

A multilevel modelling approach to investigating the predictive validity of editorial decisions: do the editors of a high profile journal select manuscripts that are highly cited after publication?

Lutz Bornmann,

Max Planck Society, Munich, Germany

Rüdiger Mutz,

ETH Zurich, Switzerland

Werner Marx and Hermann Schier

Max Planck Institute for Solid State Research, Stuttgart, Germany

and Hans-Dieter Daniel

ETH Zurich and University of Zurich, Switzerland

[Received April 2009. Final revision December 2010]

Summary. Scientific journals must deal with the following questions concerning the predictive validity of editorial decisions. Is the best scientific work selected from submitted manuscripts? Does selection of the best manuscripts also mean selecting papers that after publication show top citation performance within their fields? Taking the journal *Angewandte Chemie International Edition* as an example, this study proposes a new methodology for investigating whether manuscripts that are most worthy of publication are in fact selected validly. First, the influence on citation of the accepted and rejected but then published elsewhere manuscripts was appraised on the basis of percentile impact classes scaled in a subfield of chemistry and, second, the association between the decisions on selection and the influence on citation of the manuscripts was determined by using a multilevel logistic regression for ordinal categories. This approach has many advantages over methodologies that were used in previous research studies on the predictive validity of editorial selection decisions.

Keywords: Chemistry; Editorial decisions; Multilevel modelling; Peer review; Predictive validity; Reference standard

1. Introduction

When selecting manuscripts for publication in a scientific journal, the referees' central aims are to help editors to select the most appropriate manuscripts for the journal and to improve the quality of manuscripts that are accepted for publication (Gosden, 2003). According to Marsh and Ball (1991) this selection procedure is of utmost importance to the academic community. It is one of the most highly regarded and frequently used procedures for evaluating the merit of not only scientific manuscripts but also grant proposals and fellowship applications (Bornmann,

Address for correspondence: Lutz Bornmann, Max Planck Society, Hofgartenstrasse 8, D-80539 Munich, Germany.

E-mail: bornmann@gv.mpg.de

2011). As the selection of manuscripts based on peer recommendations is so central to what and where manuscripts are published, 'it is essential that it is carried out well and professionally' (Hames (2007), page 2). Scientific journals that all use this selection procedure must deal with the following questions concerning the predictive validity of the selection decisions. Are in fact the best scientific works selected from the manuscripts that are submitted? Does selecting the best manuscripts also mean selecting papers that after publication show top citation performance within their fields?

However, to answer the questions about predictive validity there is no easy way of quantifying which scientific work is 'better' than another (Giske, 2008). The quality of research cannot be measured *directly* (Evidence Ltd., 2007). In a *Nature* 'Web focus' article, Jennings (2006) thus commented:

'the most important question is how accurately the peer review system predicts the longer-term judgments of the scientific community. One way to address this would be through citation data.'

Also van Raan (2005) and Harnad (2008) recommended appraising the merit of this system by using bibliometric data. For this reason, most of the previous studies examining the predictive validity of selection decisions in science measured the quality of the decisions based on citation data.

In a comprehensive research project, we investigated the validity of the journal peer review process at *Angewandte Chemie International Edition* (ACIE) and conducted a citation analysis for manuscripts that were accepted by the journal or rejected but published elsewhere. Our analyses of the citation counts for the submissions at ACIE showed that the editorial decisions made by ACIE have high predictive validity (Bornmann and Daniel, 2008a, b): on average, accepted manuscripts have clearly higher citation counts than manuscripts that were rejected but published elsewhere. These results suggest for ACIE that the publication decisions correspond on average to the scientific influence of the manuscripts. In the present study, we propose a new approach towards investigating the predictive validity of selection decisions. The first step in this new approach is to appraise the effect on citation of accepted and rejected but then published elsewhere manuscripts based on percentile impact classes scaled in a subfield of chemistry, and the second step is to determine the association between the selection decisions and the later effect on citation of the manuscripts in a multilevel logistic regression for ordinal categories. As we show in what follows, this approach has many advantages over methodologies that were used in previous research on the predictive validity of selection decisions.

ACIE is an international journal of the German Chemical Society (Gesellschaft Deutscher Chemiker, Frankfurt am Main, Germany) and is published by Wiley (Weinheim, Germany). It introduced a peer review system in 1982, primarily in conjunction with one of the types of document that are published in the journal, 'communications', which are short reports on work in progress or recently concluded experimental or theoretical investigations. ACIE is one of the prime chemistry journals in the world, with a higher annual journal impact factor (JIF) than the JIFs of comparable journals (at 10.879 in the 2008 *Journal Citation Reports Science Edition*). JIFs are published by Thomson Reuters (Philadelphia, USA) and are a measure of the 'average' and fast response of the scientific community to a paper in a journal (see Bornmann, Leydesdorff and Marx (2007)). What the editors of ACIE look for most of all is the best research in the various areas of chemistry. Submissions that referees deem to be of high quality are selected for publication: for most submissions, a manuscript is published only if two external referees rate the results of the study that is reported in the manuscript as important and also recommend publication in the journal (Bornmann and Daniel, 2009a, 2010).

With our proposed multilevel logistic modelling approach towards investigating the predictive validity of selection decisions, we follow the study by Jayasinghe *et al.* (2003), which proposed a

multilevel modelling approach to the data analysis for examination of the reliability and fairness of peer review (of grant proposals). Alongside reliability and fairness, predictive validity is the third quality criterion for professional evaluations (Daniel, 1993; Marsh *et al.*, 2008).

2. Previous research on the predictive validity of selection decisions

According to our search of the literature, up to now only six empirical studies have been published on the level of predictive validity that is associated with editorial decisions. All six studies were based exclusively on citation counts (and/or JIFs) as a criterion of validity. The editors of the *Journal of Clinical Investigation* (Wilson, 1978) and the *British Medical Journal* (Lock, 1985) undertook their own investigations into the question of predictive validity. Daniel (1993), Bornmann, Marx, Schier, Thor and Daniel (2010) and Opthof *et al.* (2000) examined the editorial decisions at ACIE, *Atmospheric Chemistry and Physics* and *Cardiovascular Research* respectively, and McDonald *et al.* (2009) looked at the *American Journal of Neuroradiology*. All the six studies showed that the editorial decisions of acceptance or rejection for the various journals appear to reflect quite a high degree of predictive validity, when citation counts and/or JIFs are used as a criterion of validity.

Also, in the area of the selection of fellowship applications or grant proposals, some studies have been published recently on the predictive validity of selection decisions based on citation counts and/or JIFs. These studies examined whether papers by applicants whose proposals were accepted for funding by foundations were more frequently cited than papers by applicants whose grant proposals had been turned down. Whereas the studies by Bornmann and Daniel (2005, 2006) on the Boehringer Ingelheim Fonds (Heidesheim, Germany, which is an international foundation for awarding long-term fellowships to post-graduate researchers in biomedicine), by Bornmann *et al.* (2008b) on the European Molecular Biology Organization (Heidelberg, Germany) and by Reinhart (2009) on the Swiss National Science Foundation (Bern) confirmed the predictive validity of the selection decisions, the studies by Melin and Danell (2006) on the Swedish Foundation for Strategic Research (Stockholm) and by Hornbostel *et al.* (2009) on the Emmy Noether Programme of the German Research Foundation (Bonn) found no significant differences between accepted and rejected grant or fellowship applicants. van den Besselaar and Leydersdorff (2007) reported mixed results in a study on the Council for Social Scientific Research of the Netherlands Organization for Scientific Research (Den Haag) (see here also Bornmann, Leydersdorff and van den Besselaar (2010)).

Hence, in the area of the selection of fellowship or grant proposals, in contrast with the area of selection of manuscripts, the findings are mixed. This can be attributed mainly to the fact that very different study designs and statistical procedures were used. Despite the differences, most of the studies on the predictive validity of selection decisions on manuscripts or proposals used statistical methods that strictly speaking should not be used with bibliometric data.

3. Aims and objectives of the present investigation

In the present study, we take the ACIE journal as an example and present an approach that has important methodological advantages over research that has been done to date on the predictive validity of selection decisions (in this connection, see also the recommendations by Goldstein and Spiegelhalter (1996)). The advantages pertain mainly to the following seven points.

- (a) Lock (1985) attempted to estimate the validity of manuscript evaluation on the basis of the JIF for journals that published manuscripts that had previously been rejected by the

British Medical Journal, but the analyses on predictive validity should not be conducted with the JIFs for the journals in which the individual accepted and rejected but then published elsewhere manuscripts appeared, i.e. the JIF should not be used as an impact measure for one publication in this journal (Seglen, 1997). The JIF is only a very rough measure for determining predictive validity, because all the contributions in a journal are characterized by an average value (Braun *et al.*, 2007). For this reason, we determined in this study how frequently the individual manuscripts that had been accepted for publication and the manuscripts that had been rejected but then published elsewhere were cited after being published.

