Project no.	824574
Project acronym:	GRANteD
Project title:	Grant Allocation Disparities from a Gender Perspective
Instrument:	H2020-SwafS-2018-2020/H2020-SwafS-2018-1
Start of the project:	01.01.2019
Duration:	50 months
Work Package:	4.1
Title of deliverable:	Conceptualizing, measuring and developing indicators
Due date of deliverable:	Month 36
Organisation name of lead contractor for this deliverable:	Teresa Mom consultancy bv
Other partners involved:	CSIC, DZHW, FPS
Document version:	V5d2

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824574.

Abstract

This deliverable reports the work done in Task 4.1 on developing several core variables for the model. Firstly, an alternative approach is proposed and tested for measuring scholarly performance with bibliometric data. It avoids using individual indicators but is based on the latent dimensions underneath the indicator space. Secondly, we further developed the indicators for measuring researcher independence, based on earlier work. However, the available indicators are very resource intensive and therefore not useful for large scale samples. We here show how independence can be measured in large samples. Thirdly, we investigated whether gender differences in awards and prizes may be a form of gender bias, and the case investigated confirmed that. The implication is that awards and prizes are not only an important measure of merit, but can lead to indirect bias in grant decision-making. It is therefore important to include a variable representing awards and prizes in the model also as a mediator. Fourthly, we explored the role of team characteristics, and sketch how the new collected data can be used to include team characteristics in the model. Finally, as the development of quality indicators has proceeded slower than expected, we focus in this deliverable on the conceptual and measurement aspects of indicator development. That part of the work in WP4.1 is also a more general contribution to the quality of bibliometric and other indicators.

Name	Organisation	Contribution
Peter van den Besselaar	TMC	Main author D4.1 Co-author annexes 1, 2, 3, 4, 5, 6 Co-author working papers 1, 2, 5 (not included here)
Charlie Mom	TMC	Co-author D4.1 Co-author annexes 1, 3, 4, 5. Co-author working papers 1, 2, 5 (not included here)
Laura Cruz-Castro	CSIC	Co-author D4.1 Co-author working paper 4 (not included here)
Luis Sanz Menendez	CSIC	Co-author D4.1 Co-author working paper 4 (not included here)
Torger Möller	DZHW	Co-author D4.1 Co-author annexes 3. Co-author working paper 3 (not included here)
Ulf Sandström	FPS	Co-author annexes 1, 2.

Dissemination level	
PU	Public

The GRANteD consortium consists of six partners:

Teresa Mom Consultancy bv (TMC) (Netherlands) - Task 4.1 leader,

Joanneum Research Forschungsgesellschaft mbH (JR) (project coordinator, Austria),

Orebro University (ORU) (Sweden),

Agencia Estatal Consejo Superior De Investigaciones Cientificas (CSIC) (Spain),

Deutsches Zentrum Fur Hochschul- Und Wissenschaftsforschung Gmbh (DZHW) (Germany),

Research Policy Sweden Inc (FPS) (Sweden).

Keywords: gender differences, indicators, independence, research teams, academic careers, measurement

Preface

This deliverable reports the work done in WP4, task 4.1, on the development of core indicators for the project. It consists of a main report, based on several papers and working papers.

As the work underlying this deliverable was organized as a set of partly parallel activities, each of these activities resulted in a published paper or a working paper. Some of the working papers have in the meantime been (partly) published and are included in this version as annexes.

This main report of deliverable D4.1 summarizes the results of the various (working) papers and discusses the implications for the project.

Version information:

Version V3 (July 14, 2021) was distributed in the consortium to collect comments and suggestions.

Version V4 (October 14, 2021) was reviewed by two members of the Scientific Advisory Board,

- professor Renata Siemieńska
- professor Jacques Mairesse

The meeting with the SAB members was held on November 3rd, 2021

Version V5d is the final version (December 23, 2021) approved by the EC review (March 2022).

Version *V5d2* is the public version (April, 1, 2023), which does not include the underlying working papers, but does include the published versions of those papers as annexes. The main text has been edited in order to improve the language, but contains no substantial changes compared to the approved version V5d.

Table of contents

Annexes

Working papers (not included in this report)

Working paper 1 (section 2.2): Charlie Mom & Peter van den Besselaar, *Team size, performance, and gender*, (2021)

Working paper 2 (section 2.3): Charlie Mom & Peter van den Besselaar, *Large scale measuring of independence* (2021)

Working paper 3 (section 2.3): Torger Möller, *Researcher independence and funding decisions* (2021).

Working Paper 4 (section 2.6): Laura Cruz-Castro & Luis Sanz-Menéndez Quality of the Proposal. *Conceptualizing, measuring and constructing Research Quality (in research project applications)*, (2021)

Working paper 5 (section 1): Peter van den Besselaar & Charlie Mom, *The dataset for Annexes 3, 4, and 5, and for Working papers 1, and 2* (2021)

Working paper 6 (section 2.5 and annex 5): Peter van den Besselaar & Charlie Mom, *How to measure gender differences in academic careers* (2021).

Chapter 1 Introduction

The goal of Task 4.1 was to further develop useful and valid indicators that can be used for the study of gender differences in grant allocation and in career decisions. The indicators relate to different Work Packages:

- WP3 and WP4: Glass ceiling, leaky pipelines, and gender differences in academic careers (Section 2.5);
- WP4: How to measure researcher's independence in large scale studies (Section 2.3); the role of teams (Section 2.2);
- WP7: Do we need survey items on awards in the applicant survey? (Section 2.4);
- WP3, WP4, WP9: Bibliometric indicators (Section 2.1) that will be used in the overall models developed in WP4 and used in the data analysis in WP3, WP4, and WP9.

Most of the indicators were already foreseen in the proposal and mentioned in the task description, such as an indicator for *proposal quality*, an *upscaled independence indicator*, and indicators for *team science*, whereas others are the result of the literature review in D1.1[1], such as the need to include *other merit variables* than bibliometric indicators only, the *valid use* of bibliometric indicators, and an indicator for measuring the *glass ceiling*.

When writing the proposal, it was expected that the literature on measuring *proposal quality* would progress enough to become useful for empirical studies such as GRANteD. However, we had to conclude that the work on indicators for the quality of grant proposals is still in a relatively early stage. Therefore, it was decided that the work on quality indicators would be restricted to theoretical, conceptual and methodological contributions (Section 2.6). In this deliverable D4.1, we present the results of a series of activities to improve and further develop useful and valid indicators.

The underlying papers have been (and will be) presented at several conferences (ISSI 2019 in Rome, Gender Summit 2019 in Amsterdam, STI 2022 in Granada, ISSI 2023 in Bloomington) and published in the proceedings of those conferences or available as preprint.

Overview

The main role of the (bibliometric) indicators in GRANteD is to include factors in the analysis that play a role in peer review, and therefore could account for the gender differences in decisions on research funding and careers: differences in merit, performance or reputation. As was argued in the previous work in GRANteD (Cruz-Castro & Sanz Menendez 2019; van den Besselaar et al. 2020) it is necessary to conceptually and empirically differentiate between *gender differences* and *gender bias* as outcomes of (selection) processes. If big differences between males and females are found in the outcome data, we should not conclude or assume that there is gender bias. We always need to account for *competing explanations* or, if we are constructing empirical models, control for other relevant factors, most importantly those representing *merit*. At the minimum, if we model gender differences we consider *gender bias* as the residual effect, after controlling for other factors, like merit, preferences, reputation, etc. This in order to replace *naïve residualism* (J.R Cole, 1979) with at least *sophisticated residualism* (S. Cole & Fiorentine 1991).

[1] Cruz□Castro & Sanz□Menéndez 2019.

Many bibliometric indicators are available (Wilgaard et al. 2014), all directly measuring some property of publications (e.g. how often cited) and of authors (e.g. how many publications), and if applied correctly always considering the context of the scientific domains and countries. Although one can argue that these indicators may reflect some underlying quality dimension, there are no good reasons to expect that they are individually a reliable measured for that dimension. Therefore, we propose to interpret bibliometric indicators as 'items', and that one needs several items to construct reliable measures of unobservable characteristics such as the researcher's quality. In section 2.1.1, we show how this can be done using a factor analysis of the (Scival) indicators available in our dataset, resulting in three underlying quality dimensions: total impact, relative impact, and journal impact. Scale analysis show that the underlying items create reliable scales.

