Journal of Informetrics 12 (2018) 518–533

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Regular article

Predicting research excellence at the individual level: The importance of publication rate, top journal publications, and top 10% publications in the case of early career mathematicians

Jonas Lindahl

INFORSK, Department of Sociology, Umeå University, Umeå, SE-90187, Sweden

ARTICLE INFO

Article history:
Received 9 May 2017
Received in revised form 16 April 2018
Accepted 16 April 2018

Keywords:
Excellence
Productivity
Early career
Dominance analysis
Mathematics
Highly cited
Bibliometric indicator
Journal prestige

ABSTRACT

The purpose of this study was to examine the relationship between publication rate, top journal publications and excellence during the first eight years of the career, and how well publication rate, top journal publications and highly cited publications during the first four years of the career can predict whether an author attain excellence in the fifth to the eighth year. The dataset consisted of publication track records of 406 early career mathematicians in the sub-field of number theory collected from the MathSciNet database. Logistic regression and dominance analysis was applied to the data. The major conclusions were (1) publication rate had a positive effect on excellence during the first eighth years of the career. However, those who publish many articles in top journals, which implicitly require a high publication count, had an even higher probability of attaining excellence. These results suggest that publishing in top journals is very important in the process of attaining excellence in the early career in addition to publishing many papers; and (2) a dominance analysis indicated that the number of top journal publications and highly cited publications during the first four years of the career were the most important predictors of who will attain excellence in the later career. The results are discussed in relation to indicator development and science policy.

1. Introduction

Bibliometric indicators are increasingly used in science policy to identify and stimulate research excellence by allocating resources on the basis of publication and citation statistics (Abramo, Cicero, & D'angelo, 2014; Danell, 2011). The focus on excellence as an outcome criterion in evaluations and selection procedures in academia have motivated research on how to measure and identify excellent research and those who produce it (see e.g., Costas & Noyons, 2013). Two important questions in the context of incentivizing for the production of excellent research in academia are (1) What publication behaviors are favorable for the production of excellent research; and (2) Can we predict who will produce excellent research in the future on the basis of past performance?

Numerous scientometric studies have shown a strong correlation between publication rate (i.e., number of publications) and impact (i.e., number of citations). Recent studies have extended this line of research to the relationship between publi-

E-mail address: jonas.lindahl@umu.se

https://doi.org/10.1016/j.joi.2018.04.002
1751-1577/

cation rate and the production of excellent research (i.e., highly cited papers) at the individual level and found moderate to high correlation between publication rate and the production of excellent research (Abramo et al., 2014; Larivière, Dorta-González, & Costas, 2016; Sandström & van Den Besselaar, 2016). In general, this line of research suggests that those who publish more also publish more excellent research (i.e., that the publication rate of an author is a central driver for the production of excellent research in academia). There is an ongoing debate in the scientometric community concerning the policy implications of these findings and the degree to which performance-based evaluation systems favoring publication volume have an overall positive or negative effect on science (see Waltman (2017) for a more in depth discussion on this topic).

Another line of research has examined if future excellence can be predicted on the basis of information from a researcher's past publication record. This line of research is motivated by the use of bibliometric indicators in contexts of selecting for excellence. When we are, for example, hiring a researcher or allocating funding, we are interested in providing resources to those who will make the best use of them. It is therefore relevant to examine how well an indicator in a previous time period can be used to predict who will attain excellence in a future time period. Danell (2011) tested two bibliometric indicators as predictors of excellence and found that a citation-based indicator performed better in predicting future research excellence than an indicator based on total publication volume. Haveman and Larsen (2014) tested a large number of indicators among early career astrophysicists and found that a citation-based indicator was the only one that significantly distinguished between excellent researchers and a control group. The results from these two studies suggest (1) that it is possible to use information of the publication track record of a researcher to predict future excellence and (2) that the best predictions are made with citation-based indicators.

Previous research indicate that one of the most important predictors of the impact of research articles is the prestige of the journal in which the articles are published (see e.g., Callaham, Wears, & Weber, 2002; Didegah & Thelwall, 2013; Judge, Cable, Colbert, & Rynes, 2007; Peng & Zhu, 2012; Yu, Yu, Li, & Wang, 2014). In a recent study by Bornmann and Williams (2017), the authors examined a large dataset of early career researchers and found that the citation based indicator Journal impact factor can predict future research performance (as defined by a citation based indicator). This is an interesting result since (1) journal metrics are often used by, e.g., funding agencies and research managers, to assess the performance of researchers in the early career and to make selections for future funding or job opportunities (Bornmann & Williams, 2017); (2) early career researchers have relatively short publication track records and CVs which makes selection more difficult; and (3) it is difficult to use citation based indicators in the early career due to the required citation windows (see e.g., Bensman, Smolinsky, & Pudovkin, 2010).

The data in this study consisted of the first eight years of the careers of 406 mathematicians active in the subfield of number theory. The aim of this study was two fold, (1) to contribute to the line of research examining the relationship between publication rate and excellence (see e.g., Larivière et al., 2016; Sandström & van Den Besselaar, 2016) by including publications in top journals as a potential third variable effecting the outcome, and (2) to contribute to the line of research on predicting future research excellence on the basis of past performance (see e.g., Bornmann & Williams, 2017; Danell, 2011; Haveman & Larsen, 2014) by introducing top journal publications as a potential predictor of future excellence in addition to the dimensions of total publication volume and impact. For these aims, the following research questions were formulated:

1. What is the relationship between publication rate, top journal publications, and attaining excellence during the first eight years of the career?
2. What is the importance of publication rate and publications in top journals in predicting who will attain excellence during the first eight years of the career?
3. Can we predict who will produce excellent research during the fifth to the eighth year on the basis of information on publication rate, top journal publications, and highly cited publications in the first four years of the career?
4. Which indicator is more important in predicting who will attain excellence during the fifth to the eighth year of the career – publication rate, top journal publications, or highly cited publications?

With the first two questions, I am examining whether the publication strategy or behavior of publishing in top journals, in addition to publication volume, might be an important predictor of who will attain excellence during the first eight years of the career. In the third and fourth question, I am interested in examining which publication behaviors during the first four years might be the most important in predicting whether an author will attain excellence in the fifth to the eight year of their career.

1.1. Field of study: mathematics and number theory

The object of analysis in this study is early career mathematicians active in number theory, a subfield in mathematics situated in pure mathematics. Mathematics is a discipline with specific characteristics that differ from many other fields (Abramo et al., 2011; Rousseau, 1988). In general, external resources are less important, publication volumes are smaller, collaboration is lower (Dubois, Rochet, & Schlenker, 2014), and papers have shorter reference lists and receive relatively fewer citations per paper compared to other disciplines (American Mathematical Society, 2015b; Bensman et al., 2010). Another characteristic is that the accumulation of citations takes a longer time in mathematics than in most other fields

(American Mathematical Society, 2015b; Bensman et al., 2010). If not taken into consideration, some of these characteristics might lead to unwanted biases in the analyses.

Some have examined whether citation analysis might be inappropriate in mathematics due to field-specific views on scientific quality and the significance of scientific contributions. Stern (1978) hypothesized that the most prestigious mathematicians are generally not the most highly cited ones. However, an empirical test of this hypothesis showed that top mathematicians were twice as highly cited as a randomly selected control group (Stern, 1978). Korevaar (1996) showed that expert assessments of the quality of publications and journals in mathematics correspond very well with citation-based indicators.

Finally, mathematics is a relatively large discipline (Smolinsky & Lercher, 2012). In 2013, 19,000 active PhDs were employed in academia in the US, while computer science had 8400 PhDs, psychology had 36,300 PhDs, and the physical sciences had 44,900 PhDs (National Science Board NSB, 2016).

2. Method

This section consists of three subsections. First the data collection is presented, followed by a subsection presenting the overall research design and the variables used in the study. In the last subsection, the two main methods are presented – logistic regression and dominance analysis.

2.1. Data collection

The dataset used in this study is based on the same dataset that was used in Lindahl and Danell (2016) and consists of article publication track records of 406 authors in number theory retrieved from the MathSciNet (MSN) database. The authors were selected on the basis of the following three criteria: (1) at least one published article in class 11 (i.e., Number theory) in the Mathematics Subject Classification (MSC) scheme between 2000 and 2003 (the start year of 2000 was chosen because the MSN citation index covers the period 2000 to the present); (2) that the publication career of an author was at least eight years to increase the likelihood that the authors were active researchers during the time span and to ensure a minimum citation window of almost five years (i.e., the citation counts were collected in December 2015); and (3) that the share of articles belonging to the MSC class in the track record of an author was equal to or larger than the share of any other MSC class found in that author's track record. The third criterion was used to decrease research field heterogeneity in the sample (Costas & Noyons, 2013).

