Journal of Informetrics 9 (2015) 826–838

Contents lists available at ScienceDirect

Journal of Informetrics

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

Early career grants, performance, and careers: A study on predictive validity of grant decisions

Peter van den Besselaar[a,*], Ulf Sandström[b]

[a] Department of Organization Science & Network Institute, VU University Amsterdam, De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands

[b] Royal Institute of Technology, INDEK, KTH, Stockholm & Orebro University, Business School, Orebro, Sweden

ARTICLE INFO

Article history:
Received 24 March 2015
Received in revised form 12 July 2015
Accepted 12 July 2015

Keywords:
Grant allocation
Predictive validity
Academic careers
Career grants
Gender differences

ABSTRACT

The main rationale behind career grants is helping top talent to develop into the next generation leading scientists. Does career grant competition result in the selection of the best young talents? In this paper we investigate whether the selected applicants are indeed performing at the expected excellent level—something that is hardly investigated in the research literature.

We investigate the predictive validity of grant decision-making, using a sample of 260 early career grant applications in three social science fields. We measure output and impact of the applicants about ten years after the application to find out whether the selected researchers perform ex post better than the non-successful ones. Overall, we find that predictive validity is low to moderate when comparing grantees with all non-successful applicants. Comparing grantees with the best performing non-successful applicants, predictive validity is absent. This implies that the common belief that peers in selection panels are good in recognizing outstanding talents is incorrect. We also investigate the effects of the grants on careers and show that recipients of the grants do have a better career than the non-granted applicants. This makes the observed lack of predictive validity even more problematic.

1. Introduction

An important question about peer and panel review is the predictive validity: does the *post*-performance of selected researchers or selected projects legitimize their selection? Do the selected researchers perform better than those that were not selected? But why would peer and panel review of grant applications successfully select the best applicants and projects? Basically, this is based on Merton's sociological theory about scientific norms (Merton, 1973), which are expected to regulate researchers' behavior. According to this functionalist theory, scientific norms such as CUDOS[1] (should) govern the science system, and social factors that interfere with these norms should be avoided through a correct organization of selection processes. Then one may expect reviewers and panel members collectively trying to select the best researchers and the best proposals. Of course, one needs to realize that the decision-making process remains uncertain (Cole, 1992) — also

* Corresponding author.
E-mail addresses: p.a.a.vanden.besselaar@vu.nl (P. van den Besselaar), ulf.sandstrom@indek.kth.se (U. Sandström).

[1] According to Merton, scientists should follow the norms of "Communism, Universalism, Disinterestedness and Organized Skepticism" (Merton, 1973 [1942]).

http://dx.doi.org/10.1016/j.joi.2015.07.011

when decision makers are governed by the mentioned scientific norms. But one would expect a reasonable level of correct decisions, and consequently a reasonable predictive validity.

On the other hand, peer review and panel decision-making are social processes. Nepotism and sexism have been shown to characterize the system (Wennerås & Wold, 1997; Sandström & Hällsten, 2008; Sandström & Wold, 2015). Social–psychological theories about decision-making are relevant in this context. A whole set of social and psychological factors do influence academic decision-making, as is supported by much experimental research (Olbrecht & Bornmann, 2010; Van Arensbergen, van den Besselaar, & van der Weijden, 2014a), as well as by anthropological studies (Lamont, 2009): self presentation (Lamont, 2009), conformity, stereotyping (token-theory; Kanter, 1977), group-think (Esser, 1998), but also self interest and group interests (Tindale, Meisenhelder, Dykema-Engblade, & Hogg, 2001), as well as political coalitions. Furthermore, it has been shown that the organization of the selection process influences the dynamics and the outcomes (Langfeldt, 2001), such as having or not having an interview (Van Arensbergen & Van den Besselaar, 2012). These different mechanisms can be expected to work simultaneously, and which dominate depends on the context. It may also differ between research fields, and between different funding instruments. For example, one would expect that when resources are scarce, and the amount of high qualified and suitable candidates is high compared to the available grants, the probability that extra-scientific criteria come in increases, and interests and power may be more important. But when resources are larger and the success ratio is higher, norm-oriented behavior may dominate.

Unfortunately, data to investigate predictive validity are scarce — it is notoriously difficult to get data about *rejected* applicants and applications. Not surprisingly, a recent review of research on peer review (Bornmann, 2011) could only identify a handful of studies about the predictive validity of grant peer review (Armstrong, Caverson, Adams, Taylor, & Olley, 1997; Bornmann & Daniel, 2006; Bornmann, Wallon, & Ledin, 2008; Bornmann, Leydesdorff, & van den Besselaar, 2010; Van den Besselaar & Leydesdorff, 2009; Hornbostel, Böhmer, Klingsporn, Neufeld, & von Ins, 2009). Recently, a few other studies have been published, indicating the growing interest in the subject (Reinhart, 2009; Campbell et al., 2010; Benda & Engels, 2011; Neufeld & von Ins, 2011; Neufeld & Hornbostel, 2012; Neufeld, Huber, & Wegner, 2013; Van Leeuwen & Moed, 2012; Cabezas-Clavijo, Robinson-García, Escabias, & Jiménez-Contreras, 2013; Mutz, Bornmann, & Daniel, 2014; Decullier, Huot, & Chapuis, 2014; Danthi, Wu, Shi, & Lauer, 2014; Kaltman et al., 2014; Saygitov, 2014). The studies cover a variety of countries (US, Canada, Netherlands, Germany, France, Spain, Austria, Switzerland, Russia), the European Union (ERC) and other international organizations (EMBO). Interestingly, half of the studies were published in 2013 and 2014, showing that the issue gets more attention recently. In Table 1, we have summarized the studies on a few dimensions, relevant to our analysis.

What do these studies show? A first observation is that all studies relate the grant decision to publication and citation indicators, and to this end many use the journal impact factor or the *h*-index. In this paper, we also limit ourselves to publications and citations, and we discuss this methodological decision in Section 3.

Second, while most studies focus on granted versus non-granted *researchers*, there are others that focus on granted and non-granted research *projects*. The latter studies are to some extent problematic, as non-granted projects are difficult to study, specifically as they are not always carried out, and when carried out difficult to identify. The studies focusing on projects [16–18 in Table 1] indeed do not compare granted projects with non-granted, but try to establish whether the panel scores correlate with post performance in terms of publications and citations. Study [15] is different as it compares ex-ante panel evaluation scores with ex-post panel evaluation scores. This study only has evaluation scores for the granted projects, and the researchers had to 'impute' scores for the non-funded projects. Here the positive conclusions about predictive validity completely depend on the assumptions in the imputation procedure. Although the study is presented as an empirical one, the contribution is in fact methodological, and therefore we discard it here. The three other studies either do not find a correlation between review scores and performance [17, 18] or a weak correlation [16], suggesting many type-2 errors (very good projects that were not funded).