- (b) To compare the rates of citation of accepted and rejected but then published elsewhere manuscripts, no arithmetic means should be calculated. However, in most of the studies that were described in Section 2 arithmetic average rates of citation were calculated. There are dangers in the orientation to the measures of central tendency: in the face of non-normally distributed citation data, the arithmetic mean value can give a distorted picture of the kind of distribution (Bornmann *et al.*, 2008a), ‘and it is a rather crude statistic’ (Adler *et al.* (2009), page 1). As the distribution of citation frequencies is usually right skewed, distributed according to a power law (Adler *et al.*, 2009), arithmetic average rates of citation show mainly where publications with high citation rates are to be found. According to Evidence Ltd. (2007), ‘where bibliometric data must stand-alone, they should be treated as distributions and not as averages’ (page 10). We followed this recommendation in this study.
- (c) All things being equal, a manuscript is more likely to be cited if it is published in a reputable journal rather than in a journal with low reputation (Bornmann and Daniel, 2008c; Larivière and Gingras, 2010; Shatz, 2004). Mostly, the journal that is the basis of an evaluation study concerning the predictive validity is a high profile journal (e.g. ACIE or the *British Medical Journal*), whereas the journals in which rejected manuscripts are published elsewhere mostly not. If in an evaluation study higher citation impact values are found for accepted than for rejected but then published elsewhere manuscripts, this finding could be not (only) the result of higher manuscript quality (established by peer review and documented through the decision to publish) but instead the higher reputation of the journal. To determine the effect of the editorial decision on citation adjusted for journal reputation, in the present study we included in the analyses a measure for the reputation of the journals in which accepted and rejected but then published elsewhere manuscripts were published.
- (d) As the aim of the present study on ACIE was to evaluate a high impact journal in the field of chemistry, our focus in the data analysis using a multilevel modelling approach was on highly cited papers. Determining highly cited papers (or also lowly cited papers) is possible only on the basis of field-specific reference standards (Aksnes, 2003). A reference standard is informative about the expected citation frequency within a certain field. Radicchi *et al.* (2008) found, for example, that papers in general biology (the subject category of the journal where the paper appears; see Thomson Reuters) are cited on average 14.6 times, whereas papers in developmental biology are cited on average 38.7 times (see here also Bornmann and Daniel (2009b)). Measured relative to these values, 20 citations for a paper in general biology mean a higher impact than 20 citations for a paper in developmental biology.

Most of the studies that were described in Section 2 used citation counts but not reference standards to consider different expected frequencies of citation within the fields.

In this study the performance of manuscripts that were submitted to ACIE, accepted and rejected but then published elsewhere, was compared with subfield-specific reference standards. For this, Vinkler (1997) recommended calculating relative subfield citedness, R_w , where w refers to 'world': R_w relates the citation counts that are obtained by the papers evaluated to the citation counts that are received by the papers published in journals that are dedicated to the respective (sub)field.

As Vinkler's (1997) definition of R_w indicates, the determination of research fields in most studies of research evaluation is based on a classification of journals into subject categories that were developed by Thomson Reuters (Bornmann *et al.*, 2008a).

'The Centre for Science and Technology Studies (CWTS) at Leiden University, the Information Science and Scientometrics Research Unit (ISSRU) at Budapest, and Thomson Scientific [now Thomson Reuters] itself use in their bibliometric analyses reference standards based on journal classification schemes'

(Neuhaus and Daniel (2009), page 221). Each journal is classified as a whole to one or several subject categories. In general, this journal classification scheme proves to be of great value for research evaluation. But its limitations become obvious in the case of multidisciplinary journals such as *Nature* or *Science* (see, for example, Glänzel *et al.* (1999)) and highly specialized fields of research (e.g. Glänzel *et al.* (2009), Kostoff (2002) and Schubert and Braun (1996)). Papers that appear in multidisciplinary journals cannot be assigned exclusively to one field, and for highly specialized fields no adequate reference values exist (Opthof and Leydesdorff, 2010). To overcome the limitations of journal classification schemes, Neuhaus and Daniel (2009) proposed an alternative reference standard that is paper by paper based (see also Neuhaus *et al.* (2009)). We follow that proposal in the present study. In contrast with a reference standard based on journal sets, where all papers in a journal are assigned to one and the same field, with the alternative reference standard every publication is associated with a single principal field or subfield entry that makes clear the most important aspect of the work (see here also Kurtz and Henneken (2007) and Pendlebury (2008)).

- (e) In educational and psychological testing, percentile rank scores are used widely as a standard for comparison, to judge a person's test scores (intelligence test scores, for example) on the basis of a comparison with the percentiles of a calibrated sample (see Jackson (1996)). Percentile rank scores usually involve ranking the units, which in the present study are papers published in a certain chemical subfield, in ascending order according to a criterion, here citation counts (see Rousseau (2005) for an example of the use of percentiles describing journal impact). Next, the frequencies with which papers with a certain citation count are found are accumulated successively across all papers (papers with citation count 0, 1, 2,...). The percentile rank score amounts to the fraction of the cumulative frequencies of the total number of all papers (Bornmann, 2010; Bornmann, Mutz and Daniel, 2007).

Particularly in bibliometric analysis the use of percentile rank scores for evaluative purposes is very advantageous (see Evidence Ltd. (2007) and Plomp (1990)), as no assumptions have to be made about the distribution of citations, i.e. the scores are applicable also to the (usually) skewed distributions of bibliometric data (see above). In the present study we used percentile rank scores for chemical (sub)fields to assess the effect on citation of manuscripts that were accepted by ACIE or rejected but then published elsewhere. Through the use of percentile rank scores the accepted and rejected but published elsewhere manuscripts can be assigned directly to unambiguous impact classes.

- (f) To conduct a comparison with field-specific reference standards, previous studies on the predictive validity of selection decisions (see Section 2) determined the difference between accepted and rejected but then published elsewhere manuscripts in each individual field on which there were papers in the publication set. The findings consisted usually of a number of different single results—i.e. smaller and larger relative differences in influence between the two groups, leaving the problem of integrating these single results into one statement about the quality of a peer review process. Another difficulty is that no significance tests were performed for the pairwise comparisons to determine the significance of the relative differences in influence between accepted and rejected but then published elsewhere manuscripts. As evaluation studies on the predictive validity of selection decisions seek to clarify whether the best research is selected *overall*, what was missing is a convincing overall result stating whether the overall difference in effect on citation between accepted and rejected but then published elsewhere manuscripts is statistically significant. In the present study, that overall result was determined by using a regression or fixed effects model.
- (g) For the statistical analysis for testing the validity of editorial decisions, no single level models, like regression analysis or analysis of variance, should be used. The data are usually clustered: accepted or rejected but then published elsewhere manuscripts are published in certain fields (or subfields). The validity of editorial decisions at a journal might vary between fields. Since usually different staff editors are responsible for different fields (Gölitz, 2005), more valid decisions could be made on the submissions in one field than the decisions that are made on submissions in another field. This produces heterogeneity in the regression parameters. If with heterogeneity in regression parameters the overall effects are generalized across different fields, an aggregation bias would result. Principally, heterogeneity in parameters could be analysed by using single-level models by including statistical interactions between editorial decision and fields in the regression model. However, if, as with the multidisciplinary journal ACIE, the manuscripts can be assigned to more than 60 different chemical subfields, this would lead to a considerable inflation of parameter estimates.

Multilevel or mixed effects models can take into account both the problem of dependence among observations within clusters and the problem of heterogeneity in regression parameters.

In the present study, using a multilevel logistic model for ordinal categories, we examine the predictive validity of editorial decisions at ACIE, seeking answers to the following research questions.

- (i) Are there any overall differences in citation counts between accepted and rejected but published elsewhere manuscripts across all chemical (sub)fields to which the individual manuscripts can be assigned and, if so, how large are the overall differences? Is the work of the ACIE staff editors successful overall, i.e. predictively valid?
- (ii) Does the predictive validity of the decisions that are made by ACIE staff editors vary between the different chemical subfields? Are the editors in one field in a better position to make more valid selection decisions than in another field?
- (iii) Do differences in citation counts between accepted and rejected but published elsewhere manuscripts still remain, if the effects are ‘controlled’ for the reputation of the journal in which the manuscripts are published? Since, in comparison with the journals in which the manuscripts that are rejected by ACIE were then published, ACIE is a high profile journal, a higher effect on citation found for accepted manuscripts than for rejected but

then published elsewhere manuscripts could be not (only) the result of higher quality of manuscript but instead the higher reputation of the journal.

- (iv) Does the predictive validity of the editorial decisions vary across different citation impact classes of the later published manuscripts? For example, can the staff editors at ACIE validly reject those manuscripts that later (i.e. after publication elsewhere) show poor citation performance (and the other way around)?

4. Methods

4.1. Database for the present study

For the investigation of the editorial decisions at ACIE we used information on 1899 manuscripts that were reviewed in the year 2000. The information was taken from archived material that was stored by the publisher, Wiley. Of the 1899 manuscripts, 46% ($n = 878$) were accepted for publication as communications in ACIE, and 54% ($n = 1021$) were rejected. A search in the databases the Scientific Citation Index (Thomson Reuters) and Chemical Abstracts (Chemical Abstracts Service, Columbus, USA) revealed that, of the 1021 rejected manuscripts, 959 (94%) were then published in 136 other (different) journals: 723 in 21 journals (more than nine manuscripts published in each) and 236 in 115 journals (fewer than 10 published in each). 50 or more rejected manuscripts were published later in each of the following journals: *Chemical Communications* ($n = 119$), *Organic Letters* ($n = 91$), *Journal of the American Chemical Society* ($n = 70$), *Tetrahedron Letters* ($n = 60$) and *Organometallics* ($n = 50$). The results in Bornmann and Daniel (2008a) showed that the authors of about 75% of the rejected manuscripts did not change or only marginally changed the content of the manuscripts for publication elsewhere. About a quarter of the rejected manuscripts were changed to a substantial extent, or the content of the rejected manuscript was published together with other research results. The assessment of the extent of changes in the rejected manuscripts for the publication elsewhere was carried out by a senior scientist with a doctoral degree in chemistry in collaboration with other members of our research team.

4.2. Citation analysis

For accepted and rejected but then published elsewhere manuscripts, we determined the total number of citations for a fixed time window of 3 years after the publication year.