In section 2.1.2 we discuss a wider set of indicators (not only bibliometric ones) and propose a distinction between performance and reputation indicators, which will play a role in the analysis of gender inequality and gender bias in next steps in WP4. We also acknowledge that other reputational dimensions (and their consequences) cannot be properly measured with bibliometric indicators only (Espeland & Sauder 2016).

Modern science is increasingly a team activity: 'team science' (Bozeman & Youtie 2017), and we discuss this in Section 2.2. There we address the question whether team characteristics have an effect on individual performance, and through that indirectly on the probability of receiving grants. This relates to a methodological and measurement issue regarding the interaction between individuals and groups recognized as the "peer effect" (Angrist 2014) that need to be taken into account in empirical analysis.

A core question in the GRANteD project is whether *men and women profit differently* from being in a team, and whether this is this related to e.g. team size and team diversity. In a study on 108 teams (*section 2.2.1)*, we found that gender variety of teams has no effect on the team performance. This is as such an interesting result as that may release gender diversity policies from discussions about research quality. What has not yet been investigated is whether male and female team members profit differently from team science, which may be addressed in a follow-up analysis. As the study includes many variables, but a low N (108), the other study reported in *Section 2.2.2* is based on a large sample of about 900 teams. We created a large dataset consisting of about 3000 junior researchers and about 800 senior researchers, working together in small teams between 2 and 5 people. For about 900 teams we know the size, the gender composition, and we have bibliometric scores of all team members. We present a first analysis[2] and found in our sample no systematic relationship between gender, team size, and performance of the junior researchers. However, team size correlates with the average number of coauthors (including coauthors not belonging to the team), and the latter variable will be taken into account in the analyses.

A question that always comes up when discussing team science is how to account for the individual contributions of team members. This issue is addressed through the measurement of *researcher independence* in Section 2.3. In an earlier study we developed the independence indicator (Van den Besselaar & Sandstrom 2019) and showed that it may indeed work: high scores on traditional bibliometric variables like productivity and impact should be reinterpreted taking independence into account. That helps to distinguish researchers who have contributed to new ideas and new topics from those researchers who are productive and impactful but mainly due to the collaboration with

[2] Further analyses will be done for a later version to be published, including team composition in terms of gender and academic ranks, and the quality of the senior researchers.

their former supervisors or with other very productive colleagues. As calculating the developed independence indicator is rather resource-intensive for large studies, we here develop alternative approaches that can be used on large scale datasets, based on our earlier work. In *section 2.3.1* we propose two related indicators that require much less resources to calculate: Firstly the share of (fractionally counted) publications of a researcher where none of the supervisors is a co-author. And secondly the (lack of) overlap between the topics covered by a researcher and the topics covered by the supervisor(s). For a sample of some 900 PhD students, we show that the indicators can be calculated with reasonable effort, using some scripts developed for this task. *Section 2.3.2* addresses the question to which extent researcher's independence is taken into account in the funding decisions of a prestigious early career grant in Germany. A distinction is proposed between *independence from* and *independence to*, as *Independence from* a postdoctoral supervisor is seen as a condition for developing a certain degree of *independence to* conduct one's own research agenda. On the one hand, it can be demonstrated that *independence from* the supervisor increases over time. On the other hand, the data do not show any significant influence of independence on funding decisions. The results are discussed against the background of the underlying data quality, selection bias, the specific funding program and procedure, and the basic independence model. The section concludes with an outlook on further research including a more differentiated independence concept.

If at all, studies on gender differences in grant allocation use as *merit variables* mainly bibliometric indicators. This is a step forward compared to not including any merit variables, but not sufficient as grant reviewers do take more merit dimensions into account than only productivity and impact (Van Arensbergen et al. 2014) such as earlier grants, and awards and prizes. But as is the case for bibliometric indicators, also for other merit indicators the question has to be answered whether these indicators are (gender) biased themselves. For example, if women would have on average less research time than men do, then bibliometric indicators may reflect this task difference, which indirectly may have a gender effect on grant decisions. The same holds for awards and prizes. If so, this may add an additional level of (indirect) gender bias in grant decision-making. Therefore one should decompose gender bias in grant allocation into *direct* bias, caused by the decision-making process within the funding organization, and a component of *indirect* gender bias which comes from outside of the grant decision process and emerges from e.g. bias in award decisions that influence grant decision making. Consequently, it is an important question whether men and women score differently on merit variables and why, and we test that here for *awards and prizes*. In the large-scale case described in *Section 2.4*, we do find a strong gender bias in receiving the highest award for one’s PhD thesis (in the Netherlands this is *cum laude*). This finding influenced the GRANteD design, as the decision was made to include in the *applicant survey* (WP7) questions about received prices and awards. This analysis is important also because different merit dimensions are used in the selection criteria (see for example Hug & Aeschbach 2020), and reviewers have different preferences on how to evaluate merit (Cruz-Castro & Sanz Menéndez 2021). Moreover, almost every funding scheme includes “worth” criteria, and not just “merit” to assess the applicants and applications (see Cruz-Castro & Sanz-Menendez 2019).

In *section 2.5*, we address the issue how to measure the *glass ceiling*, which is needed when studying gender effects in academic careers (in this project in WP3). We argue that a cohort approach enables to link the share of women appointed as full professor to the share of qualified female candidates in the relevant cohort. An analysis of Dutch data shows that in most fields women still face a glass ceiling: less women are appointed as full professor than one would expect based on the availability of qualified candidates. Furthermore, the preliminary analysis suggests that also the leaky pipeline is stronger for women than for men, although the differences are mainly visible in the very early career (up to five years after the PhD). *Section 2.5* also shows some initial results of the work for WP3 (how do grants affect the career?)[3]. In another GRANteD deliverable (D3.3)[4], we will analyze the data in depth, using an event history approach for those researchers that received the PhD between 2000 and 2005, six cohorts whose career until now is long enough to be appointed as associate or full professor. The data collected for this analysis (and for the work presented in parts of sections 2.2, 2.3, 2.4, and 2.5) are described in working paper 5.

Finally, the need for an indicator *measuring the quality of project proposals* has been formulated over the last few years, and efforts have been made to develop those (see for example Langfeldt et al. 2020). However, the *quality* of a grant proposal is a highly contextual variable defined by the funding agencies and the involved reviewers. We had expected that over the years between writing the GRANteD proposal in 2017/2018 and the work reported here, some possible operational candidate indicators would have become available to further develop and apply in this project. However, progress on this topic has been slow, and there is still not something available that looks as a solution. Still, the implicit approach has been to assume that peer reviewing is capable of identifying quality or excellence (e.g. Cabezas-Clavijo et al. 2013) and that the degree of coherence can be checked with bibliometric indicators. Therefore, the problem of measuring proposal quality is discussed in a systematic way in *Section 2.6* to outline further work which should bring indicator development forward. Scientific quality, not as ex post outcome, but as an on-time product is resulting from an interaction process between the product (and the author) and the evaluators. This approach, emerging within *cultural sociology*, could be of interest to address the issue. As a summary, the reflection brought into the debate stresses the need to establish better links between theory development, conceptualization and measurement; lessons that could be extended to indicators development in science studies overall. Indicator development should be linked to the concepts they represent and be embedded in a theoretical framework that clarifies the purpose of the indicators.

[3] The model developed in D2.1 consists of various parts, which together are needed for a full understanding of gender bias in grants and careers: (i) Explaining application behavior; (ii) explaining grand success; and (iii) explaining career success. Section 2.5 – using a new dataset created for doing the work in T4.1 and D4.1 - is a contribution to the third part of the model. The first part of the model requires data about non-applicants.

[4] Sandström et al. (2022) Deliverable D3.3, GRANteD project.

Chapter 2 Main findings

2.1 Bibliometric indicators

Bibliometric indicators are widely used within the science system, although the discussion about the meaning and use of the indicators has intensified.[5] These indicators are intended to measure aspects of scientific quality, a concept that covers a variety of dimensions. As we will argue, only some of those dimensions are to a certain extent covered by the indicator toolbox, and even those not always correctly.