Third, quite a few studies correlate the granting decision in fact with *past performance*, and not post performance [1–5]. In some cases, citations were counted until a date after the grant was received, but these citations related only to publications written before the application [e.g., 2]. These studies on the relation between past performance in terms of citations and publications do point in several directions. Sometimes one finds a positive correlation between receiving grants and performance, sometimes not. As the number of studies is low, it is difficult to see patterns. Here we formulate the assumption that at least two characteristics of these studies are important. First of all, the selection of the contrast groups is crucial. The more restrictive the group of non-granted is, the less we expect to find performance differences with the successful applicants. Second, the higher the success rate, the larger the probability of finding a different performance between the granted group and the non-grated group.

Fourth, a number of papers do relate funding to post performance. These studies differ in several aspects, especially in terms of the post performance period. Most cover only a short period, such as only the project period, or the project period plus one year [6–9], some cover a slightly longer period, such as three years after the project ended [10–12], and others have a reasonable long period up to 9 years [13, 14]. This may influence the findings considerably. With short periods, one may not have captured the full effects of the grant, but with a long period, the effects one may measure could easily be influenced by e.g., other grants obtained in the meantime. The probability of these other influences may differ depending on the career phase of the researchers involved. We expect that early career researchers have fewer opportunities for additional funding, and this group may therefore be more suitable for studies on predictive validity of grant selection procedures.

Table 1
Overview of the relevant studies.

	References	Country	Post grant period	Unit	Contrast with	Field	Success rate	Performance indicators	Field normalized	Predictive validity	Remarks
1	Bornmann (2006)	DE (BIF)	(past performance)	Researcher	All rejected	Basic biomedicine	16%	Publications, citations	No	Positive relation	
2	Van den Besselaar (2009)	NL (NWO)	(past performance) Citations until a few years after application	Researcher	All rejected; best rejected	Social sciences	12%	Publications, citations	Analysis by discipline	All: positive relation; best rejected: no effect	
3	Neufeld (2011a)	EU (ERC)	(past performance)	Researcher	All rejected	All fields	12%	Publications, citations	Yes	Very small effect	Life sciences: impact factor
4	Neufeld (2011b)	DE (DFG)	(past performance)	Researcher	All rejected; best rejected	Medicine; Physics; Chemistry; Biology	41%; 59%; 53%; 57%	Publications, citations	Discipline normalized citations	All: positive relation; best rejected: no effect in chem and in bio	
5	Cabezas-Clavijo (2013)	Spain (ANEP)	(past performance)	Researcher	All rejected	All fields	63%	Publications, citations, journal impact	Analysis by discipline	Granted score slightly higher; differences by discipline	Publication in high impact journals
6	Campbell (2010)	Canada	1 year after start grant until 1 year after grant end	Researcher	All rejected; all researchers in field	Oncology	No info available	Average of relative citations	Yes	All rejected: positive relation; all researchers: positive relation	
7	Hornbostel (2009)	DE (DFG)	Maximum of 4 years after application	Researcher	All rejected	Medicine; Physics	41%; 59%	Publications, citations	By discipline	Medicine: no relation; Physics: positive relation	Career is predicted much better
8	Van Leeuwen (2012)	NL (NWO)	Between 1 and 4 years after application	Researcher	All rejected; all researchers in field	Math & Astronomy; Chemistry; Earth & Life sciences	32%; 33%;32%	Publications, citations based indicators	Yes	Positive relation	Nr of applications by researcher has effect
9	Reinhard (2009)	Switzerland (SNSF)	Publications until 4 years after application; cit 5 year	Researcher	All rejected	Chemistry; Medicine	55%	Publications, citations	No	Positive relation	
10	Saygitiv (2014)	Russia	Publications until 5 years after application	Researcher	All rejected	Medicine	24%	Publications, citations	NO	No relation	Experience researcher has effect

11	Bornmann (2008)	International (EMBO)	Publ 4–8 years after application	Researcher	All rejected	Life sciences	15%	Publications, citations	Yes	Positive relation	About 40% type 2 errors
12	Bornmann (2010)	Intern (EMBO) & NL (NWO)	Publ 3–5 years after application	Researcher	All rejected; best rejected	Life sciences; Economics, Psychology	14%; 19%; 30%	Publications, citations, H-index	Analysis by discipline	All: positive relation; best rejected: no effect	Many type 2 errors
13	Van den Besselaar (2013)	NL (NWO)	Publications until 9 years after applications	Researcher	All rejected; best rejected	Economics; Behavioral sciences; Psychology	18%; 13%; 30%	Publications, citations, cit/publ	No	All: between strong and zero correlation; best rejected: between very negative and zero correlation	Depends on field: the lower the successrate, the lower predicted validity
14	Deculier (2014)	France (PHCR)	Publications until 9 years after applications	Researcher	All rejected	Clinical research	21%	Publications, citations	No information	No effect on output	Funding: more likely that project started
15	Mutz (2014)	Austria (FWF)	Ex post peer review after completion of project	Project	Versus rejected (no real data)	all	20%	Peer review, imputed data	No	Positive effect of funding	
16	Gallo et al. (2014)	US (AIBS)	Publ 1–8 years after application	Project	No contrast group: review score versus performance	Biology	11%	Publications, citations	Publication year normalized	Positive correlation	Many type 2 errors
17	Danthi (2014)	US (NIH)	Publ 3–9 years after application	Project	No contrast group: review score versus performance	Circulation research	No info	Variety of citation based indicators	Yes	No correlation	
18	Kaltman (2014)	US (NIH)	Publ 3–9 years after application	Project	No contrast group: review score versus performance	Circulation research	No info	Variety of citation based indicators	No	No correlation	Past performance has stronger effect

The findings are in the majority of studies positive: researchers with a grant do have on average a higher post performance. Despite this positive relation, not all selected applicants score higher than all rejected ones. A significant number of false positives and false negatives is reported, implying that – in the reviewed studies – between one third and two third of all decisions can be considered wrong if one accepts the deployed performance criteria as meaningful. False negatives (type-2 errors) may be the more important, as this refers to researchers with high post performance that were not funded. If there are high percentages of type-2 errors, large numbers of important ideas may be missed by the panel review (Bornmann et al., 2010; Gallo et al., 2014). Above that, those studies that use more restricted contrast groups (the best performing not-granted applicants) do not find an effect of grants on post-performance, or sometimes even a negative relation [e.g. 13].