‘Fixed citation windows are a standard method in bibliometric analysis, in order to give equal time spans for citation to articles published in different years, or at different times in the same year’

(Craig *et al.* (2007), page 243). The citation searches for the present study were conducted on the basis of Chemical Abstracts, which is a comprehensive database of publicly disclosed research in chemistry and related sciences (see <http://www.cas.org/>). In the citation searches we included self-citations, because it is not expected that the number of self-citations varies systematically for the accepted and rejected but then published elsewhere manuscripts. Citations were searched for all the manuscripts that had been reviewed by ACIE in the year 2000 and published between 2000 and 2005 by ACIE or elsewhere ($N = 1834$). For three rejected manuscripts that were published elsewhere in 2006 a citation window of 3 years (1 year after publication up to the end of 2009) was not available. More manuscripts could be included in the present study than in Bornmann and Daniel (2008a, b), because a 3-year citation window was available for a larger number of manuscripts in 2009 than was available at the time of the earlier studies.

4.3. Reference standards

For evaluation studies in the field of chemistry, Neuhaus and Daniel (2009) proposed building

reference values based on publication and citation data that refer to the subject areas of Chemical Abstracts (see also van Leeuwen (2007)). For Chemical Abstracts, Chemical Abstracts Service categorizes chemical publications into 80 different subject areas (chemical fields, called 'Chemical Abstracts sections'). Every publication becomes associated with a single principal entry that makes clear the most important aspect of the work (Daniel, 1993). In contrast with the journal sets that are provided by Thomson Reuters, Chemical Abstracts sections are assigned paper by paper (Bornmann *et al.*, 2008a). According to Neuhaus and Daniel (2009),

'the sections of *Chemical Abstracts* seem to be a promising basis for reference standards in chemistry and related fields for four reasons: (1) the wider coverage of the pertinent literature, (2) the quality of indexing, (3) the assignment of papers published in multidisciplinary and general journals to their respective fields, and (4) the resolution of fields on a lower level (e.g. mammalian biochemistry) than in journal classification schemes (e.g. biochemistry and molecular biology). The proposed reference standard is transparent, reproducible and overcomes some limitations of the journal classification scheme of Thomson Scientific'

(pages 227–228).

For the present study, to set reference values we used publication and citation data for 63 of the 80 Chemical Abstracts sections. For each of these 63 sections, we have in the sample at least one accepted or rejected but then published elsewhere manuscript. For a total of 1827 manuscripts there is a section in Chemical Abstracts; for seven manuscripts there is no entry. As the accepted and rejected but then published elsewhere manuscripts appeared mostly between 2000 and 2002 and there was a fixed 3-year citation window for each publication, we based the reference values for a Chemical Abstracts section on the publications for the year 2001 and the citations of these publications in the years 2002–2004. The manuscripts accepted and rejected by ACIE were published mainly as communications or research articles. For that reason the reference standards should also be set on the basis of papers of these types of document. Because Chemical Abstracts Service 'does not provide a distinct document type for research articles' (Neuhaus and Daniel (2009), page 226), the reference values for this study were generated by excluding publications with non-relevant types of document, such as conference proceedings and reviews.

4.4. Establishing the reference values with the percentile rank procedure

Using the publication and citation data for each Chemical Abstracts section, reference values were set in order to allow comparisons, on a common scale, of the citation counts of the manuscripts accepted by ACIE or rejected but then published elsewhere and assigned by Chemical Abstracts Service to individual Chemical Abstracts sections (see here also Godin (2007) and Thomson Reuters (2008)). The reference values were computed by using a percentile rank procedure. First, the citations X_i that were received by the i th publication within n publications published in a given Chemical Abstracts section were gathered. Then the publications were ranked in increasing order

$$X_1 \leq X_2 \leq \dots \leq X_n,$$

where X_1 and X_n denote the number of citations received respectively by the least and most cited publication. Finally, in each Chemical Abstracts section, each individual publication was assigned a percentile rank based on this distribution. If, for example, a single publication within a Chemical Abstracts section had 50 citations, and this citation count was equal to or greater than the citation counts of 90% of all publications in the section, then the percentile rank of this publication would be 90. The publication would be in the 90th percentile.

Within each Chemical Abstracts section the percentiles were grouped in six percentile rank classes (see the scale values in parentheses):

- (a) the *bottom 25%* (papers with a percentile less than the 25th percentile),
- (b) *50–75%* (papers within the [25th; 50th[percentile interval),
- (c) *25–50%* (papers within the [50th; 75th[percentile interval),
- (d) *10–25%* (papers within the [75th; 90th[percentile interval),
- (e) *5–10%* (within the [90th; 95th[percentile interval) and
- (f) the *top 5%* (papers with a percentile equal to or greater than the 95th percentile).

With this classification, top performance, or highly cited, papers are those papers within or above the 95th percentile, i.e. papers in the top 5%. In the present study we followed Glänzel and Schubert's (1992) recommendations when setting these percentile rank classes (for example it should be sufficiently large to guarantee that the selected items form a real 'elite').

In evaluative bibliometrics there is uncertainty regarding what percentile rank a paper should have to be considered highly cited. According to Tijssen *et al.* (2002) and Tijssen and van Leeuwen (2006), highly cited papers are those among the top 10% of the most cited papers—i.e. papers in or greater than the 90th percentile (see also Lewison *et al.* (2007)). In the essential science indicators Thomson Reuters classifies as highly cited, papers that belong to the top 1% of papers world wide, taking into account the field and year of publication. However, according to Evidence Ltd. (2007), this 1% threshold for highly cited papers makes the metric not very suitable for the evaluation of institutions:

'There is no doubt that highly cited papers are associated with exceptional research, but the metric is a poor index of more general research activity. The threshold is so high that for many fields there would be few UK institutions that had more than a handful of papers in the index'

(page 18). In this study, the publications with the highest citation impact are those among the top 5% within their field.

On the basis of the limit values (citation counts) of the percentile rank classes that were set for each Chemical Abstracts section in the present study, each individual accepted or rejected but then published elsewhere manuscript was assigned to one of the six percentile rank classes (from the top 5% to the bottom 25%). By assigning the manuscripts to one of the six classes, the

Table 1. Limit values (citation counts) and proportions (row percentages) of the percentile impact classes for the Chemical Abstracts section 'Inorganic chemicals and reactions' by accepted and rejected but then published elsewhere manuscripts

	N	Results for the following percentiles:						Total
		Bottom 25%	50–75%	25–50%	10–25%	5–10%	Top 5%	
Limit values (citation counts)		1	2	5	10	18	24	
Accepted manuscripts	156	1.3	10.9	19.9	30.8	5.8	31.4	100
Rejected but then published elsewhere manuscripts	131	3.8	16.8	31.3	22.9	13.0	12.2	100
Total	287	2.4	13.6	25.1	27.2	9.1	22.6	100

influence of each individual publication is measured according to a six-point ordinal evaluation scale that is standard for all publications (6, top 5%; . . . ; 1, bottom 25%) and is assessed with regard to the size of the impact reached within the 3-year citation window.

As an illustration, Table 1 shows the limit values for the individual percentile rank classes for the Chemical Abstracts section ‘Inorganic chemicals and reactions’ and the proportions of accepted and rejected but then published elsewhere manuscripts within the individual classes. 31.4% of the accepted manuscripts that were assigned to this section were among the top 5% as opposed to only 12.2% of the rejected manuscripts. However, with 12.2%, more rejected manuscripts are in the top 5% than we would expect for randomly chosen papers. For randomly chosen papers, we would expect this value to be 5%.

Table 2 presents for 10 Chemical Abstracts sections with the highest number of manuscripts in the data set the limit value for the top 5% of the papers. This value ranges from 18 (‘Carbohydrates’) to 36 citations (‘General organic chemistry’). Between 9.4% (‘Organometallic and organometalloidal compounds’) and 26% (‘Heterocyclic compounds (one hetero atom)’) of the manuscripts are among the top 5% of papers in a section. The number of top 5% manuscripts accepted or rejected by ACIE for the total of 10 Chemical Abstracts sections shows that, for six, the ACIE staff editors accepted for publication (clearly) more manuscripts belonging after publication to the top 5% papers than they rejected. No clear relationship is found for only four sections (‘Heterocyclic compounds (one hetero atom)’, ‘Physical organic chemistry’, ‘General organic chemistry’ and ‘Organometallic and organometalloidal compounds’) regarding the percentage proportion of accepted or rejected top 5% manuscripts to respectively all accepted or rejected manuscripts.

4.5. Reputation of a journal as a covariate

Differences in effect on citation of accepted or rejected but then published elsewhere manuscripts can be—as we described above—the effect of not only the editorial decision of rejection or acceptance but also the reputation of the journals in which the manuscripts were published. For this reason, reputation should be included in the regression model as a covariate. The JIF is usually used as an indicator for the reputation of a journal. But, because the JIF of a journal is dependent on the expected citation rate in the (sub)field in which the journal publishes papers, JIFs of journals in different (sub)fields are not comparable with one another (Archambault and Larivière, 2009). As we described above, the manuscripts that were accepted by ACIE or rejected but then published elsewhere were assigned by Chemical Abstracts Service to 63 different Chemical Abstracts sections. To be able to compare the reputation of journals across fields, Bordons and Barrigon (1992) recommend calculation of the normalized journal rank position NJP. For this, the ordinal position in the (sub)field is determined for each journal as follows: in the *Journal Citation Reports* each journal is assigned to at least one subject category (for example, the journal *Chemistry & Industry* is assigned to ‘Chemistry, applied’ and the journal *Aldrichimica Acta* to ‘Chemistry, organic’). If the journals within a subject category are sorted by JIF, the result for each journal is its ordinal position within the category. As the individual subject categories contain a different number of journals (for example, in the 2008 *Journal Citation Reports* there are 61 journals in the category ‘Chemistry, applied’ and 55 in ‘Chemistry, organic’), to calculate NJP the ordinal position is divided by the total number of journals. Thus, a higher journal reputation is associated with smaller values. If the journal in which an accepted or rejected but then published elsewhere manuscript was published was assigned by Thomson Reuters to several subject categories, for the present study we used the subject category in which the journal had the smallest value. The covariate is centred on the grand mean, to facilitate the

