Factors Affecting Citation Rates of Research Articles

Natsuo Onodera*
Professor Emeritus, the University of Tsukuba, 1-2, Kasuga, Tsukuba, Ibaraki 305-8550, Japan.
E-mail: nt.onodera@y5.dion.ne.jp

Fuyuki Yoshikane
Faculty of Library, Information and Media Science, the University of Tsukuba, 1-2, Kasuga, Tsukuba, Ibaraki 305-8550, Japan.
E-mail: fuyuki@slis.tsukuba.ac.jp

* Correspondence author

The purpose of this study is to examine whether there are some general trends across subject fields regarding the factors that affect the number of citations of articles, especially focusing on those factors that are not directly related to the quality or content of articles (extrinsic factors). For this purpose, from six selected subject fields (condensed matter physics, inorganic and nuclear chemistry, electric and electronic engineering, biochemistry and molecular biology, physiology, and gastroenterology), original articles published in the same year were sampled out (*n* = 230‐240 for each field). Then, the citation counts received by the articles in a relatively long citation window (6 and 11 years after publication) were predicted by negative binomial multiple regression (NBMR) analysis for each field. Various article features about author collaboration, cited references, visibility, authors' achievement (measured by past publications and citedness), and publishing journals were considered as the explanatory variables of NBMR. Some generality across the fields were found regarding the selected predicting factors and the degree of significance of these predictors. The Price index was the strongest predictor of citations and number of references was the next. The effect of number of authors and authors' achievement measures were rather weak.

Introduction

The application of citation data to research evaluation has attracted a lot of attention in recent years (Moed, 2005). The concept is to obtain a quantitative measure of the importance of an article using its citation rate. Expanding this approach by grouping articles according to researchers (i.e., the authors of the articles), research groups, research institutions, and countries would make it possible to conduct evaluations of individuals and groups as well as articles.

There are many criticisms regarding the use of citation data for research evaluation (Lindsey, 1989; MacRoberts & MacRoberts, 1987; 1989; 1996; 2010). However, it is undeniable that citation rate gives the most appropriate statistical indicator measuring an aspect of importance of research (the degree of impact or utilization of articles) among those presently available. Hence, citation data can be used as important information for research evaluation provided that it is done carefully and its limitations are considered (van Raan, 1996; Moed, 2005). It should be noted, of course, that research must be evaluated from various aspects, and citation rates provide valuable data as one of these aspects. The measures based on citations are not objective indicators of evaluation themselves but complementary information for subjective peer review.

Even if it is generally accepted that the citation count of an article is an effective measure of its importance, an individual article's count does not always agree with the assessment of the article. Many studies have demonstrated that the correlation between citation rate and score of peer evaluation is moderate; that is, the correlation coefficient is approximately 0.4–0.6 (Oppenheim, 1997; Rinia, van Leeuwen, van Vuren, & van Raan, 1998; Aksnes, 2006; Abramo, D'Angelo, & Di Costa, 2011; Mryglod, Kenna, Holovatch, & Berche, 2013). The reasons for this correlation are as follows:

(a) The citation rate is a measure of only one aspect of research (impact or utility).

(b) Citations do not always positively refer to the cited articles.

(c) The citation count of an article is influenced by various “extrinsic” factors not directly related to the content or quality of the article.

Point (a) is an important matter that has to be kept in mind when using citation data for research evaluation, as mentioned above. Concerning point (b), the reported ratios of negative citations are generally low; Moravcsik and Murugesan (1975), Chubin and Moitra (1975), and Krampen, Becker, Wahner, and Montada (2007) reported rates of 13%, 3%–6%, and <1%, respectively. Therefore, this is not a serious problem from a statistical viewpoint.

The purpose of this study is related to point (c). It is well known that the citation rate of an article is influenced by the subject field, country where the journal is issued, type of article (original article, short report, review, etc.), and language. Therefore, it is almost meaningless to simply compare articles of different types that belong to different fields based on their raw (not normalized) citation counts. In addition, as discussed in the next section, many studies have been conducted about various factors that may influence the citation rate of an article. However, there is still no consensus about which factors significantly affect citation rate. This is in part because most existing studies focus on a single factor (or multiple factors as mutually independent); hence, they could not consider interactions among different factors. Another reason is that some studies considering integratedly multiple factors (using multiple regression models in general) restrict the subject or source of the sample articles to a specific area, hence, the generality of the conclusions might be limited.

In this study, we investigated the variation in citation rates among articles and their dependence on numerous factors for articles of the same type (original articles) published in the same year in several journals (English language only) for several different fields. Using negative binomial multiple regression analysis, we attempted to determine the contribution of each factor. If the analyses of several different fields result in some common tendencies, then it is expected that a reference citation rate would be given for an article with a set of factor values, and that the citation data could be more accurately applied to research evaluation.

Literature Review

As described above, it is well known that the citation rate of an article depends on the field, type, and language of the article. In this section, addressing other features of articles, we briefly review the results of studies that have investigated whether these features influence the citation rates of articles. This review first examines the studies investigating many potential influencing factors integratedly (most of these studies use multiple regression analysis), and then discusses the influence of individual factors. We restricted the range of the review to citation rates of individual articles rather than aggregates of articles on particular authors or research groups, which meets the objective of this research.

Citation analyses considering integratedly various potentially influencing factors

Although numerous features have been investigated as potential factors that influence the citation rate of articles, most of the studies conducted to date have focused on a single factor or have considered multiple factors as mutually independent. Therefore, even when a correlation does exist between a factor and citation rate, it is not possible to exclude the possibility that the correlation is due to confounding of other factor(s). For example, the number of authors, the number of institutions, the article length, and the number of references have been reported in many studies as positively correlated to the citation rate of articles. However, it is worth noting that these factors may be positively correlated to each other, meaning that it is likely that only some of the factors have significant correlations with citation rate when each factor is assessed separately.

Multiple regression analysis is the most commonly used approach to separate the effects of individual factors (independent variables) and to identify the factors that are significant with respect to citation rate. In this subsection, we outline several studies using this method.

Studies on main determinants of citation rates of articles. Peters and van Raan (1994) noted that almost no studies have investigated a broad spectrum of factors to identify fewer factors that primarily determine citation rates, despite the many studies on factors influencing citation scores. From this viewpoint, they investigated the extent to which various factors influence the number of citations of articles in the field of chemical engineering. They selected eighteen internationally reputed scientists from the field and counted the number of citations received within five years after publication by each of the articles (n = 226) published by those scientists between 1980 and 1982.

Multiple regression analysis using 14 factors as explanatory variables showed that the highly significant explanatory variables were as follows (in decreasing order of partial correlation coefficient): (a) the scientist's rank according to the number of articles published between 1980 and 1982, (b) number of references, (c) language, (d) reputation of the publishing journal, (e) influence weight (Narin's indicator of journal influence), and (f) Price index. The following four variables were also significant: (g) CA Section, (h) number of pages, (i) number of authors, and (j) nationality of the scientist. Interestingly, the scientist's rank, which had extremely high explanatory power, did not show a significant relation with the number of citations when the correlation between these two variables was simply calculated, which indicates the importance of integrated analysis considering various factors that may influence the citation rates of articles.

Didegah and Thelwall (2013) investigated the determinants of citation rates in their study using over 50,000 articles published from 2007 to 2009 in the field of nanoscience and nanotechnology. They selected eight factors as the independent variables, two of which had not been considered until that time –

internationality of the publishing journal and internationality of references (both were measured by Gini coefficient for the geographic distribution of authors). Then, they carried out zero-inflated negative binomial regressions for four article sets (for each of publication years and for 3 years together). The journal impact factor (JIF) and impact of references (the mean citations of referenced publications) were the most strongly influencing factors, and number of references, internationality of references, and number of institutions of affiliation were also significant predictors for all article sets. On the other hand, number of authors was shown to have little influence. Number of countries of affiliation and journal's internationality tended to give a negative effect on citations, which might be related to dominance of the USA in the research in this field.

What is dominant among the features of author, journal, and article? There have been some studies that divide potential factors that influence the citation rates of articles into the factors of author, journal, and article itself, and examine which factor is the most dominant. The earliest study of this kind is by Stewart (1983), wherein he tried to predict the citation count received by each of 139 articles published in 1968 in the field of geoscience using multiple regression models including many author and article variables. He concluded that the article features were more important than the author features from a comparison of the coefficient of determination (R^2) among the regression with author variables only, article variables only, and all variables. The article variables that contributed to high citations were number of references, article length, time from acceptance to publication, and (to a lesser extent) the recency of references. Moreover, several dummy variables on the subject or type of article (e.g., relevance to plate tectonics) had a significant influence on citations. On the other hand, significant author variables included average citations per article published in the past by the author(s) and the proportion of authors with a university affiliation. The number of authors was not a significant predictor.

Walters (2006) used negative binomial regression analysis to predict the citation counts of 428 articles published in 12 prime-psychology journals in 2003, with nine explanatory variables including author, article, and journal characteristics. The results revealed significant positive effects of the average citations of the first author's past publications, the first author's nationality (whether the USA or not), and whether or not it was a review article. Multi-authorship and the journal impact were significant, while the first author's gender and occupational affiliation, article length, and the subject of the article (correctional/criminological) were not significant. From these results, Walters suggested that the author characteristics might be more powerful for citation prediction than the journal and article characteristics.

Haslam et al. (2008) analyzed the citation counts of 308 articles published in three major journals of social-personality psychology in 1996. Thirty potential factors affecting citations were classified into four groups (characteristics of author, institution, article organization, and research approach). Multiple regression analyses were performed at two stages: first, using the characteristics within each of the four groups as the explanatory variables and second, using nine variables that were significant in the first four analyses. The main factors that increased citations were (a) high productivity (number of past publications) of the first author, (b) existence of a co-author with higher productivity than the first author, (c) high journal prestige, (d) more pages in the article, (e) more references, and (f) recency of references. Aggregate productivity of the authors other than the first author, competitive grant support, length of the article title, and whether it was a theoretical/review article were significant at the first stage, but not at the second stage.

Peng and Zhu (2012) used 18,580 social science articles about internet studies. They also carried out two-stage multiple regression analyses, using article characteristics (including author characteristics) as explanatory variables at the initial stage and adding journal characteristics at the second stage. The results indicated a stronger effect of the journal characteristics, especially the JIF. Significant predictors, however, included some article characteristics such as article length, number of authors, topical popularity (measured as the number of internet-related words in the abstract), the proportion of highly-cited publications in references, and active years of the first author.

"Signals" bringing quick attention to an article. Van Dalen and Henkens (2001; 2005) examined which factors influence the citation impact of articles in the field of demography to determine whether the factors shown to influence the citation impact in natural sciences are also applicable to the social sciences field. They especially focused on the roles of author and journal reputation as "signals" that brought quick attention to an article. They counted citations received by each of 1,371 articles in this field with citation windows of 5 years (van Dalen & Henkens, 2001) and 10 years (van Dalen & Henkens, 2005) after publication and developed several negative binomial regression models using characteristics of authors, visibility, content, and publishing journal as the explanatory variables.

The variables regarding journal reputation, such as the JIF, journal circulation number, and reputation of the editorial board (measured by the average number of citations obtained by the editorial board members) showed extensive influence on the citation rate of articles, while the influence of the variable regarding author reputation, which was measured by the accumulated number of citations obtained by the author (the author with the highest accumulated citations in case of co-authored articles), was significant but less influential. Other variables that showed highly significant associations with the citation rate of articles were article type (notes or comments were less cited than normal articles), number of pages, regional focus of the article (articles focusing on the USA or Europe were highly cited), and the language of the journal. Additionally, the author's nationality, number of authors, and the position of the article in the journal issue showed moderately significant relations to the article's citation rate.

Other studies. In addition to those mentioned above, some studies investigate integratedly factors that potentially influence the citation rate of articles. These studies focused on the effect of a specific factor on citations, taking various controlling variables into account.

To investigate whether the peer review system of refereed journals fulfills its objective to select superior work, Bornmann and Daniel (2008) compared the citation rates between 878 accepted articles and 959 articles that were initially rejected but later accepted by another journal, both of which were submitted to *Angewandte Chemie International Edition (ACIE)* in 2000. The results of negative binomial regression analysis controlling the possible effect of various influencing factors revealed that the accepted articles had an advantage over the rejected articles by 40–50% in the average number of citations. Regarding the control factors, the language of the article (English or other) and author's status (number of authors in the article listed at ISIHighlyCited.com) were statistically significant. Furthermore, articles about organic, physical/inorganic, and macromolecular chemistry had significantly more citations than those about applied chemistry and biochemistry (subject classification according to CA Sections). The number of authors was statistically significant only in a short citation window (three years after publication). Article length was not statistically significant, probably because the articles used in this study were from the "Communications" of *ACIE*, which generally publishes short articles.

Lokker, McKibbon, McKinlay, Wilczynski, and Haynes (2008) tested the predictability of citation counts of clinical articles that met basic criteria for critical appraisal from data obtained within three weeks of publication. The explanatory variables in multiple regression analysis included those about quality assessment of the articles and publishing journals as well as the usual bibliographic attributes. The significant predictors of citations were the average relevant score by raters (the average newsworthiness score was not significant) and selection by EBM synoptic journals with regard to article quality; number of databases indexing the journal and the proportion of articles recorded in EBM synoptic journals with regard to journal quality; and number of authors, number of references, being a multicenter study, and being a therapy article with regard to bibliographic attributes.

Under the hypothesis that open access (OA) articles have a higher citation impact than non-open access (NOA) articles because of biases toward OA (e.g., self-selection by the authors), Davis, Lewenstein, Simon, Booth, and Connolly (2008) conducted a randomized controlled trial (RCT) that compared the citation rates between randomly-assigned OA and NOA articles. Negative binomial regression analysis including many control variables revealed no evidence that OA articles received higher rates of citation than NOA articles. Among the control variables considered, the number of authors, inclusion of author(s) from the USA, the JIF, and number of references were significant, while article length was not significant.

Intending to argue against the accepted view that internationally co-authored articles have a higher citation rate compared to domestic ones, He (2009) applied negative binomial regression analysis to articles by biomedical researchers in New Zealand using several co-authorship variables and controlling variables as the explanatory variables. The results revealed that adding one author from the same institute brought an effect on increase of the citation count of an article comparable to adding one foreign (outside New Zealand) author. Among the controlling variables used, more cited references boosted the citation count; the significance of h-index depended on the model used, and the length and type of article was not significant.