MSN has a few favorable features for studying mathematics at the individual level. (1) MSN is a comprehensive database for mathematics with a global coverage. (2) Each author in the MSN database has a unique author ID that is connected to a unique publication profile. The mapping of author names to real persons is robust from 1985 and is conducted by an author name disambiguation algorithm and manually by professional indexers (American Mathematical Society, 2014). Thus, the problem with author name ambiguities (Smalheiser & Torvik, 2009) is to a large extent solved in MSN (American Mathematical Society, 2014). (3) All articles in MSN are classified according to the MSC scheme by professional indexers, and the MSC classes can be used to delineate subfields in mathematics (Dubois et al., 2014). (4) MSN is a citation database that covers the period of 2000 to the present. The citations in the MSN citation database are extracted from the reference lists of journals included in the so-called reference journals list that consists of more than 600 of the most important mathematical journals (Borjas & Doran, 2012). An advantage of using citation data from the MSN database rather than, e.g., the citation indices of Web of Science, is that MSN has better coverage in mathematics. Borjas and Doran (2012) showed only a 50 percent match rate between MSN and the citation indices of Web of Science. According to Ramos and Sarrico (2016), the Web of Science subject categories of Mathematics and Applied Mathematics had an estimated coverage of approximately 60 percent. To compare the coverage of MSN and Web of Science for the sample used in this study, I matched the documents in the MSN sample with the Web of Science citation indices. I found a match rate of 64 percent. If we define the MSN database as the gold standard for coverage in mathematics, Web of Science had a 36 percent lower coverage in number theory than the MSN database (see Appendix A for a description of the matching procedure). In the context of using publication-based indicators as predictors of research excellence, 36 percent missing values could potentially bias the results. I used the citation counts of the 64 percent matching documents between the MSN database and the citation indices of Web of Science to validate the MSN citation database. A correlation analysis showed a Pearson's r^2 of 0.93 between the citation counts of the documents from the MSN database and the citation counts of the matching documents from Web of Science. The high correlation indicates that the citation distributions of MSN and Web of Science are very similar in the case of authors who are active in number theory.

Two limitations with the citation index in the MSN database are that the citation counts are based on the reference lists of articles that are published in a subset of the most important mathematical journals included in the MSN citation database reference journal list (American Mathematical Society, 2016) and that interdisciplinary citations are not counted (i.e., citations from publications in other disciplines than mathematics). However, contrary to interdisciplinary citation databases, e.g., Web of Science, which have a low coverage in mathematics, the use of MSN reduces the missing data considerably. The high coverage of mathematics in MSN, the high correlation with the Web of Science citation counts, and the reduction of missing data values in the sample suggest that the use of MSN might provide a more accurate picture of

the predictability of excellence in mathematics and the subfield of number theory than a larger interdisciplinary citation database.

2.2. Design and variables

The dataset comprises the first eight years of the publication career of the 406 number theorists. The publication career of an author begins with the first publication in the MSN database (Costas & Noyons, 2013). The design of the study can be said to consist of two different models with different dependent variables and time frames:

1. The synchronic model: This model consists of an examination of the relationship between publication rate, top journal publications, and excellence during the first eight years of the career. The aim of this analysis is to examine to what degree the publication behavior of an author, i.e., publication volume and publications in top journals, might be important for the production of excellence. Thus, the synchronic model comprises the first eight years of the publication career.
2. The predictive model: This model is constructed to examine the importance of publication rate, top journal publications, and top 10% publications during the first four years of the career in predicting who will attain excellence during the fifth to the eighth years of their career. The purpose with this analysis is to see whether the publication behaviors that affect excellence are predictable over time. Thus, with the predictive model I examine how information from the publication track record in the first four years (i.e., period 1) can predict who will attain excellence during the fifth to the eighth year of their career (i.e., period 2).

The dependent variables were binary and indicated if an author could be defined as excellent. A binary dependent variable was chosen to make the research design more similar to decision making in science policy and management at the individual level, where the aim is to identify and distribute resources to individuals (e.g., approving applications or employing postdocs) that makes the best use of the resources according to the goals of the decision maker (i.e., attaining excellence). These decision are all binary.

A common standard for determining excellence at the individual level has not yet been developed within the scientometric community (Costas & Noyons, 2013). In this study an excellent researcher in the predictive model was defined as an author who had published at least one articles during the fifth to the eighth years of their publication career that could be considered excellent (Costas & Noyons, 2013). In the synchronic model, a researcher was defined as excellent if he or she had at least two excellent article during the first eighth years of their publication career. The higher threshold of two excellent articles in the synchronic model was chosen on the basis of the longer time period. A threshold of at least two excellent articles indicates continuity in the publication activity of excellent articles of an author.

In accordance with recommended best practice, research excellence was operationalized by the percentile-based indicator "top 10%" (Bornmann, 2013, 2014). I followed the approach for online identification of the top *k*% papers in the results page of a citation database interface as suggested by Ahlgren et al. (2014). The MSN database interface was used to construct the dependent variable. I adjusted the top 10% indicator for publication year, document type (i.e., journal articles), and whether an article was published in a sub-field oriented towards pure or applied mathematics (Smolinsky & Lercher, 2012) on the basis of the MSC scheme. A publication was defined as oriented towards pure mathematics if it had been classified with an MSC code between 00 and 58, and it was defined as oriented towards applied mathematics if it had been classified between 60 and 97. A document was defined as excellent if it had a citation count above the 90th percentile in a publication reference set (i.e., the retrieved documents ranked by citation rate in the results page in the MSN interface) adjusted for document type, publication year, and sub-field orientation. Approximately 13.3% of the articles were defined as top 10% during the first eight years of the career. This percentage was somewhat higher than the theoretical definition of top 10%, suggesting that the articles published by the authors included in the sample might have higher citation rates than the reference group as a whole. The proportion of authors that produced at least two top 10% articles during the first eight years of their career (i.e., the synchronic model) was 0.24, and the proportion that produced at least one top 10% article during the first four years (i.e., the predictive model) was 0.26.

Four predictors was constructed for the synchronic and the predictive models:

(1) Scientific publication rate was operationalized as the number of MSN journal articles during the first eight years in the synchronic model and during the first four years in the predictive model (coding: *Publication rate*). The total output of the 406 mathematicians consisted of journal articles, proceedings, and books. Because the focus in this study is scientific publications, and the main publication channel in mathematics is peer-reviewed journals (American Mathematical Society 2015a), the document types of proceedings and books were excluded from the sample.
(2) Publications in prestigious journals (coding: *Top journal publications*) were operationalized as the number of articles published in journals with high ranking according to the citation-based indicator of source-normalized impact per paper (SNIP). I downloaded an excel file from CWTS Journal Indicators that contained a list of all journals indexed in the Scopus database between 1999 and 2014 (CWTS, 2015). The CWTS Journal Indicators list contained the journal name with corresponding print-ISSN, e-ISSN, and SNIP values for each year. Each article in the dataset was matched on the basis of print-ISSN, e-ISSN, and full journal title against the journal list provided by CWTS Journal Indicators to obtain a SNIP value. A high-prestige journal was defined as a journal with a SNIP value equal to or above the 75th percentile in

Table 1
Descriptive statistics for the four non-binary predictors.

	M	1st	2nd	3rd	Min	Max	1	2	3
1. Publication rate									
First 8 years	7.7	4	6	10	1	53			
First 4 years	3.8	2	3	5	1	24			
2. Top journal publications									
First 8 years	2	0	1	3	0	20	0.557		
First 4 years	1.1	0	1	2	0	7	0.498		
3. Top 10% publications									
First 8 years	–	–	–	–	–	–	–	–	
First 4 years	0.5	0	1	2	0	7	0.380	0.597	
4. Collaboration									
First 8 years	0.8	0.4	0.8	1.1	0	3.5	0.007	0.019	–
First 4 years	0.7	0	0.6	1	0	4	−0.041	−0.025	−0.014

Note. *$P < 0.05$; M = Mean, 1st = 1st quartile, 2nd = Median, 3rd = 3rd quartile; 1, 2, and 3 = Pearson's r^2 correlation coefficients.