Table 2. Limit values (citation counts) for the top 5% papers within a Chemical Abstracts section and number of accepted or rejected but then published elsewhere manuscripts (absolute and as percentages) among the top 5% in a section for 10 sections with the highest number of manuscripts in the data set (sorted by the percentages of the top 5% manuscripts)

Chemical Abstracts section	Total number of manuscripts	Number of accepted / rejected manuscripts	Limit value for top 5%	Number of top 5% manuscripts		Number of top 5% manuscripts accepted by ACIE		Number of top 5% manuscripts rejected by ACIE	
				Absolute	%	Absolute	% of accepted	Absolute	% of rejected
Heterocyclic compounds (one hetero atom)	50	18 / 32	24	13	26.0	5	27.8	8	25.0
Inorganic chemicals and reactions	287	156 / 131	24	65	22.7	49	31.4	16	12.2
Physical organic chemistry	125	63 / 62	23	27	21.6	13	20.6	14	22.6
Chemistry of synthetic high polymers	58	30 / 28	25	11	19.0	9	30.0	2	7.1
Biomolecules and their synthetic analogs	58	33 / 25	25	10	17.2	9	27.3	1	4.0
Carbohydrates	88	40 / 48	18	15	17.1	12	30.0	3	6.3
Benzene, its derivatives, and condensed benzenoid compounds	81	28 / 53	31	12	14.8	8	28.6	4	7.5
General organic chemistry	95	50 / 45	36	13	13.7	7	14.0	6	13.3
Amino acids, peptides, and proteins	59	19 / 40	26	7	11.9	5	26.3	2	5.0
Organometallic and organometalloidal compounds	244	119 / 125	22	23	9.4	13	10.9	10	8.0

interpretation of the results (Hox, 2010). By including this covariate, we can explain both the citation impact variability at the level of the individual manuscripts and the intercept variance at the level of the Chemical Abstracts sections.

4.6. Statistical procedure

The manuscripts that were accepted by ACIE or rejected but then published elsewhere are clustered in different Chemical Abstracts sections. In a multilevel framework, j ($j = 1, \dots, N$) denotes the level 2 units (here Chemical Abstracts sections) and i ($i = 1, \dots, n_j$) the level 1 units, which here are the manuscripts, nested within each level 2 unit. The dependent variable is denoted by Y_{ij} with C ordered response categories ($c = 1, 2, \dots, C$) referring to the percentile rank class in which a manuscript (belonging to a section) was assessed. The editorial decision for a single manuscript i in section j serves as explanatory variable x_{1ij} and is coded 1 (accepted manuscript) or 0 (rejected manuscript but published elsewhere).

In a multilevel or mixed effects model the regression (intercept and slope) from percentile rank classes on editorial decisions is allowed to vary randomly across Chemical Abstracts sections (random effects). Whereas the intercept variance component represents differences in effect on citation of the manuscripts that were rejected by ACIE and then published elsewhere, the slope variance component represents different effects of the editorial decision on the percentile rank of a manuscript. If the variance component of slopes is statistically significant, then the predictive validity of the editorial decisions is not constant or fixed but instead varies across different Chemical Abstracts sections. If the regressions do not vary across the sections, the fixed effects part of the model gives an answer to the validity of the editorial decisions over all sections.

In a generalized linear mixed model the dependent variable with ordered categories (here percentile rank classes) must be related to the mixed effects model by using a certain link function (e.g. logit or probit). In this study, we prefer a logit function, as logits represent log-transformed odds ($P/(1 - P)$) which are relatively easy to interpret.

In the multilevel logistic model for ordered categories, according to Bauer (2009) or Hedeker (2008) the cumulative probabilities P_{ijc} ($= P(Y_{ij} \leq c)$) of a manuscript i in section j are modelled across the C percentile rank classes as follows:

$$\log\left(\frac{P_{ijc}}{1 - P_{ijc}}\right) = \gamma_c - (\mathbf{x}'_{ij}\hat{\boldsymbol{\beta}} + \mathbf{z}'_{ij}\mathbf{u}_j) \quad (c = 1, \dots, C - 1), \quad (1)$$

where $\boldsymbol{\beta}$ is the vector of fixed effects parameters of the explanatory variables (intercept x_{0ij} , slope x_{1ij} , covariates), \mathbf{u}_j is the vector of random effects (random intercept and slope of the editorial decision variable in section j) of the design vector of covariates \mathbf{z}_{ij} . The random effects are assumed to be independently and normally distributed with $\mathbf{u}_j \sim N(\mathbf{0}, \Sigma_{\mathbf{u}})$, where $\Sigma_{\mathbf{u}}$ is the variance-covariance matrix of the random effects. The meaning of the γ_c -parameter becomes clear in the context of a latent variable version of the model. For citation counts we can assume that the observed percentile rank classes Y_{ij} are based on a latent continuous variable y_{ij}^* (quality dimension) with an arbitrary fixed scale:

$$y_{ij}^* = \mathbf{x}'_{ij}\hat{\boldsymbol{\beta}} + \mathbf{z}_{ij}\mathbf{u}_j + \varepsilon_{ij}, \quad (2)$$

where ε_{ij} are random individual disturbances (level 1), which are assumed to follow a logistic distribution (mean = 0; $\sigma_{\varepsilon}^2 = \pi^2/3$) according to the logistic link function. The latent variable y_{ij}^* and the observed variable Y_{ij} are linked by a threshold concept. The model contains $C - 1$ monotonically and strictly increasing thresholds γ_c (i.e. $\gamma_1 < \gamma_2 \dots < \gamma_{C-1}$). The first threshold

γ_1 is arbitrarily set to 0 (with $\gamma_0 = -\infty$ and $\gamma_C = \infty$), which is why only $C - 2$ threshold parameters are estimated. The continuous latent variable y_{ij}^* and the ordinal outcome variable Y_{ij} are related as follows: $Y_{ij} = c$ if $\gamma_{c-1} \leq y_{ij}^* < \gamma_c$. If, for example, the value y_{ij}^* of a manuscript i in Chemical Abstracts section j exceeds the threshold of percentile class $c = 4$ but remains below the threshold of percentile class $c = 5$, then the rank class of the manuscript is $Y_{ij} = 5$. The model formula above (equation (1)) with the minus sign allows an interpretation as is customary with regression models: positive coefficients of the covariate indicate that, as the values of the covariates increase, so do the probabilities that a manuscript is in a category higher than c . However, other formulations are also possible (see Fielding *et al.* (2003)).

The conditional (unit-specific) probabilities of being in various response categories are calculated by using the following formula with cumulative probabilities:

$$P(Y_{ij} = c) = P_{ijc} - P_{ijc-1} = \frac{1}{1 + \exp[-\{\gamma_c - (\mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\mathbf{u}_j)\}]} - \frac{1}{1 + \exp[-\{\gamma_{c-1} - (\mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\mathbf{u}_j)\}]}.$$

The proportional odds model presupposes that the effect of the covariate is the same across the cumulative logits. As a violation of this assumption is quite common, a non-proportional odds model with respect to the fixed effects should be calculated, also. Here, the assumption of equal parameter estimates in each category is relaxed.

With regard to the models that are presented in the following section we would like to point out a fundamental problem when comparing different models: as the latent scale y_{ij}^* in this study was set arbitrarily with mean 0 and variance $\pi^2/3$, there is a scaling problem, if the level 1 variance is explained by introducing level 1 covariates into the model (Liang and Zeger, 1986). As the level 1 variance remains constant, the scale of the latent variable y_{ij}^* and the model parameter are implicitly rescaled. This rescaling makes it difficult to compare nested models. Fielding (2004) and Bauer (2009) developed procedures to rescale model parameters in a way that model comparisons are possible. Bauer (2009) extended the single-level approach by Winship and Mare (1983, 1984) to a multilevel approach by putting the parameter estimates on a common scale to facilitate model comparisons. Bauer (2009) showed in a simulation study that his procedure works very well. Therefore, we used Bauer's (2009) rescaling procedure and the rescaled parameters are presented in what follows. The variance of the transformed common latent scale is 1.0.

In this study, the regression analyses were conducted with SAS[®] 9.1, using the procedure PROC NLMIXED (Hedeker and Gibbons, 2006; Littell *et al.*, 2006; Mutz *et al.*, 2004) (see Appendix A).

5. Results

5.1. Null model

The base for the interpretation of the multilevel logistic model for ordinal categories, the results of which are presented in what follows, is the null model (Table 3), which contains only the intercept as fixed effect and the variance of the random intercepts across Chemical Abstracts sections ($\hat{\sigma}_{u_0}^2$) as random effect. The variance component ($\hat{\sigma}_{u_0}^2 = 0.17$) is statistically significant by using a deviance test (Hox, 2010) with a p -value divided by 2 ($\chi^2(1) = 33.7$; $p < 0.05$). The quality of the manuscripts submitted (measured *ex post* by citations) to the ACIE differs between the sections. However, this difference is relatively small (5%) in comparison with the variability within the sections (95%). The variance partition coefficient (Browne *et al.*, 2005) is only 0.05 ($= \hat{\sigma}_{u_0}^2 / (\hat{\sigma}_{u_0}^2 + 3.29)$). Despite this low coefficient, multilevel analysis should still be preferred to single-level analysis (Hox, 2010) to avoid biased standard errors.