Here we do not aim to further develop the concept of quality in its various dimensions (contributing new knowledge; independence[6]; leadership of research teams; creating societal impact; supervising PhD students; teaching; etc.), nor do we aim at developing ways of measuring all these quality dimensions. We start here from the commonly used indicators, and address two in our view important aspects of the indicator discussion, that is (1) the *validity* and (2) the *reliability* of the indicators. Validity relates to the question what the indicators do measure, which is especially important when discussing indicators like the journal impact, or the H-index: The first is only indirectly related to the individual performance, and the second is not field normalized, and not restricted in terms of the period covered. Reliability relates to how we should measure the various quality dimensions, as there are quite some alternative indicators around. In section 2.1.1 we propose to focus not on the individual indicators but on the underlying dimensions, as this increases validity and reliability.[7]

We argue in section 2.1.2 that the distinction between performance and reputation should be made, when using bibliometric indicators. *Individual performance* is measured by indicators that say something about the individual researcher's scholarly work, such as the (fractional counted) number of publications, or the (field normalized) number of top cited papers. *Reputation* indicators on the other hand are more indirectly related to the work of the scholar, such as the Impact Factor of the journals one publishes in, the ranking of organizations in one's network, and the number of earlier grants.

The results of the analysis in these sections are relevant for the GRANteD project, as (i) its helps to distinguish better between reputation and performance, and (ii) it enables the use of more reliable bibliometric variables in the quantitative analyses.

[5] We do not aim to review the literature here. Overviews are available, like Glanzel et al 2019, and recently also studies have been published that report about the opinions of researchers about the use of indicators (Cruz-Castro & Sanz-Menendez 2021), and that report about the use in practice (Van den Besselaar & Sandström 2020). Finally, well known reports (Wilsdon et al. 2015) and statements (Hicks et al. 2015) about 'proper use' have attracted quite some attention, and a long discussion exists on the possible negative (perverse) effects of indicator use (Butler 2003). However, the empirical evidence for that is questionable (Van den Besselaar et al. 2017).

[6] But we do develop a new indicator for another understudied quality dimension: The researcher's independence (section 2.3).

[7] Another methodological problem, not addressed here, refers to the *aggregation problem* – units of observation and levels of analyses- already highlighted more than 50 years ago (e.g. Hannan 1971)

2.1.1 Bibliometric indicators and the underlying performance and reputation dimensions

Many bibliometric indicators have been developed, although most of them all measure the same things: productivity, impact, and (international) collaboration (co-authoring). Wilgaard et al. (2014) already distinguished 108 different variants, and since the number has grown even more. However, there is a lack of indicators for many relevant merit dimensions such as received awards (section 2.4) and developed independence (section 2.3). The literature comparing the various indicators and arguing which ones are better than others, is also abundant but not at all conclusive.

We would argue that most indicators measure an underlying dimension but cannot be conceived as a direct measurement of e.g. impact or quality. Therefore, we suggest treating bibliometric indicators as items measuring an underlying (latent) performance or reputation dimension, and that we should use conventional statistical procedures to assess the items and the scale they may form.

Let's give an example. We used Scopus data, and the bibliometric indicators available in Scival for a large set of researchers in some study. We won't discuss here the details of those indicators (see: Elsevier, *Research metrics guidebook*), as we are interested in constructing scales to measure the underlying quality dimensions. We included ten bibliometric indicators, and first calculated the z-scores at field level (as field normalization). Principal Axis Factoring (PAF) with oblique rotation and Kaiser normalization identified three underlying dimensions (Table 1).

Table 1: The bibliometric performance indicators.

Pattern matrix	Relative impact (1)	Total impact & output (2)	Journal impact (3)
PP10% FN	0.980		
PP10% frac FN	0.950		-0.105
Average FWCI	0.719		0.135
PP10%	0.670		0.218
Citations per paper	0.593		0.205
Fractional papers	-0.233	0.919	
P10% frac FN	0.320	0.801	-0.123
Fractional citations	0.215	0.709	
SJR			0.942
SNIP average	0.134		0.703

Extraction Method: Principal Axis Factoring. Rotation Method: Oblimin with Kaiser Normalization.
Rotation converged in 5 iterations.
PP10% = Share 10% top cited papers
P10% = Number 10% top cited papers
SH = Share
Frac = fractional counted

Which factors can we distinguish?

- The first factor measures *relative impact*, the impact relative to the total output. This factor consists of items like 'share of top 10% highly cited papers', the average citations per publication (C/P), and the average FWCI. The Cronbach alpha – a measure of reliability - of this scale is 0.914.
- The second factor measures *total impact*, and the following items loaded high on this component: Publications (fractionally counted), Citations (fractionally counted) and the fractional count of top 10 field normalized highly cited papers. The Cronbach alpha of this scale is 0.873.

- The third factor measures *journal impact*, and on this component the SNIP and SJR loaded. The related Cronbach alpha is 0.851.

In terms of the distinction introduced in the previous section, the second factor can be conceived as measuring the past performance of researchers, whereas the third measures a part of reputation. The meaning of the first factor is somewhat less clear, as it measures the share of 'good' output among all output. This is ambiguous, as a researcher who contributed only to one highly cited paper will score higher on this indicator than someone who contributed to 50 papers, of which are 30 of very highly cited.

We also tested the same approach including the full counts of the various variables (publications, citations, 10% papers and FN 10% papers) instead of only the fractional counts, but the results were very similar. These four additional variables loaded on the same component as their fractionalized counterparts, so they do not influence the results of the Factor Analysis, and the Cronbach Alphas did hardly change.

The factor scores were saved using 'regression' and the three resulting scales can be used as past performance measures. The correlations between Factor 1 and Factor 2 and between Factor 1 and Factor 3 are moderate, and between Factors 2 and 3 the correlation is low (table 2).

Table 2: Factor correlations

Factor	1	2
2	0.349	
3	0.463	0.162

There is a positive relation between total output and the number of highly cited papers, in line with our earlier findings (Sandström & van den Besselaar 2017) that authors publishing more papers (Factor 2) have a higher chance of having a top cited paper (Factor 1). But at the same time some researchers will be 'lucky' and have one (early career) paper which becomes top cited. For these authors the relative impact component starts high and would be expected to drop off: as they publish more papers their share of top papers should come down, so the correlation between Factors 1 and 2 is not very high. Generally, one can expected that researchers with *high shares* of top papers will be on the lower end of productivity (in terms of total output). High output and impact do not go together with average high journal impact, as prolific authors need to use a larger spectrum of journals to get their work published. We will come back to this issue in one of the following sections.

For further validating the approach, we used the same method on two other datasets, and this resulted in the same three components, which suggest that the approach is robust. The advantage of using the underlying factors to measure performance and reputation over using single indicators is clear: the stability and reliability of the measurement is strongly improved.

2.1.2 Recognition through performance and reputation[8]

In the Credibility Cycle (Latour and Woolgar 1986), recognition is the step that follows publications and that precedes money – the resource that enables a new round of research – which then may result in publications and recognition Figure 1). This is a somewhat traditional view on the research process, insofar recognition is not only based on publications.

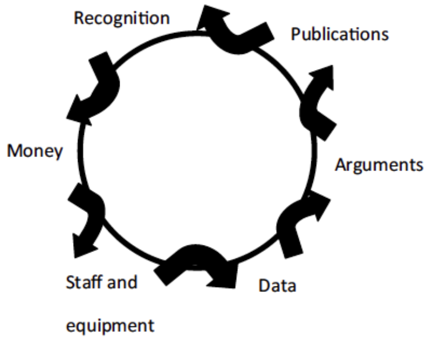


Figure 1: The Credibility Cycle
(Latour and Woolgar 1986)

As the various disciplines have different forms of social and intellectual organization (Whitley 2000), scholars in various fields may depend less on their peers, and more on other audiences for recognition and funding. For the Credibility Cycle this means that the phase "publications" should be combined with other sources of recognition, such as innovations (e.g., patents), policy reports, and contributions to societal problem solving and to the public debate. More recently, also outputs related to other tasks of scholars, including teaching, community service, and knowledge transfer are claimed to contribute to recognition – the extent to which needs further investigation. Another more recent phenomenon that may modify the credibility cycle is that money (grants) is not anymore solely an effect of recognition and an input for research, as receiving a (prestigious) grant is more and more seen as a performance and bringing directly additional recognition (Van Arensbergen et al 2014; Van den Besselaar et al 2018).