This brings us to the issue how predicted validity should be measured. Most studies use datasets that include the long tail of low performing rejected applicants. These studies compare successful applicants with *all* unsuccessful applicants. In this way, it is rather trivial to show that the successful applicants *on average* outperform the non-successful, in terms of past performance (before the grant application) as well as in terms of post-performance (a few years after the application). The long tail of low performing rejected applicants negatively influences the mean and median of the rejected group. This is typically the case, as it is well known that the performance distribution is very skewed. If, as is the intention of the career grants under study, the successful applicants should be in the top 10% of the researchers in their field, the panel is expected to select applicants that are in the (vertical) top of the distribution. Performance data will be subjected to random variation, and the presumed best (and selected) applicants are not necessary those with the highest scores: more is not always better (Waltman, van Eck, & Wouters, 2013). But given the shape of the distribution, even with random error the scores of the best should be in the higher part of the distribution. Therefore the *group* of successful applicants is expected to score on average higher than the *group* of best performing non-successful applicants. Consequently, in terms of contrast groups we suggest that one should compare the successful applicants with a same size group of best performing non-successful applicants. If the selection process works well, the successful should on average outperform this set of good performing rejected applicants.

As suggested above, differences between success rates play a role (Gallo et al., 2014). If success rates are low, the cut-off threshold between grant and no-grant is in the steep part of the performance distribution, and the best non-successful applicants are also high performers. In contrast, with high success rates, even the best performing non-successful applicants are more toward the tail of the performance distribution, making the probability of false negatives lower. Indeed, the available evidence suggests when success rates are higher, predictive validity is also higher: the successful applicants then also outperform the equally large group of non-successful applicants, as a large part (Van den Besselaar, 2013) or the whole (Reinhart, 2009) low performing tail is included in the 'best performing' contrast group.

2. Research questions

In this paper we go beyond earlier studies by combining the following characteristics: (i) Mid-career researchers and the advanced researchers do have a variety of funding possibilities. It is hardly possible to control for that, as information lacks about grants researchers may have obtained elsewhere. This is not the case for early career researchers, who have more or less a single funding source available: early career grants. By restricting the analysis to the early career program (here: the Dutch VENI program), we as good as possible have disjoint sets of successful and unsuccessful applicants, not influenced by other grants. However, as far as other grants have been acquired, we assume (Merton: the Matthew effect) that the successful applicants have profited more from that than the non-successful. (ii) Many studies only cover a rather short period of *post-performance*. That holds also for our own previous studies (Bornmann et al., 2010): the projects ran in the periods 2004–2007, 2005–2008 and 2006–2009. As the time between having results and having those published is fairly long in the social sciences, and citing that work takes some more time, our earlier mid-2009 output and citation data hardly measure real post performance. Now we are six years later, and therefore we now may measure post performance more reliable. (iii) Comparisons are generally made in terms of average numbers of publications and citations (Bornmann, 2011). As distributions are skewed, we use here also non-parametric statistics, next to general linearized models for multivariate analysis. (iv) We compare the successful applicants with all non-successful applicants and with the set of best performing non-successful applicants, a method introduced earlier (Van den Besselaar & Leydesdorff, 2007, 2009) and explained above. (v) We use advanced *size-independent* field-normalized indicators to measure performance, as this is up to now the standard in bibliometrics. However, we also deploy a few *size-dependent* citation indicators, as there are good reasons to prefer those above the size-independent (Sandström & Wold, 2015). (vi) If the best applicants are selected by the panels under study, one would expect that the successful applicants also have a faster academic career. This is in fact one of the goals of the early career grant schemes. We test that too in this paper.

We will therefore answer two research questions in this paper: (i) Do the selected applicants show a higher scholarly performance than the non-selected, some nine years after the application: were the best researchers selected? And (ii) do the selected applicants have achieved higher positions after nine years: did the grant help in achieving higher positions?

3. Data and method

Our sample consists of some 400 researchers that have submitted proposals to an early career grant program of a social science council in the Netherlands between 2003 and 2005. From the total set of young researchers, we selected here the young researchers in psychology, in behavioral & educational research, and in economics. This selection was made because for these fields the Web of Science is covering academic output relatively well, enabling us to use a bibliometric approach to performance measurement. *Relatively well* does not mean that all output is covered; here it means that the WoS indexed journals are considered within relevant communities as the most important publication venues. We discuss this choice more in detail below.

This selection resulted in a set of 262 researchers, of which 19 are deleted because of missing data—we could not trace them anymore. This leaves us 104 early career economics researchers (73% male), 48 behavior and education researchers (38% male), and 91 psychology researchers (44% male). Almost half, 45%, of the researchers in the sample are female. The researchers are in our sample as they applied for an early career grant in 2003 (24%), 2004 (38%), or 2005 (38%).

The application data consist of:

- Name
- Year of application (as proxy for academic age)
- University
- Field and discipline
- Reviewer scores (between two and five per application)
- Panel decision

For career information we searched for all the 262 researchers on the web, and for almost all (243) we could find a homepage and/or a full CV. This provides us with the:

- The *current position* (early 2014) is measured on an ordinal scale with values from 10 (teacher) to 16 (full professor); in fact this are the codes used for the positions in universities' job structure. About 20.6% of the cases was early 2014 full professor (score = 16), 28.3% was associate professor (score = 14), 27.5% assistant professor (score = 12), 2.1% senior researcher (score = 11)[2], 14.2% researcher (score = 11), and 3% of the applicants was in teaching positions (score = 10). The remaining 4.3% has a career outside academia. The small number of applicants that went to positions outside academia is not included in the analysis, as it is difficult to integrate their positions into the academic rank system. Furthermore, those that left academia stopped publishing, so we do not have scores on the independent variables.
- The level of *mobility*: 47% of the sample showed no mobility, 32% showed national mobility, and 21% showed international mobility.
- An overview of the applicants' *publications* — which was used to validate the performance data as downloaded from the WoS.

In order to make an adequate bibliometric dataset, we retrieved publication data from the Web of Science (*SCI-expanded, Social Science Citation Index; Arts and Humanities Citation Index*) using the following query:

> AU = last_name first–initial* AND (CU = Netherlands OR CU = *country name*) AND DT = (article OR letter OR note OR proceedings paper OR review) and PY = 2001–2012

'Country name' in the query refers to countries where the researcher has worked according to his or her CV. The data were manually cleaned, by comparing the found WoS records with the publication lists found on the Web. In this way we could delete papers that were authored by others with the same name. In cases where we missed papers from the publication list, we searched for the missing titles in WoS, and added it to the set. Generally, missing papers was due to the fact that authors used different initials. As well known, disambiguation and entity resolution are time consuming. But it creates a reliable data set, which is needed given sample size involved. With the BMX tool (Sandström & Sandström, 2009) the following field normalized citation indicators were calculated:

- *P*: number of publications, full counting
- Frac *P*: number of publications, fractional counting based on author shares
- NCSf: field normalized citation score, until 2014 (size-independent)
- NCSf2y: field normalized citation score, 2 years window (size-independent)
- $\sum$FracNCSf: sum of field normalized citation score, in the *core journal set* (size-dependent)
- TOP *x*%: share of publications in the set of (1%, 5%, 10%, 25% and 50%) highest cited publications, field normalized (size-independent)
- $\sum$TOP10%: sum of TOP 10% publications, field normalized in the CWTS *core journal set*[3] (size-dependent).