Table 3. Results of three multilevel logistic regressions for ordinal categories†

Term	Results for null model			Results for proportional odds model			Results for non-proportional odds model			Results for non-proportional odds model with journal reputation		
	Estimated parameter	Estimate	Standard error	Estimate	Standard error	Rescaled	Estimate	Standard error	Rescaled	Estimate	Standard error	Rescaled
<i>Fixed effects</i>												
Threshold 2	γ_2	1.35‡	0.09	1.38‡	0.09	0.71	1.18‡	0.09	0.60	1.29‡	0.10	0.64
Threshold 3	γ_3	2.57‡	0.10	2.64‡	0.10	1.36	2.41‡	0.11	1.23	2.65‡	0.12	1.32
Threshold 4	γ_4	3.77‡	0.11	3.92‡	0.11	2.01	3.38‡	0.13	1.88	4.02‡	0.14	2.00
Threshold 5	γ_5	4.44‡	0.11	4.62‡	0.12	2.38	4.37‡	0.14	2.23	4.74‡	0.16	2.36
Intercept	β_0	2.91‡	0.13	2.57‡	0.13	1.32	2.35‡	0.13	1.20	2.89‡	0.15	1.44
Slope	β_1			1.00‡	0.10	0.51						
Slope 1	β_{11}						2.00‡	0.30	1.03	1.15‡	0.31	0.57
Slope 2	β_{12}						1.09‡	0.15	0.56	0.35‡	0.17	0.17
Slope 3	β_{13}						0.97‡	0.12	0.50	0.36‡	0.13	0.18
Slope 4	β_{14}						0.96‡	0.12	0.49	0.45‡	0.13	0.23
Slope 5	β_{15}						0.94‡	0.14	0.48	0.46‡	0.14	0.23
Covariate	β_2									-4.57‡	0.42	-2.27
<i>Random effects</i>												
μ_{0j}	$\sigma_{\mu_0}^2$	0.17‡	0.06	0.17‡	0.08	0.05	0.15‡	0.08	0.05	0.20‡	0.08	0.05
μ_{1j}	$\sigma_{\mu_1}^2$			0.04	0.06	0.01	0.06	0.07	0.02	0.04	0.05	0.01
	$\sigma_{\mu_{01}}^2$			0.06	0.05	0.02	0.08	0.05	0.02	0.06	0.05	0.01
ε_{ij}	σ_{ε}^2	3.29	—	3.29	—	0.87	3.29	—	0.87	3.29	—	0.81
$-2 \log(L)$			6121.6		5996.6			5979.0				5839.1

†Threshold 1 is fixed at 0 ($\lambda_1 = 0$). $N_j = 1827$ ($N = 1824$ for the non-proportional odds model with journal reputation, as for three manuscripts no NJP was available). The variance component σ_{ε}^2 is fixed at 3.29.
‡ $p < 0.05$ (Wald test).

5.2. Proportional odds model

The proportional odds model in Table 3 assumes that the effect of acceptance or rejection by the ACIE editors on the later influence on citation of the publications is the same across all cumulative logits $\log\{P_{ijc}/(1 - P_{ijc})\}$ of the six-point percentile scale. The fixed part in Table 3 represents the overall regression on the editors' decisions (acceptance, 1, or rejection, 0) within a Chemical Abstracts section. As conditional or unit-specific and not population-average models were calculated, fixed effects are to be interpreted as effects for an 'average section' with zero random effects (Raudenbush *et al.*, 2004). The random part in Table 3 represents the variability of the editorial decision effects over different sections (expressed as variances of intercepts and slopes).

The estimate of β_1 (1.00) in the fixed part gives, over all Chemical Abstracts sections, the effect of the editors' decisions on the percentile rank of the accepted or rejected but then published elsewhere manuscripts. There appears to be strong evidence that the odds of being in a percentile class c or below $P(Y_{ij} \leq c)$ are smaller for manuscripts accepted for publication in ACIE (equation (1)) than for manuscripts rejected by ACIE and then published elsewhere with an estimated odds ratio of $\exp(\beta_1) = \exp(1.00) = 2.72$. Expressed in probabilities $P(Y_{ij} = c)$ instead of cumulative probabilities $P(Y_{ij} \leq c)$, manuscripts accepted for publication by ACIE are clearly more rarely in the lower percentile rank classes (bottom 25% or 50–75%) and clearly more frequently in the upper percentile rank classes (5–10% or top 5%) than manuscripts rejected by ACIE and then published elsewhere. For example, the conditional probability that an accepted manuscript will be in the percentile rank class top 5% (rank 6) is $P(Y_{ij} = 6 \mid \text{editor's decision} = \text{acceptance}) = 0.255$ with an approximate 95% confidence interval [0.211, 0.299]. This probability is clearly greater than for a manuscript that is rejected by ACIE and then published elsewhere: $P(Y_{ij} = 6 \mid \text{editor's decision} = \text{rejection}) = 0.115$ [0.091, 0.138]. The probability that a manuscript accepted by ACIE will be in the percentile rank class 50–75% (rank 2) is $P(Y_{ij} = 2 \mid \text{editor's decision} = \text{acceptance}) = 0.102$ [0.079, 0.125] and is thus lower than the probability for a manuscript that was rejected by ACIE and then published elsewhere: $P(Y_{ij} = 2 \mid \text{editor's decision} = \text{rejection}) = 0.231$ [0.193, 0.270]. The findings concerning the random part of the model estimation show (see Table 3) that the variability of the effect of the editorial decisions at the level of Chemical Abstracts sections is not statistically significant ($\hat{\sigma}_{u1}^2 = 0.04$, not significant) by using a deviance test which compares a model including random-intercept and slope variance components (but not a covariance component) with a model not including the slope variance component, where the p -value is divided by 2 ($\chi^2(1) = 0.06$; $p \geq 0.05$). As the general fixed effect of the editorial decision is the same across all Chemical Abstracts sections ($\beta_1 = 1.00$) and the covariance σ_{u01}^2 between the intercept and the slope is not statistically significant also, the editorial decisions seem to have the same predictive validity across all different sections. However, there is a statistically significant intercept standard deviation ($\hat{\sigma}_{u0}^2 = 0.17$)—using a deviance test which compares a model with and without random intercepts, where the p -value is divided by 2 ($\chi^2(1) = 47.3$; $p < 0.05$)—that does not differ from the parameter in a null model in the rescaled version (column 'scaled' in Table 3). This means that not only does the quality of manuscript vary from section to section but also the conditional probability of the editorial decisions.

For the calculations of the following conditional probabilities the fixed effects parameters and the empirical Bayes estimates of the random effects u_{0j} were used: with respect to the upper percentile class (top 5%) for the Chemical Abstracts section 'General biochemistry' (in this section manuscripts with the lowest quality are submitted to ACIE; minimum of u_{0j}), the conditional probability for accepted manuscripts $P(Y_{ij} = 6 \mid \text{editor's decision} = \text{acceptance})$ amounts to 0.092 [0.038, 0.145], and the conditional probability for rejected but then published elsewhere manuscripts $P(Y_{ij} = 6 \mid \text{editor's decision} = \text{rejection})$ amounts to 0.037 [0.014, 0.058].

For the section ‘Inorganic analytical chemistry’ (in this section manuscripts with the highest quality are submitted to ACIE; maximum of u_{0j}), the conditional probability of the accepted manuscript $P(Y_{ij} = 6 \mid \text{editor's decision} = \text{acceptance})$ amounts to 0.471 [0.266, 0.677], and the conditional probability for the rejected but then published elsewhere manuscripts $P(Y_{ij} = 6 \mid \text{editor's decision} = \text{rejection})$ amounts to 0.252 [0.098, 0.406].

5.3. Non-proportional odds model

In the non-proportional odds model (see Table 3) the assumption is relaxed that the effect of the editorial decision is the same across all cumulative logits of the six-point percentile scale. Depending on the quality of the manuscripts submitted to ACIE (measured *ex post* by citations) differences in the predictive validity of the editorial decisions are assumed. Comparing the two deviances ($-2 \log(L)$) with 5996.6 (proportional odds) and 5979.0 (non-proportional odds), the significant likelihood ratio test rejects the proportional odds assumption ($\chi^2_{LR}(4) = 17.6$; $p < 0.05$). Table 3 shows that there are differences between the proportional and non-proportional model for the fixed effects: the slopes in the various rank classes are different, mainly slope 1 (2.00, rescaled 1.0). The effect of the editorial decision in the lowest rank class ($Y_{ij} = 1$, bottom 25%) is significantly higher than in the other rank classes. For manuscripts submitted to ACIE with the lowest quality, the editors come to decisions with the highest predictive validity. The conditional probability that a manuscript accepted for publication by the ACIE editors will be in the percentile rank class bottom 25% ($Y_{ij} = 1$) is clearly lower ($P = 0.013$ [0.006, 0.021]) than the probability for a manuscript that was rejected but then published elsewhere ($P = 0.087$ [0.066, 0.108]).

Fig. 1 shows a graphic representation of the logistic cumulative probability plot $P(Y_{ij} \leq c)$ for the non-proportional odds model: for each percentile class c , Fig. 1 shows the cumulative probability that a manuscript will be in a certain class or in one of the classes below. The full