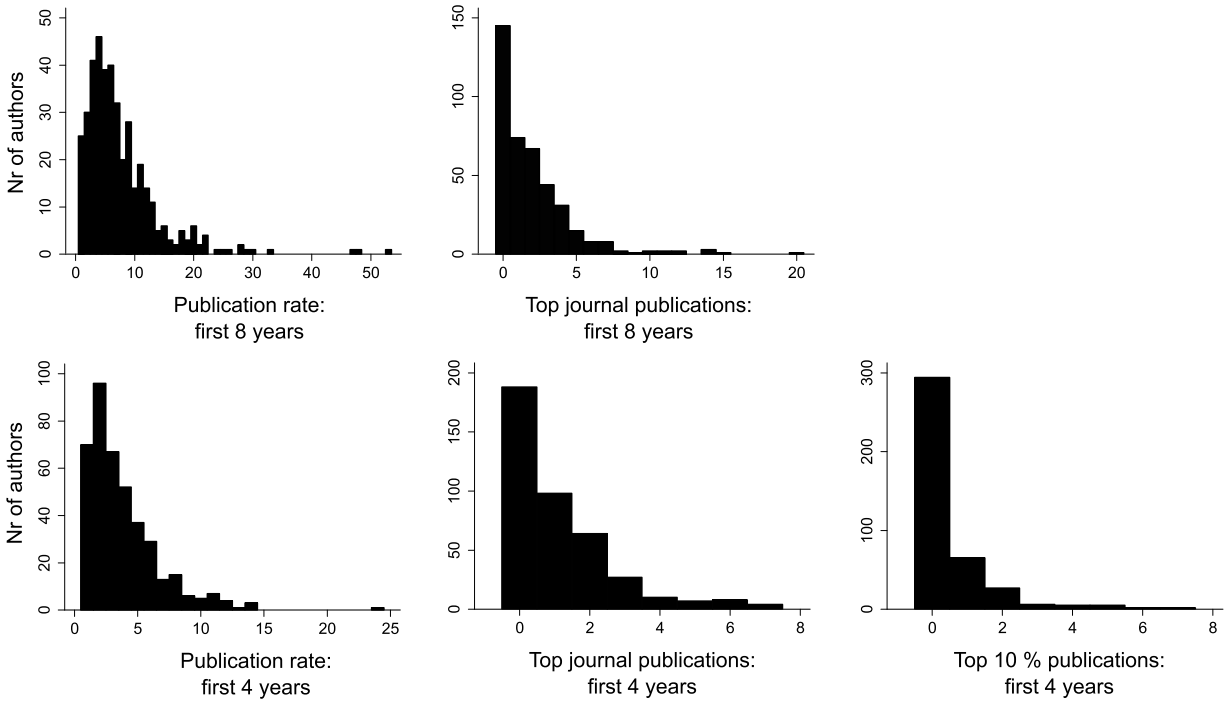

Fig. 1. Frequency distribution of the five predictors based on publication counts. The first row represents frequencies for the first eight years of the career (i.e., the synchronic model). The second row represents the first four years of the career (i.e., the predictive model).

the CWTS Journal Indicators list (CWTS, 2015; Waltman et al. 2012) in the publication year of the article during the first eight years in the synchronic model and during the first four years in the predictive model. I calculated one percentile for each year (1999–2014). The SNIP values were calculated on the basis of the revised SNIP indicator (Waltman et al., 2012).

(3) Several past studies have found a positive relationship between collaboration and research performance in general (Franceschet & Costantini, 2010), and more specifically from the point of view of scientometrics, a positive relationship between the number of co-authors and the citation rate of a paper (see e.g., Glänzel, 2002; Persson, Glänzel, & Danell, 2004). A collaboration variable was therefore constructed as a potential control variable in the analyses. Collaboration was operationalized as the average number of co-authors per publication during the first eight years in the synchronic model and during the first four years in the predictive model (coding: *Collaboration*).

(4) To operationalize an author's ability to produce excellent research in period 1, I constructed a predictor consisting of the number of top 10% articles of an author during the first four years of their career (coding: *Top 10% publications*).

Descriptive statistics for the non-binary predictors are presented in Table 1 and Fig. 1.

2.3. Data analysis

In this sub-section I define the concept of predictor importance and introduce the two main statistical methods in this study – logistic regression analysis and dominance analysis.

2.3.1. Logistic regression analysis

Logistic regression analysis was used to examine and describe the relationship between the binary dependent variable and both single and multiple predictors (Hosmer & Lemeshow, 2000). Consider k predictors denoted by a vector $x' = (x_1, x_2 ..., x_k)$. The log odds of the multiple logistic regression model are given by the following equation:

$$g(x) = ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \quad (1)$$

where $g(x)$ denotes the log odds of the model, ln denotes the natural logarithm, and $\pi(x)$ denotes the conditional probability that an author is a producer of at least two top 10% papers. More specifically, the logistic regression analysis was used to examine and describe the change in the probability of being a producer of at least two top 10% articles given a unit change in *Publication rate*, *Top journal publications*, or *Top 10% publications*.

2.3.2. Predictor importance

The concept of predictor importance is relevant in situations when we have, for example, many different bibliometric indicators and want to know how important these different indicators are relative to each other in terms of predicting the dependent variable. There are at least three different definitions in the literature (Tonidandel & LeBreton, 2010). According to the first definition, the importance of a predictor is viewed in terms of statistical significance (Tonidandel & LeBreton, 2010). The second definition of predictor importance refers to the practical utility of a predictor (Tonidandel & LeBreton, 2010). An assessment of the practical utility usually takes into account the effect size and the circumstances of the particular situation in which the predictor might be useful (Tonidandel & LeBreton, 2010). The main concern in these definitions of predictor importance (i.e., statistical significance and practical significance) is usually the effect of a predictor by itself or the effect of a predictor in combination with other predictors (Azen & Budescu, 2003). A third definition – the one used in this study – of predictor importance, can be found within the framework of dominance analysis where predictor importance is determined by the contribution each predictor makes to the model fit by itself and in combination with the other predictors (Azen & Budescu, 2003). This definition is referred to as the relative importance of a predictor. The relative importance of a predictor, as defined in this study, is not concerned with statistical significance or directly with the practical utility of a predictor. Compared with the above-stated definitions, dominance analysis has a more general definition of predictor importance that is suitable when the main interest is to rank-order predictors according to their importance in predicting an outcome, e.g., bibliometric indicators in the context of predicting research excellence (Azen & Budescu, 2003).

For example, if we want to know whether *Publication rate* or *Top journal publications* is more important in predicting future research excellence in a model that also contains *Top 10% publications*, we might find that *Top journal publications* contributes more to prediction in combination with the other predictors than does *Publication rate*. This result does not necessarily mean that *Top journal publications* by itself is more important than *Publication rate* by itself or in any other subsets of the model (Azen & Budescu, 2003). However, if it can be established that *Top journal publications* is the better predictor than *Publication rate* in all subset models, we could be more confident in defining *Top journal publications* as the more important predictor of the two (Azen & Budescu, 2003).

2.3.3. Dominance analysis

The goal of dominance analysis is to examine the relative importance of predictors, to determine patterns of dominance between predictors, and to rank order the predictors accordingly. Dominance analysis can answer questions such as whether *Publication rate* or *Top journal publications* is more important or better at predicting research excellence. To analyze relative importance, a common approach in the literature has been to use standardized regression coefficients. However, it has been shown that such methods do not work well when the predictors are correlated (Azen & Traxel, 2009). Dominance analysis solves the problem of correlated predictors by focusing the analysis on the additional contribution to the model fit across all sub-models in a regression model (Azen & Traxel, 2009). Bibliometric indicators are often highly correlated (see e.g., Bornmann, Mutz, & Daniel, 2008; Costas & Bordons, 2007), and dominance analysis is therefore a suitable method to analyze the relative importance of the predictors in this study (see Table 1 for correlations between the predictors in this study).

In dominance analysis, the relative importance for a given predictor is determined by its additional contribution to the model fit across all sub-models in the full model (i.e., by itself and in combination with the other predictors). Consider a regression model with P predictors. Dominance analysis defines the additional contribution of any given predictor to a given subset model as the change in the model fit (e.g., R^2, a model fit measure that indicates the proportion of variance that is explained in the dependent variable by the predictors in the model; Azen & Traxel, 2009) when the predictor is added to the model (Azen & Traxel, 2009). To exemplify this, the additional contribution of the predictor X_1 to a model containing X_2 and X_3 is defined as the difference between the model with X_1 X_2 X_3 and the model with X_2 X_3 (i.e., model X_1 X_2 X_3–model X_2 X_3).

Table 2
Log odds, standard errors, *p*-values, 95% confidence intervals, and R^2_M for the bivariate and multiple logistic regression models. .

				95% CI		R^2_M
	Log odds	Standard error	*p*-value	Lower	Upper	
Bivariate model						0.143
Publication rate	0.159	0.024	<0.001*	0.113	0.205	
Intercept	−2.521	0.243	<0.001*	−2.998	−2.045	
Bivariate model						0.306
Top journal publications	0.718	0.084	<0.001*	0.554	0.882	
Intercept	−2.868	0.254	<0.001*	−3.366	−2.370	
Full model						0.323
Publication rate	0.066	0.024	0.007*	0.018	0.114	
Top journal publications	0.651	0.087	<0.001*	0.481	0.821	
Intercept	−3.264	0.305	<0.001*	−3.861	−2.666	

Note. *$p < 0.05$.