Publications have many properties that may help to increase recognition. It could be the number of publications (productivity), the number of highly cited papers, the number of citations received (impact), the size and quality of the co-author network, the international nature of the co-author network, the journal impact (reputation of the journal), and so on. Furthermore, when recognition comes into play in e.g., grant selection procedures, also other signs of recognition may play a role – which can be found in the CV of the applicant: reputation or performance of the organizations the applicant has worked and/or collaborated with or is going to work, the reputation of the PhD supervisor, and the amount of earlier acquired grants (see above). These contributing factors partly relate to the performance of the applicant and partly to the reputation, a distinction already made by Merton (1973), and that is also relevant at the level of research organizations and universities (Paradeise and Thoenig 2013). Using this approach, Van den Besselaar et al. (2019) classified the bibliometric indicators into two groups: performance indicators and reputation indicators (Table 3).

[8] This section is based on Annex 1, which was published in the Proceedings 17th International Conference on Scientometrics & Informetrics ISSI 2019, pp 2065-2069, September 2-5, 2019, Sapienza University of Rome, Italy

Table 3. Indicators for performance and for reputation

Indicator	category
Fractional number of papers	Performance
Total number of citations	Performance
Number of top 10% highly cited papers	Performance
Field weighted citation impact (FWCI)	Performance
Impact of the journal a paper is published in	Reputation
Median of the ranking of the organizations a researcher collaborated with	Reputation
Quality of the host organization	Reputation
Number of international coauthors	Reputation
Earlier grants	Ambiguous

Performance was defined as the outcomes of the scholarly work of a researcher within a relevant (recent) period, as well as the impact of those publications. This can be measured using a variety of indicators, such as the number of (fractionally counted) papers, the citation impact, and the number of top cited papers. These indicators say something about the individual researcher's scholarly work. Reputation indicators point at something else, that is the accumulated standing of a researcher within his/her community (of practice). These are more indirectly related to the work of the scholar, such as the Impact Factor of the journals one publishes in, the ranking of organizations in the researcher's network, and the earlier grants (s)he acquired. For example, even if a paper is hardly cited, if it is published in a high impact journal it adds to the reputation of a researcher ("he/she has again a paper in that top journal"), although if it hardly adds to the performance record as the peer environment has not taken it up as an important contribution.

In a study of a prestigious grant of a leading research funder, Van den Besselaar et al. (2019) found that in the life sciences reputation indicators predicted the grant success more than the performance variables did. Restricting the analysis to the group of very good researchers (the granted and a similar large group of the best performing non-granted), they found the granted applicants score significantly better than the best-performing non-granted on two reputation indicators, and equal on four reputation indicators. In contrast, the best-performing non-granted applicants scored better than the granted applicants on two of the four performance indicators and equal on the two others.

The conclusion is that in the life sciences reputation seems more important than performance for acquiring funding. For some of the social sciences, economics and psychology, we observed the same pattern. On the other hand, in physics and engineering the pattern is different and there performance seems to be dominant over reputation. For studying grant allocation processes, it is therefore necessary to distinguish between performance and reputation, and this also may have a gender dimension.

2.2 Research teams and team performance

As research is increasingly a team activity, and as in all fields publications are coauthored by a growing number of authors, research teams themselves have become an important research topic. Questions have been asked as whether team characteristics, such as team size, team composition and diversity, and team culture do affect the team outcomes, and in what way. These are also important questions for individual researchers, as the success of teams where one has been a member of does influence the individual performance and reputation, and through that e.g. grant application success and career success. The latter questions are important for the GRANteD project: If team characteristics are influencing the performance and if men and women tend to work in different types of teams, then it would be relevant to take the collaboration background of grant applicants into account when investigating the prevalence of gender bias in grant allocation and careers – the topics of the GRANteD project.

A team may be defined as a group of people working together, which is not the same as publishing together. But the delineation of teams is not unproblematic as it may be related to self-identification, to formal rules which may differ between countries and organizations, and above that teams most probably change over time. In the two studies presented here, this is solved in different ways.

2.2.1 Performance of research teams – results of a study of 107 European groups[9]

In the first study, we address the claim that that gender diversity is beneficial to science, using data on gender diversity in research teams and data on research performance at the team level. The study is based on 107 research teams including 1272 individuals. The membership of the teams was determined in a variety of ways, including the description by a team leader and a survey to all mentioned team members. The temporal dimension was not taken into account, as the moment of data collection was decisive.

Three types of teams were distinguished in the sample: 1) the PI-led group with a number of postdoc's and PhD's; 2) the research teams with about 2-4 seniors and junior personnel including several postdocs and PhD students; and 3) large-scale initiatives or consortia of research teams (Stokols et al. 2008).

We focus on the first two types of teams and use the following aspects of research performance:

- field adjusted volume of publications (total production)
- productivity per principle investigator (PI)
- field normalized citation-scores (totals per team and per PI)

Furthermore, a variety of team characteristics are included: team composition, team dynamics, team culture, and a variety of dimensions related to gender diversity. Figure 2 summarizes the model, based on an earlier study by Verbree et al. (2015). The analysis was done using generalized linear models (negative binomial model). Multi-collinearity may be a significant source of error, but does not seem to occur in the analysis as no strong change in regression coefficients occurs when

[9] This section is based on Ulf Sandström & Peter Van den Besselaar, Performance of Research Teams: results from 107 European groups, *Proceedings 17th International Conference on Scientometrics & Informetrics ISSI 2019*, pp 2240-2252. September 2-5, 2019, Sapienza University of Rome, Italy

including new variables in the model. We include independent variables block by block (see Figure 2), and keep from each block only the (marginally) significant variables (p <0.20) for the final model.

The main findings are:

- Team context variables seem to have no effect on the performance variables.
  - University versus public research organization
  - Basic research versus applied orientation
- Team culture variables also do not have an effect on performance in this case study.
  - Team and working climate
  - Leadership style
- The same holds for gender diversity variables
- Several team composition variables have no effect on the performance variable in this case study, such as
  - Gender of the leadership
  - Geographic dispersion of the team
- Some other team composition variables do have an effect on some or all of the performance variables:
  - Single PI-led teams tend to perform lower than more-PI led teams
  - Coverage of competencies has a positive effect on performance
  - Time available for publishing results has a positive effect on performance
  - Share of temporary staff has a marginally significant positive effect on performance, as has the average number of years of team membership
  - Team size is positively related to performance