In order to measure predictive validity, we will compare the performance of successful applicants (S), best performing non-successful applicants (BPNS), and all rejected applicants (NS), in terms of the different above-mentioned indicators. Comparison is done using the SPSS22 procedure *non-parametric tests* and *Anova*. Best performing non-successful applicants are defined in terms the size-dependent $\sum$TOP10%. The distribution is in Fig. 1.

[2] As the number of senior researchers is very low, we take those together with the researchers category.

[3] For the core journal set see Waltman and Van Eck (2013).

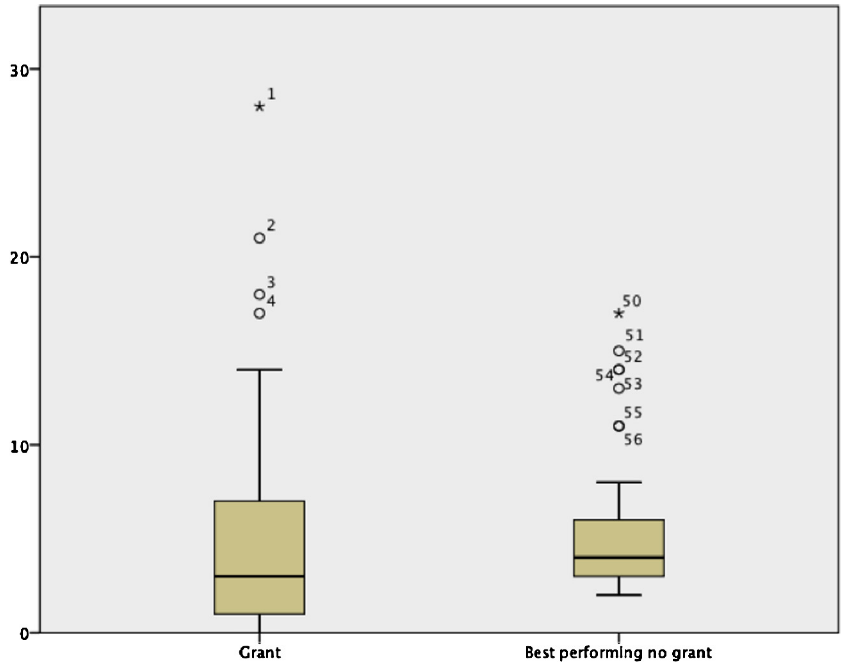

Fig. 1. Distribution of number top 10% cited papers for two groups of applicants (N = 49 for each of the two groups).

Generalized Linear Models was used for investigating the effect of grants on careers. As the dependent variable is ordinal, we use the ordinal response (multinomial) model with a cumulative logit link function. Several relevant covariates are included in the analysis (see below).

We are aware (i) that performance may include more publications than WoS-indeed articles, (ii) that performance includes also non-publication outcomes, such as economic and societal impact, and (iii) that the decision is also based on the proposal itself and not only on past performance of the applicant. Basically, there is a practical reason to use WoS-data and most researchers in this field do, also for the social science fields we analyze here (Bornmann et al., 2010). However, we also have positive reasons to use the indicators selected. (i) According to interviews with panel members and reviewers, the publications and received impact is very important – if not dominant – in the selection process of early career researchers, especially in the first rounds when most applicants are rejected (Van Arensbergen, van der Weijden, & van den Besselaar, 2014b). (ii) In the fields under study – for sure in the Netherlands – journal articles are considered as the main publication channels, much more important than other types of output. (iii) The selection procedure under study took place in 2003–2005, when e.g., societal impact was not taken into account yet. It is taken into account in the more recent period, also for the career grants we study. The effect of the more recent panel scores for societal relevance on the outcome of the selection process has been studied elsewhere, and this is almost zero (Van den Besselaar, 2011). (iv) As we study post performance, the quality of the proposal is taken into account in the performance data—under the assumption that better projects (at least in the time frame we deploy here) result in more publications and citations. (v) We do not discuss indicators for the evaluation of individuals, but compare groups—and as explained above the group of grantees is expected to score (substantially) better than the group of non-granted applicants.

4. Findings

4.1. Predictive validity

Table 2 shows the mean scores on the various indicators for the three groups. The selected applicants have a much higher output in terms of papers than the group of all non-selected ones—and the differences are statistically significant. This holds for the full count and for the fractional count. The successful applicants also have a higher score on the various citation-based indicators than the non-successful ones—the differences are smaller and only a few of the differences are statistical significant. As performance variables are generally skewed, we also use non-parametric tests. Comparing the granted applicants with all non-granted creates a clear picture: the former group has a statistically significant higher performance than the latter in almost all performance indicators used.

The picture changes when restricting the comparison to the *best performing non-successful* applicants who are (size-dependent) defined in terms of the sum of the (field normalized) top 10% cited papers in the core journal set. The granted researchers are slightly more productive and have a slightly higher mean and a slightly higher median for papers and

Table 2
Success by post performance.

		49 successful applicants versus all non-successful applicants			49 successful applicants versus 49 best performing non-successful applicants		
		Mean	Median	Distribution mean rank	Mean	Median	Distribution mean rank
Papers (integer)	Success	26.4*	20*	161.44*	26.4	20	51.09
	No success	11.9	8	105.17	22	19	47.91
Papers (fractional)	Success	7.7*	7.3*	161.62*	7.7**	7.3	54.1
	No success	4	3.5	105.12	5.9	5.6	44.9
NCSf core	Success	1.4**	1.1*	136.16**	1.4	1.1	38.51
	No success	1.1	0.96	111.9	1.9**	1.7*	60.49*
NCSf3y	Success	1.4	1.2	134.31**	1.4	1.2	40.98
	No success	1.2	1.1	112.39	1.9**	1.7**	58.02*
$\sum$FracNCSf#	Success	12.5*	8.9*	159.47*	12.5	8.9	47.29
	No success	4.9	3.1	105.69	10.7	9.8	51.71
TOP 1%	Success	0.02	0	128.62***	0.02	0	42.58
	No success	0.01	0	113.9	0.04	0.02*	56.42*
TOP 5%	Success	0.08	0.04*	134.03**	0.08	0.04	36.42
	No success	0.06	0	112.46	0.18*	0.15*	62.58*
TOP 10%	Success	0.17	0.13**	137.63**	0.17	0.13	38.78
	No success	0.13	0.06	111.51	0.28*	0.27*	60.22*
$\sum$TOP10% core#	Success	4.7*	3.0*	156.33*	4.7	3	43.06
	No success	1.7	1	106.53	5.6	4	55.94**
TOP 25%	Success	0.38***	0.38***	132.27***	0.38	0.38	41.35
	No success	0.32	0.28	112.28	0.48*	0.49*	57.65*
TOP 50%	Success	0.67*	0.68	135.28**	0.67	0.68	44.53
	No success	0.58	0.59	112.13	0.73***	0.73	54.47***