Fu and Aliferis (2010) developed a supervised learning model using the SVM (support vector machines) algorithm for predicting long-term citation counts (10 years after publication) of articles. Applying this model to 3,788 articles about internal medicine sampled from eight general medical journals published between 1991 and 1994, they selected effective variables from many content-based features (terms extracted from the title, abstract, and indexed MeSH of each article) and nine bibliometric features. Within the selected variables, they further identified significant variables using logistic regression analysis that differentiated between more highly-cited and less-cited articles. Among the bibliometrics features, only the JIF and accumulated number of citations obtained by the last author were significant, while the publication type and accumulated number of citations obtained by the first author were not significant in the final logistic regression. The effective content-based features varied greatly according to the threshold value of the citation count in the logistic regression.

The studies described in this subsection are summarized in Table 1. (Chen (2012) in this table is introduced in “Quantitative relations between citation rates and measures of the quality or content of articles” in the next subsection.) Although each of these studies yielded interesting results, the conclusions of these studies do not have generality because the sample articles used were restricted to a specific field (Stewart, 1983; Peters & van Raan, 1994; van Dalen & Henkens, 2001; van Dalen & Henkens, 2005; Walters, 2006; Haslam et al., 2008; Lokker et al., 2008; Fu & Aliferis, 2010; Peng & Zhu, 2012), to one (or a few) specific journal(s) (Bornmann & Daniel, 2008; Davis et al., 2008）, or to articles by authors from a specific nation (He, 2009).

In this research, using only “extrinsic” factors that do not directly associate to the quality or content of articles as explanatory variables, we aim to find the factors affecting citation rate of articles common to several different fields to examine whether there are general tendencies among fields.

Table 1 Outline of studies considering integratedly various citation-influencing factors.

				Potential factors influencing citations of articles[b)]																	
Work	Target field	Sample size (n)	Analysis[a)]	#Au-thors	#Insti-tutions[c)]	#Coun-tries[d)]	#Refer-ences	Recency of refs	Other features of refs	#Fig-ures	#Tables	Article length	Author's affili-ation	Author's national-ity	Author's produc-tivity	Author's cited-ness	Other status of author	Subject/ content	Quality of research	Journal impact	Other journal features
Bornmann & Daniel (2008)	Chemistry	1,837	NBMR	B								C					A	A	A		
Chen (2012)	Four topics	1,300 - 6,800 for a topic	NBMR, ZINBMR	A			A					C							A		
Davis et al. (2008)	Physiology	1,619	NBMR	A			A					C		A				B		A	
Didegah & Thelwall (2013)	Nano sci-tech	50,162	ZINBMR	C	A	C	A		A											A	B
Fu & Aliferis (2010)	Internal medicine	3,788	SML, LogMR	C	C								C		C	A		B		A	
Haslam et al. (2008)	Social-personality psychology	308	LMR (LogC)	C			A	A		C	C	A	C	C	A			C	B	A	
He (2009)	Biomedicine	1,860	NBMR	A		A	A					C				B	C				
Lokker et al. (2008)	Clinical medicine	1,261	LMR (RootC)	A			A					C		C				A	A	A	A
Peng & Zhu (2012)	Internet research	18,580	LMR (Root(C/y))	A					A			A		A			A			A	B
Peters & van Raan (1994)	Chemical engineering	226	CMR	B			A	A				B		B	A			B		A	
Stewart (1983)	Geoscience	139	LMR (logC)	C			A	B				A	B			A	C	A			
Van Dalen & Henkens (2001)	Demography	1,371	NBMR	B								A		A		A		A		A	A
Walters (2006)	Prime psychology	428	NBMR	B								C		A		A		C		B	

a) LMR:Linear multiple regression, CMR:Categorical multiple regression, NBMR:Negative binomial multiple regression, ZINBMR:Zero-inflated negative binomial multiple regression, LogMR:Logistic multiple regression, SML:Supervised machine learning LogC, RootC, and Root(C/y) show the independent variable of LMR.

b) A:Strong or definite predictor B:Weak predictor or predictive power dependent on the model, C:Not-significant or negative predictor

c) Including domestic interinstitutional collaboration and number of co-authors of other domestic institutions

d) Including international collaboration and number of foreign co-authors

Analyses of the individual factors potentially influencing citations

This subsection summarizes the findings that have been reported about the main factors that potentially influence citations, which include those from the integrated analyses mentioned in the preceding subsection and those from other studies that have investigated the relations between citation rates and specific factors.

Does collaboration boost the citation rate of articles? There have been many studies investigating the relationship between citation rate and the number of authors of an article. These studies, at least partially, rely on the hypothesis that adding authors to an article leads to more citations of the article because the authors would have different scientific influences. This idea was expanded to examine the relationship between the number of collaborative institutions or countries in articles and the citation rates of the articles. Such research would be motivated by questions about whether joint research among different research institutions or countries contributes to an increased impact of the research.

The studies by Basu and Levinson (2005), Figg et al. (2006), and Sooryamoorthy (2009) focused on the relationship between citation rates of articles, and the number of authors and their institutions and/or countries. Their results showed that in many cases multiple authors, institutions, or countries have a positive effect on the citation rate of an article.

Persson, Glänzel, and Danell (2004) investigated the relationship between the number of authors of articles and the citation rate received by the articles in two publication years, 1980 and 2000, to test their hypothesis that the inflationary tendency of co-authorships in the last two decades is (at least) one cause of the increase in references (consequently, citations) per article. For both years, a clear relation was shown that adding one author to an article resulted in adding 0.6 citations on average in the 3-year citation window after publication. They also found that the average number of citations for articles with the same number of authors increased by 8 from 1980 to 2000, which implies that the spread of research collaboration is not the only cause of growth of citations during this period.

In addition to these studies, many studies have reported a positive correlation between the number of authors and the citation rate of articles (Aksnes, 2003a; Leimu & Koricheva, 2005; Bornmann & Daniel, 2006; Davis et al., 2008; Lokker et al., 2008; Sin, 2011; Chen, 2012; Peng & Zhu, 2012; Fanelli, 2013; Rigby, 2013). However, some studies using multiple regression analysis with numerous explanatory variables demonstrated that the ability of the number of authors to predict the citation impact of articles is weak (Peters & van Raan, 1994; van Dalen & Henkens, 2001; Walters, 2006; Bornmann & Daniel, 2008) or insignificant (Stewart, 1983; Fu & Aliferis, 2010).

Analyzing eight journals that publish many articles and also have a high JIF, Hsu and Huang (2011) showed that the probability that an article with more authors gains more citations than an article with fewer authors is not as high as expected, i.e., 53– 65% depending on the journal, although statistically the more authors an article has, the more citations it tends to receive. On the basis of their observation that there is no significant relation between the number of authors and the citation rate within article sets in which only the articles by highly-cited authors were extracted, Levitt and Thelwall (2009) suggested that the apparent positive correlation between them seen in a mixed article may reflect a positive correlation between the average number of co-authors and the average citation count of individual authors.

Articles with international collaboration, which have multiple author affiliation countries, have been suggested to be more highly cited than those with local or domestic collaboration (Katz & Hicks, 1997; van Raan, 1998; Persson et al., 2004; Sooryamoorthy, 2009; Sin, 2011; Peclin, Juznic, Blagus, Sajko, & Stare, 2012; Ibanez, Bielza, & Larranaga, 2013; Bordons, Aparicio, & Costas, 2013). For example, Katz and Hicks (1997) reported that adding one foreign co-author increased citations by 1.6 per article on average, while adding one co-author from the same or different domestic institution resulted in an increase of only 0.75 citations. In contrast, He (2009) suggested that grouping articles into international, national, and local categories used in many of the previous studies may understate the contribution of local or national co-authorship to the citation impact compared with that of international co-authorship. This is because the average number of authors in the international group would be larger than that in the national or local group, and the number of authors would positively correlate with citation rate. He assigned the three collaboration variables (numbers of foreign, domestic, and local co-authors) to each article in his sample (1,860 articles published by 65 biomedical scientists at a university in New Zealand) and indicated through negative binomial regression analysis that the effect of adding one local co-author on the citation impact was comparable to that of adding one international co-author. (Domestic collaboration was not significantly associated with the citation impact.)

Table 2 summarizes the studies mentioned here excluding those included in Table 1.

Table 2 Outline of studies investigating attributes on collaboration as citation-influencing factors.

Work	Target field	Sample size (n)	Analysis [a]	Influencing factors [b]		
				#Authors	#Institutions [c]	#Countries [d]
Aksnes (2003a)	Natural sciences	46,849	Simple comparison	A		A
Basu & Levison (2005)	Astronomy & astrophysics	95,186	LMR (Log(C+1))	A	A	C
Bordons et al. (2013)	Pharmacology & Pharmacy	1,971 and 2,858 (Two samples)	CMR	A	A	A
Bornmann & Daniel (2006)	Biomedicine	1,586	NBMR	A		
Fanelli (2013)	Hypthesis-testing	2,545	NBMR	A		
Figg et al. (2006)	Medicine	164 - 886 (Six samples)	LR (LogC)	A	A	
Hsu & Huang (2011)	Natural sciences	10,000 - 15,000 (Eight samples)	LR	A		
Ibáñez et al. (2013)	Computer science	ca. 20,000	Mann-Whitney test; Kruskal-Wallis test	C		A
Katz & Hicks (1997)	General	c.a. 376,000	LR	A	A	A
Leimu & Koricheva (2005)	Ecology	228	t-test; ANOVA; Correlation	A		
Peclin et al. (2012)	Natural sciences	5,263	ANOVA		B	A
Persson et al. (2004)	General	All WoS articles in 1980 and 2000	LR	A		A
Rigby (2013)	Biochemistry	3,596	LMR (Log(C+1))	A		C
Sin (2011)	Library & inf science	7,489	LogMR	A	C	A
Slyder et al. (2011)	Geography & Forestry	213	t-test; ANOVA; Correlation	C		
Sooryamoorthy (2009)	General	11,196	LR (LogC)	A	C	A
van Raan (1998)	Astronomy	2,090	Simple mean comparison		C	A

a) LR:Linear single regression, LMR:Linear multiple regression, CMR:Categorical multiple regression, NBMR:Negative binomial multiple regression, LogMR:Logistic multiple regression
LogC shows the independent variable of LMR.

b) A:Strong or definite predictor B:Weak predictor or predictive power dependent on the model, C:Not-significant or negative predictor

c) Including domestic interinstitutional collaboration and number of co-authors of other domestic institutions

d) Including international collaboration and number of foreign co-authors

Do articles determine the journal impact, or is the reverse true? It is natural that the citation rate of an article is positively associated with the citation impact of the journal in which it is published, but it is debatable whether the quality or importance of an article determines the journal impact indicators such as JIF (Seglen, 1994), or the reputation of a journal attracts citations to articles in that journal (van Dalen & Henkens, 2005).

As mentioned in the preceeding subsection, Van Dalen and Henkens (2001; 2005) showed that the journal reputation measures such as the JIF, the average number of citations obtained by the editorial board members, and the circulation numbers had a strong positive influence on the citation impact. Some other studies described in the preceding subsection took the JIF (or other impact indicator) of the journal publishing an article as one of the most important factors to increase the citation rate of that article (Peters & van Raan, 1994; Davis et al., 2008; Fu & Aliferis, 2010; Peng & Zhu, 2012 Didegah & Thelwall, 2013). However, the explanatory power of the JIF was not as strong by Walters (2006).

Callaham, Wears, and Weber (2002), Aksnes (2003a), Bornmann and Daniel (2006), Slyder et al. (2011), and Wang, Yu, and Yu (2011) also reported an association between the citation rate of articles and the JIF of the journals in which they were published. Moreover, Lariviere and Gingras (2010b) used a unique method of comparing 4,532 pairs of "duplicate" articles with the same title, the same first author, and the same number of references published in two different journals; they reported that the article published in a higher-impact journal obtained on average twice as many citations as its counterpart published in a lower-impact journal. The obvious difference in citation count between identical articles is strongly suggestive of the halo effect of journal prestige on the scientific impact of articles.

Table 3 summarizes the studies mentioned here excluding those included in Table 1.

Table 3 Outline of studies investigating journal attributes as citation-influencing factors.

Work	Target field	Sample size (n)	Analysis [a)]	Influencing factors [b)]	
				JIF	Other citation impact
Aksnes (2003a)	Natural sciences	46,849	Simple comparison	A	
Bordons et al. (2013)	Pharmacology & Pharmacy	1,971 and 2,858 (Two samples)	CMR		A
Bornmann & Daniel (2006)	Biomedicine	1,586	NBMR	A	
Callaham et al. (2002)	Emergency medicine	204	CMR	A	
Ibáñez et al. (2013)	Computer science	ca. 20,000	Mann-Whitney test; Kruskal-Wallis test	A	
Larivière & Gingras (2010b)	General	4,532 pair	t-test	A	
Slyder et al. (2011)	Geography & Forestry	213	t-test; ANOVA; Correlation	A	

a) CMR:Categorical multiple regression, NBMR:Negative binomial multiple regression,

b) A:Strong or definite predictor B:Weak predictor or predictive power dependent on the model, C:Not-significant or negative predictor

Is there a halo effect of authors, institutions, or countries? There have been many discussions about the halo effect on scientific impact suggesting that articles written by authors with good reputations or high achievement levels or authors affiliated with famous institutions attract more citations than those written by other authors.

The measures of reputation or achievement of authors include indicators that are based on the number of publications in the past (published before the publication of the article under consideration), the citation count received by the past publications (before the publication of the article under consideration), active years, and present status. Most of the studies mentioned in the preceding subsection have used such indicators as the explanatory variables of their prediction model. Peters and van Raan (1994) and Haslam et al. (2008) determined past publications to be an effective predictor of citations, but Fu and Aliferis (2010) did not find it to be significant. As the indicators of the past citations, aggregated citations (Fu & Aliferis, 2010), average citations per article (Stewart, 1983; Walters, 2006), h-index (He, 2009; Wang et al., 2011; Wang, Yu, An, & Yu, 2012), and the proportion of authors appearing in the ISI Highly Cited list (Bornmann & Daniel, 2008) were significant predictors in all cases. Some reports claimed that articles by senior authors tended to receive higher citations (Slyder et al., 2011; Peng & Zhu, 2012), whereas others contradicted the claim (Stewart, 1983; He, 2009).