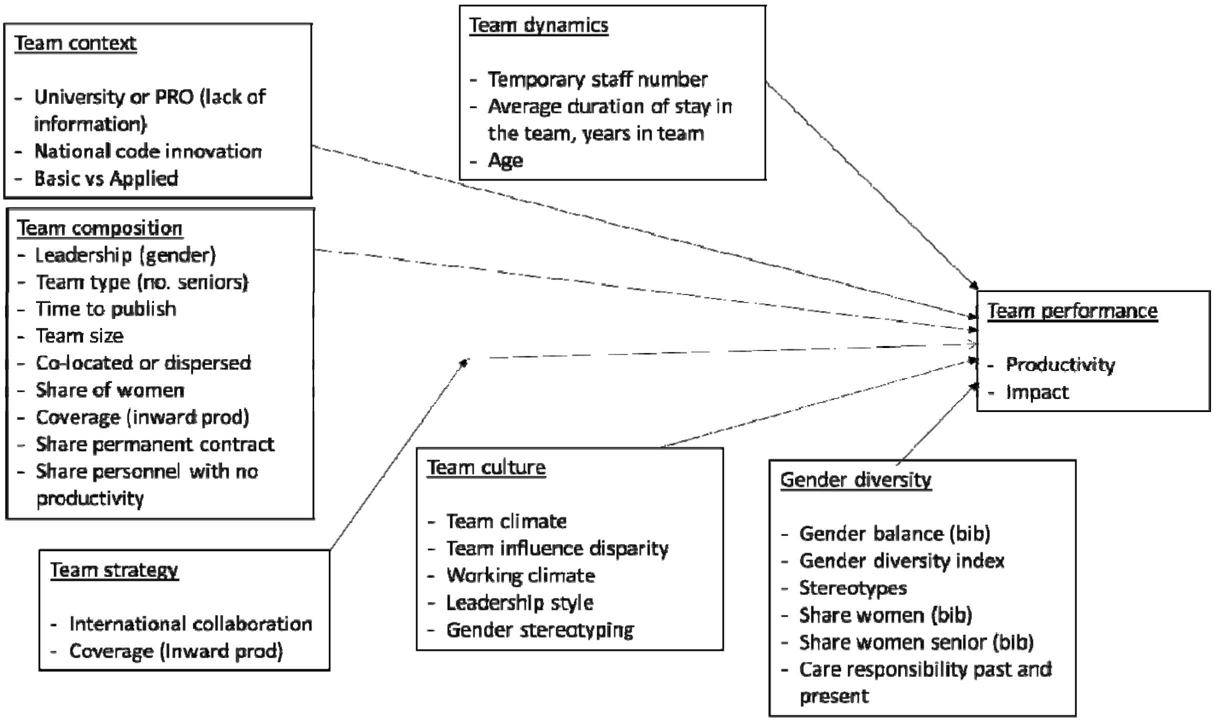


Figure 2: The model with its different components

(based on: Verbree, Horlings, Van der Weijden, Groenewegen, Van den Besselaar 2015)

The overarching research question in this study was whether gender diversity makes a difference for research team performance. However, *the gender variables seem to be without significant effects* on the dependent variables production, productivity, and impact. This holds for the GEDII composite gender diversity indicator – which may so due to the problematic nature of composite indicators – but also for other gender-related variables.[10] However, although gender diversity does not result in higher performance, it also does not imply lower performance, despite significant performance differences at the individual level between female and male researchers.[11]

Furthermore, the type of organization does not seem to matter, as does the difference between co-located and distributed teams, and the orientation towards basic research or applied research. Finally, team culture variables did not have a significant effect.

The findings indicate firstly that teams with several experienced researchers tend to have a larger paper output and a higher productivity than teams of single PIs. Secondly, we found that team composition matters. When the senior team members complement each other in terms of skills and resources and when they share understanding of the work to be done, the effect on performance is positive. Thirdly, also trivial characteristics matter, e.g. time for writing publications should not be forgotten. In all, we conclude with the following: For productivity and citation impact there is one thing that is crucial and that is to construct research teams that have the capacity to combine and use different competencies in a creative way.

The study has a few limitations. Firstly, the study is based on a small sample (108 teams), which had the advantage that many variables could be measured but which also has a disadvantage in terms of the lower robustness of the findings. Secondly, the temporal dynamics of the teams have not been taken into account. Even when the cross-sectional analysis does not show effects of gender diversity (and other team characteristics) on team performance, a longitudinal analysis of the relation between *changes in diversity* and *changes in performance* may result in a different picture.

In the next study, we have solved these issues, however at the expense of the number of independent variables. We did a larger scale study of teams with the same start position, comparable duration, and the same clearly defined target. Some preliminary results of the study are summarized in the next section.

2.2.2 Performance of research teams – results of a large-scale study

For the GRANteD project, the performance of applicants is an important variable to explain grant application success, and the question is to what extent performance of applicants is itself dependent on other (gender related) variables. One of those variables is whether men and women work in different team contexts, and whether that influences their scholarly performance (and through that the probability of success in grant applications).

This second study is based on a large sample of small teams, with the size between 2 and 5 members. One of the team members is a junior researcher, and the others are senior researchers at various levels (full professor or associate professor). Often all are collocated in the same

[10] We tested this elsewhere.

[11] If gender diversity does not influence team performance, it may influence within team performance differences. For example, in more gender balanced teams, performance differences between male and female researchers may be smaller than in other teams.

organization, but this is not necessarily the case. The teams have a similar goal: that is to do research on a specific topic with as output a set of articles and the PhD dissertation of the junior researcher. Given this goal, the team's members work together for about four[12] on somewhat more years. But the start and end of teams is clear. The senior researchers may be in more than one team, but that is not (yet) taken into account here. The *output* of the team is defined as all papers with the junior researcher as (co-author). The context is the Dutch PhD system, which is in rather uniform ways implemented in the different research universities.

We know for all these teams the discipline[13], the output of the research and its impact[14] - based on bibliometric indicators – and the gender of the team members. [15] However, we have not yet included analyzed the effect of gender composition on the performance variables, which will be done in a later phase. As teamwork and team sizes may be different between disciplines, we perform the analyses on the discipline level.

Table 1 shows the distribution of teams by size and discipline (which is operationalized here as *faculty*). We include at this moment only those disciplines where journal articles are the main communication medium, and therefore we exclude the more qualitative social sciences, humanities, law, and religion studies & theology. Furthermore, we include only those disciplines with a large enough N, which are according to Table 4: (i) natural sciences, mathematics & computer science; (ii) Earth and life sciences, (iii) behavioral and movement sciences and (iv) biomedical sciences, and therefore we exclude economics and dentistry.

Table 4. Team size by faculty (2000-2014)*

Faculty / Team size	2	3	4	5	Total
Science, math, computer science	118	362	122	13	615
Earth and life sciences	54	252	182	51	539
Economics	44	164	59	2	269
Behavioral and movement sciences	31	180	132	32	375
Dentistry	4	28	22	14	68
Medicine	70	459	591	316	1436
Social sciences	25	100	57	8	190
Humanities	66	154	34	1	255
Law	33	62	17	1	113
Religion studies & theology	30	105	20	1	156
Total	475	1866	1236	439	4016

* Over the years, the average team size increases, and this especially true for the years we did not include (due to the need to measure citation impact): 2015 and after. In the analysis, only the fields (colored blue) with enough (>300) teams are included.

Table 4 also shows that the number of teams with five members is in the natural sciences and in earth and life sciences (almost) zero, and therefore we exclude that category for these two

[12] There is variation in the duration of the PhD trajectory, the period can be longer or (sometimes) shorter.

[13] We did not yet take into account the possible interdisciplinarity of the team. This can be measured by taking the disciplines of the various senior members into account.

[14] The output related to the team project are the PhD thesis and 7 years publications.

[15] Details about the dataset are in Working paper 5.

disciplines. Table 4 includes all teams, but we did not (yet) collect bibliometric data for all teams: For the field of medicine we included for the time being only half of the teams[16].

We investigate whether team performance differs by size. Team performance variables included here are (i) the average number of coauthors on the team's publications, (ii) the number of (full counted) papers; (iii) the number of research topics covered by the teams (iv) the absolute impact and output of the teams, (v) the average journal impact, and (vi) the level of independence of the junior team member. Please note that we left out all teams where the junior member (the PhD student) had an average number of coauthors of twenty or more. These are mainly teams within those parts of physics where most papers have author numbers of several hundreds to several thousands, and there publishing practices and team compositions are very different. In fact, these about 60 teams have to be considered as parts of much larger teams. Table 5 shows for the dependent variables the means for each of the four disciplines included in this analysis. The average scores for number of coauthors, number of full papers, and number of topics are about equal for three of the four disciplines, and only medicine shows considerable higher scores on these three variables. The scores on the absolute impact and journal impact variables are standardized at the faculty-level, so their averages should be zero. Table 5 shows that this is also true for our sample, suggesting that out sample is not different from the total population that was used for the construction of the variables and the standardization.

Table 5: Mean scores

Faculty	Authors	Full papers	Topics#	Total impact**	Journal impact**	Indepen-dence*
Science, math, computer science	3.7	6.4	2.5	0.025	-0.038	26.1%
Earth and life sciences	4.5	6.6	2.9	-0.001	-0.015	27.8%
Behavioral and movement sciences	3.2	6.9	2.6	-0.002	-0.026	21.2%
Medicine	6.1	13.5	4.7	-0.030	-0.018	24.5%

N = 2193; for independence N = 1144 (due to limited bibliometric data on supervisors).
* Papers without supervisors as percentage of all papers (fractionally counted). See section 2.3 for a further explanation of how independence can be measured.
** See section 2.1.1 for the way the performance variables are calculated.
# How the topics variable was defined and calculated, is described in section 2.3.