* Significance <0.01.
** Significance <0.05.
*** Significance <0.10.
\# Size-dependent.

fractionalized papers — but most differences are not significant. But for the size-dependent and size-independent citation indicators, the non-successful applicants perform better, and have a higher mean, median and mean rank distribution than the granted applicants — and most of these differences are statistically significant. These findings imply that the grant decisions have no predictive validity: the granted researchers overall do not show a better performance than the best performing non-granted, but on average a lower one.

If the *panels* were not very good in selecting the future best performing researchers, how did the *reviewers* do? In the selection procedure under study, panels get review reports with scores from external reviewers. The panels are expected but not obliged to use those scores. We correlate the average review score with the performance scores of the applicants. This is done for all 233 applicants, and for the restricted set of 49 successful plus 49 best performing non-successful. Table 3 shows the results. Obviously the peer review score correlates low (citation based indicators) to moderate (publications) with post performance, when taking all applicants into consideration. Focusing only at the granted and best non-granted, the correlations are low: positive with publications, and negative with all citation-based indicators. Obviously, the reviewers have similar problems as the panel to recognize the researchers that will have the higher impact.

4.2. Impact of the grants on careers

As shown, the grant selection procedure did not predict future performance. Whether the lack of predictive validity is a problem depends on the effects of such important career grant systems. Grants are resources, and should help researchers to improve performance—which we did not find. However, increasingly prestigious grants are not only input but also output. Receiving a grant is seen as a *proof* of quality and an outcome as such. Grants, and especially prestigious career grants, have high symbolic value and may help grantees with their academic career (Hornbostel et al., 2009). We test whether that is the case. Past positions were generally postdoc. Current positions are between full professor and teaching staff. We distinguish six levels: full professor, associate professor, assistant professor, senior researcher, researcher, and teacher.

Table 3
Post performance by peer review scores.

Correlation with review score	Applicants All[#]	S & BPNS[##]
Publications full count	0.284**	0.123
Publications fractional count	0.283**	0.173
Citations field normalized—Core	0.136*	−0.133
Citations, three years window, field normalized	0.097	−0.162
Sum citations, field normalized, size-dependent	0.274**	−0.105
Share in top 1% cited papers	0.057	−0.102
Share in top 5% cited papers	0.153**	−0.151
Share in top 10% cited papers	0.165**	−0.112
Total papers in top 10% cited papers, core set, size-dependent	0.292**	−0.069
Share in top 25% cited papers	0.152*	−0.006
Share in top 50% cited papers	0.204**	−0.024

* Sign <0.05.
** Sign <0.01.
# All 233 applicants.
49 successful (S) and 49 best performing non-successful (BPNS).

Fig. 2 shows that the granted applicants have a much better academic career, as they more often reached full professor positions (37%) than the non-granted researchers (15%) in the period under consideration. The same holds for the associate professor rank (36% versus 26%). The granted applicants hardly stay in the lower academic ranks, whereas this often happens with the non-successful applicants. In fact, the granted applicants score on average a full rank higher (almost associate professor) than the non-granted (almost assistant professor). The pattern does not change when we look only at the best performing non-successful applicants. So in this sense, the career grants do what they are expected to do.

Elsewhere we showed that in our sample of early career researchers, male and female applicants perform on average about equally (Van Arensbergen & van den Besselaar, 2012; Van den Besselaar and Sandström, under review). Men were slightly more successful (22.4%) than female researchers (17.3%) in the grant applications, but the difference is not statistically significant. However, the impact of grants is different for men and women as Fig. 3 illustrates. This is especially the case when not having obtained a career grant. With and without the grant, men did get more often a full professor position. And although male and female researchers in our sample with a grant are equally often associate professor, women without a grant are much less often associate professor than men without a grant, but they remain much more frequent researcher.

4.3. Explaining careers

Section 4.2 suggests that career success correlates with success in getting prestigious grants and with gender. To go beyond the bivariate analysis, we test a multivariate model explaining career success by (i) performance; two indicators are used in the analysis — field normalized citations and number of publications; (ii) obtaining prestigious career grants, as

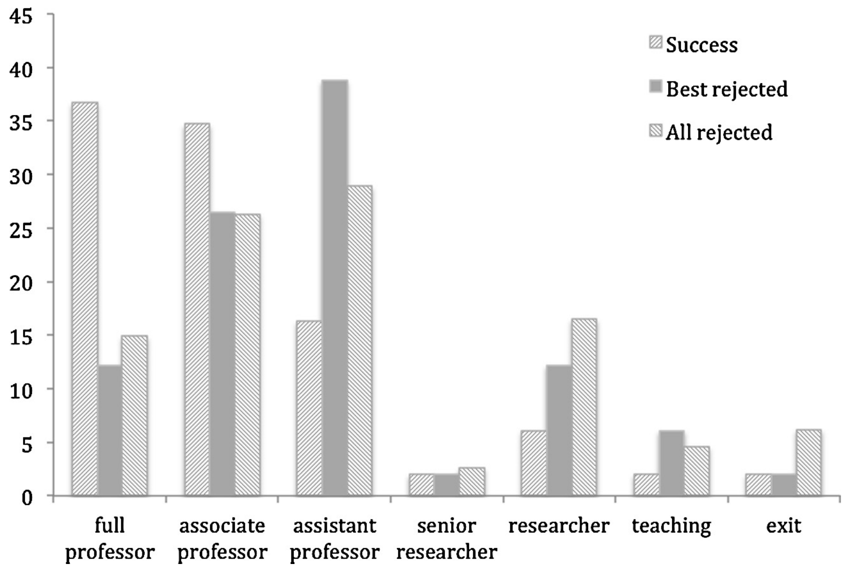

Fig. 2. Career by grant-success (percentages).