Danell (2011) investigated whether the citation rate of an article in the future can be predicted from the author's previous publication number and citation rate using two article sets of limited subject areas (episodic memory and Bose-Einstein condensation). Using quantile regression models, he found that the previous citation rate was a significant predictor at most quantiles of the dependent variables (future citation rate) and was more significant at higher quantile values, while the previous publication number was not significant at most quantiles except in some quantiles near the median.

As an indicator of the status of the institution with which the author is affiliated, Leimu and Koricheva (2005) and Fu and Aliferis (2010) used the rank given by the Academic Ranking of World Universities (ARWU), and Stewart (1983) used the past publications by the institution, but these indicators were found to provide no or very weak influence on citations.

Regarding bias toward particular countries in citation rates, it has been suggested that articles by authors in a few highly productive countries, such as the USA, tend to acquire higher numbers of citations because authors tend to favorably cite articles by other authors from their own country. Using several sets of hypothesis-testing articles in the field of ecology, Leimu and Koricheva (2005) indicated that the annual citation rate of articles was positively associated with authors from English-speaking nations compared with non-English-speaking nations and with US authors compared with European authors. Cronin and Shaw (1999) showed that in the field of library and information science, the proportion of uncited articles was lower in the case of a first author from the USA, UK, or Canada than from other countries. There have been other reports demonstrating an association of articles by authors from the USA or western/northern Europe with higher citation rates (van Dalen & Henkens, 2001; van Dalen & Henkens, 2005; Basu & Lewison, 2005; Walters, 2006; Davis et al., 2008; Sin, 2011; Peng & Zhu, 2012). On the other hand, Peters and van Raan

(1994), Haslam et al. (2008), and Lokker et al. (2008) reported that the affiliation country of authors was not an important factor for predicting the citation rate of articles.

Pasterkamp, Rotmans, de Kleijn, and Borst (2007) examined the relationship of the affiliation countries of corresponding authors of articles published in 1996 in six cardiovascular journals to those of corresponding authors of references cited by those articles, and indicated that authors cited articles by authors from their own country as much as 32% more frequently than expected, even excluding author self-citations. The bias toward self-country citations was observed for all countries and in all journals that were investigated. Schubert and Glänzel (2006) also reported evidence of the tendency for self-country citations. However, some studies denied this tendency based on a modified method for calculating self-country (or self-language) citation rate (Bookstein & Yitzhaki, 1999; p.291–300 in Moed, 2005).

Table 4 summarizes the studies mentioned here excluding those included in Table 1.

Table 4 Outline of studies investigating attributes on author's status as citation-influencing factors.

Work	Target field	Sample size (n)	Analysis [a)]	Influencing factors [b)]				
				Author's affiliation	Author's nationality	Author's productivity	Author's citedness	Author's career
Bornmann & Daniel (2006)	Biomedicine	1,586	NBMR					A
Basu & Levison (2005)	Astronomy & astrophysics	95,186	LMR (Log(C+1))		A			
Cronin & Shaw (1999)	Library and inf science	716	Chi-squre test		A			
Danell (2011)	Two topics	(a) 728 (b) 1,450	Quantile regression			B	A	
Leimu & Koricheva (2005)	Ecology	228	t-test; ANOVA; Correlation	B	A			
Slyder et al. (2011)	Geography & Forestry	213	t-test; ANOVA; Correlation					A

a) LMR:Linear multiple regression, NBMR:Negative binomial multiple regression
Log(C+1) shows the independent variable of LMR.

b) A:Strong or definite predictor B:Weak predictor or predictive power dependent on the model, C:Not-significant or negative predictor

Quantitative relations between citation rates and measures of the quality or content of articles. Although quality and content are the most crucial factors determining the citation rate of an article, quantitative analysis of this is difficult. There have, however, been several studies examining the relation between the number of citations of articles and their quality ranking or subject classes. Although this study mainly focuses on the citation-influencing factors not directly related to the quality or content of articles, these studies are briefly discussed below.

Some studies have examined whether the citation impact of hypothesis-testing articles differs depending on the testing method or test results. These studies analyzed the correlation between the citation rate and the following attributes: sample size or the type of subjects (Callaham et al., 2002; Leimu & Koricheva, 2005; Lortie, Aarssen, Budden, & Leimu, 2013); presence/absence of a control group or randomization (Callaham et al., 2002); positiveness/negativeness of the results or support/rejection of the hypothesis (Callaham et al., 2002; Leimu & Koricheva, 2005; Fanelli, 2013); and strength of statistical significance (Leimu & Koricheva, 2005). In many cases, however, the correlation was either not significant or, if present, weak.

Some of the studies mentioned in the preceding subsection involved features concerning article's content as the explanatory factor in their multiple regression models. These features include topic terms of medical articles (Fu & Aliferis, 2010); research design of clinical medicine research (Lokker et al., 2008); subfields and research methods of geoscience research (Stewart, 1983); and themes of demographic articles (van Dalen & Henkens, 2001; 2005).

A considerable number of studies have assessed the connection between an article's citation rate and the peer evaluation it received. The indicators of peer evaluation examined in the studies were as follows: the results of peer reviews of manuscripts submitted to a journal (Bornmann & Daniel, 2008; Patterson & Harris, 2009; Bornmann, Schier, Marx, & Daniel, 2011); acquisition of competitive funding mentioned in the acknowledgments (Cronin & Shaw, 1999; Haslam et al., 2008; Rigby, 2013); and self-evaluation by the authors (Aksnes, 2006). Some studies addressing articles of clinical medicine analyzed the dependence of citation rate on the score of clinical relevance and newsworthiness (Callaham et al., 2002; Lokker et al., 2008), methodological rigor (Akcan, Axelsson, Bergh, Davidson, & Rosen, 2013), and whether they were abstracted by EBM synoptic journals (Lokker et al., 2008). However, these indicators mentioned here could not be obtained from the articles, except for funding information in the acknowledgments.

It is difficult to represent the quality of an article by a quantitative measure that does not rely on self-evaluation or peer review. A recent study by Chen (2012) is notable in this regard. Chen proposed to represent the potential (or value) of an article in terms of the degree to which it alters the intellectual structure of the state-of-the-art (an ability of "boundary-spanning") and to measure this ability by three metrics quantifying the change in the existing intellectual network structure: (a) modularity change rate, (b) cluster linkage, and (c) centrality divergence. Using these three "intrinsic" attributes and three traditional "extrinsic" attributes (number of authors, number of references, and number of pages) as the explanatory variables of negative binomial regression analysis, he predicted the citation rates of articles in several document sets in different fields. The results revealed that the cluster linkage was a much stronger predictor than the three extrinsic variables and that the centrality divergence might also have a boundary-spanning ability, although its predicting power was somewhat unstable.

Table 5 summarizes the studies mentioned here excluding those included in Table 1.

Table 5 Outline of studies investigating attributes on research quality/content as citation-influencing factors.

Work	Target field	Sample size (n)	Analysis [a)]	Influencing factors [b)]			
				Peer evaluation	Self-evaluation	Funding acquisision	Method /results
Akcan et al. (2013)	Clinical medicine	192	Kruskal-Wallis test; Rank correlation	C			
Aksnes (2006)	General	1549	Rank correlation		B		
Bornmann et al. (2011)	Atmospheric science	315	Chi-squre test	A			
Callaham et al. (2002)	Emergency medicine	204	CMR	B			B
Cronin & Shaw (1999)	Library and inf science	716	Chi-squre test			C	
Fanelli (2013)	Hypthesis-testing	2545	NBMR				B
Leimu & Koricheva (2005)	Ecology	228	t-test; ANOVA; Correlation				C
Lortie et al. (2013)	Ecology and evolutionary biology	1332	Generalized linear model				C
Patterson & Harris (2009)	Physics in biomedicine	1095	Correlation	A			
Rigby (2013)	Biochemistry	3596	LMR (Log(C+1))			B	

a) LMR:Linear multiple regression, CMR:Categorical multiple regression, NBMR:Negative binomial multiple regression

Log(C+1) shows the independent variable of LMR.

b) A:Strong or definite predictor B:Weak predictor or predictive power dependent on the model, C:Not-significant or negative predictor

Other potential factors that might influence citation rate. As shown in the preceding subsection, several studies that used multiple regression analysis considering various citation-influencing factors included number of references as one of the explanatory variables and found it to be a significant predictor of citations (Stewart, 1983; Peters & van Raan, 1994; Davis et al., 2008; Haslam et al., 2008; Lokker et al., 2008; He, 2009; Didegah & Thelwall, 2013). Many other studies have demonstrated that articles with a greater number of references tend to be cited more often (Chen, 2012; Rigby, 2013; Bordons et al., 2013).

Although more specific characteristics of references, such as the ratio of self-citations and the age and subject distributions, also appear to be related to the number of article citations, they have rarely been considered in previous research, probably because of the difficulty in obtaining data. However, a few studies have included the recency of the references as one of the latent factors in the multiple regression model predicting citation rates. Stewart (1983) and Peters and van Raan (1994) took the proportion of references within 3 and 5 years (Price index), respectively, as the recency measure, and both found these variables to be a moderate predictor. Haslam et al. (2008) demonstrated that the newer the mean year of references, the more citations were obtained by the article.

Regarding the other characteristics of references, Peng and Zhu (2012) and Didegah and Thelwall (2013) showed that an article whose references have a higher impact (a greater ratio of highly-cited documents in, or higher mean citations of, the references) have a tendency to acquire higher citations. Lariviere and Gingras (2010a) analyzed the effects of interdisciplinarity on citation impact of articles indexed in Web of Science in 2000 in 14 subject areas, defining the indicator of interdisciplinarity of an article as the percentage of its cited references of subject areas other than that of the article. The pattern of dependence of the citation rate of articles on their degree of interdisciplinarity was different in each subject area, but in all subject areas the citation rate of articles became low at both extremes of high and low interdisciplinarity. Didegah and Thelwall (2013) showed internationality of references was one of the significant predictors in their analysis of determinants of citation rates in the field of nanoscience and nanotechnology.

Some studies indicated a positive association between the number of citations and article length (number of pages) (Stewart, 1983; Peters & van Raan, 1994; Leimu & Koricheva, 2005; van Dalen & Henkens, 2001; van Dalen & Henkens, 2005; Haslam et al., 2008; Peng & Zhu, 2012), while others showed no significant correlation (Walters, 2006; Davis et al., 2008; Slyder et al., 2011; Rigby, 2013). On the other hand, Lokker et al. (2008), He (2009), and Chen (2012) showed a negative correlation between article length and citations, but it could be due to influences of other explanatory variables used in their multiple regression models.

Table 6 summarizes the studies mentioned here excluding those included in Table 1.

Table 6 Outline of studies investigating other attributes as citation-influencing factors.

Work	Target field	Sample size (n)	Analysis a)	Influencing factors b)		
				#Refer-ences	Interdis-ciplinarity of references	Article length
Bordons et al. (2013)	Pharmacology & Pharmacy	1,971 and 2,858 (Two samples)	CMR	A		
Larivière & Gingras (2010a)	General	All WoS articles in 2000	Simple comparison		B	
Leimu & Koricheva (2005)	Ecology	228	t-test; ANOVA; Correlation			A
Rigby (2013)	Biochemistry	3596	LMR (Log(C+1))	A		C
Slyder et al. (2011)	Geography & Forestry	213	t-test; ANOVA; Correlation			C

a) LMR:Linear multiple regression, CMR:Categorical multiple regression

Log(C+1) shows the independent variable of LMR.

b) A:Strong or definite predictor B:Weak predictor or predictive power dependent on the model, C:Not-significant or negative predictor

Questions about self-citations. Do self-citations largely contribute to highly-cited articles? Moreover, are articles co-authored by large numbers of authors cited many times because the authors will each cite these articles? If the answers to these questions are affirmative, then self-citations should be excluded in analyses of factors influencing the citation impact of articles. However, studies by Aksnes (2003b) and Glänzel, Thijs, & Schlemmer（2004) indicated that the ratio of self-citations to non-self-citations tended to decrease with an increase in total citations received by articles from their analyses of longitudinal data of a large quantity.

Furthermore, Glänzel & Thijs（2004) and Aksnes (2003b) also observed that the more authors of an article there are, the lower is the self-citation ratio. From these results, it appears that self-citations need not be excluded in analyses, at least in the case of statistical analyses from a macroscopic view.

Purpose of this Research

The purpose of this study is to examine whether there are general trends across subject fields regarding the factors that affect the number of citations of articles and the extent to which each factor influences citation rate. We focus on those factors that are not directly related to the quality or content of articles. For this purpose, a systematic analysis is done for several selected subject fields separately to examine whether some common features about principal influential factors across fields are noticeable. We limit the sample articles to original papers published in English journals because citation rate is known to be dependent on the type and language of articles.

On the basis of these considerations, the following strategy is adopted in this research:

(1) several journals (all using English language only) are selected in each of several different fields;

(2) from the selected journals, original articles published in the same year are sampled;

(3) for each of the selected subject fields, several negative binomial multiple regression models are examined alternatively;

(4) the number of citations received by the articles is set as the response variable;

(5) a wide range of factors that potentially influence citation rates are used as the explanatory variables; and,

(6) journals are included as dummy explanatory variables depending on the effect of the journals' citation impacts.

Using this method, it is possible to separate the effects of the potential factors on the number of citations and to evaluate the contribution of each factor.

Because the factors directly related to the quality or content of articles are not considered, the models obtained are not expected to predict citation rates with very high accuracy. This study aims not to obtain a model with high explanatory power, but rather to determine a baseline of the citation rate expected from bibliometric factors. If a common baseline can be found across different subject fields, the deviation of each article from this baseline would be regarded as a more adequate indicator of the impact of the article than those reported to date.

Our multiple regression models include the following factors as explanatory variables, considering the results of the studies described in the preceding section "Literature Review."