The main findings are:

1. The number of coauthors increases with team size in all fields but is, apart from M&B, not large. There is no gender difference.
2. Larger teams do not cover more topics than smaller teams, and for Medicine, the number of topics covered even decreases by team size. This suggests that in Medicine, smaller teams allow more for broader exploration of research questions than larger teams, and that may imply that smaller teams have higher levels of innovativeness. The analysis does not indicate a gender difference.
3. Larger teams also do not seem to be more productive than smaller teams, and also here the effect of team size on paper production is for medicine negative. Some gender differences emerge, as women seem to have some advantage of larger teams: Output of women goes somewhat more up than for men, or (for medicine) declines less than for men.

[16] The reason of this is simple: the resources needed to collect the data. By the way, the sample of teams we collected data for is representative for the whole discipline in terms of years of PhD and gender.

4. The effect of team size on impact seems field dependent: in the natural sciences it fluctuates, and in the earth and life sciences, impact increases somewhat with team size. In movement and behavioral science, and in medicine, impact declines with increasing team size. Differentiating between men and women shows that women win more and lose less than men when team size increases, just as in with productivity (point 3 above).
5. Larger teams do not score higher in terms of average journal impact, and the findings for men and women differ (slightly) without any pattern.
6. We also tested whether independence of the early career researcher differs with team size (see section 2.3). Achieving researcher's independence is a main goal of the science system. Of course, in our case the junior researchers are still in an early career phase – up to three years after the PhD. But we are interested whether we see independence emerging, and whether that depends (among others) on team size. We find that independence declines with an increase of team size. This holds equally for men and women, apart for Earth and Life Sciences where for women independence first increases with team size and then again decreases. Splitting the data for men and women does not change the picture.

In this study we investigated whether the size of teams correlates with the size of the co-author network, with team productivity, with total impact, with journal impact, and with the junior researchers' independence. The overall conclusion is that larger teams do not score better in terms of these performance variables than the smaller teams. The only advantage of larger teams seems to be the larger number of coauthors, which suggests that larger teams may create larger networks. For the rest, the larger transaction costs of bigger teams seem to outweigh the possible advantages that were expected to result from pooling more human resources, different disciplinary perspectives, and more available skills. The general trend towards larger research teams may in many cases not be beneficial, at least not for the type of teams we studied.

The preliminary results presented here can be extended in several ways related to the model of the GRANteD project:

- Adding data about the gender of the supervisors. Would more women in the team have an effect on performance, and does this differ for male and female early career researchers? We can do that now only partly, as we have data on the gender of the first supervisor, but not (yet) of the other supervisors. In Task 4.3 of this project we may do this as test of a part of the models.
- Does team size in the early phase of the career have an effect on early grant success, and possibly on the career (e.g., on the leaky pipeline), and can gender differences be observed? If so, early team size can be a useful variable in testing whether grants have an effect on the career. This can be done after the data on grant applications success have been added to the dataset.
- Due to the size of the dataset, it also enables causal modeling, by e.g. a matched pairs design.

2.3 Measuring researcher independence

The concept of independence

In the discussions about team science, the question of how to measure individual performance within teams has been asked many times. One way of handling this is to measure researcher independence. For many selection procedures in science, such as for grants (ERC starting grant; German Emmy Noether program), independence is mentioned as a criterion. What is meant with independence? Within the context of science, independence cannot mean doing research 'alone': research is a collaborative activity. One may therefore define independence as the ability (in terms of skills and resources) to decide what research questions one addresses and with who to do this.[17]

The following concept of researcher independence is based on a distinction that originated from the philosophical discourse on freedom. As a result, the current distinction between negative freedom (*freedom from*) and positive freedom (*freedom to*) has been established in the context of freedom of action.[18] The negative freedom is the *freedom from* influences, especially from other actors (personal or organizational) or constraints. This freedom is "negative" because it is undetermined what a person does with the *freedom from* and how he or she acts concretely. This is where the term positive freedom comes in, as the freedom of a person to act (in a self-determined direction). Negative freedom can be understood as a precondition for positive freedom, because the *freedom from* enables the *freedom to*. The realized particular action leads thereby to the fact that other potential options are no longer possible. In this respect, actions limit the scope of further actions. These path dependencies are not perceived by the self-determined actors as a restriction of their scope of action, but as a realization of their *freedom to* go their own way. Positive freedom in this sense does not mean to keep all paths open and to live in a free-floating unboundedness. Martin Heidegger, for example, distinguished in this respect between a "freedom as unbound, *freedom from* (negative freedom)" and a "freedom as being bound to, libertas determinationis, *freedom to* (positive freedom)" (Heidegger, 1971, p. 106).

The above distinction is applied to the concept of researcher independence. In the following we distinguish between an *independence from* other actors or constraints, e.g. by the doctoral supervisor or the university in which one works, and an *independence to* develop one's own research topics or agenda and to undertake that research.

In the sense, researcher independence means both breaking ties (independence from) and entering into new ties (*independence to*). This point is important to emphasize as independence should not be interpreted as the least number of ties. Being an independent researcher does not mean doing research alone in a world of team science, but having the *freedom to* choose the collaborations that are most appropriate for one's particular research agenda. Independence should not be confused with relationshiplessness.

In the academic literature and in practical guidelines, various definitions of researcher independence can be found that more or less comprise the above distinction without making it explicit. For example, the Research Excellence Framework Guidance on Submission defines an independent

[17] See Working 3 paper for a more in-depth discussion of the distinction between independent from (e.g., a supervisor) and independent to (e.g., decide to undertake specific research activities) is developed. We leave that out in this summary, but the indicators used here can be interpreted in both ways. This theoretical issue needs further elaboration.

[18] The difference between freedom of will and freedom of action will not be discussed here. However, the concept of negative and positive freedom can also be applied to freedom of will.

researcher as someone who "undertakes self-directed research, rather than carrying out another individual's research programme" (REF Guidance, 2019, p. 118). The definition implies that *independence from* "carrying out another individual's research programme" is a prerequisite for *independence to* "undertake self-directed research".

Particularly the documents targeting academic career development refer to a process of *becoming independent* instead of *being independent*. A guidebook from a research-oriented university states: "Research degrees are training in independent research" (Code of Practice, cited by Guccione 2020). It could be summarized that from entering a university as a student, through the positions of a doctoral student, postdoc, senior scientist, and assistant/associate/full professor, the researcher independence evolves and increases. We assume that there is an increasing *independence from* and *independence to* at each stage. A co-evolution of negative and positive independence seems to be the ideal development scenario. The academic literature suggests that the question of how much academic independence is desirable cannot be answered conclusively. "It is important to note here that there is no reason to assume that beyond a minimum that guarantees the functioning of the science system, more independence is always better" (Gläser, et al., 2021). We agree with this view, but consider the distinction between *independence from* and *independence to* as useful in answering this open question, especially with regard to academic careers.

Summarizing, the distinction between negative and positive freedom (*freedom from* and *freedom to*) borrowed from philosophical discourse can be profitably applied to the question of researcher independence. It focuses on the process of becoming independent, instead of being independent, which can be analyzed in terms of researchers' career development over time, in which a co-evolution of *independence from* and *independence to* seems to be most conducive. Furthermore, defining independence in this way shows that researcher independence is not contradictory to collaborative research and team science.

Earlier work

Elsewhere, we have developed the concept of independence in relation to the PhD supervisors, as this is the initial dependent collaboration that early career researchers engage in (Van den Besselaar & Sandström 2019). After graduating, the postdoc period is then seen as the period to develop research independence, which should result in new collaborations and new research topics. We translated that concept in terms of indicators and tested those on a small sample of researchers. The following indicators were proposed:

- Becoming independent means that the role of the former supervisors should become less in the collaboration network of the early career researcher. This would imply that the position of the supervisor in the ego-network of the early career researcher becomes less prominent – which can be measured through the *eigenvector centrality* of the researcher and the *clustering coefficient* of the supervisor(s). This is in terms of the previous section: becoming "independent from" the supervisor's network and "independent to" create an own network.
- Becoming independent also means that publications increasingly do not have the former *supervisor(s) as coauthor* – which can be measured by dividing the number of the publications of the researcher without supervisors as coauthor by all publications of the researcher. In terms of the previous section: becoming "independent from" the supervisor and "independent to" publish.