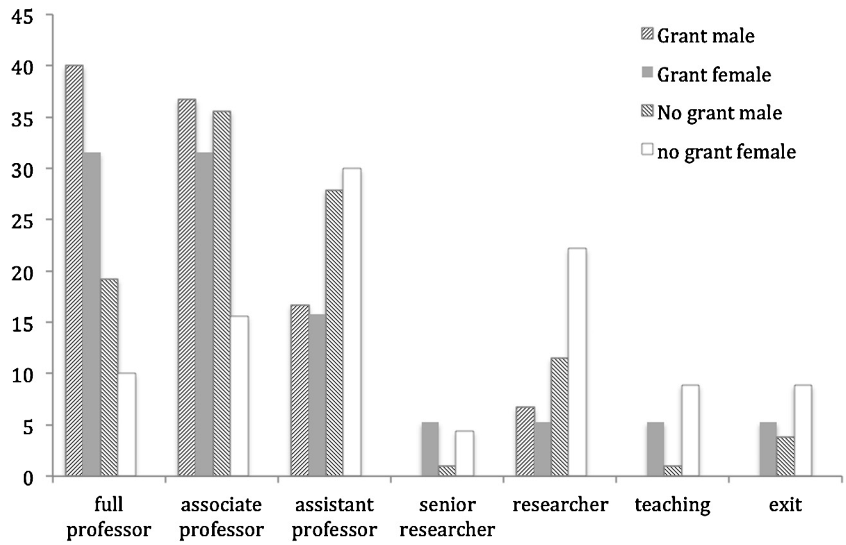

Fig. 3. Career by grant-success and gender (percentages)

they have increasingly a symbolic value and may therefore effect selection decisions; (iii) gender, as gender bias is expected to play a role in hiring processes; (iv) the year of applying for the career grant, as the earlier a career starts the more time there has been to get promotion; (v) mobility, as this is often expected to have a positive effect on careers; and finally (vi) discipline, as the labor market may be different between disciplines, as may hiring habits.

We restrict the data set to those that remained within higher education and research, as leaving the system for a position elsewhere is a different dimension than the career levels within the system. Above that, the independent variables cannot be meaningfully compared for the group of 'leavers'. The resulting sample size is 233. The selected GLM model gives a very good fit of the model: the test gives a deviance of 550.592 with 874 degrees of freedom ($p = 0.999$; non-significant means a good fit). The descriptive statistics are in Table 4.

Table 5 shows the result of the generalized linear model estimates. To start with the contextual factors: Indeed, the three context variables behave as expected, but the effects are not all statistically significant. With respect to the year of application: the two older group (2003 and 2004) groups have achieved a higher average position. This is easy to understand, as they have had more time than the younger group. Mobility seems to have some effect on the career, but the *B*-parameter is rather small, and not statistically significant (we consider $p < 0.10$ as significant). But, the parameters are as expected: national mobility scores better than no mobility and international mobility scores slightly better than national mobility.

Table 4
Descriptive statistics, variables included in the GLM analysis.

variable	Value	*N*	Percent	
Rank	Teacher	7	3.20	
	Researcher	38	17.10	
	Assistant professor	63	28.40	
	Associate professor	66	29.70	
	Full professor	48	21.60	
Application year	2003	53	23.90	
	2004	85	38.30	
	2005	84	37.80	
Gender	Male	124	55.90	
	Female	98	44.10	
Success	Grant	49	22.10	
	No grant	173	77.90	
Discipline	Economics	94	42.30	
	Behavior	42	18.90	
	Psychology	86	38.70	
Mobility	No mobility	108	48.60	
	National mobility	68	30.60	
	International mobility	46	20.70	
Variable	Minimum	Maximum	Mean	Std. deviation
Full P	1	99	15.5	15.3
NCSf2y	0	8.8	1.2	0.86

Table 5
Career success by gender, grant success, performance, mobility, year and discipline[*]

Parameter		B	Std. error	95% Wald confidence interval		Hypothesis test			Exp(B)
				Lower	Upper	Wald Chi-2	df	Sig.	
Threshold position:	10	−1.426	0.5523	−2.508	−0.343	6.665	1	0.01	0.24
	11	0.969	0.4573	0.072	1.865	4.485	1	0.034	2.634
	12	2.726	0.4855	1.775	3.678	31.536	1	0	15.277
	14	4.533	0.5382	3.479	5.588	70.953	1	0	93.083
Male versus female		0.793	0.2785	0.247	1.339	8.106	1	0.004	2.21
Grant versus no grant		0.643	0.3412	−0.026	1.312	3.552	1	0.059	1.902
Publications (integer)		0.071	0.0129	0.046	0.096	30.67	1	0	1.074
Citations (NCSf3y)		0.273	0.1633	−0.047	0.593	2.792	1	0.095	1.314
Mobility	None vs. international	−0.148	0.3514	−0.837	0.541	0.177	1	0.674	0.863
	National vs. international	−0.057	0.374	−0.79	0.676	0.024	1	0.878	0.944
Application year	2003 vs. 2005	0.404	0.3339	−0.251	1.058	1.462	1	0.227	1.497
	2004 vs. 2005	0.498	0.2949	−0.08	1.076	2.855	1	0.091	1.646
Discipline	Economics vs. psychology	1.704	0.3361	1.046	2.363	25.72	1	0	5.498
	Education vs. psychology	−0.545	0.3701	−1.27	0.181	2.166	1	0.141	0.58
(Scale)		1[b]							

Model: probability distribution: multinomial; link function: cumulative logit; goodness of fit (Chi^2): deviance = 550.592; df = 874, dependent variable: position early 2014.

[b] Fixed at the displayed value,

[*] Only those applicants that remained in the higher education and research system: N = 233.

Third, discipline has a significant effect, and the career goes faster in the economics field and slowest in the field of behavior and education.

The analysis suggests that gender (p = 0.010) and grants (p = 0.059) both have a strong effect on obtained position, in the way that women and non-successful applicants score much lower on career. The effect of gender is slightly stronger than of receiving a grant. Performance has a positive effect on career too, citations (p = 0.091) as well as publications (p = 0.000).

5. Conclusions and discussion

In this paper we aim to answer two questions. The first question is whether the selection process under study is characterized by predictive validity; or in other words, do the selected researchers on average outperform the best performing non-selected after a period of almost ten years?