In dominance analysis, the relative importance of a predictor has three different definitions: (a) a predictor, X_i, is considered to completely dominate another predictor, X_j, if its additional contribution to the model fit in every possible subset model is greater than the additional contribution of the other predictor; (b) a predictor is considered to conditionally dominate another if its average additional contribution within each model size is greater than that of another predictor; and (c) a predictor is considered to generally dominate another if its average conditional contribution over all model sizes is greater than that of the other predictor (Azen & Traxel, 2009). Complete dominance is the strongest definition of predictor importance. If, for example, *Publication rate* is consistently a better predictor than *Top journal publications* both when considered by itself and in any subset of the other predictors in the full model, *Publication rate* is viewed as the more important predictor of the two. However, it might also be of interest to examine conditional and general dominance. A desirable property of general dominance is that it sums to the model fit measure used in the full model (e.g., R^2) containing all predictors and can be interpreted as the amount of variance (both shared and unique) a predictor explains in the dependent variable (Azen & Traxel, 2009). As such, the general dominance can be interpreted as a measure of the relative effect size of a predictor in terms of a predictor's contribution to model fit (Tonidandel & LeBreton, 2010).

To determine the relative importance in terms of complete, conditional, and general dominance in this study, an R^2 analogue for logistic regression was needed. According to Azen and Traxel (2009), an R^2 analogue for logistic regression should be assessed on the basis of the following four criteria: (1) Boundedness, the measure should take values between 0 and 1; (2) Linear invariance, the measure should be invariant to nonsingular linear transformations of the variables; (3) Monotonicity, the measure should not decrease with additional predictors; and (4) Intuitive interpretability, the measure should agree with the scale of the intermediate values of R^2 in the linear regression.

In this study I used McFadden's measure, R^2_M (Azen & Traxel, 2009; Mcfadden, 1974), which satisfies all four of the above-mentioned criteria. McFadden's measure can be defined as:

$$R^2_M = 1 - \frac{ln(L_M)}{ln(L_0)} \tag{3}$$

where L_0 denotes the likelihood of the intercept model (i.e., an empty model), L_M denotes the likelihood of the fitted model (i.e., the intercept and predictors), and *ln* denotes the natural logarithm. As an R^2 analogue, R^2_M has an intuitive interpretation as a proportional reduction in error measure where the intermediate values (i.e., the values between 0 and 1) agree with the scale of the ordinary R^2 (Azen & Traxel, 2009).

3. Results and discussion

In this section the results from the logistic regressions and the dominance analysis are presented. The results for the synchronic model are presented first (i.e., the relationship between *Publication rate*, *Top journal publications*, and excellence during the first eight years), then the results from the predictive model are presented (i.e., how well *Publication rate*, *Top journal publications*, and *Top 10% publications* during the first four years can predict whether a researcher will produce excellent research in the fifth to the eighth year of their career).

3.1. The synchronic model: examining the relationship between publication rate, top journal publications, and excellence during the first eight years

3.1.1. Logistic regression

A logistic regression analysis was applied to examine the relationships between *Publication rate*, *Top journal publications*, and the dependent variable during the first eight years of the publication career. The *Collaboration* control variable did not have an effect in any of the models and was therefore excluded from all models. Two bivariate models and a multiple model

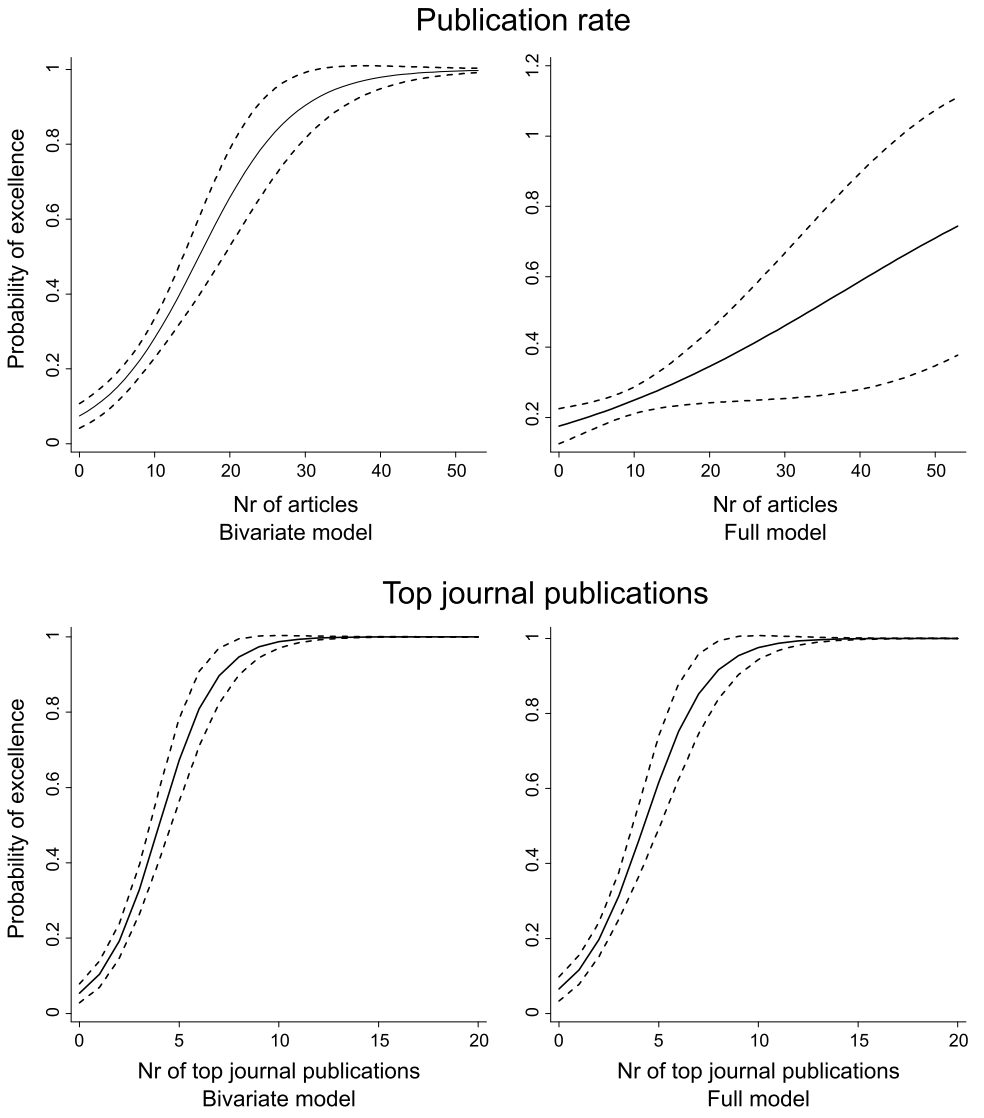

Fig. 2. Two plots for each predictor. The plot to the left shows the bivariate models. The plot to the right shows the full models. The y-axis denotes the probability of producing at least two top 10% articles, and the x-axis shows the predictor values.

(i.e., the full model, Table 2) were tested. The results are summarized in Table 2, where the coefficients for the log odds, standard errors, *p*-values, 95% confidence intervals, and the model fit measure R^2_M are reported.

The log odds coefficients in Table 2 for *Publication rate* and *Top journal publications* were statistically significant ($p < 0.05$) in both the bivariate and the full models. The log odds coefficients were transformed to predicted probabilities and plotted to get an indication of the size of the effects in the bivariate and full models. The following equation was used to transform the log odds to probabilities:

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}} \tag{4}$$

where $g(x)$ denotes the log odds, e denotes the natural logarithm, and $\pi(x)$ denotes the conditional probability of producing at least two top 10% articles in the given time period.

Fig. 2 consist of four plots, two plots for each predictor – one plot for the effect of the predictor in the full model and one plot for the effect in the bivariate model. The y-axis denotes the probability of producing at least two top 10% articles, and the x-axis denotes the values of the predictor. Each plot contains 95% confidence intervals.

As can be seen in Fig. 2, both *Publication rate* and *Top journal publications* seem to have an increasing marginal effect until the higher values where the effect levels off (i.e., the incremental increase in the probability of producing at least two top 10% article becomes larger up to a point as the values of *Publication rate* and *Top journal publications* increase). The positive effect of *Publication rate* on attaining excellence is in line with previous research (see, e.g., Larivière et al. 2016; Sandström

Table 3
Dominance analysis for *Publication rate* and *Top journal publications*.

Subset Model	R^2_M	Additional contribution	
		PR	TJP
0 predictor (intercept) average		0.143	0.306
PR	0.143		0.181
TJP	0.306	0.018	
1 predictor average		0.018	0.181
PR+TJP	0.323		
Overall average		0.080	0.243
% of R^2_M		24.80	75.20

Note. PR = subset model with only *Publication rate*, TJP = subset model with only *Top journal publications*.

& van Den Besselaar, 2016). However, the plot for the full model (Fig. 2, upper right) indicates that the effect of *Publication rate* was quite heavily reduced in the full model, while the effect of *Top journal publications* remained strong (Fig. 2, lower right).