- Factors regarding collaboration: number of authors; number of institutions; number of affiliation countries.
- Factors regarding author's reputation: number of articles published by the first author before publication of the target article; number of citations that the articles had received by the time of publication of the

target article; active years of the first author before publication of the target article.

- Factors regarding cited references: number of references; Price index (ratio of references within the last five years before the citation occurred).
- Factors regarding visibility of the articles: article length (normalized number of pages); number of figures; number of tables; number of mathematical equations; journal in which the article is published (dummy variable).

Details of these explanatory variables, and also of the response variable, are presented in the following section.

Data Sources and Methods

To achieve the purpose mentioned above, we tried to identify the primary factors affecting citation rates of research articles using the citation frequency data obtained from sample articles published in the same year. We analyzed the factors influencing citations for each of six subject fields to investigate whether a prediction model with some generality could be found across the fields.

Target fields and sampled articles

The following six subject fields were selected as targets. We will use the abbreviations shown in parentheses in the descriptions hereafter.

- Condensed Matter Physics (*CondMat*)
- Inorganic and Nuclear Chemistry (*Inorg*)
- Electric and Electronic Engineering (*Elec*)
- Biochemistry and Molecular Biology (*Biochem*)
- Physiology (*Physiol*)
- Gastroenterology (*Gastro*)

From journals included in the Journal Citation Reports (JCR) Subject Category corresponding to each of these fields, four journals were chosen as the sources from which the articles were sampled, using the JCR Science Edition 2004. The followings were considered when selecting the journals:

- Journals to which only one subject category is assigned (that is, journals with more than one subject category were not selected)
- Journals of English-language only
- Journals with both top-ranked and moderately ranked impact factors in each field
- Journals that are not concentrated in one or two publishing countries in each field

Using Web of Science (WoS), we randomly sampled 50–60 research articles (“articles” as classified by WoS) published in 2000 from the individual journals selected. We excluded proceeding papers, short articles (2 pages or less), and articles in which the author (AU) or affiliation (C1) data were lacking. The 24 journals selected (4 per field) and the numbers of sample articles are shown in Table 7.

The sampling method above should eliminate the influences on citation rates by publication year, article type, and language.

Table 7 Selected subject fields and journals.

Subject Field		Journal Title [a]	Publishing Country	Sampled papers
Condensed Matter Physics	(CondMat)	Physical Review B	USA	55
		Journal of Physics - Condesed Matter	GBR	56
		European Physical Journal B	DEU	60
		Physica B	NLD	59
Inorganic and Nuclear Chemistry	(Inorg)	Inorganic Chemistry	USA	53
		Journal of the Chemical Society - Dalton Transactions	GBR	54
		Inorganica Chimica Acta	CHE	60
		Transition Metal Chemistry	NLD	60
Electric and Electronic Engineering	(Elec)	IEEE Transactions on Microwave Theory and Techniques	USA	59
		IEEE Transactions on Circuits and Systems I - Fundamental Theories and Applications	USA	60
		Signal Processing	NLD	59
		IEE Proceedings - Circuits, Devices and Systems	GBR	51
Biochemistry and Molecular Biology	(Biochem)	Journal of Biological Chemistry	USA	60
		Journal of Molecular Biology	USA	60
		European Journal of Biochemistry	GBR	60
		Journal of Biochemistry (Tokyo)	JPN	60
Physiology	(Physiol)	Journal of General Physiology	USA	60
		Journal of Physiology - London	GBR	58
		Pflugers Archive European Journal of Physiology	DEU	58
		Japanese Journal of Physiology	JPN	60
Gastroenterology	(Gastro)	Gastroenterology	USA	59
		Gut	GBR	56
		American Journal of Gastroenterology	USA	58
		Journal of Gastroenterology	JPN	60

a) The journal titles at the time of 2000, although some were changed after that.

Obtaining citation frequency data

The citation frequencies received by the sample articles were measured in October 2006 and December 2011 using WoS, and therefore, the length of the citing window is 6–7 years and 11–12 years, respectively. The citation frequencies corresponding to these two citing windows are hereafter called *C*6 and *C*11.

These citation frequencies include self-citations[1]. As described in the last part of the preceding section, the possibility that inclusion of self-citations biases the results of such a macroscopic analysis as this research is not high. However, it has been reported that self-citations tend to be concentrated in a short period after publication (Aksnes, 2003b); therefore, we use considerably long citing windows (more than five years after publication).

Obtaining data about factors potentially affecting citation rates

In this study, the following attributes were considered as the factors potentially affecting the citation frequency of the sample articles.

- Authors' collaborative degree
  - (a) *Authors*: Number of authors of the article
  - (b) *Institutions* (*Insts*): Number of institutions the authors are affiliated with
  - (c) *Countries*: Number of countries where the institutions are located
- Cited references
  - (d) *References* (*Ref*): Number of references cited in the article
  - (e) *Price*: Price index (percentage of the references whose publication year is within 5 years before the publication year of the article).
- Article's visibility
  - (f) *Figures*: Number of figures in the article
  - (g) *Tables*: Number of tables in the article
  - (h) *Equations* (*Eqs*): Number of numbered equations in the article
  - (i) *Length*: Number of normalized pages of the article
- Authors' past achievements
  - (j) *Published articles* (*Publ*): Number of articles published by the first author of the article up to the year 2000
  - (k) *Cited*: Number of citations received by the published articles (*Publ*) up to the year 2000
  - (l) *Age*: Active years (elapsed years from the year of the first article publication to the year 2000) of the first author
  - (m) *Rate of publication* (*RatePubl*): Number of articles published per annum by the first author during his/her active years (= *Publ*/*Age*)
  - (n) *Median of the number of citations* (*MedCites*): Median of the number of citations received per annum by each published article

  The attributes (j), (k), and (l) are collectively called "cumulative achievement indicators," and the attributes (m) and (n) are called "efficient achievement indicators."
- Publishing journal
  - (o) *Jnl*-1; *Jnl*-2; *Jnl*-3: Dummy variables representing the individual journals publishing the articles

Values of these attributes for the sample articles were acquired according to the procedures discussed below.

Data on *Authors*, *Insts*, *Countries*, *Refs*, and *Price* were obtained from the downloaded WoS records. For *Authors* and *Insts*, the numbers of entries in the AU and C1 fields of WoS, respectively, were counted. *Countries* were measured from country names described at the end of the C1 entries[2]. The values of *Price* were obtained from the reference list in the WoS CR field of the article by counting references with publication years between 1996 and 2000 (*i.e.,* within 5 years before publication of the article).

Figures, *Tables* and *Eqs* were directly counted from the original documents. *Figures* included figures, charts, diagrams, and pictures. *Length* was defined as the number of converted pages under normalization of 6,400 characters per page, and the value for an article was determined by measuring, from sampled pages, the average characters per page of the journal publishing the article.

We faced a difficultly in calculating the values of authors' past achievements (*Publ, Cited*, *Age*, *RatePubl*, and *MedCites*) because of the existence of homonym authors. The results of the WoS search with the author names of the sample articles during 1970–2000[3] were found to be contaminated by a large number of homonym authors' articles. To eliminate these articles, we developed a model for author disambiguation based on the similarity of each retrieved article to its originating article (the sample article whose author name was used for the search) and extracted the retrieved articles to be discriminated as "true" (Onodera et al., 2011). With this procedure, we obtained data about *Publ*, *Age*, and *RatePubl* for only the first authors of the 1,395 sample articles, as time did not permit us to collect data for all the authors (about 6,000 individuals).

Data on the articles that cited "true" retrieved articles until 2000 were purchased from Thomson Reuters to calculate the values of *Cited* and *MedCites*.

The values of *Publ*, *Cited*, *RatePubl*, and *MedCites* were calculated using full counting and fractional counting (giving each author a credit equal to the inverse of the number of authors); however, only the results that used fractional counting will be shown hereafter because a significant difference was not found between the two counting methods (fitness to the negative binomial regression was somewhat better with fractional counting).

Negative binomial multiple regression analysis

We used negative binomial multiple regression (NBMR) analysis to investigate the extent to which the citation rates of articles are influenced by the individual potential factors introduced in the preceding subsection. The NBMR analysis has been demonstrated to successfully work for predicting citations in several studies (van Dalen & Henkens, 2001; van Dalen & Henkens, 2005; Walters, 2006; Bornmann & Daniel, 2008; Davis et al., 2008; He, 2009; Chen, 2012; Didegah & Thelwall, 2013), because the citation frequency as a response variable is a non-negative integer, its distribution is remarkably skewed, and the variance is usually larger than the mean. A linear multiple regression (LMR) model with a logarithm of citation frequency (in many cases, log (C+1)) as the response variable has also been frequently utilized

(Stewart, 1983; Basu & Lewison, 2005; Figg et al., 2006; Davis & Fromerth, 2007; Davis, 2009; Haslam et al., 2008). However, we adopted the NBMR model because it provided us with results much better than those of the LMR model (see the subsection "Comparison of fitness of NBMR to LMR" in the "Discussion" section).

In the NBMR analysis, the value of the response variable y_i for a case i is supposed to be subject to negative binomial distribution, as follows (here, $\Gamma(\cdot)$ is a gamma function):

$$\Pr(y_i = k) = \frac{\Gamma(k+\theta)}{\Gamma(\theta)\Gamma(k+1)}\left(\frac{\theta}{\mu_i+\theta}\right)^{\theta}\left(\frac{\mu_i}{\mu_i+\theta}\right)^{k}. \qquad (1)$$

The expected value (μ_i) of y_i is estimated from the following regression equation:

$$\ln(\mu_i) = \beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\ldots+\beta_p X_{ip}\,. \qquad (2)$$

Estimated values of the partial regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$ and parameter θ are given on the basis of the input data $\{ X_{i1}, X_{i2}, \ldots X_{ip}; y_i \}$. The value of θ is supposed to be independent of i.

In this study, the NBMR analysis was performed for each of the six subject fields, considering that not only the distribution of citation frequencies but also the distributions of the attributes' values used as the explanatory variables differed by field. However, we expect the results to have some generality across fields, as stated in the section "Purpose of this Research."

The response variable y_i in Equation (2) is *C*6 or *C*11. The explanatory variables $X_{i1}, X_{i2}, \ldots X_{ip}$ are the attributes (a)–(o), introduced in the preceding subsection. We set three regression models (Models A, B, and C) that were different from each other regarding selection of the explanatory variables about the authors' achievements, as follows:

Model A: *Authors*, *Insts*, *Countries*, *Refs*, *Price*, *Figures*, *Tables*, *Eqs*, *Length*, *Publ*, *MedCites*, *Jnl*-1, *Jnl*-2, and *Jnl*-3.

Model B: *Authors*, *Insts*, *Countries*, *Refs*, *Price*, *Figures*, *Tables*, *Eqs*, *Length*, *Cited*, *RatePubl*, *Jnl*-1, *Jnl*-2, and *Jnl*-3.

Model C: *Authors*, *Insts*, *Countries*, *Refs*, *Price*, *Figures*, *Tables*, *Eqs*, *Length*, *Age*, *RatePubl*, *MedCites*, *Jnl*-1, *Jnl*-2, and *Jnl*-3.

The reasons these three models were examined will be described later (see the subsection "Some preliminary analysis" in the "Results" section).

Considering that the degree of citedness differs by journal, we introduced dummy variables representing the journals in which the individual sample articles were published into the explanatory variables of the NBMR analysis. As described earlier, the sample articles were extracted from four journals in each subject field. Hence for articles from a "baseline" journal, all values of the three dummy variables—*Jnl*-1, *Jnl*-2, and *Jnl*-3—were set at 0, and for articles from the other three journals, a value of 1 was given to one of the three dummy variables (corresponding to the journal) and the value of 0 was given to the other two. The baseline journal was taken as the one having the lowest average citation frequency in each field.

The Advanced Regression Model of SPSS/PASW Version 18 was used to perform the NBMR analysis. Variable selection was not chosen in regression, and variables showing a significant relation to *C*6 or *C*11 were identified from the regression results.

Results

Some preliminary analysis

Means and standard deviations of the variables. The means and standard deviations of the two response variables (*C*6 and *C*11) and 14 explanatory variables are shown in Table 8. While the means of *C*6 differ by a maximum factor of 4.5 among the fields, the factor decreases to 2.6 for *C*11. This is because the ratio *C*11/*C*6 of the *Elec* field, the field with the lowest degree of citedness, is considerably higher (greater than 3) compared to those of the other five fields (less than 2).

Table 8 Means and standard deviations (in parentheses) of the response and explanatory variables.

Field	CondMat	Inorg	Elec	Biochem	Physiol	Gastro
n	230	227	229	240	236	233
C6	10.7 (12.3)	10.7 (11.4)	5.6 (8.6)	21.3 (19.3)	15.8 (14.8)	25.0 (32.0)
C11	19.5 (24.0)	19.0 (22.4)	17.3 (30.9)	35.0 (33.2)	26.4 (25.3)	45.8 (67.6)
Authors	3.37 (2.06)	4.27 (1.97)	2.68 (1.52)	4.88 (2.13)	4.14 (2.16)	6.61 (2.99)
Insts	1.86 (0.86)	1.70 (0.80)	1.44 (0.71)	1.81 (0.85)	1.55 (0.73)	1.67 (0.95)
Countries	1.39 (0.56)	1.29 (0.55)	1.17 (0.43)	1.28 (0.54)	1.18 (0.44)	1.15 (0.42)
Refs	27.3 (13.1)	34.3 (18.6)	18.9 (17.2)	41.7 (16.1)	36.7 (15.4)	31.0 (14.3)
Price	32.4 (19.6)	26.6 (15.8)	33.0 (21.0)	40.1 (18.8)	36.6 (18.1)	33.9 (20.4)
Figures	5.89 (3.79)	4.83 (2.88)	8.63 (4.88)	6.49 (2.31)	6.66 (3.24)	3.55 (2.39)
Tables	0.87 (1.56)	2.98 (1.88)	1.18 (1.59)	1.45 (1.45)	0.91 (1.33)	1.74 (1.58)
Eqs	13.27 (17.46)	2.13 (4.22)	19.25 (20.53)	0.80 (3.13)	1.24 (3.93)	0.00 (0.00)
Length	7.28 (3.18)	7.10 (2.72)	8.28 (3.43)	10.70 (3.40)	9.94 (4.11)	6.20 (2.16)
Publ	6.41 (12.68)	10.74 (29.01)	3.58 (5.17)	2.54 (4.66)	3.48 (5.39)	7.69 (19.04)
Cited	35.2 (226.6)	98.1 (492.3)	6.1 (17.7)	26.6 (75.3)	24.7 (60.6)	53.9 (315.1)
Age	7.09 (6.76)	8.73 (8.40)	4.61 (5.93)	5.41 (5.94)	5.90 (6.77)	8.36 (7.43)
RatePubl	0.85 (1.13)	0.92 (1.40)	0.69 (0.96)	0.38 (0.40)	0.57 (0.55)	0.86 (1.09)
MedCites	0.074 (0.197)	0.072 (0.119)	0.046 (0.182)	0.171 (0.259)	0.122 (0.269)	0.049 (0.134)

Among the explanatory variables, *Eqs*, *Publ*, and *Cited* show especially large variations in the mean values across the fields. For instance, the mean of *Eqs* is 10–20 for the *CondMat* and *Elec* fields, while none of the sample articles in the *Gastro* field includes equations. Compared to *Publ* and *Cited*, which are the cumulative achievement measures of authors, *RatePubl* and *MedCites*, as the efficient achievement measures, show considerably less variation in the mean values among the fields.