The results of an earlier analysis (Van den Besselaar, 2013) did suggest that predictive validity was absent. However, that conclusion was based on citation-based indicators that were not field-normalized. Here we repeated the analysis using a series of size-independent and size-dependent and field-normalized indicators. Only in terms of output, grantees score slightly higher, but that may indeed be the effect of more resources (the grant) and not of higher quality. Comparing the granted researchers with the set of best performing non-successful applicants, the grantees have on average less impact after about a ten years period. These results suggest that the panel decisions indeed have no predictive validity as far as scholarly performance is concerned. We also tested whether the peer review alone would have been a better predictor, but that is not the case. The review scores hardly correlate with the various post performance variables. Despite the fact that panel members are convinced that real talents are easily recognized (Van Arensbergen et al., 2014b), our current results suggest the opposite. This study adds to the growing evidence that predictive validity of panel and peer review of grant applications is difficult to establish.

Second, we wanted to test whether the career grants do support the career of the grantees, as intended. The results indicate that receiving a prestigious early career grant indeed seems to have a strong influence on careers, as grantees have a two to three times bigger chance to become full professor, and a much lower chance to remain in one of the lower academic ranks. On average, the granted applicants reached a full rank higher than the others within the ten years period under study. In this sense, the funding instrument works as intended. Using a multivariate model, we could show that apart from obtaining grants, the achieved career level is influenced by gender. Finally, academic performance – publication output as well as citation impact – has a positive impact on the career. Overall, this suggests that non-academic factors such as the symbolic value of grants, and the gender of the candidate do impact academic careers in the early phase.

Our study has limitations. (i) We assumed – but could not measure – that the early career researchers in our sample only had the VENI grant or no grant. This may be not completely true. However, the Matthew effect would predict that the VENI-grantees would have been more successful in obtaining other grants than the non-grantees. In that case, this would in fact strengthen our conclusions. (ii) One may dispute how the comparison between the successful and best performing non-successful applicants should be done, and how to interpret the differences. One possible argument against our method would be that we may have in the successful group also those with – due to random variation – low scores leading to an underestimation of their quality, but in the non-successful group only those with – again partly due to random variation – the highest scores that overestimate their quality[4]. However, the results presented in Table 2 suggest that this problem may be less serious than expected. The best performing non-successful applicants not only score higher in terms of the indicator that was used to define the group (the number of top 10% highly cited papers) but also in terms of almost all other performance indicators. As correlation between these indicators is moderate to high (but not very high), this in our view cannot be explained by random variation. Nevertheless, it would be useful to experiment with different definitions of the contrast groups, and we may do this in a follow-up paper. (iii) We only use scholarly output and impact to compare grantees and non-grantees, and in a restricted form: only WoS indexed publications. Other types of output are not irrelevant, but in the fields under consideration, the performance data we used are dominant. (iv) Other (non-scientific) criteria play also a role in career decisions, as argued elsewhere (Van Arensbergen et al., 2014b). However, if these criteria would make the gender effect spurious, they should be distributed uneven between male and female researchers, to an extent we consider unlikely (Van den Brink & Benschop, 2012). Also here additional research is useful.

Related to the last point, it is crucial to expand the repertoire of bibliometric indicators to cover also other dimensions that play a role in selection processes. Up to now the bibliometric community has focused far too much on many variants of output and impact indicators only. A good candidate for new indicators would be scholarly independence, which is often mentioned as a core quality dimension in academic job and grant decisions (Van Arensbergen et al., 2014b). Elsewhere we showed how bibliometric data might be used for developing an *independence indicator* (Van den Besselaar et al., 2012; Van den Besselaar and Sandström, under review).

Overall, our results suggest that the current selection processes are not optimal. This study is only one case. But as the literature review suggest, also in other cases covering various countries and fields, similar results are found. The importance of the topic asks for more research. It would be good if councils would be required to have studies like this one done repeatedly and on a large scale. National science policy makers should in our view not so much interfere with the operations of councils, as that would hinder progress of science even more. But councils should be required to go for serious and first of all *independent* evaluation of their procedures. If one believes that peer and panel selection are crucial and core elements of the research system, one should at least evaluate those with large-scale studies.

Acknowledgments

The authors have benefitted from the comments of JoI editor Ludo Waltman (CWTS), Professor Thorsten Nybom (Orebro University), and of the comments of two anonymous reviewers. S. Charlie Mom contributed substantially to the data collection. We acknowledge the support of the European Research Council ERC (the *GendERC project*, grant 610706) and the Bank of Sweden Tercentenary Foundation (the *Searching for excellence and funding effects* project, grant P12-1302:1).

References

Armstrong, P. W., Caverson, M. M., Adams, L., Taylor, M., & Olley, P. M. (1997). Evaluation of the Heart and Stroke Foundation of Canada Research Scholarship Program: Research productivity and impact. *Canadian Journal of Cardiology*, *13*(May), 507–516.

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. (2011). Scientific peer review. *Annual Review of Information Science and Technology*, *45*, 199–245.

Bornmann, L., Wallon, G., & Ledin, A. (2008). Does the committee peer review select the best applicants for funding? An investigation of the selection process for two European molecular biology organization programmes. *PLoS ONE*, *3*(10), e3480.

Bornmann, L., Leydesdorff, L., & van den Besselaar, P. (2010). A Meta-evaluation of scientific research proposals: Different ways of comparing rejected to awarded applications. *Journal of Informetrics*, *4*(3), 211–220.

Benda, W. G. G., & Engels, T. C. E. (2011). The predictive validity of peer review: A selective review of the judgmental forecasting qualities of peers, and implications for innovation in science. *International Journal of Forecasting*, *27*(1), 166–182.

Cabezas-Clavijo, A., Robinson-García, N., Escabias, M., & Jiménez-Contreras, E. (2013). Reviewers' ratings and bibliometric indicators: Hand in hand when assessing over research proposals? *PLoS ONE*, *8*(6), e68258.

Campbell, D., Picard-Aitken, M., Cote, G., Caruso, J., Valentim, R., Edmonds, S., et al. (2010). Bibliometrics as a performance measurement tool for research evaluation: The case of research funded by the National Cancer Institute of Canada. *American Journal of Evaluation*, *31*, 66–83.

Cole, S. (1992). *Making science: Between nature and society*. Cambridge, MA: Harvard Univ. Press.

Danthi, N. S., Wu, C. O., Shi, P., & Lauer, M. (2014). Percentile ranking and citation impact of a large cohort of NIHLBI-funded cardiovascular R01grants. *Circulation Research*, *114*, 600–606.

Decullier, E., Huot, L., & Chapuis, F. R. (2014). Fate of protocols submitted to a French national funding scheme: A cohort study. *PLoS ONE*, *9*(6), e99561. http://dx.doi.org/10.1371/journal.pone.0099561

Esser, J. K. (1998). Alive and well after 25 years: A review of groupthink research. *Organizational Behavior and Human Decision Processes*, *73*(2–3), 116–141.