While interpreting these results we have to keep in mind that *Publication rate* and *Top journal publications* are related in such a way that the number of publications in top journals are dependent on the publication rate of an author because the set of top journal publications is a subset of the total number of publications. An author can never have more top journal publications than the maximum number of articles. For an author to have, for example, four top journal publications, that author needs at least four published articles. This dependency suggests that treating *Publication rate* and *Top journal publications* as covariates that are correlated to the dependent variable and to each other in the multiple logistic regression model could lead to somewhat counterintuitive results. For example, if the effect of *Publication rate* completely disappears when *Top journal publications* is controlled for, it would seem like *Publication rate* has no effect on the production of excellence even though the value of the *Top journal publications* predictor is dependent on the value of *Publication rate*. *Publication rate* may therefore be viewed as a necessary condition for attaining excellence.

If we look at Fig. 2 (lower right graph) we can see that the probabilities of attaining excellence increases much between the values of 3 and 8, e.g., an author with seven top journal publications also have a value of at least 7 on the publication rate predictor. We should also consider that even authors that have an active strategy to publish in top journals won't publish in top journals every time. In this dataset the authors who had published at least one top journal publication had, on average, 1.9 articles in non-top journals for each top journal publication, and approximately one top journal publication in three published articles. Thus, authors that publish many top journal publications also tend to publish many non-top journal publications. The correlation coefficient between *Publication rate* and *Top journal publications* is relatively high at 0.557 (Table 1). While the effect of *Publication rate* seem to disappear to a large extent when we are controlling for *Top journal publications*, the strong increasing marginal effect of *Top journal publications* (Fig. 2, lower right) implies the importance of a high publication rate. I therefore conclude from the logistic regression analysis that those with many publications in top journals, which implicitly require a high publication volume, have a higher probability of attaining excellence than those with fewer publications in top journals. Those authors with no top journal publications had very slim chances of attaining excellence. Looking at the data, only three of 145 authors with no top journal publications had at least two top 10% publications. The data contained 96 excellent authors.

3.1.2. Dominance analysis

A dominance analysis was conducted to examine the relative importance of *Publication rate* (i.e., subset model PR) and *Top journal publications* (i.e., subset model TPJ) during the first eight years of the publication career in predicting who will produce at least two top 10% publications during the fifth to the eighth years of their career. The results from the dominance analysis are summarized in Table 3. Column one of Table 3 consists of the full model and all subset models. Column one also contains averages for all different group sizes (i.e., the overall average) and the predictor percentage of the model fit measure R^2_M that was used in this study. Column two shows the R^2_M value for each model. Columns three and four represent the additional contribution of each model. The additional contribution can be understood as the unique R^2_M contribution of a predictor to the intercept model or to some of the subset models.

The R^2_M for the subset model PR, i.e., the bivariate model for *Publication rate*, was 0.143, and the R^2_M for the subset model TPJ, i.e., the bivariate model for *Top journal publications*, was 0.306 (Table 3). The R^2_M for PR+TJP (i.e., the full model) was 0.323. If we look at the dominance patterns between PR and TPJ in Table 3, we can establish complete dominance of TPJ over PR because the additional contribution to model fit of TPJ is greater in every possible subset model than the additional contribution of TP, 0.306 > 0.143. From this it follows that conditional dominance of TPJ over PR was established as well because the average additional contribution within each model size was greater for TPJ than for PR – 0.306 > 0.143, and 0.181 > 0.018.

By averaging all conditional dominance values for each predictor, we can establish general dominance (i.e., the overall average) and get a measure of the effect size of the relative importance of the predictors in the full model. The overall average for PR was (0.143 + 0.018)/2 = 0.080, and the overall average for TPJ was (0.306 + 0.181)/2 = 0.243 (Table 3). The overall average for TPJ and PR sum to the R^2_M for the full model PR+TJP, i.e., 0.080 + 0.243 = 0.323. TPJ had the highest overall average and

Table 4
Sample D_{ij} values (n = 406), Average D_{ij}, and Reproducibility Sample D_{ij} over 10,000 bootstrap samples.

Dominance pattern	i	j	SampleD_{ij}	AverageD_{ij}	Reproducibility Sample D_{ij}
Complete	PR	TJP	0	0.0002	0.9998
Conditional	PR	TJP	0	0.0002	0.9998
General	PR	TJP	0	0.0002	0.9998

contributed 75.20 percent to R^2_M, and TP had the smallest contribution at 24.80 percent. Because both TPJ had a higher overall average than PR, the general dominance of TPJ over PR could be established (Table 3).

To examine the stability of the dominance patterns if the study were to be repeated many times, I applied the bootstrap procedure suggested by Azen and Budescu (2003). This procedure generates a sampling distribution consisting of bootstrapped samples from the original sample of 406 authors on the basis of random sampling with replacement (Azen & Budescu, 2003). The dominance value, D_{ij}, for a predictor pair (e.g., *Publication rate* and *Top journal publications*) in a sample can take the values of 1 if X_i dominates X_j, 0 if X_j dominates X_i, or 0.5 if dominance cannot be established between X_i and X_j. I generated 10,000 bootstrap samples.

The results from the bootstrap procedure are presented in Table 4. The first column shows the dominance pattern of each row. Columns two and three show the predictor pair. Columns 4–6 represent the following three statistics: (1) Sample D_{ij} represents the actual dominance values in the original sample of 406 authors used in this study, (2) Average D_{ij} represents the expected dominance value of X_i and X_j in the 10,000 bootstrap samples (i.e., in the bootstrap sampling distribution) and takes a value between 0 and 1 (Azen & Budescu, 2003), and (3) Reproducibility Sample D_{ij} represents the proportion of bootstrap samples that agree with the actual dominance values in the sample of 406 authors used in this study. This proportion can be interpreted as the probability of reproducing the dominance value observed in the sample used in this study if the study were to be repeated many times (Azen & Budescu, 2003).

As can be seen in Table 4, the dominance patterns from the 10,000 bootstrap samples agree well with the dominance patterns in the original sample. The Average D_{ij} is very small at 0.0002, indicating that if the study were to be repeated many times the general dominance of TJP over PR that we observed in Sample D_{ij} would persist. The Reproducibility Sample D_{ij} indicated that for general dominance, 99.98 percent of the 10,000 bootstrap samples had the same dominance pattern as observed in Sample D_{ij}.

The following is a summary of the dominance analysis for the synchronic model. *Publication rate* was completely, conditionally and generally dominated by *Top journal publications*. Ordering the predictors by their relative importance as defined by the overall average, i.e., the percentage contribution to R^2_M, gave the following rank: (1) *Top journal publications*, and (2) *Publication rate*. Given the definition of predictor importance in this study, this result indicated that *Top journal publications* was relatively more important than *Publication rate* in predicting who will produce at least two top 10% papers during the eight first years of the career.

However, we have to remember that the number of top journal publications are dependent on the total number of publications, e.g., an author who has published five top journal publications has a *Publication rate* of at least five. Authors with more top journal publications have a higher publication volume. The overall conclusion from the logistic regression and the dominance analysis in the synchronic model is that the publication volume of an author is a necessary and important condition for producing excellent research. However, the higher relative importance of *Top journal publications* suggest that publication behaviors related to publishing in top journals are very important in the process of attaining excellence among number theorists in the early career in addition to the publication volume.

3.2. The predictive model: examining the importance of publication rate, top journal publications, and top 10% publications in predicting future excellence

3.2.1. Logistic regression

Multiple logistic regression analysis was used to examine how well *Publication rate*, *Top journal publications*, and *Top 10% publications* during the first four years of the career could predict who will produce at least one top 10% article in the fifth to the eighth year of their career. The purpose was first and foremost to describe the direction of the relationships and the size of the effects. A summary of these results can be found in Table 5. The multiple model (i.e., the full model) consisted of the predictors *Publication rate*, *Top journal publications*, and *Top 10% publications*. Each bivariate model consisted of one of these predictors and the dependent variable (Table 5).

An examination of the log odds coefficients in Table 5 revealed that *Publication rate*, *Top journal publications*, and *Top 10% publications* were all statistically significant ($p < 0.05$) in the bivariate models. However, in the full model only *Top journal publications* and *Top 10% publications* maintained statistically significant effects.

The log odds were transformed to predicted probabilities (see Eq. (3)) and plotted (see Fig. 3). As can be seen in Fig. 3, the three bivariate logistic regression models show quite similar positive relationships with the dependent variable. Similar to the synchronic model, the plots indicated that the effect of *Publication rate* was reduced after controlling for the other two predictors in the full model (Fig. 3). The strong positive effect of *Top 10% publications* on attaining excellence is in agreement with previous findings (see e.g., Danell, 2011; Haveman & Larsen, 2014).

Table 5
Logistic regression analysis for the bivariate and full models.