- Finally, an independent researcher is expected to enter in *new research topics*, different from those of the environment where the PhD degree was obtained. To calculate this, we retrieved all publications of the supervisor(s) and the researcher from WoS, did some topic mapping of the set of publications, and then calculated the share of independent topics. This is in terms of the previous section: becoming "independent from" the supervisor's themes and "independent to" develop an own research agenda.

The test on a small sample of researchers in their late thirties showed that calculating independence in this way made a difference: the researchers with the higher level of independence had a more successful academic career than the others, even when the other scored higher on productivity and impact.

A disadvantage of the approach is that the calculation of the indicators is time-consuming, which holds especially for the collaboration network indicator and for the topic-indicator. The goal of this part of T4.1 is to try to upscale the independence indicator in order to make it usable for large scale studies. For the collaboration we do not have an alternative yet, and that is something for future work. For the topical independence we developed an easier to calculate alternative. This results in the following two indicators for independence:

1. Independent publications: We produced a script that splits the publications of a young researcher into two groups: those with and those without any of the supervisors as coauthor. The query basically looks in the author-name or Scopus authors-ID-field which contains the names or the IDs of all the coauthors. If one of the names or IDs belongs to one of the supervisors, this paper is counted as 'dependent'. In case none of the supervisor IDs is found in the mentioned field, the paper is counted as 'independent'. One can do this for full papers and for fractional paper counts. And it can be done for different periods in the career, in order to be able to investigate the development of independence over time. Basically, this leads to the formula:

$$\text{Independence} = (\text{Number independent publications})/(\text{All publications}) \quad (1)$$

2. Independent topics: A second script was developed to find all research topics in which the former PhD student has published without any of the supervisors as coauthor. Also here we use Scopus Scival data, and one has the choice between about 40 fields of research, some 1,500 topic clusters, and about 96,000 topics[19]. We decided to do this analysis at the level of *topic clusters*, as the fields of research are not sufficiently fine-grained, but the topics are too fine-grained (among others as we would have more topics than papers in the analysis).

Technically, this means that we need to identify the set of topic clusters in which the former PhD student has been publishing, and then identify the subset of those topic clusters in which one or more of the supervisors have been publishing (with or without coauthoring with the former PhD student). The remaining subset (so the topic clusters in which the supervisors are not visible) can be counted as the independent topic clusters. However, one could argue that in cases where the former PhD student published on a topic cluster without the supervisor as coauthor *before* the supervisor published in that topic cluster, that those topic clusters should also be counted as independent. We indeed include both sets of topic clusters in the *independent topic cluster set*. We did not use a threshold for the period between the former PhD student entering a topic cluster and

[19] Source: https://www.elsevier.com/solutions/scival/features/topic-prominence-in-science

a supervisor entering that topic cluster. Again this can be done for different periods in the career. The indicator would be:

$$\text{Independence} = (\text{Number of independent topics}) / (\text{all topics}) \qquad (2)$$

We applied the derived indicators on two cases, both of early career researchers during the PhD period and the early postdoc period.

We do not expect to find much independence in the very early career, as this is a period where the researcher is still in the educational phase, even in systems where PhD students are employees (like in the Netherlands). In one of the cases, however, we not only calculated the independence indicators for the PhD period[20], but also for the whole career (up to 2018). One may expect an optimal career pattern where a researcher is very close to (and dependent on) the supervisors during the PhD period, as that may lead to the strongest learning effect, and then becoming independent in the later part of the career. Becoming independent too early may even be disadvantageous, as it may hinder a steep learning curve in the early career phase which is in fact meant for that.

2.3.1 Measuring independence in the Emmy Noether case (Germany)[21]

The following analyses rely on data from the Emmy Noether Program of the German Research Foundation (DFG) in which researcher independence is a central review criterion. The program is open to academics from all fields who have completed their doctorate with an outstanding result and submitted an excellent research proposal within four years after the date of dissertation. They should have published demanding articles in internationally renowned journals and have already proven their scientific independence and substantial international research experience in the postdoctoral period (Böhmer, Hornbostel & Meuser, 2008, p. 18). The data cover funding decisions between 2000 and 2006 in biology, chemistry and physics, and consist of 328 Emmy Noether (predominantly male) applicants (table 6).

N=328	male	female	sum
granted	47.0%	9.8%	56.7%
not granted	32.6%	10.7%	43.3%
sum	79.6%	20.4%	

Table 6: All applicants (population)

N=107	Male	female	sum
granted	65.4%	7.5%	72.9%
not granted	23.4%	3.7%	27.1%
sum	88.8%	11.2%	

Table 7: Used sample

The success rate in the period under consideration is relatively high (56.7%, table 6), which is probably mainly due to the formal entry requirements (see above, in particular, that an application had to be submitted up to four years after the dissertation). After easing the rules, the success rate dropped to the current 20% (DFG 2021). In the basic population (table 6), the average age of the applicants was 32.6 years. The doctorate was awarded at an average age of 29.1 years. On average, applicants submitted their proposal 3.5 years after receiving their doctorate. Women are on average half a year younger at the time of the doctorate and half a year older at the time of application.

[20] Defined as the three years before until the three years after the year of obtaining the PhD degree.

[21] This section is based on Working paper 3.

The composition of the sample used (table 7) differs from the population (table 6). The data we have on the Emmy Noether program do not include the year of the doctorate nor information on the doctoral thesis. This information had to be obtained by other means. In the first step, the doctoral theses were searched via the German National Library, which lists all dissertations in Germany. In the second step, the names of the supervisors were investigated, through (where available) (i) the metadata from the German National Library, (ii) names of supervisors from the online publications of the doctoral thesis or (iii) from an internet search of the applicants. The first two methods yielded hardly any hits, and also the internet search hardly found names of the supervisors, and if so, usually only that of the first supervisor. This procedure led to a selection bias, since persons who left the scientific system could not be found on the Internet or did not have names of their supervisors on their homepages. All data required for the analysis could only be found for 107 researchers. Successful male applicants were overrepresented in this sample (table 7).

For the remaining 107 applicants all publications were retrieved from Scopus by using the Scopus author IDs. Often, researchers have several Scopus author IDs, usually a main ID covering the majority of publications and one or more additional IDs cover only a few publications (Aman 2018; Kawashima & Tomizawa 2015). Where possible, the IDs were merged.

In the following analyses, the first (publication) *independence* indicator was applied (see above, both full and fractional count). The value of the independence indicator ranges from 0 (no independent publication) to 1 (all publications without the supervisor). The indicator is calculated for three periods: (i) publications up to and including the year of the PhD thesis, (ii) publications up to and including the year of application, and (iii) publications up to four years after application.

The first analysis shows that researchers become more independent over the years (figure 3). The researcher *independence* is lowest at the end of the doctorate and increases step by step thereafter. This seems to fulfill a central assumption of the independence model, namely that researchers become more and more independent over time. It is remarkable that the full and fractional counts hardly differ, and the correlation between full and fractional count is 0.95.

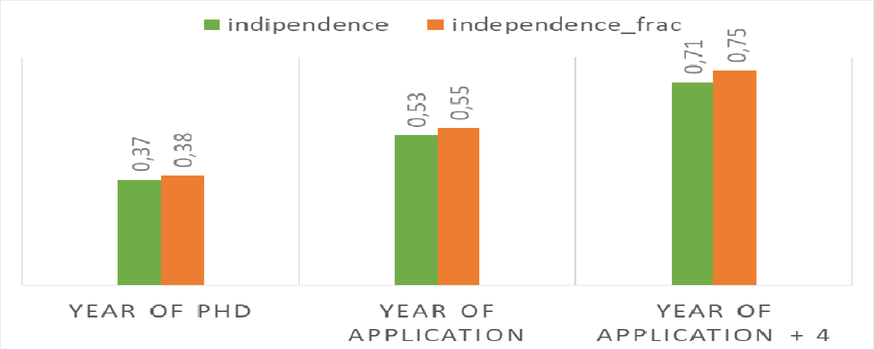


Figure 3: Independence indicator for publications in three time periods

Subsequently, different logistic regression models were calculated, in which the time periods were also varied, as shown in figure 3. The dependent variable is grant success and the independent variables are independence, gender and the subject of the applicant. None of the models returned significant results. Table 8 shows an example of the logistic regression model for publications up to the year of application. Due to the non-significant results, this will not be interpreted further.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.6014	0.5845	2.740	0.00615
Independence (full count)	-0.7082	0.7309	-0.969	0.33260
Gender female	-0.3082	0.6675	-0.462	0.64423
Subject Chemistry	-0.2865	0.5601	-0.512	0.60893
Subject Physics	-0.2627	0.5522	-0.476	0.63425

Table 8: Logistic regression: Grant success as a function of independence (full count & publications until the year of application), gender and subject (Biology, Chemistry, Physics) of the applicants

The findings suggest that independence has no effect on the grant decision, although *independence from* the supervisor increases over time (Figure 3), and this is consistent with the criteria the research funder has formulated.

