assigned to main area (C), 'design/conception.' Still, at least one criterion or reason in this main area was found in 43 of the studies. Similar to 'writing/presentation,' about half of the criteria and reasons in area (C) are stated in a negative form. This means that the design and conception of the study reported on in the manuscript is frequently criticized in the peer review process. As Table 1 (C) shows, grouped under 'design/conception' are criteria and reasons referring to correct and logical conceptual framework as well as to the adequacy of the research design. Further criteria and reasons here are the internal consistency of a study, the plausibility of the research design with regard to the research question, the quality of sampling, the generalizability of the results, and replicability.

A total of 72 criteria and reasons from 34 studies were assigned to main area (D), 'method/statistics' (each of the 34 studies contained on average approximately two criteria or reasons in this area). The 'method/statistics' area contains criteria and reasons that refer to the correctness, appropriateness, and newness of methods or statistical analyses. Also found here are criteria and reasons pertaining to the quality of operationalization of key constructs and to the measurement of data (see Table 1 (D)). It is noticeable that almost half of the criteria and reasons in the main area 'method/statistics' are stated in the studies in a very general form, such as "methodology" or "statistical analyses." In other main areas, the greater part of the criteria and reasons is clearly stated in a more precise form, as is the case with the next main area (E) shown in Table 1, 'discussion of results.' A total of 45 criteria and reasons (from 31 studies) were assigned to 'discussion of results.' The criteria and reasons pertain mainly to whether the conclusions drawn in a manuscript are objective, correct, and properly based in the results. A few of the criteria and reasons address also the existence and clarity of a "take-home message" in the manuscript or the breadth or depth of the discussion section.

Main area (F), 'reference to the literature and documentation,' contains a total of 27 criteria and reasons that were found in 26 of the studies (see Table 1 (F)). The criteria and reasons under 'reference to the literature and documentation' have to do with whether the research study reported in the manuscript is embedded in the frame of the relevant literature. The criteria and reasons pertain to the up-to-date review of the literature cited and the thoroughness of the author's review of the literature.

The next main area (G), 'theory,' contains 24 criteria and reasons and is therefore similarly low in the number of assigned criteria as is main area (F), 'reference to the literature and documentation' (see Table 1). The criteria and reasons under 'theory' are

concerned with whether the manuscript contributes toward theory development or whether the theory underlying the research study seems complete and sound.

Whereas the main areas (and the criteria and reasons assigned to those areas) presented above pertain to the content of manuscripts and the research on which manuscripts report, we found in some of the studies (n=11) criteria and reasons that refer to the reputation or institutional affiliation of the authors (see Table 1 (H)). The criteria and reasons under main area (H), 'author's reputation/institutional affiliation,' address in the main the scholarship demonstrated in the manuscript and the reputation of the authors in their research areas.

The last main area listed in Table 1 is our main area of interest in the present study. The area considered here in the assessment of manuscripts is, namely, whether the authors of a manuscript follow ethical guidelines (main area (I), 'ethics'). The main area 'ethics' captures criteria and reasons that pertain to the compatibility of a manuscript (or the compatibility of the research behind a manuscript) with scientific or disciplinary ethics. Only 10 of the total of 572 criteria and reasons found in the studies examined can be assigned to this main area (see Table 1 (I)), and most of them have to do with the problem of multiple publication – that is, the practice of reporting the results of a single definable body of research in more than one publication (that is, in repeated reports of the same work, or in fractional reports) (see BORNMANN & DANIEL, [2007], HUTH, [2000]). *Not one* of the criteria and reasons in the main area 'ethics' refers to possible fabrication of the data on which the findings in the manuscript are based. The possibility of scientific misconduct by the author is apparently not a relevant issue in manuscript assessment.

The importance of the individual main areas in manuscript assessment

The categorization of criteria and reasons provides an overview of the main areas considered by editors and referees when assessing manuscripts in the peer review process. But which of these areas do editors and referees find to be of greater and lesser importance when assessing manuscripts? This is the question that we investigated in the second step of our study. Included in this analysis were the 38 of the total 46 studies that provided completely quantitative data on the criteria and reasons (for example, frequencies of specific reasons given in reviews for rejection of manuscripts for publication). Based on this quantitative information, we determined for each of the studies the importance of the different main areas considered in manuscript assessment.