				95% CI		
	Log odds	Standard error	*p*-value	Lower	Upper	R^2_M
Bivariate model						0.051
Publication rate	0.190	0.041	<0.001*	0.110	0.270	
Intercept	−1.843	0.210	<0.001*	−2.254	−1.432	
Bivariate model						0.102
Top journal publications	0.516	0.082	<0.001*	0.356	0.677	
Intercept	−1.743	0.166	<0.001*	−2.069	−1.417	
Bivariate model						0.095
Top 10% publications	0.704	0.124	<0.001*	0.460	0.948	
Intercept	−1.482	0.140	<0.001*	−1.757	−1.207	
Full model						0.129
Publication rate	0.059	0.048	0.215	−0.034	0.153	
Top journal publications	0.304	0.104	0.003*	0.100	0.508	
Top 10% publications	0.423	0.141	0.003*	0.147	0.699	
Intercept	−1.944	0.217	<0.001*	−2.370	−1.518	

Note. *$p < 0.05$.

Table 6
Dominance analysis for *Publication rate*, *Top journal publications*, and *Top 10% publications*.

Subset Model	R^2_M	Additional contribution		
		PR	TJP	T10
0 predictors (intercept) average		0.051	0.102	0.095
PR	0.051		0.056	0.059
TJP	0.102	0.006		0.024
T10	0.095	0.015	0.030	
1 predictor average		0.010	0.043	0.041
PR+TJP	0.108			0.021
PR+T10	0.110		0.019	
TJP+T10	0.126	0.003		
2 predictors average		0.003	0.019	0.021
PR+TJP+T10	0.129			
Overall average		0.022	0.055	0.053
% of R^2_M		16.75	42.45	40.80

Note. PR = subset model with only *Publication rate*, TJP = subset model with only *Top journal publications*, T10 = subset model with only *Top 10% publications*.

We have to remember that the number of top journals and top 10% publications are dependent on the total number of publications. The effect of *Publication rate* seem to disappear while controlling for *Top journal publications* and *Top 10% publications*. However, the strong increasing marginal effect of *Top journal publications* and *Top 10% publications* (Fig. 3, middle right, lower right) implies the importance of a high publication rate. Similar to the synchronic model, authors who publish many top journal publications and many top 10% publications tend to have a high publication rate. The authors had approximately one top journal publication in two published articles, and one top 10% publication in two published articles. I conclude from the logistic regression analysis that those with many publications in top journals and many top 10% publications, which implicitly require a high publication volume, have a higher probability of attaining future excellence than those with fewer top journal, and top 10% publications.

3.2.2. Dominance analysis

A dominance analysis was conducted to examine the relative importance of *Publication rate* (i.e., subset model PR), *Top journal publications* (i.e., subset model TPJ), and *Top 10% publications* (i.e., subset model T10) during the first four years of the publication career in predicting who will produce at least one top 10% publication during the fifth to the eighth years of their career. The results from the dominance analysis are summarized in Table 6.

The R^2_M for the subset model PR, i.e., the bivariate model for *Publication rate*, was 0.051; the R^2_M for the subset model TPJ, i.e., the bivariate model for *Top journal publications*, was 0.102; and the R^2_M for he subset model T10, i.e., the bivariate model for *Top 10% publications* was 0.095 (Table 6). The R^2_M for PR+TJP+T10 (i.e., the full model) was 0.129. If we look at the dominance patterns between PR and TPJ in Table 6, we can establish complete dominance, conditional, and general dominance of TPJ over PR because the additional contribution to model fit of TPJ is greater in every possible subset model than the additional contribution of TP. The pair T10 and PR had a similar pattern where complete, conditional, and general dominance could be established for T10 over PR. Complete and conditional dominance could not be established for the predictor pair T10 and TPJ. TPJ had the highest overall average and contributed 42.45 percent to R^2_M, T10 had the second largest contribution at 40.80 percent, and TP had the smallest contribution at 16.75 percent (Table 6).

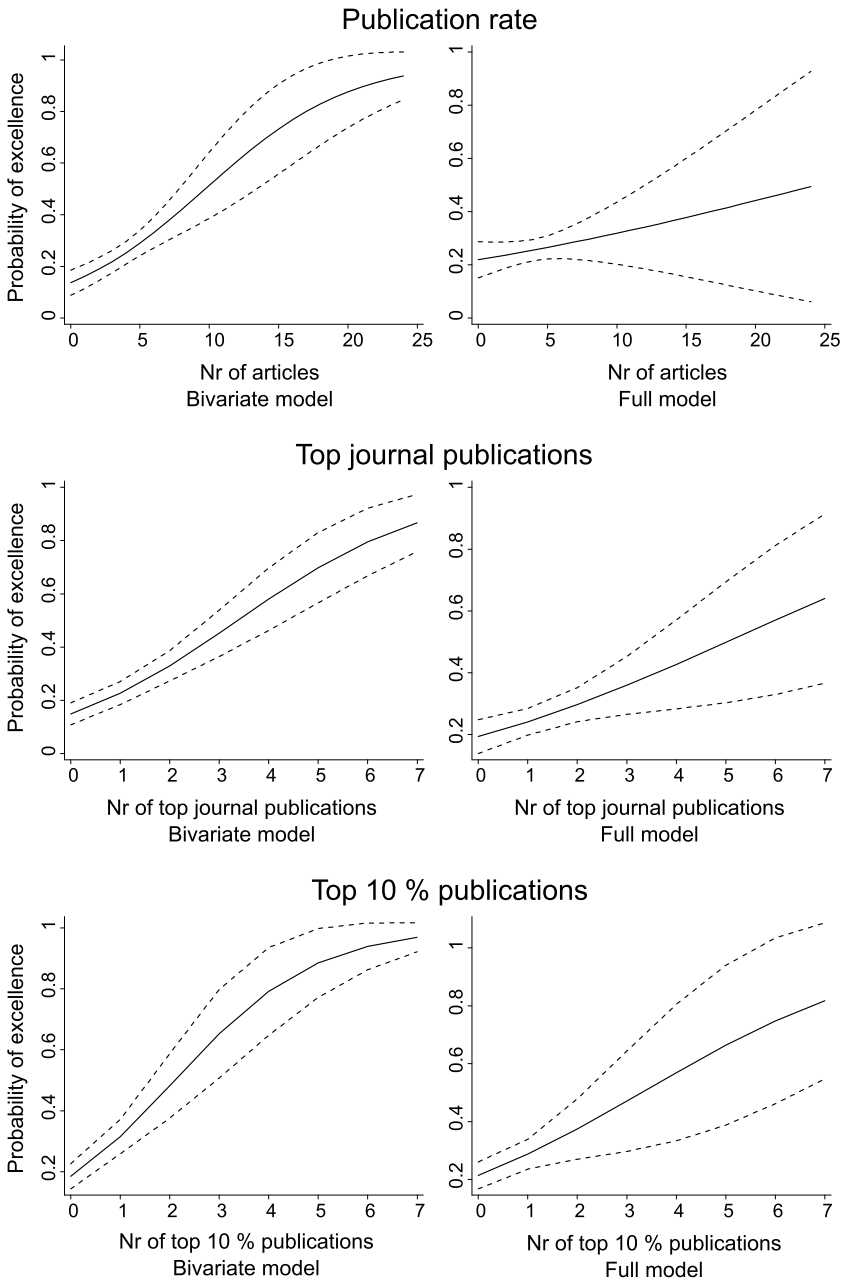

Fig. 3. Two plots for each predictor. The plots to the left show the bivariate models. The plots to the right show the full models. The y-axis denotes the probability of producing at least two top 10% articles, and the x-axis consists of the predictor values.

In the context of using bibliometric indicators as decision support tools, it is interesting to specifically examine the additional contribution (i.e., the unique contribution) of a predictor to a given model to see whether the new predictor actually adds anything to the model in terms of predicting the outcome. Suppose that the contribution of 11.91 percent to R^2_M by PR is to a large extent an additional contribution (i.e., unique contribution) to PR+TJP+T10 (i.e., the full model), then PR might be useful in a decision-making context even if the actual contribution is quite small. However, if the contribution of 11.91 percent does not add anything to the given model, PR would be redundant because it does not add anything in predicting who will produce at least one top 10% articles in period 2.

In Table 6 we can see that PR has an additional contribution of 0.006 to R^2_M in TPJ, 0.015 to T10, and less than 0.003 to TJP+T10. The additional contribution of PR to any of the subset models is very small, indicating that *Publication rate* is more or less redundant in a model with only *Top journal publications*, a model with only *Top 10% publications*, or a model with both *Top journal publications* and *Top 10% publications*. It is also interesting to look at the additional contribution of

Table 7
Sample D_{ij} values (n = 406), Average D_{ij}, and Reproducibility Sample D_{ij} over 10,000 bootstrap samples.