As shown here, the distributions of the response and explanatory variables are significantly different among fields. This suggests that it is inappropriate to use sample articles from multiple subject fields to investigate factors affecting citations and such investigations should be done within a limited field.

Correlations of the response variables with the explanatory variables. Prior to the NBMR analysis, we examined the correlations between the response variables (*C*6 and *C*11) and the individual explanatory variables. The Spearman's rank correlation coefficients (ρ) of *C*6 with the explanatory variables are shown in Table 9. This table mentions the coefficient values only if they are significant ($p < 0.05$). Similar results were obtained for the response variable *C*11.

Table 9 Spearman's rank correlation coefficients between *C*6 and the explanatory variables.
(Only correlation coefficients whose *p* value is less than 0.05 are shown.)

Variable	**CondMat**	**Inorg**	**Eng**	**Biochem**	**Physiol**	**Gastro**
Authors		**0.204 ****				
Insts		**0.199 ****	**0.171 ****			**0.246 ****
Countries			**0.134 ***			**0.283 ****
Refs	**0.254 ****	**0.395 ****		**0.312 ****	**0.494 ****	**0.382 ****
Price	**0.376 ****	**0.357 ****	**0.188 ****	**0.555 ****	**0.488 ****	**0.392 ****
Figures		**0.375 ****		**0.153 ***	**0.442 ****	**0.132 ***
Tables						**0.153 ***
Eqs						
Length	**0.138 ***	**0.439 ****	**0.130 ***	**0.363 ****	**0.601 ****	**0.349 ****
Publ			**0.215 ****			**0.239 ****
Cited	**0.190 ****		**0.277 ****			**0.242 ****
Age			**0.190 ****	**-0.169 ****	**-0.225 ****	
RatePubl			**0.191 ****		**0.166 ***	**0.285 ****
MedCites	**0.219 ****		**0.196 ****	**0.178 ****		

Refs, *Price*, and *Length* positively correlate at a significant level in all, or almost all, fields. *Insts*, *Figures*, *Cited*, *RatePubl*, and *MedCites* show a significant positive correlation in half or more fields. (*Age* is also significant in three fields but may not have a definite tendency since the ρ-value is positive in some cases and negative in others.) The absolute ρ-value is less than 0.4 in most cases and 0.6 at maximum; hence, a very strong correlation is not found between the response and explanatory variables.

Correlations between the explanatory variables. Table 10 demonstrates the extent to which a correlation exists within an individual pair of explanatory variables, showing the numbers of fields in which the Spearman's rank correlation coefficient is significant ($p < 0.05$).

Table 10 Correlations between the explanatory variables.

Figures mean the numbers of fields for which the Spearman's rank correlation coefficient is significant (p < 0.05), and figures in parentheses mean the number of fields for which the significant correlation coefficient is negative.

	Authors	*Insts*	*Countries*	*Refs*	*Figures*	*Length*	*Price*	*Tables*	*Eqs*	*Publ*	*Cited*	*Age*	*RatePubl*
Insts	6												
Countries	6	6											
Refs	1	3	2										
Figures	3 (1)	3 (2)		5									
Length	3 (1)	3 (1)	1	6	6								
Price	1	2	1	2	1	2							
Tables	3		1	1	3 (1)	3	3 (3)						
Eqs	2 (2)		1	3	2	4	2 (2)	3 (1)					
Publ	1 (1)	2					1	1 (1)					
Cited	1	3					2	1		6			
Age	3	2				1 (1)	3 (3)	1		6	6		
RatePubl		1		1	1	1	2		1	6	6	5	
MedCites		1				1	4 (1)	1 (1)		5	6	6	3

There is a definite positive correlation within the following three variable groups, each of which is enclosed with a bold line in Table 10: the group of five variables on authors' achievements (*Publ*, *Cited*, *Age*, *RatePubl*, and *MedCites*); the group of three variables on collaborative degree (*Authors*, *Insts*, and *Countries*); and the group of three variables on article visibility (*Refs*, *Figures*, and *Length*). Within the first group, however, the fields of a significant correlation are relatively less between the two efficient achievement variables, *RatePubl* and *MedCites*. In addition to the above combinations, *Length* shows a significant positive correlation with *Tables* and *Eqs* in more than half the fields. On the other hand, *Price* has a tendency to negatively correlate with *Tables*, *Eqs*, and *Age*.

Further details for the combinations of variables showing a significant correlation in more than half the fields are shown in Table 11. The three cumulative achievement variables (*Publ*, *Cited*, and *Age*) have the strongest correlation with each other, thus their ρ-values are greater than 0.7 for most cases, while the correlation between each of these variables and each of the two efficient achievement variables (*RatePubl* and *MedCites*) is not as strong, except for that between *Publ* and *RatePubl*. Within the groups (*Authors*, *Insts*, and *Countries*) and (*Refs*, *Figures*, and *Length*), moderately strong correlations (ρ = 0.5–0.7) are observed in many cases, but a stronger correlation is not found.

Although not a small number of explanatory variables are found to have a significant correlation with the response variable as shown in Table 9, these explanatory variables are thought to be not always significant in the NBMR analysis because there exist associations among the explanatory variables in many cases as described here. Thus, either *Refs* or *Length* (or both) might not be selected as a predictor of the citation frequency in the NBMR analysis, even though both show a significant correlation with *C*6 and *C*11 in all fields, since there is also a considerably strong correlation between the two variables.

It should be avoided to use a regression model involving variables having a strong correlation with each other because of the possibility of the problem of multicollinearity. Accordingly, we decided not to simultaneously include explanatory variables whose ρ-values are greater than 0.7 in most fields in a regression model:

- More than one of *Publ*, *Cited*, and *Age* is not included in the model

- *Publ* and *RatePubl* are not included together in the model

Subject to this decision, we designed the three regression models, Models A, B, and C (as mentioned in the subsection “Negative binomial multiple regression analysis” of the “Data and Methods” section). *MedCites* was not included in Model B despite the above-mentioned conditions not prohibiting its inclusion, because *Cited*, similar to *MedCites* in nature, was one of the explanatory variables of Model B. Variables within the groups (*Authors*, *Insts*, and *Countries*) and (*Refs*, *Figures*, and *Length*) were not separated, as the ρ-values of the variables within those groups are less than 0.7 in any case.

Table 11 Spearman's rank correlation coefficients for the explanatory variable pairs which show significant correlations in many fields.

Variable pair	CondMat	Inorg	Eng	Biochem	Physiol	Gastro
(***Authors***, ***Insts***)	**0.555 ****	**0.550 ****	**0.441 ****	**0.474 ****	**0.436 ****	**0.169 ****
(***Authors***, ***Countries***)	**0.376 ****	**0.324 ****	**0.243 ****	**0.274 ****	**0.281 ****	**0.134 ***
(***Insts***, ***Countries***)	**0.647 ****	**0.627 ****	**0.613 ****	**0.498 ****	**0.541 ****	**0.487 ****
(***Refs***, ***Figures***)	**0.113**	**0.412 ****	**0.196 ****	**0.342 ****	**0.497 ****	**0.287 ****
(***Refs***, ***Length***)	**0.548 ****	**0.634 ****	**0.478 ****	**0.614 ****	**0.711 ****	**0.705 ****
(***Figures***, ***Length***)	**0.494 ****	**0.693 ****	**0.486 ****	**0.660 ****	**0.763 ****	**0.598 ****
(***Eqs***, ***Length***)	**0.421 ****	**0.099**	**0.352 ****	**0.160 ***	**0.376 ****	**-** [a]
(***Price***, ***MedCites***)	**0.273 ****	**-0.169 ***	**0.122**	**0.196 ****	**0.069**	**0.146 ***
(***Publ***, ***Cited***)	**0.821 ****	**0.862 ****	**0.731 ****	**0.811 ****	**0.818 ****	**0.845 ****
(***Publ***, ***Age***)	**0.731 ****	**0.779 ****	**0.816 ****	**0.762 ****	**0.749 ****	**0.638 ****
(***Publ***, ***RatePubl***)	**0.687 ****	**0.785 ****	**0.835 ****	**0.706 ****	**0.706 ****	**0.743 ****
(***Publ***, ***MedCites***)	**0.308 ****	**0.321 ****	**0.126**	**0.328 ****	**0.238 ****	**0.167 ***
(***Cited***, ***Age***)	**0.747 ****	**0.841 ****	**0.817 ****	**0.825 ****	**0.756 ****	**0.735 ****
(***Cited***, ***RatePubl***)	**0.442 ****	**0.553 ****	**0.470 ****	**0.495 ****	**0.467 ****	**0.473 ****
(***Cited***, ***MedCites***)	**0.622 ****	**0.625 ****	**0.538 ****	**0.691 ****	**0.570 ****	**0.473 ****
(***Age***, ***RatePubl***)	**0.180 ****	**0.335 ****	**0.544 ****	**0.334 ****	**0.259 ****	**0.073**
(***Age***, ***MedCites***)	**0.310 ****	**0.461 ****	**0.383 ****	**0.435 ****	**0.283 ****	**0.262 ****

** 1% significant, * 5% significant

a) In Gastro field, Eqs = 0 for all sample articles.

Results of NBMR

Goodness of fit of regression: comparison among the models. There are several measures of goodness of fit of the NBMR model (See Chap. 4 (p.85‐113) of Long, 1997). From those measures, the Akaike information criterion (*AIC*) and the adjusted pseudo coefficient of determination (pseudo R_c^2) are selected to compare among the three models (Table 12).

The smaller the *AIC*, or the larger (the nearer to 1) the pseudo R_c^2, the better is the fit of the model. Determining which model is better, as seen in Table 12, is dependent on the field, but a large difference is not seen among the models in every field. Model C appears slightly better than the other two models because of the higher stability across the fields.

Table 12 Goodness of fit measures for three regression models.

(a) AICs (Akaike information criteria)

Field	Response variable: *C*6			Response variable: *C*11		
	Model A	Model B	Model C	Model A	Model B	Model C
CondMat	1511.3	1511.2	1515.6	1783.5	1782.5	1787.1
Inorg	1466.4	1475.9	1470.9	1721.7	1728.3	1725.8
Eng	1214.0	1211.4	1211.9	1687.2	1685.5	1686.8
Biochem	1811.4	1814.9	1814.1	2054.3	2059.1	2056.8
Physiol	1613.9	1613.2	1612.1	1872.4	1871.1	1868.2
Gastro	1814.7	1811.0	1813.4	2088.1	2084.9	2087.5

(b) Adjusted pseudo R-squared measures

Field	Response variable: *C*6			Response variable: *C*11		
	Model A	Model B	Model C	Model A	Model B	Model C
CondMat	0.261	0.261	0.251	0.236	0.239	0.228
Inorg	0.332	0.304	0.322	0.318	0.299	0.310
Eng	0.276	0.285	0.287	0.283	0.288	0.288
Biochem	0.441	0.433	0.438	0.424	0.413	0.421
Physiol	0.531	0.533	0.537	0.483	0.485	0.493
Gastro	0.487	0.495	0.492	0.488	0.495	0.492

Factors affecting citations: comparison among the models. Table 13 shows the number of fields in which each explanatory variable is significant in estimation of the response variable *C*6 or *C*11 for the three models, wherein these numbers were counted when the significant probability *p* for the regression coefficient of the explanatory variable is less than 0.1.

Table 13 Significance of the explanatory variables for predicting the citation counts.

Figures mean the numbers of fields for which the regression coefficient is significant ($p < 0.1$), and figures in parentheses mean the number of fields for which the significant regression coefficient is negative.

Explanatory variable	Response variable: *C*6			Response variable: *C*11		
	Model A	Model B	Model C	Model A	Model B	Model C
Authors	2	3	2	3(1)	2	2
Insts		1(1)	2(2)	1	1	2(1)
Countries	1	1	1	1	1	1
Refs	3	3	3	3	3	3
Price	6	6	6	6	6	6
Figures	1	1	1	2	2	2
Tables	1	1	1	1	1	1
Eqs	1(1)	1(1)	1(1)	1(1)	2(2)	1(1)
Length	1	1	1	2(1)	2(1)	2(1)
Publ	2(1)	-	-	2(1)	-	-
Cited	-		-	-		-
Age	-	-		-	-	1(1)
RatePubl	-	2	4(1)	-	2	3(1)
MedCites	1	-	2	2	-	2

Price is the explanatory variable showing the most obvious effect in all fields, predicting higher citations if it has a higher value. *Refs*, *Authors*, and *Figures* also have a positive influence on citation frequency in some fields. These features are not dependent on the models.

Of the authors' achievement measures, for which the selection of the explanatory variables was made according to the regression models, the three cumulative achievement measures (*Publ*, *Cited*, and *Age*) are found not to influence citations in most cases. On the other hand, the two efficient achievement measures (*RatePubl* and *MedCites*) have a positive influence in some fields. Therefore, Model C, including both these variables, is preferable to the other two.