The exact procedure that we used to determine the importance of the main areas can be illustrated taking the example of the study conducted by HOWARD & WILKINSON [1998]. Table 2 presents an overview in table form of the procedure in the case of that study. HOWARD & WILKINSON [1998] mention six reasons editors of the *British Journal of Psychiatry* stated on a questionnaire for rejecting the manuscripts at the initial

(editorial) decision without sending these manuscripts to external reviewers for assessment. For our analysis, in a first step we ranked the six reasons from 1 (frequently stated) to 6 (seldom stated) according to the frequency with which they were stated, as reported by HOWARD & WILKINSON [1998] (the main areas to which the reasons were assigned in the present study are shown in the parentheses): rank 1 (39.6%): "paper was too specialized" (main area A, 'relevance of contribution'), rank 2 (21.5%): "paper was considered to be unoriginal" (main area A: 'relevance of contribution'), rank 3 (20.8%): "paper was poor methodologically" (main area D: 'method/statistics'), rank 4 (7.4%): "paper was rejected because of its subject matter" (main area A: 'relevance of contribution'), rank 5 (6.7%): "paper was a case report" (main area C: 'design/conception'), and rank 6 (4.0%): "paper was written in an inappropriate format" (main area B: 'writing/presentation').

Table 2. Determination of the importance of the main areas illustrated taking the example of the study by HOWARD & WILKINSON [1998]

Reason	Frequency of the reason for rejection stated (in percent)	Rank of the reason	Main area of the reason	(Average) rank per main area	Rank of a main area	Main areas grouped in higher-order categories
"paper was too specialized"	39.6	1	'relevance of contribution'	2.33	1	Upper third of the rankings (high importance)
"paper was considered to be unoriginal"	21.5	2	'relevance of contribution'			
"paper was poor methodologically"	20.8	3	'method/ statistics'	3	2	Middle third of the rankings (medium importance)
"paper was rejected because of its subject matter"	7.4	4	'relevance of contribution'			
"paper was a case report"	6.7	5	'design/ conception'	5	3	Lower third of the rankings (low importance)
"paper was written in an inappropriate format"	4.0	6	'writing/ presentation'	6	4	

Note. The gray shadings of the cells indicate the assignment of reasons and main areas to the three categories (high, medium, and low) for the degree of importance of the main areas considered in manuscript assessment.

Whereas the main areas 'method/statistics,' 'design/conception,' and 'writing/presentation' are represented in the ranking order with one reason for rejection each from the HOWARD & WILKINSON [1998] study, three reasons for rejection as reported in that study can be assigned to 'relevance of contribution' (see Table 2). In order to obtain one single ranking also for the main area 'relevance of contribution,' we calculated the average of the three ranks 1, 2, and 4 of the reasons (see above) (arithmetic mean: 2.33).

We used the same procedure as in the example of the study by HOWARD & WILKINSON [1998] for all of the studies to create a ranking list in order to determine the importance of the different main areas considered in manuscript assessment. One ranking list could be created for 27 of the total of 38 studies, and more than one ranking list could be created for 11 studies (for example, some studies captured both reasons for rejection and review criteria).

As the category system for reasons and criteria in the present study contains nine main areas, the ranking list of the main areas in one study can contain up to nine ranking places. As it turned out, however, for many of the studies there were fewer than nine place rankings (such as for [HOWARD & WILKINSON, 1998]), because the criteria and reasons in the studies did not always cover all of the main areas considered in manuscript assessment. Because a direct comparison of main area rankings in the different studies is possible only if all nine place rankings are occupied, our next step in the analysis was to group the place rankings for each study into three, high-order categories: (category 1) the main area is of high importance (upper third of the rankings), (category 2) the main area is of medium importance (middle third of the rankings), and (category 3) the main area is of low importance (lower third of the rankings). Taking the example of the HOWARD & WILKINSON [1998] study, this means that 'relevance of contribution' (average rank 2.33) was assigned to category 1, 'method/statistics' (rank 3) to category 2, and 'design/conception' (rank 5) and 'writing/presentation' (rank 6) to category 3 (see Table 2). In order to be able to say something about the importance of a main area considered in manuscript assessment across all of the studies (or rankings), we counted, in a third and last step of our analysis, how frequently a main area fell into category 1, 2, or 3.

Table 3 shows the frequency distributions found. The distributions show how frequently a main area can be categorized as having high, medium, or low importance based on the place rank in the rankings. For each main area, in addition to the frequency distribution, Table 3 shows an average place ranking on a scale from 1 (high importance) to 3 (low importance). As can be seen in Table 3, the average place rankings of the nine main areas vary between 1.7 ('discussion of results') and 1.8 ('theory') and 2.7 ('ethics') and 2.9 ('author's reputation/institutional affiliation').