Dominance pattern	*i*	*j*	Sample D_{ij}	Average D_{ij}	Reproducibility Sample D_{ij}
Complete	PR	TJP	0	0.069	0.878
	PR	T10	0	0.082	0.887
	TJP	T10	0.5	0.514	0.134
Conditional	PR	TJP	0	0.069	0.878
	PR	T10	0	0.082	0.887
	TJP	T10	0.5	0.514	0.134
General	PR	TJP	0	0.035	0.965
	PR	T10	0	0.064	0.936
	TJP	T10	1	0.521	0.521

TPJ to T10 and PR+T10, and T10 to TPJ and PR+TPJ. As can be seen in Table 6, TPJ does not seem to add much to T10 (additional contribution = 0.024) or to PR+T10 (additional contribution = 0.019). Similarly, T10 does not seem to add much to TPJ (additional contribution = 0.024) or to PR+TPJ (additional contribution = 0.021). This result suggests that TPJ and T10 are basically interchangeable in predicting who will produce at least two top 10% publications in period 2, i.e., TPJ could replace T10 and vice versa without losing predictive power in the model.

I applied a bootstrap procedure to examine the stability of the observed dominance patterns (Azen & Budescu, 2003). I generated 10,000 bootstrap samples. The results from the bootstrap procedure are presented in Table 7. As can be seen in Table 7, the dominance patterns from the 10,000 bootstrap samples agree quite well with the dominance patterns in the original sample. The indication of indeterminacy of general dominance in the case of the predictor pair TJP and T10 is strengthened by the bootstrap procedure where the Sample D_{ij} is 1, but the Average D_{ij} is 0.514, indicating that if the study were to be repeated many times general dominance would not be established. A reasonable interpretation seems to be that TPJ and T10 are approximately equally important predictors (given the definition of predictor importance in this study) of who will produce at least two top 10% articles during the fifth to the eighth year of their career. For the predictor pair PR and TJP, the Reproducibility Sample D_{ij} indicated that for general dominance, 96.5 percent of the 10,000 bootstrap samples had the same dominance pattern as observed in Sample D_{ij}. The Reproducibility Sample D_{ij} was 93.6 percent for the PR and TJP pair.

The following is a summary of the dominance analysis. *Publication rate* was completely, conditionally, and generally dominated by *Top journal publications* and *Top 10% publications*. Ordering the predictors by their relative importance as defined by the overall average gave the following rank: (1) *Top journal publications*, (2) *Top 10% publications*, and (3) *Publication rate*. Given the definition of predictor importance in this study, this result indicated that *Top journal publications* was more important than *Publication rate* in predicting who will produce at least two top 10% papers later in their career. Examining the additional contribution indicated that *Publication rate* was redundant in all models containing either *Top journal publications* or *Top 10% publications*, indicating that if we have information on *Top journal publications* and/or *Top 10% publications* of an author, the additional information of the number publications would not provide much additional help in predicting whether this author will attain excellence or not during period 2. Strengthened by the bootstrap analysis, I conclude that *Top journal publications* and *Top 10% publications* contributed an equal percentage to R^2_M and neither of them added much to the other suggesting that these two predictors were, more or less, interchangeable in predicting excellence in this study. Similar to the logistic regression analysis it is important to recognize the implied importance of *Publication rate* given the inherent dependencies between the variables. I conclude that publications in top journals and high impact publications have a high importance in predicting who will attain future excellence, which in turn implies the importance of the overall publication output.

4. Conclusions

The first and second research questions were:

(1) *What is the relationship between publication rate, top journal publications, and attaining excellence during the first eight years of the career?*
(2) *What is the importance of publication rate and publications in top journals in predicting who will attain excellence during the first eight years of the career?*

To answer these questions, I conducted a logistic regression analysis and a dominance analysis examining the relationship between *Publication rate*, *Top journal publications*, and excellence in the first eight years of the publication career and the relative importance of these two variables in predicting who will attain excellence. The results from these analyses generally agree with previous research indicating that publication rate has a positive effect on the production of excellent research (see e.g., Larivière et al., 2016; Sandström & Van Den Besselaar, 2016). An important contribution of this study is the high relative importance of publishing in top journals and the indication that those who publish many articles in top journals, which implicitly require a high publication count, have a higher probability of attaining excellence. These results suggest

that publishing in top journals is very important in the process of attaining excellence in the early career in addition to *publication volume*.

Cole and Cole (1967) classified scientists on the basis of their publication strategy into four types: (1) prolific scientists, i.e., researchers with high publication rate and high impact; (2) mass producers, i.e., researchers with high publication rate but low impact; (3) selective scientists, i.e., researchers with intermediate to low publication rate but high impact; and (4) silent scientists, i.e., researchers with low publication rate and low impact. Based on the results in the synchronic model, it could be the case that the difference between prolific scientists and mass producers might partially be explained by the propensity to publish in top journals. This could be investigated in future research.

The third and fourth research questions were:

(3) *Can we predict who will produce excellent research during the fifth to the eighth year on the basis of information on publication rate, top journal publications, and highly cited publications in the first four years of their career?*
(4) *Which indicator is more important in predicting who will attain excellence during the fifth to the eighth year of their career – publication rate, top journal publications, or highly cited publications?*

To answer the second and third questions I conducted a logistic regression analysis and a dominance analysis. The results from the dominance analysis for predicting future excellence indicated that *Publication rate* was completely, conditionally, and generally dominated by *Top journal publications* and *Top 10% publications*. Ordering the predictors by their relative importance (i.e., contribution to model fit) gave the following rank: (1) *Top journal publications*, which contributed 44.51 percent to model fit; (2) *Top 10% publications*, which contributed 43.58 percent; and (3) *Publication rate*, which contributed 11.91 percent. Given the definition of predictor importance in this study, this result indicated that *Top journal publications* was the most important indicator in predicting who will produce at least one top 10% paper during the fifth to the eighth year of their career.

Examining the additional contribution indicated that *Publication rate* was redundant in all models containing *Top journal publications* and/or *Top 10% publications*, indicating that if we have information on *Top journal publications* and/or *Top 10% publications* of an author, the additional information of the number publications will not provide much help in predicting whether this author will attain excellence or not in period 2. This does not mean, however, that *Publication rate* is of no importance at all in terms of predicting future excellence. Such a conclusion would be counter-intuitive because at least one publication is a necessary condition for an author to have at least one top journal or top 10% publication. The dominance analysis indicated that *Top journal publications* and *Top 10% publications* were interchangeable in terms of predicting research excellence (i.e., it would not matter much if we used a model based on only *Top journal publications*, only *Top 10% publications*, or both, at least given the definition of predictor importance used in this study).

The logistic regression indicated that the effect of *Publication rate* disappears in the full model when *Top journal publications* and *Top 10% publications* are controlled for. Taking into consideration the results of the dominance analysis, the logistic regression, and the inherent dependencies between the variables, I conclude that those authors with many publications in top journals and/or many top 10% publications, which implicitly require a high publication volume, have higher probability of attaining excellence than those with fewer top journal, and top 10% publications.

Researchers with a track record of publishing in top journals and publishing excellent research are more likely to produce excellent research in a future time period. This result is in line with previous research (see e.g., Bornmann & Williams, 2017; Danell, 2011; Havemann & Larsen, 2015). The results also suggest that the *Top 10% publications* predictor and the *Top journal publications* predictor were more or less equally important in terms of predicting future excellence. It is often difficult to use citation-based indicators as decision support tools in the early career due to the required citation windows (see e.g., Bensman et al., 2010). Does this finding suggest that top journal publications might be an alternative for citation-based excellence indicators in contexts of selecting for excellence in the early career? The answer to that question depends on the mechanism behind the observed relationship between top journal publications and excellence or high citation rates. If authors with a higher ability to produce excellent research tend to publish their findings in journals with higher prestige and authors with a lower ability to produce excellent research tend to publish their research in journals with lower prestige, it might be reasonable to reward authors that publish in top journals. If, on the other hand, authors who publish in top journals are more likely to become highly cited due to unwanted Matthew effects and if the observed effects of top journal publications are a consequence of such biases (Merton, 1968), it would seem problematic to use top journal publications to incentivize for excellence. More research is needed, both concerning mechanisms and potential biases (e.g., Matthew effects) and the practical usefulness associated with such indicators as decision support tools in specific decision scenarios (Lindahl & Danell, 2016).

Finally, the data set in this study comprises one subfield in mathematics. As such the generalizability is somewhat limited. It could be the case that mathematics is a field where top journals have a more significant role in discriminating between excellent and non-excellent research than they have in other fields of research, as discussed by Dubois et al. (2014). We should also take into consideration the early career perspective. It might not be the case that the effects of publication rate, top journal publications, and top 10% publications on excellence are constant over the career. A task for future research could be to further examine these variables in other fields and in other career phases.