Good predictors of citation rates: results of Model C. From the results mentioned above, Model C appears to be more appropriate compared to Models A and B in terms of both goodness of fit and selected explanatory variables. Accordingly, we will describe the results using Model C.

Table 14 The *x*-standardized regression coefficients ($s_j\beta_j$'s) of NBMR for the Model C.

(a) Response variable: *C*6

	CondMat	Inorg	Eng	Biochem	Physiol	Gastro
n	230	227	229	240	236	233
θ	1.44	2.15	1.25	2.95	3.11	1.90
$s_j\beta_j$						
Authors	0.174 *	0.052	-0.115	0.116 *	-0.012	0.096
Insts	-0.161 +	0.119	0.166	0.003	0.083	-0.107 +
Countries	-0.074	-0.099	-0.040	-0.030	-0.053	0.107 +
Refs	0.273 **	0.062	0.190 *	0.122 *	0.089	0.103
Price	0.331 **	0.233 **	0.250 **	0.391 **	0.354 **	0.251 **
Figures	-0.048	0.083	0.036	-0.018	0.141 +	0.099
Tables	-0.007	-0.085	-0.085	0.015	-0.050	0.281 **
Eqs	-0.033	-0.129 *	-0.025	0.009	-0.042	-
Length	0.079	0.239 *	0.124	0.094	0.134	-0.150
Age	0.056	-0.035	0.008	-0.009	-0.071	-0.005
RatePubl	0.099 +	-0.112 +	0.194 *	0.025	0.058	0.118 *
MedCites	-0.002	0.144 **	0.102	0.071 +	0.041	-0.022
Dummy1	0.254 **	0.122	0.549 **	0.123 *	0.453 **	0.333 **
Dummy2	0.049	0.000	0.628 **	0.274 **	0.462 **	0.792 **
Dummy3	0.338 **	0.028	0.101	0.303 **	0.264 **	0.627 **

** 1% significant, * 5% significant, + 10% significant

(b) Response variable: *C*11

	CondMat	Inorg	Eng	Biochem	Physiol	Gastro
n	230	227	229	240	236	233
θ	1.27	1.86	1.04	2.58	2.44	1.69
$s_j\beta_j$						
Authors	0.175 *	0.052	-0.143	0.136 **	-0.019	0.095
Insts	-0.169 +	0.085	0.175	-0.006	0.113 +	-0.098
Countries	-0.094	-0.092	-0.066	-0.021	-0.043	0.162 *
Refs	0.213 **	0.046	0.343 **	0.123 *	0.102	0.099
Price	0.313 **	0.200 **	0.225 **	0.393 **	0.314 **	0.199 **
Figures	-0.065	0.108	0.095	-0.034	0.139 +	0.165 +
Tables	0.022	-0.012	0.008	0.044	-0.039	0.362 **
Eqs	0.000	-0.130 *	-0.074	0.032	-0.079	-
Length	0.123	0.256 *	0.137	0.085	0.148	-0.236 *
Age	0.053	-0.020	0.041	0.024	-0.107 *	-0.018
RatePubl	0.115 +	-0.125 *	0.164 *	0.010	0.073	0.099
MedCites	-0.012	0.126 *	0.066	0.081 +	0.026	-0.025
Dummy1	0.248 **	0.096	0.577 **	0.125 *	0.407 **	0.335 **
Dummy2	0.040	0.006	0.534 **	0.270 **	0.411 **	0.814 **
Dummy3	0.325 **	-0.004	-0.002	0.315 **	0.205 **	0.648 **

** 1% significant, * 5% significant, + 10% significant

Table 14 shows the results of the NBMR analysis applied to the sample articles of each of the six subject fields. In this table, the *x*-standardized regression coefficients of the explanatory variable *j*, which are the partial regression coefficients β_js multiplied by the standard deviations of the variable s_js, are reported so that relative strength of influence on the response variable can be compared among the explanatory variables. The estimated values of the parameter θ are also shown in this table.

Table 14 demonstrates that the effect of the individual explanatory variables on the two response variables, *C*6 and *C*11, is very similar.

Price is revealed to be the most important influencing factor on citations in terms of high significance and greatness of the *x*-standardized regression coefficient. Next *Refs* is a strong, or moderately strong, predictor in half the fields. For *Authors*, *Figures*, *RatePubl*, and *MedCites*, there are some fields in which they have a positive significant influence on the response variable, with the exception of *RatePubl*, which is a negative significant predictor of *C*6 and *C*11 in the *Inorg* field. The estimated citation frequencies are considerably affected by the journal in which the articles were published in five fields except the *Inorg* field.

For other explanatory variables a definite effect on citations is not demonstrated. The regression coefficients for *Insts* and *Length* are significant in more than one field, but their values are positive in some cases and negative in the others.

In Table 15, the citation predictabilities of the variables used in this research are compared with those demonstrated by the studies mentioned in Table 1 considering integratedly various variables.

Table 15 The predictive power of various factors demonstrated by several multiple regression analysis studies.

Work	Target field	Sample size (n)	Potential factors influencing citations of articles [a]												
			#Authors	#Insti-tutions [b]	#Coun-tries [c]	#Refer-ences	Recency of refs	#Figures	#Tables	#Equa-tions	Article length	Author's produc-tivity	Author's citedness	Other status of author	Journal impact
Bornmann & Daniel (2008)	Chemistry	1,837	B								C			A	
Chen (2012)	Four topics	1,300 - 6,800 for a topic	A			A					C				
Davis et al. (2008)	Physiology	1,619	A			A					C				A
Didegah & Thelwall (2013)	Nano sci-tech	50,162	C	A	C	A									A
Fu & Aliferis (2010)	Internal medicine	3,788	C	C								C	A		A
Haslam et al. (2008)	Social-personality psychology	308	C			A	A	C	C		A	A			A
He (2009)	Biomedicine	1,860	A		A	A					C		B	C	
Lokker et al. (2008)	Clinical medicine	1,261	A			A					C				A
Peng & Zhu (2012)	Internet research	18,580	A								A			A	A
Peters & van Raan (1994)	Chemical engineering	226	B			A	A				B	A			A
Stewart (1983)	Geoscience	139	C			A	B				A		A	c	
Van Dalen & Henkens (2001)	Demography	1,371	B								A		A		A
Walters (2006)	Prime psychology	428	B								C		A		B
This research	Six fields	227 - 240 for a field	B	C	C	A	A	C	C	C	C	B	B	C	A

a) A:Strong or definite predictor B:Weak predictor or predictive power dependent on the model, C:Not-significant or negative predictor

b) Including domestic interinstitutional collaboration and number of co-authors of other domestic institutions

c) Including international collaboration and number of foreign co-authors

Accuracies of prediction of the citation frequencies for individual articles. Figure 1 shows the relationship between the observed value of $C6$ of the ***i***-th article ($C6_i$) and its expected value (μ_i) predicted by the NBMR analysis for two subject fields using Model C. These fields, *Physiol* and *CondMat*, show the best and the worst pseudo R_c^2, respectively (see Table 12b). A tendency is seen that the greater μ_i is, the larger is the residual ($C6_i$ - μ_i). In the NBMR analysis, the residual is believed to become larger, roughly saying, in proportion to μ_i, since the predicted value is not μ_i but ln (μ_i) as shown in Equation (2). Therefore, the relationship between μ_i and ($C6_i$ - μ_i)/μ_i (we call this quantity "relative residual") is demonstrated in Figure 2 for the same two fields. From this figure it is understood that the relative residual is roughly independent of μ.

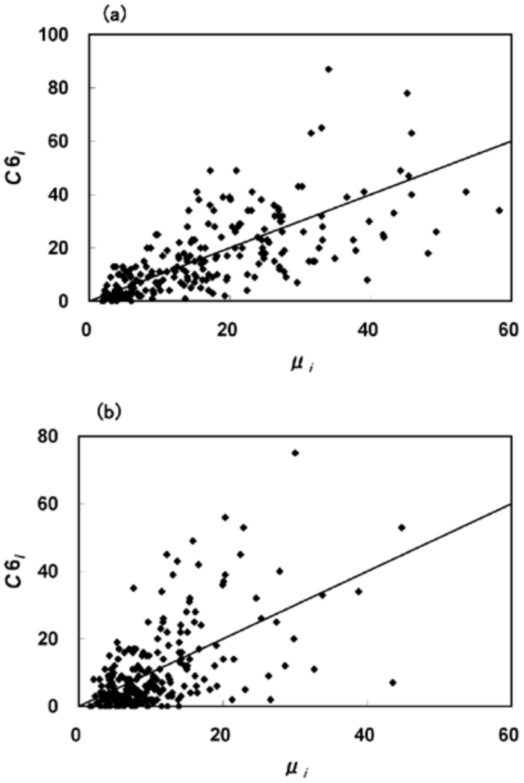

Figure 1. Relations between observed $C6_i$'s and their mean predicted values (μ_i's) by NBMR.
(a) Physiology field (b) Condensed Matter field

Figure 2 shows that the observed citation frequency of a considerable number of the articles is more than double its expected value, that is, ($C6_i$ - μ_i)/μ_i > 1. Although accuracy of the prediction does not seem to be good from the figure, it should be noted that μ_i does not directly predict $C6_i$, but does the expected value of the negative binomial distribution for $C6_i$, as understood from Equations (1) and (2).

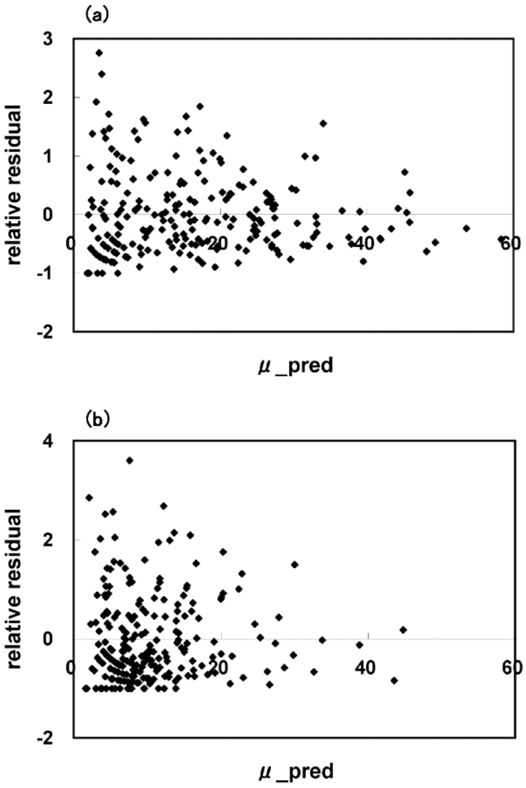

Figure 2. Relations between mean predicted values (μ_i's) and relative residuals.

The relative residual of $C6_i$ is ($C6_i$ - μ_i) / μ_i.

(a) Physiology field (b) Condensed Matter field

The frequency distributions of *C*6 and *C*11 predicted from the NBMR analysis were obtained by the following procedure:

(a) calculating the probability distribution of the citation frequency Pr($C6_i = k$) or Pr($C11_i = k$) for each article *i* based on Equation (1)

(b) summing up the probability distributions of all articles

The predicted frequency distributions obtained are compared to the observed distributions in Figure 3. This figure shows the case of the *Inorg* field, for which the fitness of the NBMR model is moderate in the six fields. Although the observed distributions fluctuate considerably, the predicted distributions fit their smoothed curves well.

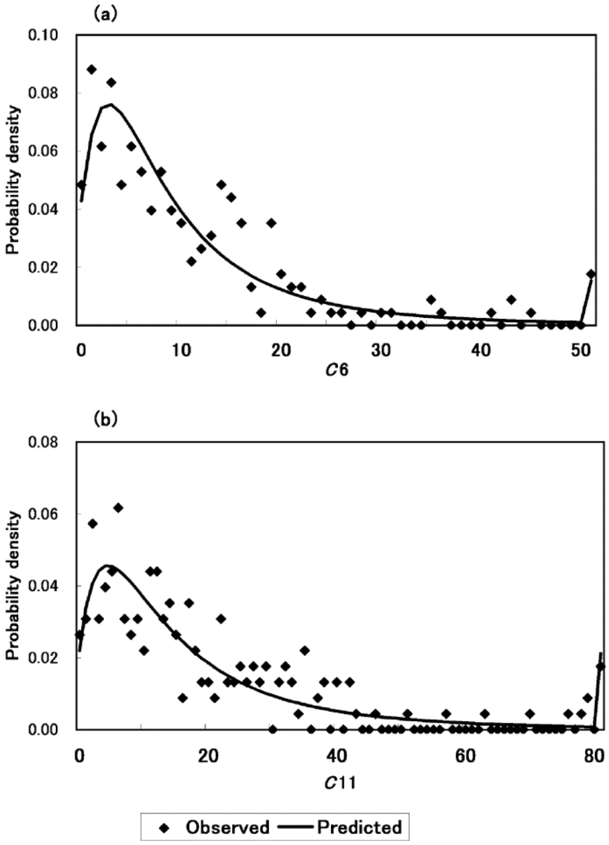
(a)
0.10
0.08
0.06
0.04
0.02
0.00
Probability density
0
10
20
30
40
50
C6
(b)
0.08
0.06
0.04
0.02
0.00
Probability density
0
20
40
60
80
C11
◆ Observed
Predicted

Figure 3. Comparison of the probability distribution of citation counts predicted by NBMR with the observed distribution.

Cases of the regression by the Model C for Inorganic and Nuclear Chemistry field.

(a) Distribution of *C*6 (b) Distribution of *C*11

Discussion

Important factors influencing the citations of articles

In the six fields, we were able to predict, with acceptable accuracy, the citation frequency of an article within the 6 years (*C*6) and 11 years (*C*11) after its publication, with 3–5 significant predictors. The pseudo R_c^2 in the NBMR analysis was 0.25–0.54 for *C*6 and 0.23–0.50 for *C*11, depending on the fields.

The significant predicting factors were common to some extent across the fields and almost the same between *C*6 and *C*11 in a field.

Price index (*Price*) was found to be the strongest influencing factor on citations. A few studies have taken notice of this kind of attribute (the recency measure of the references) as an influencing factor on citations (Stewart, 1983; Peters & van Raan, 1994; Haslam et al., 2008). These studies found a moderate positive correlation between the recency of references and the citation count of articles, but any of their results were based on a relatively small sample (*n* <~ 300) taken from a single subject field. It is a noticeable finding in this study that the Price index is very important in every subject field when considering factors influencing citation rates.