Table 3. High, medium, and low importance attached to main areas considered in manuscript assessment in journal peer review (in row percent). In addition to the relative frequencies for the pace ranking of main areas in ranking lists, the table shows for each main area an average ranking (arithmetic mean). The main areas are listed in the table in ascending order by mean

Main areas considered in manuscript assessment	Importance of the main area			
	high (1)	medium (2)	low (3)	mean
'discussion of results' (*n**=30)	50	27	23	1.7
'theory' (*n**=24)	46	29	25	1.8
'design/conception' (*n**=42)	38	26	36	2.0
'method/statistics' (*n**=37)	27	49	24	2.0
'relevance of contribution' (*n**=51)	24	37	39	2.2
'writing/presentation' (*n**=49)	24	27	49	2.2
'reference to the literature and documentation' (*n**=23)	17	40	43	2.3
'ethics' (*n**=12)	8	17	75	2.7
'author's reputation/institutional affiliation' (*n**=9)	0	11	89	2.9

The table shows, for example, that in one-half of a total of 30 ranking lists (50%) 'discussion of results' is a highly important main area in manuscript assessment; in 23% of the ranking lists it was of low importance. On a scale from 1 (high importance) to 3 (low importance), this main area on average reached a ranking of 1.7 across 30 ranking lists.

* *n* refers to the number of ranking lists in which a main area appears. For some of the studies investigated, more than one ranking list could be created.

χ^2 (16, *n*=277) = 38.10, $p < 0.001$ (based on 10,000 sampling tables), Cramér's V = 0.26.

While the importance of 'discussion of results' can be categorized as high in 50% of the rankings, it can be categorized as low in 23% of the rankings. The distribution for 'theory' is similar to the one for 'discussion of results:' in 46% of the rankings it can be categorized as high and in only 25% of the rankings as low.

In contrast to 'discussion of results' and 'theory,' the importance of the main areas 'author's reputation' and 'ethics' can be categorized as high in none and in only one ranking; in at least two-thirds of the rankings in which these main areas appear, they are categorized as being of low importance. Seen overall, the differences in the distribution of low, medium, and high importance between the nine main areas are highly significant statistically, χ^2 (16, *n*=277) = 38.10, $p<0.0001$ (based on 10,000 sampling tables), and – judged following criteria by COHEN [1988] – the effect size is medium, Cramer's V=0.26 (see Table 3). A comparison of the results in Table 3 with the results in Table 1 reveals that while the studies lists comparatively many criteria and reasons assigned to the main areas 'relevance of contribution' (*n*=148) and 'writing/presentation' (*n*=143), with regard to their importance in manuscript assessment the two main areas play a rather subordinate role for the most part (in only 24% of the rankings can they be categorized as having high importance, see Table 3).

Discussion

In journal peer review, editors of scientific journals use referee reports in their decision-making on whether a manuscript is fit for print or not [OFFICE OF MANAGEMENT AND BUDGET, 2004]. According to ZIMAN [2000], the entire scientific system ultimately rests on these assessment processes. However, spectacular cases of fraud that have been disclosed in recent years (such as the falsification of data by Hwang Woo Suk, the Korean stem-cell researcher) have called into serious question the validity of the journal peer review process: "The peer review system must be said to have failed, as the frauds were unveiled by people from outside the immediate process" [BAUCH, 2006, P. 408]. The present study examined the question of the extent to which editors and referees when assessing manuscripts consider the possibility of falsification of the underlying research data. To do so, we carried out a quantitative content analysis of published studies that examined criteria used by editors and referees for manuscript assessment and also their reasons for accepting or rejecting a manuscript for publication. Through the analyses we sought an answer to the questions as to (1) the main areas considered by editors and referees in preparing reviews of manuscripts (what the underlying dimensions of assessment are), and (2) what importance is attached to scientific misconduct in the assessment process. A total of 46 studies could be included in the quantitative content analysis. The studies yielded 572 different criteria and reasons, which could be grouped in nine main areas considered in manuscript assessment.