Author contribution

Jonas Lindahl: Conceived and designed the analysis.
Collected the data.
Contributed data or analysis tools.
Performed the analysis.
Write the paper.

Other contribution

None.

Appendix A

Citation data were retrieved from the citation indices in Web of Science (WoS). I used a semi-automated three-step approach to identify and retrieve citation data that could be added to the MSN dataset from the Science Citation Index Expanded (SCI-EXPANDED), Social Sciences Citation Index (SSCI), and Arts & Humanities Citation Index (A&HCI) between 1999 and 2010.

In the first step, publications were retrieved from SCI-EXPANDED, SSCI, and A&HCI on the basis of DOI numbers collected from the bibliographic records in the MSN dataset. The list of matching pairs was manually validated by comparing publication titles and authors for each pair.

In the second step, I identified and downloaded all publications in the SCI-EXPANDED, SSCI, and A&HCI that were not retrieved in the first step and that had a matching source name and publication year with at least one bibliographical record in the MSN dataset. I constructed a match-key consisting of Publication_Year + Volume + Issue + Start_Page + End_Page + ISSN to identify matching publication pairs in the MSN dataset and the dataset retrieved by source name and publication year from SCI-EXPANDED, SSCI, and A&HCI. The list of matching pairs was manually validated by comparing publication title and authors for each pair.

In the third step, I manually identified and retrieved the remaining publications that were not retrieved in steps one and two by conducting individual searches based on publication year, ISSN, volume, start page, authors, and title in SCI-EXPANDED, SSCI, and A&HCI.

The final dataset that was retrieved from the WoS Core Collection consisted of 2492 articles, notes, letters, and reviews (i.e., 64% of the 3904 journal articles in the MSN dataset).

References

Abramo, G., Cicero, T., & D'Angelo, C. A. (2011). Assessing the varying level of impact measurement accuracy as a function of the citation window length. *Journal of Informetrics*, *5*(4), 659–667.

Abramo, G., Cicero, T., & D'Angelo, C. A. (2014). Are the authors of highly cited articles also the most productive ones? *Journal of Informetrics*, *8*(1), 89–97.

Ahlgren, Per, Persson, Olle, & Rousseau, Ronald. (2014). An approach for efficient online identification of the top-k percent most cited documents in large sets of Web of Science documents. *Issi Newsletter*, *10*(4), 81–89.

American Mathematical Society. (2014). *Uniquely identifying mathematical authors in the mathematical reviews database*. (Accessed 12 February 2015). http://www.ams.org/publications/math-reviews/mrauthors

American Mathematical Society. (2015a). *The culture of research and scholarship in mathematics: Rates of publication*. (Accessed 5 January 2016). http://www.ams.org/profession/leaders/culture/RatesofPublicationfinal.pdf

American Mathematical Society. (2015b). *The culture of research and scholarship in Mathematics: Citation and Impact in Mathematical Publications*. (Accessed 5 January 2016). http://www.ams.org/profession/leaders/culture/PostdoctoralPositionsfinal.pdf

American Mathematical Society. (2016). *Citation database help topics*. (Accessed 15 March 2016). http://www.ams.org/mathscinet/help/citation_database_help_full.html#journallist

Azen, R., & Budescu, D. (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychological Methods*, *8*(2), 129–148.

Azen, R., & Traxel, N. (2009). Using Dominance Analysis to Determine Predictor Importance in Logistic Regression. *Journal of Educational and Behavioral Statistics*, *34*(3), 319–347.

Bensman, S., Smolinsky, L., & Pudovkin, A. (2010). Mean citation rate per article in mathematics journals: Differences from the scientific model. *Journal of the American Society for Information Science and Technology*, *61*(7), 1440–1463.

Borjas, G., & Doran, K. (2012). The collapse of the soviet union and the productivity of American Mathematicians. *Quarterly Journal of Economics*, *127*(3), 1143–1203. http://dx.doi.org/10.1093/qje/qjs015

Bornmann, & Williams. (2017). Can the journal impact factor be used as a criterion for the selection of junior researchers? A large-scale empirical study based on ResearcherID data. *Journal of Informetrics*, *11*(3), 788–799.

Bornmann, L., Mutz, R., & Daniel, H. D. (2008). Are there better indices for evaluation purposes than the h index? A comparison of nine different variants of the h index using data from biomedicine. *Journal of the American Society for Information Science and Technology*, *59*(5), 830–837. http://dx.doi.org/10.1002/asi.20806

Bornmann, L. (2013). How to analyze percentile citation impact data meaningfully in bibliometrics: The statistical analysis of distributions, percentile rank classes, and top-cited papers. *Journal of the American Society for Information Science and Technology*, *64*(3), 587–595.

Bornmann, L. (2014). How are excellent (highly cited) papers defined in bibliometrics? A quantitative analysis of the literature. *Research Evaluation*, *23*(2), 166–173.

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

Cole, S., & Cole, J. R. (1967). Scientific output and recognition: A study in the operation of the reward system in science. *American Sociological Review*, *32*(3), 377–390.

Costas, R., & Bordons, M. (2007). The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. *Journal of Informetrics*, *1*(3), 193–203.
Costas, R., & Noyons, E. (2013). *Detection of different types of "talented" researchers in the Life Sciences through bibliometric indicators: Methodological outline. CWTS Working Paper Series, (CWTS-WP-2013-006)* (Accessed 6 May 2015). http://www.cwts.nl/pdf/CWTS-WP-2013-006.pdf
CWTS. (2015). CWTS journal indicators. Accessed September 10 2015, http://www.journalindicators.com/methodology.
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.
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.
Dubois, P., Rochet, J. C., & Schlenker, J. M. (2014). Productivity and mobility in academic research: Evidence from mathematicians. *Scientometrics*, *98*(3), 1669–1701.
Franceschet, M., & Costantini, A. (2010). The effect of scholar collaboration on impact and quality of academic papers. *Journal of Informetrics*, *4*(4), 540–553.
Glänzel, W. (2002). Coauthorship patterns and trends in the sciences (1980–1998): A bibliometric study with implications for database indexing and search strategies. *Library Trends*, *50*(3), 461–473.
Havemann, F., & Larsen, B. (2015). Bibliometric indicators of young authors in astrophysics: Can later stars be predicted? *Scientometrics*, *102*(2), 1413–1434.
Judge, T., Cable, D., Colbert, A., & Rynes, S. (2007). What causes a management article to Be cited: Article, author, or journal? *The Academy of Management Journal*, *50*(3), 491–506.
Korevaar, J. (1996). Validation of bibliometric indicators in the field of mathematics. *Scientometrics*, *37*(1), 117–130.
Larivière, V., Dorta-González, P., & Costas, R. (2016). How many is too many? On the relationship between research productivity and impact. *Plos One*, *11*(9), e0162709.
Lindahl, J., & Danell, R. (2016). The information value of early career productivity in mathematics: A ROC analysis of prediction errors in bibliometricly informed decision making. *Scientometrics*, *109*(3), 2241–2262.
Mcfadden, D. (1974). The measurement of urban travel demand. *Journal of Public Economics*, *3*(4), 303–328.
Merton, R. K. (1968). The Matthew effect in science. *Science*, *159*(3810), 56–63.
National Science Board (NSB). (2016). *Science and engineering indicators 2016*. Arlington: National Science Foundation. NSB −2016-1
Peng, T., & Zhu, J. (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., Glänzel, 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.
Ramos, A., & Sarrico, C. (2016). Past performance does not guarantee future results: Lessons from the evaluation of research units in Portugal. *Research Evaluation*, *25*(1), 94–106.
Rousseau, R. (1988). Citation distribution of pure mathematics journals. In L. Egghe, & R. Rousseau (Eds.), *Informetrics* (pp. 249–262). Belgium: Diepenbeek, 87/88, Proceedings 1st International Conference on Bibliometrics and Theoretical Aspects of Information Retrieval.
Sandström, U., & van Den Besselaar, P. (2016). Quantity and/or quality? The importance of publishing many papers. *PLoS One*, *11*(11), e0166149.
Smolinsky, L., & Lercher, A. (2012). Citation rates in mathematics: A study of variation by subdiscipline. *Scientometrics*, *91*(3), 911–924.
Stern, N. (1978). Age and achievement in Mathematics: A case-study in the sociology of science. *Social Studies of Science*, *8*(1), 127–140.
Waltman, L., Van Eck, N., Van Leeuwen, T., & Visser, M. (2012). Some modifications to the SNIP journal impact indicator. *Journal of Informetrics*, http://dx.doi.org/10.1016/j.joi.2012.11.011
Waltman, L. (Eds.). (2017). Special section on performance-based research funding systems [Special section]. Journal of Informetrics, 11(3), 904–944.
Yu, T., Yu, G., Li, P. Y., & Wang, L. (2014). Citation impact prediction for scientific papers using stepwise regression analysis. *Scientometrics*, *101*(2), 1233–1252.