Gallo, S. A., Carpenter, A. S., Irwin, D., McPartland, C. D., Travis, J., Reynders, S., et al. (2014). The validation of peer review through research impact measures and the implications for funding strategies. *PLoS ONE*, *9*(9), e106474. http://dx.doi.org/10.1371/journal.pone.0106474

Hornbostel, S., Böhmer, S., Klingsporn, B., Neufeld, J., & von Ins, M. (2009). Funding of young scientist and scientific excellence. *Scientometrics*, *79*, 171–190.

Kanter, Rosabeth Moss. (1977). Some effects of proportions on group life: Skewed sex ratios and responses to token women. *American Journal of Sociology*, *82*(5), 965–990.

Kaltman, J. R., Evans, F. J., Danthi, N. S., Wu, C. O., DiMichele, D. M., & Lauer, M. S. (2014). Prior publication productivity, grant percentile ranking and topic-normalized citation impact of NIHLBI-funded cardiovascular R01grants. *Circulation Research*, *115*, 617–624.

[4] See also the arguments at the end of Section 1, where we discuss this issue.

Lamont, M. (2009). *How professors think: Inside the curious world of academic judgment*. Cambridge: Harvard University Press.
Langfeldt, L. (2001). Decision making constraints and processes of grant peer review, and their effect on review outcome. *Social Studies of Science*, *31*, 820–841.
Merton, R. K. (1973). The normative structure of science. In Robert K. Merton (Ed.), *The sociology of science: Theoretical and empirical investigations (Ed. Norman Storer)*. Chicago, IL: University of Chicago Press.
Mutz, R., Bornmann, L., & Daniel, H.-D. (2014). Testing for fairness and predictive validity of research funding decisions: A multi-level multiple imputation for missing data approach using ex-ante and ex-post peer evaluation data from the Austrian Science Fund. *Journal of the Association for Information Science and Technology*, http://dx.doi.org/10.1002/asi.23315 (published online)
Neufeld, J., & Hornbostel, S. (2012). Funding programmes for young scientists—Do the 'best' apply? *Research Evaluation*, *21*, 270–279.
Neufeld, J., & von Ins, M. (2011). Informed peer review and uninformed bibliometrics? *Research Evaluation*, *20*, 31–46.
Neufeld, J., Huber, N., & Wegner, A. (2013). Peer review-based selection decisions in individual research funding, applicants publication strategies and performance: The case of the ERC Starting Grants. *Research Evaluation*, *22*(3), 1–11.
Olbrecht, M., & Bornmann, L. (2010). Panel peer review of grant applications: What do we know from research in social psychology on judgment and decision-making in groups? *Research Evaluation*, *19*(4), 293–304.
Reinhart, M. (2009). Peer review of grant applications in biology and medicine. Reliability, fairness, and validity. *Scientometrics*, *81*(3), 789–809.
Sandström, U., & Hällsten, M. (2008). Persistent nepotism in peer review. *Scientometrics*, *74*(2), 175–189.
Sandström, U., & Sandström, E. (2009). The field factor: Towards a metric for academic institutions. *Research Evaluation*, *18*(3), 243–250. http://dx.doi.org/10.3152/095820209X466892
Sandström, U., & Wold, A. (2015). Centres of excellence: Reward for gender or top-level research", thinking ahead: Research, funding and the future. In *RJ yearbook 2015/16*. Stockholm: Riksbankens Jubileumsfond and Makadam.
Saygitov, R. T. (2014). The impact of funding through the FR president's grants for young scientists on research productivity: A quasi-experimental study and a brief systematic review. *PLoS ONE*, *9*(1), e86969. http://dx.doi.org/10.1371/journal.pone.0086969
Tindale, R. S., Meisenhelder, Helen M., Dykema-Engblade, Amanda A., & Hogg, Michael A. (2001). Shared cognitions in small groups. In M. A. Hogg, & S. Tindale (Eds.), *Blackwell handbook of social psychology: Group processes* (pp. 1–30). Oxford: Blackwell Publishers.
Van Arensbergen, P., & van den Besselaar, P. (2012). The selection of scientific talent in the allocation of research grants. *Higher Education Policy*, *25*, 381–405.
Van Arensbergen, P., van den Besselaar, P., & van der Weijden, I. (2014). The selection of talent as a group process; a literature review on the dynamics of decision-making in grant panels. *Research Evaluation*, *23*(4), 298–311.
Van Arensbergen, P., van der Weijden, I., & van den Besselaar, P. (2014). The notion of talent: What are the talents we are looking for in science? *Research Evaluation*, *23*(4), 273–284.
Van den Besselaar, P. (2011). *A quantitative analysis of the granting decisions in the NWO career grants program—Report commissioned by NWO, Amsterdam*. Amsterdam: NWO.
Van den Besselaar, P. (2013). More competition, better science—On the predictive validity of grant selection. In S. Hinze, S. Hinze, et al. (Eds.), *Translational twists and turns: Science as a socio-economic endeavor; Proceedings 18th STI* Berlin, (pp. 385–392).
Van den Besselaar, P., & Leydesdorff, L. (2007). *Past performance, peer review, and project selection: A case study in the social and behavioral sciences*. Den Haag: Rathenau Institute.
Van den Besselaar, P., & Leydesdorff, L. (2009). Past performance, peer review, and project selection: A case study in the social and behavioral sciences. *Research Evaluation*, *18*, 273–288.
Van den Besselaar, P., Sandström, U., & van der Weijden, I. (2012). The independence indicator. In Archambault, Archambault, et al. (Eds.), *Science & technology indicators 2012* (pp. 131–141). Montreal: OST & Science Metrix.
Van den Besselaar, P & U Sandström (under review). Gendered differences in research performance and in careers: A Longitudinal Case Study.
Van den Besselaar, P & U Sandström (under review). Measuring independence of mid-career researchers: A proposal for an individual level bibliometric indicator.
Van den Brink, M., & Benschop, Y. (2012). Gender practices in the construction of academic excellence: Sheep with five legs. *Organization*, *19*(4), p.507–p.524.
Van Leeuwen, T. N., & Moed, H. F. (2012). Funding decisions, peer review, and scientific excellence in physical sciences, chemistry, and geosciences. *Research Evaluation*, *21*, 189–198.
Waltman, L., van Eck, N.-J., & Wouters, P. (2013). Counting publications and citations: Is more always better? *Journal of Informetrics*, *7*(3), 635–641.
Waltman, L., & Van Eck, N. J. (2013). A systematic empirical comparison of different approaches for normalizing citation impact indicators. *Journal of Informetrics*, *7*(4), 833–849.
Wennerås, C., & Wold, A. (1997). Nepotism and sexism in peer review. *Nature*, *387*(6631), 341–343.