By assigning the criteria and reasons to the nine main areas – in a first step of the analysis – we obtained an overview of the areas that referees and editors consider when assessing a manuscript in peer review. As the results of the content analysis show, the criteria and reasons could be most frequently assigned to the main areas 'relevance of contribution' and 'writing/presentation;' more than half of the criteria and reasons fall into these areas. Distinctly fewer criteria and reasons could be assigned to the more strongly quality-oriented main areas 'design/conception,' 'method/statistics,' 'reference to the literature and documentation,' and 'theory.' In a second step, the present study examined what main areas take on high and low significance for editors and referees in manuscript assessment. The results of this analysis reveal a somewhat different picture than the results of the first step of the analysis: in comparison with the other main areas, 'relevance of contribution' and 'writing/presentation' can not be ascribed the highest priority in manuscript assessment. Other main areas that are more clearly related to the quality of the research underlying a manuscript emerged in the analysis more frequently as important: 'theory,' 'discussion of results,' and 'design/conception.' Contrary to the criticism that has been voiced on journal peer

review in the wake of scientific frauds making headlines (see, for example, [COUZIN, 2006]), based on these findings the quality of the research underlying the manuscript can be ascribed an important role in manuscript assessment.

However, the analyses of the present study also show that the main area 'ethics' takes on altogether little importance in journal peer review. Review criteria and reasons for rejecting or accepting a manuscript in the area of 'ethics' (and most of them referring to the problem of multiple publication) were found in fewer than one-fourth of the studies investigated. In *no* study did we find criteria or reasons that refer to possible fabrication of research data by the authors. This result is surprising not only considering the multitude of cases of fraud that have been uncovered but also considering findings by MARTINSON & AL. [2005] that indicate that scientific misbehaviors are widespread among researchers. Having elaborated 16 forms of possible scientific misbehaviors in expert talks with 51 researchers, MARTINSON & AL. [2005] sent a questionnaire to a random sample of 7,760 researchers (early and mid-career scientists), of which 3,427 responded. Overall, 33% of the respondents said they had engaged in at least one of the top ten most serious bad practices within the previous three years.

The result of our content analysis with regard to the low importance of the main area 'ethics' in journal peer review raises the question of possible reasons for this. One main reason could be the great importance of trust in science and the integrity of its practitioners [CHO & AL., 2006; RENNIE, 2003]. According to FOX [1994] "the editorial relationships (editor, author, reviewer) rest upon trust such that levels of scepticism are low and belief in the scientific ethos of truth-seeking is high" (p. 302, see also [ANON, 2006B]). COUZIN [2006] reports that one expert who told a *Science* reporter that he had reviewed the Hwang's 2005 paper stated, " 'You look at the data and do not assume it's fraud' " (p. 24). Trust is the foundation of scientific knowledge and knowledge is a collective good [BRAD WRAY, 2006]. "In securing our knowledge we rely upon others, and we cannot dispense with that reliance. That means that the relations in which we have and hold our knowledge have a moral character, and the word I use to indicate that moral relation is *trust*" [SHAPIN, 1994, P. XXV].

However, some new policies and mechanisms for the peer review system have been proposed in order to identify and prevent scientific misconduct. These recommendations question whether peer review should continue to operate on trust, and they move peer review a little closer to the audit. As LEE & BERO [2006] point out, "the Council of Science Editors recommends that journals establish data-access policies for editorial evaluation and peer review before and after publication so that the validity of the work can be verified or errors corrected." Despite the difficulties and expense of involved in reviewing underlying data, LEE & BERO [2006] suggest that having a policy of access to raw data would "hold authors more accountable for the accuracy of their data and potentially reduce scientific fraud or misconduct". Already today, some journals, including the *British Medical Journal*, make it a condition that the editors can

ask for the raw data on which the results of a manuscript submitted for publication are based [SMITH, 2006]. By the way, the *Journal of the American Society for Information Science and Technology* asks referees to look for suspicion of duplicate publication, fabrication of data or plagiarism.

Based on these new policies for the journal peer review system, we plan a content analysis of referee guidelines, covering letters and statements of editorial policies (see [ARMSTRONG, 1982]) to check whether or not fabrication of data and other forms of scientific misconduct are mentioned in these documents.

References

ANON (2006A), Ethics and fraud. *Nature*, 439 (7073) : 117–118.

ANON (2006B), Three cheers for peers. *Nature*, 439 (7073) : 118.

ARMSTRONG, J. S. (1982), Research on scientific journals: implications for editors and authors. *Journal of Forecasting*, 1 (1) : 83–104.

BAUCH, H. (2006), Fraud: anonymous 'stars' would not dazzle reviewers. *Nature*, 440 (7083) : 408.

BORNMANN, L., DANIEL, H.-D. (2007), Multiple publication on a single research study: does it pay? The influence of number of research articles on total citation counts in biomedicine. *Journal of the American Society for Information Science and Technology*, 58 (8) : 1100–1107.