The second important explanatory factor was the number of references (*Refs*), which has been reported to have a significant relation with citation rates by many existing studies (See the subsection “Other potential factors influencing citation rates” in the “Literature Review” section and Table 15).

Although there have been many studies reporting that articles with more co-authors tend to obtain higher citations, such claims have not been so strongly supported by several systematic multiple regression analyses (See the subsection “Does collaboration boost the citation rate of articles?” in the “Literature Review” section and Table 15). Also, in our NBMR analysis, the number of authors (*Authors*) was shown to be a (moderately) significant predictor in only two of the six fields, suggesting that the factor might not affect citation rates very strongly. Bornmann & Daniel (2008) reported that the correlation between the number of authors and that of citations diminished as the citing window became longer. This may apply to our case since the citing window we used was relatively long (6 or 11 years).

The influence of authors’ achievement variables on citations is discussed in the subsequent subsection.

Is there a halo effect of authors?

Several studies have claimed that an article written by author(s) with a higher performance (more publications and/or higher citations) have the possibility of receiving higher citations after its publication (see the subsection "Is there a halo effect of authors, institutions, journals, or countries?" in the "Literature Review" section).

In this study, the effect of the five indicators concerning the authors' past achievement on citations was investigated (all the indicators apply to the first author of articles). As a result, the effect of the three cumulative achievement indicators (*Publ*, *Cited*, and *Age*) was hardly found. On the other hand, the two efficient achievement indicators (*RatePubl* and *MedCites*) showed significant, but not remarkable, influence in some fields.

Our result that the efficient achievement indicators are better predictors of citations than the cumulative ones agrees with those of Danell (2011) and Hönekopp and Khan (2012). It may be because we used data only on first authors that the effect was not as apparent in our analysis as that shown by Danell (2011), who used the data on the authors of the highest performance, or Hönekopp and Khan (2012), who selected their sample from single-authored articles. The fact that our data were based on the first author only is thought to be a limitation of this study because it might weaken a halo effect of authors on citations. We attempted to perform the NBMR analysis for the sample articles of only one journal in each field using the achievement data of the most productive author in each article, but could not obtain a consistent result across fields. It may be due to the small size (n = 50–60) of the samples.

Interaction among the explanatory variables

By comparing Table 9 with Table 14, it is understood that the explanatory variables having significant correlation with the response variable (citation frequency) do not always become significant predictors in the NBMR analysis. The typical example is *Length*, which is not a significant predictor for most fields in spite of its positive correlation with citation frequency in almost all fields. As shown in Tables 10 and 11, *Length* has a positive correlation with several other explanatory variables (*Refs*, *Figures*, *Tables*, and *Eqs*) in many fields, which suggests that these explanatory variables are preferred to *Length* in the NBMR analysis.

Comparison of fitness of NBMR with that of LMR

The LMR analysis with log (C + 1) as the response variable is also frequently used for predicting citation rates, instead of the NBMR analysis used in this study. Comparison of fitness between these two analyses applied to the same sample is, however, not easy since there are few fitness measures commonly applicable to both. The variance ratio (F) usually applied to the LMR model is not available for NBMR. There is some difference in meaning between the coefficient of determination (R^2) in the LMR model and pseudo R^2 in the NBMR model. *AIC* is usable for the LMR and NBMR models, but it is questionable to simply compare the values obtained from the two methods. Therefore, we compared the results of the NBMR model to those of the LMR model applied to the same data using the following two measures:

(a) Mean square of relative residuals

The relative residual for a member (an article in our case) was introduced in the subsection “Results of NBMR” of the “Results” section. The mean square of relative residuals (*MSRR*) is the squared mean of this quantity, as follows:

$$MSRR = \sum_{i=1}^{n} \left(\frac{y_i - \mu_i}{\mu_i} \right)^2 \Big/ n . \qquad (3)$$

Here, y_i and μ_i are observed and predicted (expected) values of the response variable for the *i*-th member.

The residual (y_i - μ_i) becomes larger roughly in proportion to μ_i, as described in the subsection “Results of NBMR.” In the LMR model with log (y_i) as the response variable, the residual of y_i is also supposed to proportionally increase with μ_i. Therefore, it is appropriate to compare the goodness of fit between the LMR and NBMR models by this measure.

(b) The chi-square statistic of fitness

The predicted frequency distribution of citations was obtained from the NBMR and LMR analyses. (The method for the NBMR analysis is described in the subsection “Results of NBMR.”) For the distributions obtained, the frequencies of citations were divided into *m* regions (here, m = 10) so that the individual regions may have roughly equal expected values. When representing the observed and expected values in the region *i* ($1 \leq i \leq m$) as O_i and E_i, respectively, the chi-square statistic of fitness (χ^2) is given as follows:

$$\chi^2 = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i} . \qquad (4)$$

The *MSRR*s are compared in Table 16 between the NBMR and LMR analyses, both using the set of explanatory variables selected in Model C. The *MSRR* for the NBMR model is 1/5 to 1/2 of that for the corresponding LMR model.

Table 16 Comparison of the mean square of relative residuals (*MSRR*) between NBMR and LMR.

Field	NBMR		LMR	
	*C*6	*C*11	*C*6	*C*11
CondMat	0.78	0.86	2.34	2.72
Inorg	0.60	0.64	1.27	1.43
Eng	1.37	1.29	6.36	6.78
Biochem	0.44	0.47	0.73	0.81
Physiol	0.49	0.54	0.96	1.19
Gastro	0.64	0.62	1.56	1.48

A similar comparison on χ^2s is shown in Table 17. The χ^2 values for the NBMR model do not reject the null hypothesis that the predicted distribution is not different from the observed one ($p > 0.05$) except for one case (C11 prediction in the *Elec* field), while for the LMR model the null hypothesis is strongly rejected in all cases ($p < 0.001$).

Table 17 Comparison of the fitness chi squares (χ^2's) and their significance probabilities (p's) between NBMR and LMR.

Response varable	Field	NBMR		LMR	
		χ^2	p	χ^2	p
C6	CondMat	3.73	0.928	188.5	0.000
	Inorg	11.56	0.239	139.0	0.000
	Eng	6.34	0.501	259.5	0.000
	Biochem	7.82	0.553	55.7	0.000
	Physiol	12.82	0.171	43.9	0.000
	Gastro	3.51	0.941	89.7	0.000
C11	CondMat	5.52	0.786	194.7	0.000
	Inorg	10.31	0.326	146.9	0.000
	Eng	17.75	0.038	143.4	0.000
	Biochem	6.38	0.701	65.9	0.000
	Physiol	13.12	0.157	48.1	0.000
	Gastro	3.73	0.928	64.9	0.000

These indicate that the observed frequency distribution of citations fits far more to the distribution predicted by the NBMR model than the LMR model.

In Figure 4 the predicted distribution by the LMR model is compared to the observed one for the *Inorg* field. Comparison of this figure to Figure 3 for the NBMR model reveals the NBMR model’s obvious precedence. The NBMR analysis shows good fit, especially in the low citation frequency region.

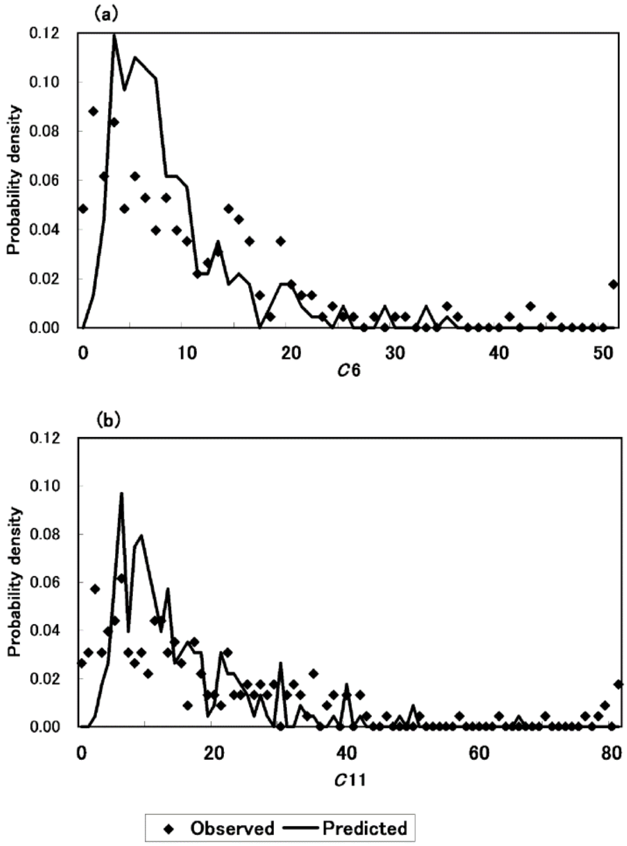

Figure 4. Comparison of the probability distribution of citation counts predicted by LMR with the observed distribution.

Cases of the regression by the Model C for Inorganic and Nuclear Chemistry field.

(a) Distribution of *C*6 (b) Distribution of *C*11

The issue of sampling

The samples used in this study consist of 50–60 research articles randomly drawn from each of four journals selected in the individual fields, as explained in the “Data Sources and Methods” section. Two issues should be mentioned concerning this point.

One of the issues is a relatively small sample size (*n* = 230–240 for each field). It was difficult to obtain a larger sample because we used, as the explanatory variables, attributes for which a considerable effort is needed to acquire the data. Although it was relatively easy to obtain data on *Authors*, *Insts*, *Countries*, and *Refs* from the data source used (WoS), the publication year of each reference of the sample articles had to be examined to get the values of *Price*. The values of *Figures*, *Tables*, and *Eqs* were counted by looking them up in the original documents. *Length* was not simply the number of pages of articles, but normalized considering the number of characters per page in each journal. The greatest effort was gaining data on the five variables of authors' achievement, which involved the search and identification of articles published earlier by the authors of the sample articles and measurement of the citation frequencies that the retrieved articles had received until the publication year (2000) of the sample articles (Onodera et al., 2011).

It is expected that more explanatory variables would be selected as a significant predictor of citations if we used larger samples. The authors' achievement indicators might become more definite predictors. However, when using samples that are too large, some explanatory variables which are not so important may be regarded as significant. In this sense, it can be said that only the variables certainly affecting citation rates were chosen as significant in this study.

Another issue involves the possibility of bias due to random sampling of a nearly equal number of articles from four journals. This issue can be divided into the following two questions:

(a) Does the distribution of citation frequencies in a randomly-drawn sample differ significantly from that in the population, considering the high skewness in the citation distributions?

(b) Is it reasonable to draw samples equally from journals that differ in size (i.e., number of published articles)?

With regard to question (a), the distribution is not systematically biased by random sampling even if the distribution is highly skewed. However, it is more likely that such a sample distribution largely deviates by chance from that in the population depending on the extent to which a few "outliers" were drawn, compared with a normal case. For the 23 journals used in this study (excluding the one from which all articles were drawn), we compared the mean citation frequency (*C*11) of the sample with that of the population (all articles published in the journals in 2000). The number of journals with a higher and lower mean than that of the population were 13 and 10, respectively, indicating that the samples are unbiased. However, one journal with the higher sample mean and two journals with lower sample means showed a significant difference ($p < 0.05$) from the population mean. The rate of journals with a significant difference (3/23) is somewhat high. As for the three journals, outliers might be either over- or under-drawn.

With regard to question (b), we designed our sampling considering the following factors. The articles should be uniformly selected not only from high-impact journals but also from relatively low-impact journals because the sample must represent the whole field. Hence we divided the journals from individual fields into four classes according to their JIF value so that the number of articles in each class might be roughly equal. We then selected one journal from each class taking into account the points mentioned in the "Data Sources and Methods" section. For example, in the *Inorg* field, the selection was made as follows:

Range of JIF	#Journals	Article share	Sampled journal	
			Name	JIF
> 3	2	23.1%	Inorg Chem	3.45
2 – 3	2	22.0%	Dalton Trans	2.93
1 – 2	3	25.0%	Inorg Chim Acta	1.55
< 1	7	29.9%	Trans Met Chem	0.86

However, equal classification failed in the *CondMat* and *Biochem* fields because only one journal (*Physical Review B* and *Journal of Biochemistry*, respectively) had a large publication share in those fields (45% and 44%, respectively). In those two fields, underestimating the contribution of these journals (with high citation impact and a large publication share) might have biased the results. However, we assume that this problem is moderated by making journals dummy variables of our regression models.

The bias could be avoided if the random sampling was conducted from all journals in a field, but we limited our target to the journals that were accessible to us because the original articles were necessary for obtaining the values of some explanatory variables (*Figures*, *Tables*, *Eqs*, and *Length*).

Setting the subject fields

The subject fields set in this study are based on the JCR Subject Categories commonly used in bibliometric research. However, our results may suggest the necessity for a more fine-grained analysis by dividing the fields into subfields. In every field studied, the Price index and the number of references were found to be significant predictors of the number of citations. As all of these three attributes are connected to citation behavior, such results are likely if it differs depending on subfields. In this regard, Moed (1989) showed some cases in which the mean values of Price index, number of references, and number of citations largely differ among subfields within the same field. Further investigations are needed to evaluate the difference in citation behavior among subfields that are narrower than the JCR Subject Categories.

Conclusion

We obtained the NBMR model explaining the citation frequencies of articles with a relatively long citing window (6 or 11 years after publication), for each of the six fields. The models for the six fields were to some extent similar regarding the selected predicting factors and the degree of significance of these predictors. Most existing studies that explain the factors influencing citation rates of articles have dealt with articles of either single subject field or mixed fields. Taking this into account, our study is original in that some generality across different fields is found regarding the important factors that influence citations.

Fitness of the NBMR model obtained in this study was not very high, but acceptable since the value of pseudo R_c^2 was 0.25–0.5. This is an expected result when considering all the explanatory variables used here were “extrinsic” factors that have no direct relation to the quality or content of the articles. The purpose of our study was not to develop a model with high fitness, but to seek a model working as a baseline of the expected citation frequency for a given article based on such extrinsic factors. The finding of generality of the significant predictors of citations across different fields is promising to develop such a baseline.