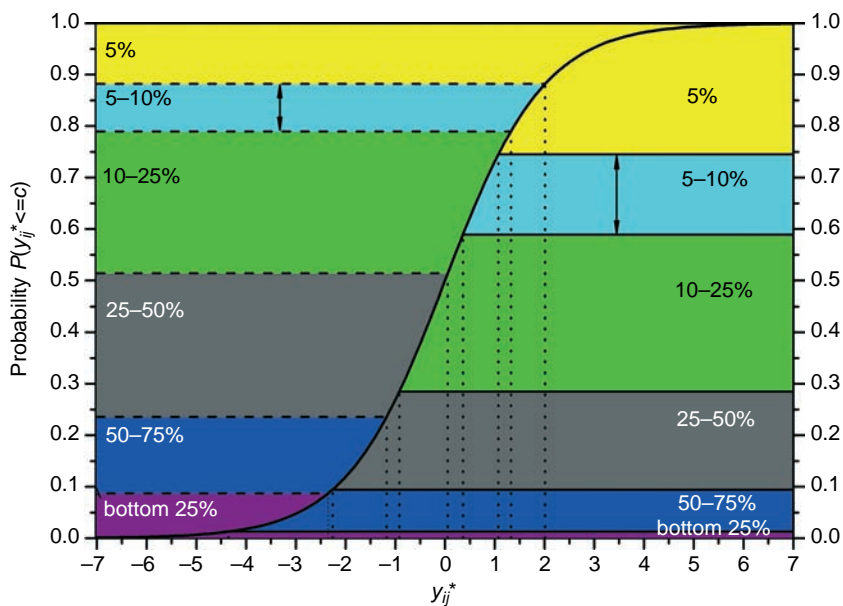


Fig. 1. Logistic cumulative probability plot (non-proportional odds model; —, accepted manuscripts; - - -, rejected manuscripts (but published elsewhere)): the probability that a manuscript is in a certain percentile impact class c is equal to the difference of the corresponding cumulative probabilities

lines in Fig. 1 show the boundaries of the percentile classes for the accepted manuscripts; the dotted lines show the boundaries of the percentile classes for the rejected but then published elsewhere manuscripts. It is clearly visible that the probabilities of being in a percentile rank class c or below $P(Y_{ij} \leq c)$ are always smaller for accepted than for rejected but then published elsewhere manuscripts. Manuscripts accepted by ACIE are significantly more rarely in the lower percentile rank classes (bottom 25% or 50–75%) and significantly more frequently in the higher percentile rank classes (top 5% or 5–10%) than manuscripts rejected by ACIE but then published elsewhere. The absolute conditional (unit-specific) probability of a manuscript being in a certain percentile class is given through the difference of the p -values of two consecutive horizontal lines (see the arrows in Fig. 1). In the high citation impact classes the p -value differences for the accepted manuscripts are larger than the differences for the rejected but published elsewhere manuscripts. In the low citation impact classes the situation is reversed.

5.4. Non-proportional odds model with reputation of the journal (NJP) as covariate

In an additional non-proportional odds model, it was tested whether the reputation of the journal (NJP) in which a rejected manuscript by ACIE was published has an effect on the percentile rank of the manuscript (see Table 3). As all accepted manuscripts were published in ACIE and thus have the same reputation, the NJP-effect is limited to the manuscripts rejected by ACIE but then published elsewhere. A statistically significant regression coefficient ($\beta_2 = -4.57$) and a decrease in the deviance ($-2 \log(L)$) from 5979.0 to 5893.1 ($\chi^2_{LR}(1) = 137.1$; $p < 0.05$) indicate a negative NJP-effect: the higher the reputation of the journal in which a rejected manuscript (by ACIE) was published, the higher its percentile rank class. The rescaled level 1 variance decreases from 0.87 (non-proportional odds without covariate) to 0.81 (non-proportional odds with covariate). As only 7% of the level 1 variance ($(0.87 - 0.81)/0.87 = 0.07$) will be explained by the reputation of the journal, the NJP-effect is not large. The rescaled variance component of the random effects hardly changes in comparison with the non-proportional odds model without covariate.

If with a Chemical Abstracts section with assumed zero random effects (conditional model) the effect of acceptance for publication in ACIE is compared with rejection and later publication in a journal with medium reputation (grand mean 0), the rescaled effect of the decision ‘acceptance’ clearly varies over the percentile rank classes between 0.17 (β_{12}) and 0.57 (β_{11}), which is comparable with the non-proportional odds model without including journal reputation. However, the effect of the editorial decision is reduced, if the rescaled regression parameters are examined. The reduction of this effect can be represented best via the cumulative probability plot (Fig. 2). Compared with Fig. 1 the overall interpretation hardly changes, but the effect of the editorial decision is smaller. The probability $P(Y_{ij} = 6 \mid \text{editor's decision} = \text{acceptance})$ of 0.255 in the model without considering journal reputation (Fig. 1) decreases to 0.199 in the model considering journal reputation (Fig. 2). In reverse order, the probability $P(Y_{ij} = 6 \mid \text{editor's decision} = \text{rejection})$ of 0.118 in the model without considering journal reputation increases to 0.135 in the model considering this for the manuscripts that are rejected but then published elsewhere.

6. Discussion, policy implications and future directions

Research evaluation is an area of increasing importance, as various scientific journals and research funding bodies need not only to assess the quality of manuscripts and applications reliably but also to examine the predictive validity of their decisions in selecting the best manuscripts

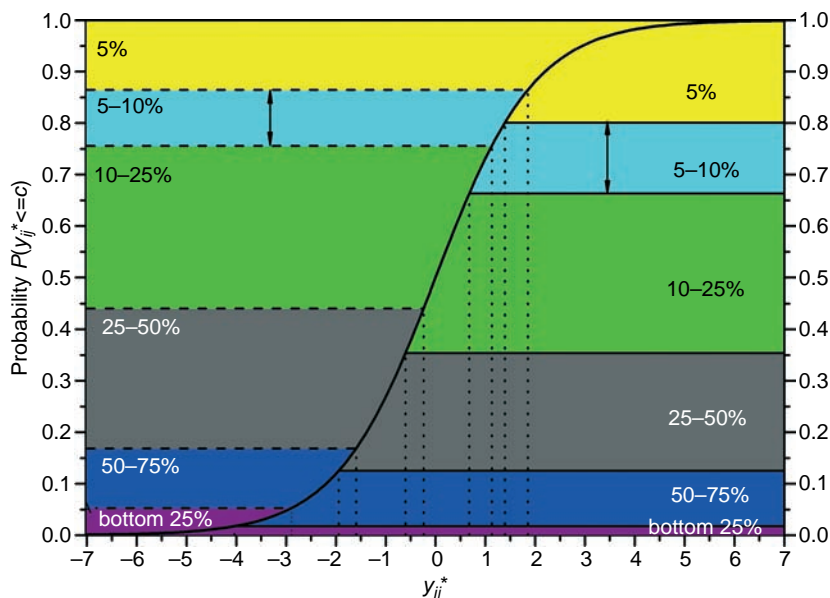


Fig. 2. Logistic cumulative probability plot (non-proportional odds model; —, accepted manuscripts; - - -, rejected manuscripts (but published elsewhere)), considering the reputation of journals where manuscripts rejected by ACIE were published elsewhere: the probability that a manuscript is in a certain percentile impact class c is equal to the difference of the corresponding cumulative probabilities

or the best applications (see here Järvelin and Persson (2008) and Pendlebury (2008)). There has been a strong focus on ‘scientific excellence’ in science policy in recent years (Kurtz and Henneken, 2007; Rons and Amez, 2008). It is expected that the manuscripts that are selected for publication by a high impact journal or the grant proposals that are selected for funding by a renowned funding body will be far above the scientific average and in the area of ‘top performance’. This focus on top quality might be justified ‘by the wish of society and politics to obtain some sort of accountability of science’ (Kurtz and Henneken (2007), page 93). If the quality demands are very high not only for planned (grant proposals) and completed (manuscripts) research but also for the selection decisions on proposals and manuscripts, then very high standards are also needed for the methodology that is used for assessing the quality of the selection procedures (see Adler *et al.* (2009)).

In the present study, we introduced a promising approach for assessing the predictive validity of editorial selection decisions. For subfields in chemistry, reference values were set based on percentile rank classes. Using the reference values it could be determined for individual manuscripts that were accepted by ACIE and manuscripts rejected but then published elsewhere whether they showed a high or low influence on citation after the editorial decision. The investigation of the predictive validity of ACIE acceptance and rejection decisions was not conducted in separate analyses for individual subfields; instead, the overall effect of the editorial decisions on the later influence of the manuscripts was determined within *one* statistical multilevel model. This study reports not only the parameter estimates of this model and the results of the statistical tests, but also the estimated (conditional, unit-specific) manuscripts’ probabilities of being in various citation impact classes were visualized. This allowed an optimal assessment of the strength of an effect that the editorial decision has.

In summary, we found the following answers to our four research questions that were listed above in Section 3.

- (a) The editorial decisions on submitted manuscripts at ACIE have predictive validity, as the decisions still have an effect on the citation counts of the later published manuscripts, even if the model controls for journal reputation. Expressed in odds, the odds (the cumulative probability of being in percentile class c or a lower rank class compared with the probability of being in a higher percentile class) increase strongly by the factor $\exp(1.15) = 3.16$, if a manuscript is accepted for publication by ACIE and not rejected and then published elsewhere with respect to the lowest citation level (bottom 25%), and moderately in the highest citation levels, for instance, by the factor $\exp(0.46) = 1.58$ in the top 5%. Our results basically indicate that ACIE acceptance greatly helps against being lowly cited (odds ratio $\exp(1.15) = 3.16$) and thereafter, in terms of higher citation levels, yields a moderate benefit (odds ratio $\exp(0.35) = 1.41$ to $\exp(0.46) = 1.58$). This basic result confirms the findings not only by Daniel (1993) and Bornmann and Daniel (2008a, b) on ACIE but also the findings of other studies published to date on the predictive validity of editorial decisions.
- (b) The predictive validity of the editorial decision does not vary statistically significantly across the different chemical subfields (Chemical Abstracts sections). Hence, the effect of the editorial decision that was described in point (a) is generalizable across all subfields.
- (c) The above reported effect of the editorial decision remains when we 'control' for journal reputation. Differences in effect on citation that were found here seem to be not the result of differences in journal reputation. However, the control of journal reputation is limited to rejected but published elsewhere manuscripts.
- (d) The predictive validity of the editorial decision is not the same for all citation impact classes. The staff editors at ACIE reject validly mainly those manuscripts that later, after publication elsewhere, show a low effect on citation.

With these answers to the four questions, all necessary information is given to assess the predictive validity of a peer review process.

In addition to the many advantages, the modelling approach that was introduced here has limitations that need to be addressed by future research studies.