BRAD WRAY, K. (2006), Scientific authorship in the age of collaborative research. *Studies in History and Philosophy of Science Part A*, 37 (3) : 505–514.

BYRNE, D. W. (1998), *Publishing Your Medical Research paper. What They Don't Teach in Medical School*, London, UK, Williams & Wilkins.

CAMPANARIO, J. M. (1998), Peer review for journals as it stands today - part 1. *Science Communication*, 19 (3) : 181–211.

CHO, M. K., MCGEE, G., MAGNUS, D. (2006), Lessons of the stem cell scandal. *Science*, 311 (5761) : 614–615.

COHEN, J. (1988), *Statistical Power Analysis for the Behavioral Sciences*, Hillsdale, NJ, USA, Lawrence Erlbaum Associates, Publishers.

COUZIN, J. (2006), ... And how the problems eluded peer reviewers and editors. *Science*, 311 (5757) : 23–24.

CYRANOSKI, D. (2006), Verdict: Hwang's human stem cells were all fakes. *Nature*, 439 (7073) : 122–123.

FLETCHER, R. H., FLETCHER, S. W. (2003), The effectiveness of journal peer review. In: F. GODLEE, T. JEFFERSON (Eds), *Peer Review in Health Sciences*. London, UK, BMJ Books, pp. 62–75.

FOX, M. F. (1994), Scientific misconduct and editorial and peer review processes. *Journal of Higher Education*, 65 (3) : 298–309.

HIRSCHAUER, S. (2004), Peer Review Verfahren auf dem Prüfstand. Zum Soziologiedefizit der Wissenschaftsevaluation. *Zeitschrift für Soziologie*, 33 (1) : 62–83.

HOWARD, L., WILKINSON, G. (1998), Peer review and editorial decision-making. *British Journal of Psychiatry*, 173 : 110–113.

HUTH, E. J. (2000), Repetitive and divided publication. In: A. H. JONES, F. MCLELLAN (Eds), *Ethical Issues in Biomedical Publication*. Baltimore, MA, USA, Johns Hopkins University Press, pp. 112–136.

LEE, K., BERO, L. (2006), Ethics: increasing accountability. What authors, editors and reviewers should do to improve peer review. Retrieved June 17, 2006, from http://www.nature.com/nature/peerreview/debate/op3.html.

MARTIN, T. J. (2006), Reactions to the Hwang scandal. *Science*, 311 (5761) : 607.

MARTINSON, B. C., ANDERSON, M. S., DE VRIES, R. (2005), Scientists behaving badly. *Nature*, 435 (7043) : 737–738.

MEADOWS, A. J. (1998), *Communicating Research*, London, UK, Academic Press.

MERTON, R. K. (1973), *The Sociology of Science: Theoretical and Empirical Investigations*. Chicago, IL, USA: University of Chicago Press.

NORMILE, D., VOGEL, G., COUZIN, J. (2006), South Korean team's remaining human stem cell claim demolished. *Science*, 311 (5758) : 156–157.

ODLING-SMEE, L., GILES, J., FUYUNO, I., CYRANOSKI, D., MARRIS, E. (2007), Where are they now? *Nature*, 445 (7125) : 244–245.

OFFICE OF MANAGEMENT AND BUDGET (2004), *Revised Information Quality Bulletin for Peer Review*. Washington, DC, USA: Office of Management and Budget.

RENNIE, D. (2003), Misconduct and journal peer review. In: F. GODLEE, T. JEFFERSON (Eds), *Peer Review in Health Sciences*. London, UK, BMJ Books, pp. 118-129.

SENSE ABOUT SCIENCE (2005), *"I Don`t Know What to Believe ..." Making Sense of Science Stories*. London, UK: Sense about Science.

SHAPIN, S. (1994), *A Social History of Truth: Civility and Science in Seventeenth-Century England*, Chicago, IL, USA, The University of Chicago Press.

SMITH, R. (2006), Peer review: a flawed process at the heart of science and journals. *Journal of the Royal Society of Medicine*, 99 (4) : 178–182.

WELLER, A. C. (2002), *Editorial Peer Review: Its Strengths and Weaknesses*, Medford, NJ, USA, Information Today, Inc.

WHITE, H. D. (2005), On extending informetrics: an opinion paper. In: P. INGWERSEN, B. LARSEN (Eds), *Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics*. Stockholm, Sweden, Karolinska University Press, pp. 442–449.

ZIMAN, J. (2000), *Real Science. What It Is, and What It Means*, Cambridge, UK, Cambridge University Press.