One of the aims of advancement in the future is an analysis of deviations of the observed citations from this baseline (expected citation frequencies based on these extrinsic predictors) for individual articles. To what attributes of articles do the deviations relate? Are the attributes intrinsic ones connected with the quality or content of articles? It is difficult, however, to strictly distinguish intrinsic attributes from extrinsic ones. The number of institutions or countries, which is used as an extrinsic variable in this study, is thought to involve some intrinsic nature provided that interinstitutional or international collaboration is connected to research quality. In addition, many references or pages may imply the width and depth of research. We assumed here that the attributes whose values can be obtained from bibliographic data were extrinsic.

Another aim of advancement is to look for "intrinsic" factors (which are closely related to the quality or content of articles) that are associated with citations. Concerning this, Chen (2012) recently proposed the "structure variation" model, supposing that the potential value of an idea conveyed in an article is measured in terms of the degree of change in the existing intellectual structure introduced by the idea (See the subsection "Quantitative relations between citation rates and measures of the quality or content of articles" in the "Literature Review" section). Based on this model, he defined some indicators on the degree of structural change using a network theory and discussed the relationship between these indicators and the citation frequency acquired by the article in the future.

It remains a difficult and complicated issue to determine the principal factors affecting citation rates of articles. A definite conclusion is not yet obtained despite much research having been dedicated to this problem. We hope this article will make some contribution to relevant literature.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 18300080. The authors appreciate useful discussions with the members of the research group, in addition to their support for data acquisition and processing. The authors express their gratitude for helpful advice from the reviewers regarding the bias that was possibly brought on by sampling, the possibility of differences in citation tendency among subfields, and the description of Literature Review.

Notes

1. Our analysis was made for both cases, including and excluding self-citations, but a significant difference affecting the results was not observed. Thus, we will describe the results of the former case.
2. Addresses ending with a US state name were read as USA. England, Scotland, Wales, and Northern Ireland were changed to UK. Hong Kong was incorporated as China.
3. Articles before 1970 were not provided by WoS available to us. However, a search period of 30 years is believed to be long enough to cover an individual's research lifetime.

References

Abramo, G., D’Angelo, C. A., & Di Costa, F. (2011). National research assessment exercises: a comparison of peer review and bibliometrics rankings. Scientometrics, 89(3), 929‐941.

Akcan, D., Axelsson, S., Bergh, C., Davidson, T., & Rosen, M. (2013). Methodological quality in clinical trials and bibliometric indicators: no evidence of correlations. Scientometrics, 96(1), 297–303.

Aksnes, D. W. (2003a). Characteristics of highly cited papers. Research Evaluation, 12(3), 159–170.

Aksnes, D. W. (2003b). A macro-study of self-citations. Scientometrics, 56(2), 235–246.

Aksnes, D. W. (2006). Citation rates and perceptions of scientific contribution. Journal of the American Society for Information Science and Technology, 57(2), 169–185.

Basu A., & Lewison, G. (2005). Going beyond journal classification for evaluation of research outputs. Aslib Proceedings, 57(3), 232–246.

Bookstein, A., & Yitzhaki, M. (1999). Own-language preference: A new measure of "Relative Language Self-Citation". Scientometrics, 46(2), 337–348.

Bordons, M., Aparicio, J., & Costas, R. (2013). Heterogeneity of collaboration and its relationship with research impact in a biomedical field. Scientometrics, 96(2), 443–466.

Bornmann, L., & 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 (3), 427–440.

Bornmann, L., & Daniel, H.-D. (2008). 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. Journal of the American Society for Information Science, 59(11), 1841–1852.

Bornmann, L., Schier, H., Marx, W., & Daniel, H.-D. (2011). Is interactive open access publishing able to identify high-impact submissions? A study on the predictive validity of Atmospheric Chemistry and Physics by using percentile rank classes. Journal of the American Society for Information Science and Technology, 62 (1), 61–71.

Callaham, M., Wears, R. L., & Weber, E. (2002). Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-revied journals. JAMA, 287(21), 2847–2850.

Chen, C. (2012). Predictive effects of structural variation on citation counts. Journal of the American Society for Information Science and Technology, 63 (3), 431–449.

Chubin, D. E., & Moitra, S. D. (1975). Content analysis of references: Adjunct or alternative to citation counting? Social Studies of Science, 5(4), 423–441.

Cronin, B., & Shaw, D. (1999), Citation, funding acknowledgment and authir natioanlity relationshups in four information science journals. Journal of Documentation, 55(4), 402–408.

Danell, R. (2011). Can the quality of scientific work be predicted using information on the author's track record? Journal of the American Society for Information Science and Technology, 62 (1), 50–60.

Davis, P. M. (2009). Author-choice apen-access publishing in the biological and medical literature: A citation analysis. Journal of the American Society for Information Science and Technology, 60(1), 3–8.

Davis, P. M., & Fromerth, M. J. (2007). Does the arXive lead to higher citations and reduced publisher downloads for mathematics articles? Scientometrics, 71(2), 203–215.

Davis, P. M., Lewenstein, B. V. Simon, D. H, Booth, J. G., & Connolly, M. J. L. (2008). Open access publishing, article downloads, and citations: randomised controlled trial. BMJ, (337), a568 (6p.).

Didegah, F., & Thelwall, M. (2013). Determinants of research citation impact in nanoscience and nanotechnology. Journal of the American Society for Information Science and Technology, 64 (5), 1055–1064.

Figg, W. D., Dunn, L., Liewehr, D. J., Steinberg, S. M., Thurnman, P. W., Barrett, J. C. et al. (2006). Scientific collaboration results in higher citation rates of published articles. Pharmacotherapy, 26(6), 759–767.

Fanelli, D. (2013). Positive results receive more citations, but only in some disciplines. Scientometrics, 94(2), 701–709.

Fu, L. D., & Aliferis, C. F. (2010). Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. Scientometrics, 85(1), 257–270.

Glänzel, W., & Thijs, B. (2004). Does co-authorship inflate the share of self-citations? Scientometrics, 61(3), 395–404.

Glanzel, W., Thijs, B., & Schlemmer, B. (2004). A bibliometric approach to the role of author self-citations in scientific communication. Scientometrics, 59(1), 63–77.

Haslam, N., Ban, L., Kaufmann, L., Loughnan, S., Peters, K., Whelan, J. et al. (2008). What makes an article influential? Predicting impact in social and personality psychology. Scientometrics, 76(1), 169–185.

He, Z.-L. (2009). International collaboration does not have greater epistemic authority. Journal of the American Society for Information Science and Technology, 60 (10), 2151–2164.

Hönekopp, J., & Khan, J. (2012). Future publication success in science is better predicted by traditional measures than by the h index. Scientometrics, 90(3), 843–853.

Hsu, J.-W., & Huang, D.-W. (2011). Correlation between impact and collaboration. Scientometrics, 86 (2), 317–324.

Ibanez, A., Bielza, C., & Larranaga, P. (2013). Relationship among research collaboration, number of documents and number of citations: a case study in Spanish computer science production in 2000–2009. Scientometrics, 95(2), 689–716.

Katz, J. S., & Hicks, D. (1997). How much is a collaboration worth? A calibrated bibliometric model. Scientometrics, 40(3), 541–554.

Krampen, G., Becker, R., Wahner, U., & Montada, L. (2007). On the validity of citation counting in science evaluation: Content analyses of references and citations in psychological publications. Scientometrics, 71(2), 191–202.

Lariviere, V. & Gingras, Y. (2010a). On the relationship between interdisciplinarity and scientific impact. Journal of the American Society for Information Science and Technology, 61 (1), 126–131.

Lariviere, V. & Gingras, Y. (2010b). The impact factor's Matthew Effect: A natural experiment in bibliometrics. Journal of the American Society for Information Science and Technology, 61 (2), 424–427.

Leimu, R., & Koricheva, J. (2005). What determines the citation frequency of ecological papaers? Trends in Ecology and Evolution, 20(1), 28–32.

Levitt, J. M., & Thelwall, M. (2009). Citation levels and collaboration within library and information science. Journal of the American Society for Information Science and Technology, 60 (3), 434–442.

Lindsey, D. (1989). Using citation counts as a measure of quality in science: Measuring what's measurable rather than what's valid. Scientometrics, 15(3/4), 189–203.

Lokker, C., McKibbon, K. A., McKinlay, R. J., Wilczynski, N. L., & Haynes, R. B. (2008). Prediction of citation counts for clinical articles at two years using data available within three weeks of publication: retrospective cohort study. BMJ, (336), 655 (6p.).

Long, J. S. (1997). Regression models for categorical and limited dependent variables. Thousand Oaks: SAGE Publications, Inc.

Lortie, C. J., Aarssen, L. W., Budden, A. E., & Leimu, R. (2013). Do citations and impact factors relate to the real numbers in publications? A case study of citation rates, impact, and effect sizes in ecology and evolutionary biology. Scientometrics, 94(2), 675–682.

MacRoberts, M. H., & MacRoberts, B. R. (1987). Testing the Ortega hypothesis: Facts and artifacts. Scientometrics, 12(5/6), 293–295.

MacRoberts, M. H., & MacRoberts, B. R. (1989). Problems of citation analysis: A critical review. Journal of the American Society for Information Science, 40(5), 342–349.

MacRoberts, M. H., & MacRoberts, B. R. (1996). Problems of citation analysis. Scientometrics, 36(3), 435–444.

MacRoberts, M. H., & MacRoberts, B. R. (2010). Problems of citation analysis: A study of uncited and seldom-cited influences. Journal of the American Society for Information Science and Technology, 61(1), 1–13.

Moed, H. F. (1989). Bibliometric measurement of research performance and Price's theory of differences among the sciences. Scientometrics, 15(5/6), 473–483.

Moed, H. F. (2005). Citation analysis in research evaluation. Dordrecht, the Netherlands: Springer.

Moravcsik, M. J., & Murugesan, P. (1975). Some results on the function and quality of citations. Social Studies of Science, 5(1), 86–92.

Mryglod, O., Kenna, R., Holovatch, Yu., & Berche, B. (2013). Absolute and specific measures of research group excellence. Scientomertics, 95(1), 115–127.

Onodera, N., Iwasawa, M., Midorikawa, N., Yoshikane, F., Amano, K., Ootani et al. (2011). A method for eliminating articles by homonymous authors from the large number of articles retrieved by author search. Journal of the American Society for Information Science and Technology, 62(4), 677–690.

Oppenheim, C. (1997). The correlation between citation counts and the 1992 Research Assessment Exercise ratings for British research in genetics, anatomy and archaeology. Journal of Documentation, 53（5）, 477–487.

Pasterkamp, G., Rotmans, J. I., de Kleijn, D. V. P., & Borst, C. (2007). Citation frequency: A biased measure of research impact significantly influenced by the geographical origin of research articles. Scientometrics, 70 (1), 153–165.

Patterson, M. S., & Harris, S. (2009). The relationship between reviewers' quality-scores and number of citations for papers published in the journal Physics in Medicine and Biology from 2003–2005. Scientometrics, 80 (2), 343–349.

Peclin, S., Juznic, P., Blagus, R., Sajko, M. C., & Stare, J. (2012). Effects of international collaboration and status of journal on impact of papers. Scientometrics, 93(3), 937–948.

Peng, Tai-Quan, & Zhu, Jonathan J.H. (2012). Where you publish matters most: A multilevel analysis of factors affecting citations of internet studies. Journal of the American Society for Information Science and Technology, 63 (9), 1789–1803.

Persson, O., Glanzel, W., & Danell, R. (2004). Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies. Scientometrics, 60(3), 421–432.

Peters, H. P. F., & van Raan, A. F. J. (1994). On determinants of citation scores: A case study in chemical engineering. Journal of the American Society for Information Science, 45(1), 39–49.

Rigby, J. (2013). Looking for the impact of peer review: does count of funding acknowledgements really predict research impact? Scientometrics, 94(1), 57–73.

Rinia, E. J., van Leeuwen, Th. N., van Vuren, H. G., & van Raan, A. F. J. (1998). Comparative analysis of a set of bibliometric indicators and central peer review criteria. Evaluation of condensed matter physics in the Netherlands. Research Policy, 27(1), 95–107.

Schubert, A., & Glänzel, W. (2006). Cross-national preference in co-authorship, references and citations Scientometrics, 69 (2), 409–428.

Seglen, P. O. (1994). Causal relationship between article citedness and journal impact. Journal of the American Society for Information Science, 45(1), 1–11.

Sin, S.-C. J. (2011). International coauthorship and citation impact: A bibliometric study of six LIS journals, 1980–2008. Journal of the American Society for Information Science and Technology, 62(9), 1770–1783.

Slyder, J. B., Stein, B. R., Sams, B. S., Walker, D. M., Beale, B. J., Feldhaus, J. J. et al. (2011). Citation pattern and lifespan: a comparison of discipline, institution, and individual. Scientometrics, 89(3), 955–966.

Sooryamoorthy, R. (2009). Do types of collaboration change citation? Collaboration and citation patterns of South African science publications. Scientometrics, 81(1), 177–193.

Stewart, John A. (1983). Achievement and ascriptive processes in the recognition of scientific articles.. Social Forces, 62(1), 166–189.

Van Dalen, H. P., & Henkens, K. (2001). What makes a scientific article influential? The case of demographers. Scientometrics, 50 (3), 455–482.

van Dalen, H. P., & Henkens, K. (2005). Signals in science – On the importance of signaling in gaining attention in science. Scientometrics, 64(2), 209–233.

van Raan, A. F. J. (1996). Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. Scientometrics, 36(3), 397–420.

van Raan, A. F. J. (1998). The influence of international collaboration on the impact of research results. Some simple mathematical considerations concerning the role of self-citations. Scientometrics, 42(2), 423–428.

Wang, M., Yu, G., & Yu, D. (2011). Mining typical features for highly cited papers. Scientometrics, 87(3), 695–706.

Wang, M., Yu, G., An, S., & Yu, D. (2012). Discovery of factors influencing citation impact based on a soft fuzzy rough set model. Scientometrics, 93(3), 635–644.

Walters, G. D. (2006). Predicting subsequent citations to articles published in twelve crime-psychology journals: Author impact versus journal impact. Scientometrics, 69(3), 499–510.