- (a) The results of this study pertain to the selection decisions of only one journal and can be generalized to other journals in only a limited way. Future research studies using our approach should therefore be conducted with a larger number of journals, so that the results are more generalizable. It would be particularly interesting for future studies to consider journals with similar reputation that are equally attractive to potential authors, such that manuscripts that are rejected by the one journal are often submitted to the other. If the pattern of results that is shown in this study could be replicated in those circumstances, it would support the generalizability of the results.
- (b) To control the effect of journal reputation on the citations of accepted and rejected but then published elsewhere manuscripts, NJPs were included in the statistical analysis as a covariate. However, two problems are connected with this procedure. For one, NJP is confounded with the editorial decision: all accepted manuscripts have the same rank position. For another, the position is mainly the result of successful editorial publication decisions: the citation counts of the later published manuscripts determine the position (Seglen, 1997). In principle, the confounding effects can be adjusted by using common methods of statistical causal analysis (e.g. propensity score matching). But, because the NJPs for the manuscripts that are accepted by ACIE are constant, the usual adjustment procedures (e.g. matching with propensity scores) cannot be applied. Perhaps, confounding problems could be made smaller by using our proposed modelling approach for journals with a similar reputation within a field. Furthermore, the selection decisions of more

than one journal with different reputations within a field could be evaluated. Rejected manuscripts from these journals will probably be published elsewhere in journals with lower as well as higher reputation.

- (c) In the sense of covariance adjustment or matching (Rosenbaum, 2002), it would be desirable to estimate the differences between accepted and rejected but then published elsewhere manuscripts independently of the effects of further covariates besides journal reputation. Bibliometric studies have demonstrated that many other factors like the number of co-authors (Wuchty *et al.*, 2007) and the length of a paper (Bornmann and Daniel, 2007) have a general effect on citation counts. However, there is no consensus between researchers in bibliometrics whether and, if so, which of these factors should definitely be included in an analysis. When a consensus is found for further factors, these could be considered in future studies.

Acknowledgements

We thank Dr Christophe Weymuth (formerly at the Institute of Organic Chemistry at the University of Zurich; now at BIOSYNTH AG in Switzerland) for the investigation of the manuscripts that had been rejected by ACIE and published elsewhere.

We thank Dr Peter Göllitz, Editor-in-Chief of ACIE, the Editorial Board of *Angewandte Chemie* and the German Chemical Society for permission to conduct the evaluation of the peer review process of the journal and we express our thanks to the members of the editorial office for their generous support during the carrying out of the study. The entire research project, which is also investigating quality assurance at open access journals, is supported by the Max Planck Society (Munich, Germany). The authors express their gratitude to the reviewers, the Joint Editor and the Associate Editor for their helpful comments.

Appendix A: SAS program for the proportional odds model

```
PROC NL MIXED data=basdat qpoints=8;
  parms b0=0.6 b1=0.6 sd0=0.5 sd1=0.5 cov01=0.5 i1=1 i2=1 i3=1 i4=1;
  Z=B0+(B1+U1)*DEC+U0;
  IF(PERCENTIL=1)THEN
    P=1/(1+EXP(-(0-Z)));
  ELSE IF(PERCENTIL=2)THEN
    P=(1/(1+EXP(-(I1-Z)))) - (1/(1+EXP(-(0-Z))));
  ELSE IF(PERCENTIL=3)THEN
    P=(1/(1+EXP(-(I1+I2-Z)))) - (1/(1+EXP(-(I1-Z))));
  ELSE IF(PERCENTIL=4)THEN
    P=(1/(1+EXP(-(I1+I2+I3-Z)))) - (1/(1+EXP(-(I1+I2-Z))));
  ELSE IF(PERCENTIL=5)THEN
    P=(1/(1+EXP(-(I1+I2+I3+I4-Z)))) - (1/(1+EXP(-(I1+I2+I3-Z))));
  ELSE IF(PERCENTIL=6)THEN
    P=1 - (1/(1+EXP(-(I1+I2+I3+I4-Z))));
  LL=LOG(P);
MODEL PERCENTIL ~ GENERAL(LL);
RANDOM U0 U1 ~ NORMAL([0,0],[VAR0,COV01,VAR1])SUBJECT=SECTIONS;
estimate 'thresh2' i1;
estimate 'thresh3' i1+i2;
estimate 'thresh4' i1+i2+i3;
estimate 'thresh5' i1+i2+i3+i4;
RUN;
```

References

- Adler, R., Ewing, J., Taylor, P. and Hall, P. G. (2009) A report from the International Mathematical Union (IMU) in cooperation with the International Council of Industrial and Applied Mathematics (ICIAM) and the Institute of Mathematical Statistics (IMS). *Statist. Sci.*, **24**, 1–28.
- Aksnes, D. W. (2003) Characteristics of highly cited papers. *Res. Eval.*, **12**, 159–170.
- Archambault, É. and Larivière, V. (2009) History of the journal impact factor: contingencies and consequences. *Scientometrics*, **79**, 635–649.
- Bauer, D. J. (2009) A note on comparing the estimates of models for cluster-correlated or longitudinal data with binary or ordinal outcomes. *Psychometrika*, **74**, 97–105.
- van den Besselaar, P. and Leydesdorff, L. (2007) *Past Performance as Predictor of Successful Grant Applications: a Case Study*. Den Haag: Rathenau Instituut.
- Bordons, M. and Barrigon, S. (1992) Bibliometric analysis of publications of Spanish pharmacologists in the SCI (1984–89): part II, Contribution to subfields other than pharmacology and pharmacy (ISI). *Scientometrics*, **25**, 425–446.
- Bornmann, L. (2010) Towards an ideal method of measuring research performance: some comments to the Opthof and Leydesdorff (2010) paper. *J. Inf.*, **4**, 441–443.
- Bornmann, L. (2011) Scientific peer review. *A. Rev. Inform. Sci. Technol.*, **45**, 199–245.
- Bornmann, L. and Daniel, H.-D. (2005) Selection of research fellowship recipients by committee peer review: analysis of reliability, fairness and predictive validity of Board of Trustees' decisions. *Scientometrics*, **63**, 297–320.
- Bornmann, L. and Daniel, H.-D. (2006) Selecting scientific excellence through committee peer review—a citation analysis of publications previously published to approval or rejection of post-doctoral research fellowship applicants. *Scientometrics*, **68**, 427–440.
- Bornmann, L. and 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. *J. Am. Soc. Inf. Sci. Technol.*, **58**, 1100–1107.
- Bornmann, L. and Daniel, H.-D. (2008a) The effectiveness of the peer review process: interreferee agreement and predictive validity of manuscript refereeing at *Angewandte Chemie. Angew. Chem. Int. Edn.*, **47**, 7173–7178.
- Bornmann, L. and Daniel, H.-D. (2008b) Selecting manuscripts for a high impact journal through peer review: a citation analysis of Communications that were accepted by *Angewandte Chemie International Edition*, or rejected but published elsewhere. *J. Am. Soc. Inf. Sci. Technol.*, **59**, 1841–1852.
- Bornmann, L. and Daniel, H.-D. (2008c) What do citation counts measure?: a review of studies on citing behavior. *J. Doc.*, **64**, 45–80.
- Bornmann, L. and Daniel, H.-D. (2009a) The luck of the referee draw: the effect of exchanging reviews. *Learn. Publ.*, **22**, 117–125.
- Bornmann, L. and Daniel, H.-D. (2009b) Universality of citation distributions: a validation of Radicchi et al.'s relative indicator $c_f = c/c_0$ at the micro level using data from chemistry. *J. Am. Soc. Inf. Sci. Technol.*, **60**, 1664–1670.
- Bornmann, L. and Daniel, H.-D. (2010) The manuscript reviewing process—empirical research on review requests, review sequences and decision rules in peer review. *Libr. Inf. Sci. Res.*, **32**, 5–12.
- Bornmann, L., Leydesdorff, L. and van den Besselaar, P. (2010) A meta-evaluation of scientific research proposals: different ways of comparing rejected to awarded applications. *J. Inf.*, **4**, 211–220.
- Bornmann, L., Leydesdorff, L. and Marx, W. (2007) Citation environment of *Angewandte Chemie*. *CHIMIA*, **61**, 104–109.
- Bornmann, L., Marx, W., Schier, H., Thor, A. and Daniel, H.-D. (2010) From black box to white box at open access journals: predictive validity of manuscript reviewing and editorial decisions at *Atmospheric Chemistry and Physics. Res. Eval.*, **19**, 81–156.
- Bornmann, L., Mutz, R. and Daniel, H.-D. (2007) The *b* index as a measure of scientific excellence: a promising supplement to the *h* index. *Cybermetrics*, **11**, no.1, paper 6.
- Bornmann, L., Mutz, R., Neuhaus, C. and Daniel, H.-D. (2008a) Use of citation counts for research evaluation: standards of good practice for analyzing bibliometric data and presenting and interpreting results. *Eth. Sci. Environ. Polit.*, **8**, 93–102.
- Bornmann, L., Wallon, G. and Ledin, A. (2008b) Does the committee peer review select the best applicants for funding?: an investigation of the selection process for two European Molecular Biology Organization programmes. *PLOS One*, **3**, no. 10, e3480.
- Braun, T., Dióspatonyi, I., Zsindely, S. and Zádor, E. (2007) Gatekeeper index versus impact factor of science journals. *Scientometrics*, **71**, 541–543.
- Browne, W. J., Subramanian, S. V., Jones, K. and Goldstein, H. (2005) Variance partitioning in multilevel logistic models that exhibit overdispersion. *J. R. Statist. Soc. A*, **168**, 599–613.
- Craig, I. D., Plume, A. M., McVeigh, M. E., Pringle, J. and Amin, M. (2007) Do open access articles have greater citation impact?: a critical review of the literature. *J. Inf.*, **1**, 239–248.
- Daniel, H.-D. (1993) *Guardians of Science: Fairness and Reliability of Peer Review*. Weinheim: Wiley.

- Evidence Ltd. (2007) *The Use of Bibliometrics to Measure Research Quality in UK Higher Education Institutions*. London: Universities UK.
- Fielding, A. (2004) Scaling for residual variance components of ordered category response in generalized linear mixed multilevel models. *Qual. Quant.*, **38**, 425–433.
- Fielding, A., Yang, M. and Goldstein, H. (2003) Multilevel ordinal models for examination grades. *Statist. Mod.*, **3**, 127–153.
- Giske, J. (2008) Benefitting from bibliometry. *Eth. Sci. Environ. Polit.*, **8**, 79–81.
- Glänzel, W. and Schubert, A. (1992) Some facts and figures on highly cited papers in the sciences, 1981–1985. *Scientometrics*, **25**, 373–380.
- Glänzel, W., Schubert, A. and Czerwon, H. J. (1999) An item-by-item subject classification of papers published in multidisciplinary and general journals using reference analysis. *Scientometrics*, **44**, 427–439.
- Glänzel, W., Thijs, B., Schubert, A. and Debackere, K. (2009) Subfield-specific normalized relative indicators and a new generation of relational charts: methodological foundations illustrated on the assessment of institutional research performance. *Scientometrics*, **78**, 165–188.
- Godin, B. (2007) From eugenics to scientometrics: Galton, Cattell, and men of science. *Soc. Stud. Sci.*, **37**, 691–728.
- Goldstein, H. and Spiegelhalter, D. J. (1996) League tables and their limitations: statistical issues in comparisons of institutional performance. *J. R. Statist. Soc. A*, **159**, 385–409.
- Gölitz, P. (2005) Who is going to read all this? *Angew. Chem. Int. Edn.*, **44**, 5538–5541.
- Gosden, H. (2003) Why not give us the full story?: functions of referees comments in peer reviews of scientific research papers. *J. Engl. Acad. Purp.*, **2**, 87–101.
- Hames, I. (2007) *Peer Review and Manuscript Management of Scientific Journals: Guidelines for Good Practice*. Oxford: Blackwell.
- Harnad, S. (2008) Validating research performance metrics against peer rankings. *Eth. Sci. Environ. Polit.*, **8**, 103–107.
- Hedeker, D. (2008) Multilevel models for ordinal and nominal variables. In *Handbook of Multilevel Analysis* (eds J. De Leeuw and E. Meijer), pp. 237–274. New York: Springer.
- Hedeker, D. and Gibbons, R. D. (2006) *Longitudinal Data Analysis*. Hoboken: Wiley-Interscience.
- Hornbostel, S., Böhmer, S., Klingsporn, B., Neufeld, J. and von Ins, M. (2009) Funding of young scientist and scientific excellence. *Scientometrics*, **79**, 171–190.
- Hox, J. J. (2010) *Multilevel Analysis: Techniques and Applications*. New York: Routledge.
- Jackson, C. (1996) *Understanding Psychological Testing*. Leicester: British Psychological Society.
- Järvelin, K. and Persson, O. (2008) The DCI index: discounted cumulated impact-based research evaluation. *J. Am. Soc. Inf. Sci. Technol.*, **59**, 1433–1440.
- Jayasinghe, U. W., Marsh, H. W. and Bond, N. (2003) A multilevel cross-classified modelling approach to peer review of grant proposals: the effects of assessor and researcher attributes on assessor ratings. *J. R. Statist. Soc. A*, **166**, 279–300.
- Jennings, C. G. (2006) Quality and value: the true purpose of peer review. What you can't measure, you can't manage: the need for quantitative indicators in peer review, doi 10.1038/nature05032.
- Kostoff, R. N. (2002) Citation analysis of research performer quality. *Scientometrics*, **53**, 49–71.
- Kurtz, M. J. and Henneken, E. A. (2007) E-prints and journal articles in astronomy: a productive co-existence. *Learn. Publishng*, **20**, 16–22.
- Larivière, V. and Gingras, Y. (2010) The impact factor's Matthew Effect: a natural experiment in bibliometrics. *J. Am. Soc. Inf. Sci. Technol.*, **61**, 424–427.
- van Leeuwen, T. N. (2007) Modelling of bibliometric approaches and importance of output verification in research performance assessment. *Res. Eval.*, **16**, 93–105.
- Lewis, G., Thornicroft, G., Szmukler, G. and Tansella, M. (2007) Fair assessment of the merits of psychiatric research. *Br. J. Psychiatr.*, **190**, 314–318.
- Liang, K.-Y. and Zeger, S. L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.
- Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D. and Schabenberger, O. (2006) *SAS for Mixed Models*. Cary: SAS Institute.
- Lock, S. (1985) *A Difficult Balance: Editorial Peer Review in Medicine*. Philadelphia: Institute for Scientific Information.
- Marsh, H. W. and Ball, S. (1991) Reflections on the peer review process. *Behav. Brain Sci.*, **14**, 157–158.
- Marsh, H. W., Jayasinghe, U. W. and Bond, N. W. (2008) Improving the peer-review process for grant applications—reliability, validity, bias, and generalizability. *Am. Psychol.*, **63**, 160–168.
- McDonald, R. J., Cloft, H. J. and Kallmes, D. F. (2009) Fate of manuscripts previously rejected by the *American Journal of Neuroradiology*: a follow-up analysis. *Am. J. Neuroradiol.*, **30**, 253–256.
- Melin, G. and Danell, R. (2006) The top eight percent: development of approved and rejected applicants for a prestigious grant in Sweden. *Sci. Publ. Poly.*, **33**, 702–712.
- Mutz, R., Guilley, E., Sauter, U. H. and Nepveu, G. (2004) Modelling juvenile-mature wood transition in Scots pine (*Pinus sylvestris* L.) using nonlinear mixed-effects models. *Ann. For. Sci.*, **61**, 831–841.

- Neuhaus, C. and Daniel, H.-D. (2009) A new reference standard for citation analysis in chemistry and related fields based on the sections of Chemical Abstracts. *Scientometrics*, **78**, 219–229.
- Neuhaus, C., Marx, W. and Daniel, H.-D. (2009) The publication and citation impact profiles of *Angewandte Chemie* and the *Journal of the American Chemical Society* based on the sections of *Chemical Abstracts*: a case study on the limitations of the Journal Impact Factor. *J. Am. Soc. Inf. Sci. Technol.*, **60**, 176–183.
- Ophof, T., Furstner, F., van Geer, M. and Coronel, R. (2000) Regrets or no regrets?: no regrets! The fate of rejected manuscripts. *Cardvasc. Res.*, **45**, 255–258.
- Ophof, T. and Leydesdorff, L. (2010) Caveats for the journal and field normalizations in the CWTS (“Leiden”) evaluations of research performance. *J. Inf.*, **4**, 423–430.
- Pendlebury, D. A. (2008) *Using Bibliometrics in Evaluating Research*. Philadelphia: Thomson Scientific.
- Plomp, R. (1990) The significance of the number of highly cited papers as an indicator of scientific prolificacy. *Scientometrics*, **19**, 185–197.
- van Raan, A. F. J. (2005) For your citations only?: hot topics in bibliometric analysis. *Measurement*, **3**, 50–62.
- Radich, F., Fortunato, S. and Castellano, C. (2008) Universality of citation distributions: toward an objective measure of scientific impact. *Proc. Natn. Acad. Sci. USA*, **105**, 17268–17272.
- Raudenbush, S. W., Bryk, A. S., Cheong, Y. F. and Congdon, Jr, R. T. (2004) *HLM 6: Hierarchical Linear and Nonlinear Modeling*. Lincolnwood: Scientific Software International.
- Reinhart, M. (2009) Peer review of grant applications in biology and medicine: reliability, fairness, and validity. *Scientometrics*, **81**, 789–809.
- Rons, N. and Amez, L. (2008) Impact Vitality—a measure for excellent scientists. In *Excellence and Emergence: a New Challenge for the Combination of Quantitative and Qualitative Approaches* (eds J. Gorraiz and E. Schiebel), pp. 211–213. Vienna: Austrian Research Centers.
- Rosenbaum, P. R. (2002) *Observational Studies*. New York: Springer.
- Rousseau, R. (2005) Median and percentile impact factors: a set of new indicators. *Scientometrics*, **63**, 431–441.
- Schubert, A. and Braun, T. (1996) Cross-field normalization of scientometric indicators. *Scientometrics*, **36**, 311–324.
- Seglen, P. O. (1997) Why the impact factor of journals should not be used for evaluating research. *Br. Med. J.*, **314**, 498–502.
- Shatz, D. (2004) *Peer Review: a Critical Inquiry*. Lanham: Rowman and Littlefield.
- Thomson Reuters (2008) *Using Bibliometrics: a Guide to Evaluating Research Performance with Citation Data*. Philadelphia: Thomson Reuters.
- Tijssen, R. and van Leeuwen, T. (2006) Centres of research excellence and science indicators: can ‘excellence’ be captured in numbers? In *Proc. 9th Int. Conf. Science and Technology Indicators, Leuven* (ed. W. Glänzel), pp. 146–147. Leuven: Katholieke Universiteit Leuven.
- Tijssen, R., Visser, M. and van Leeuwen, T. (2002) Benchmarking international scientific excellence: are highly cited research papers an appropriate frame of reference? *Scientometrics*, **54**, 381–397.
- Vinkler, P. (1997) Relations of relative scientometric impact indicators: the relative publication strategy index. *Scientometrics*, **40**, 163–169.
- Wilson, J. D. (1978) Peer review and publication. *J. Clin. Invest.*, **61**, 1697–1701.
- Winship, C. and Mare, R. D. (1983) Structural equations and path analysis for discrete data. *Am. J. Sociol.*, **89**, 54–110.
- Winship, C. and Mare, R. D. (1984) Regression models with ordinal variables. *Am. Sociol. Rev.*, **49**, 512–525.
- Wuchty, S., Jones, B. F. and Uzzi, B. (2007) The increasing dominance of teams in production of knowledge. *Science*, **316**, 1036–1039.