Three issues are important to stress here, related to the findings: (i) the underlying data quality, (ii) the funding program and procedure, and (iii) the independence model.

Data quality: The following issues came up:

- Selection bias, as women and non-granted persons are underrepresented in the sample (due to missing data on supervisors).
- In most cases only the first supervisor could be identified, so the calculation of the independence indicator is overestimating independence (as papers coauthored with second and third supervisor are counted as independent).
- In some cases not a supervisor, but a senior scientist in the lab was the main co-author and it looks like this person was the informal supervisor of the PhD student. This adds to overestimating independence.
- A solution could be to determine the *independence* indicator not by the formally first supervisor, but by the person with whom the most papers were published during the doctoral period. And it would be necessary to check what status this person has, by inspecting the publication record of that person. This would also solve the problem of finding supervisor names.

Funding program and procedure

- Although independence is an important review criterion, it does not seem to have played a role in the selection process. The proposal, the past performance of the applicant or the interview situation may have been more influential, and we may extend the analysis by including next to independence several additional variables, such as publication impact.

Independence model and independence indicator(s)

Figure 3 shows that the *independence from* the supervisor indicator is already a good indication of the increasingly developing independence of junior researchers. Other indicators for independence could be developed, to measure different aspects of independence – such as a topic independence, which is addressed in the next section. A few initial considerations about possible additional indicators are:

- *Independence from* a main (senior) co-author, instead of the first supervisor (see section on data quality above), and *independence to* create new collaborations.

- *(In)dependence from* research infrastructures / labs, and independence to enter other infrastructures or to create an own lab. Natural science research is highly dependent on infrastructures and labs. An indication of a researcher's independence could also be that an early career researcher become less dependent on the lab where the career has started and worked in different infrastructures / labs before setting up his or her own lab.
- Based on the *independence from* research infrastructure / labs, *spatial mobility* could also be considered as an aspect of researcher independence. In the Emmy Noether Program, for example, international mobility was a precondition and review criterion.

The three suggestions could expand the measurement of researchers' independence beyond the indicators already presented above.

2.3.2 Measuring independence of PhD holders[22]

A second experiment to upscale the independence indicators is based on a dataset of 5729 researcher who received their PhD between 2000 and 2018 from a Dutch university, and dataset with 3988 supervisors that supervised those PhD students. This results in 14187 PhD-student-Supervisor relations and the mean number of supervisors per PhD student is 2.48. As we use bibliometric data on the basis of published papers, we exclude in this chapter all fields (faculties) where publishing papers are not the primary mode of scientific communication. Also excluded are the most recent years, 2015 till 2018, because the papers published in this later period did not have had time to develop their citation potential, which results in bibliometric data for 2592 PhD students. Bibliometric data were also collected for the most prolific (with ≥ 4 PhD students) supervisors, in total 812. For 899 PhDs we have data about all their supervisors, and this subset has on average 2.05 supervisors (sd = 1.24).[23] We use Scopus/Scival data, because every record (paper) contains a field with *all author-identifiers*.[24] Another advantage of Scival is the *very detailed research topic classification*. The data were manually collected and cleaned.

We produced a script that splits the publications of a young researcher into two groups: those with and those without any of the supervisors as coauthor. We then calculated the independence indicator for the PhD period[25], and for the whole career (up to 2018). The latter is important, as one may expect an optimal career pattern where the researcher is very close to (and dependent on) the supervisors during the PhD period, as that may lead to the strongest learning effect, and then becoming independent in the later part of the early career. Becoming independent too early may be disadvantageous. We also use both full and fractionally counted papers, as the way of counting may also influence the independence indicators.

[22] This section is based on Working paper 2.

[23] This is lower than the average number of supervisors of the entire population. We can explain this because we only have half of the medical PhD students in our bibliometric sample, and in the medical faculty the average number of supervisors is substantially higher than in the other faculties.

[24] Several PhD students and supervisors have more than one ID, which we merged.

[25] Defined as the three years before until the three years after the year of obtaining the PhD degree.

A second script was developed to find all research topics in which the former PhD student has published but the supervisors have not (yet)[26]. Also here we use Scival data on 1,500 topic clusters (see above).

Table 9 shows some descriptive statistics of the independence variables, which suggest that the independence indicators are insensitive to the way of counting papers (full or fractional): the two PhD-period indicators are about the same, and the same holds for the two full career indicators. Independence in terms of *published papers* almost doubles from the PhD period to the later career, from 0.27 to 0.47, and one would expect that when removing those researchers from the analysis that leave the science system after obtaining a PhD degree, that the increase of independence will be even substantially larger. Independence in terms of *research topics* was very low during the PhD period (0.035), which is expected as the PhD period is about learning to do research and then one would expect a the PhD student being embedded in the research lines of the supervisors. Topical independence increases strongly in the later career.

Table 9 Share of the independent papers and topics during the PhD period and during the career

	Period	Mean	Std. Dev.	Min	Max	N
Papers	PhD period	0.2636	0.3208	0	1	925
Papers (fractional)	PhD period	0.2750	0.3296	0	1	925
Papers	Career	0.4573	0.3498	0	1	925
Papers (fractional)	Career	0.4574	0.3511	0	1	925
Topics	PhD period	0.0345	0.1186	0	1	925
Topics	Career	0.3352	0.3191	0	1	925

Table 10 shows the correlations between the six independence indicators. At the career level, the correlation between the three indicators is high to very high. In the PhD period, the correlation between the two papers-based indicators is also very high, but the correlation with the topics-based indicator is rather low, suggesting that independence goes over two stages: first own papers, then in the later career own topics. Own here does not mean single authored, but independent from the supervisors.

Table 10 Correlations between the independent indicators

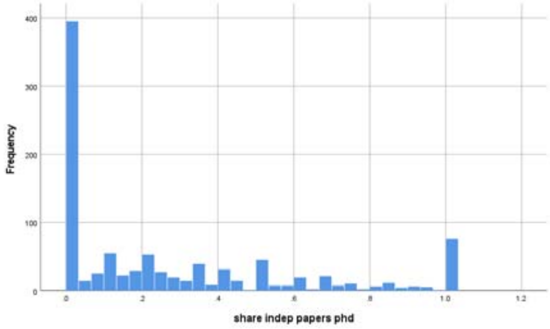
		2	3	4	5	6
career: papers	1	.981*	.594*	.689*	.682*	.173*
career: papers (frac)	2		.577*	.692*	.711*	.170*
Career: topics	3			.426*	.417*	.248*
PhD period : papers	4				.975*	.274*
PhD period: Papers (frac)	5					.266*
PhD period topics	6					

* Correlation is significant at the 0.01 level (2-tailed). N=925

Figures 4 and 5 show the frequency distributions of the first (paper based) independence indicators for the PhD period and for the career. (i) Quite a large share of the PhD students have a score of 0 on

[26] It is of course possible that the supervisors enter the topic later than the former PhD student, and without coauthoring with the former PhD student.

independence. (ii) The relative high frequency of the score 1 is to a large extent based on PhD students with only a single publication, which was published without supervisors as coauthor. (iii) Over time, early career researchers tend to become more independent, and the number of researchers that score 0 declines considerably, and all other values increase. However, quite a few remain rather dependent. This latter group may contain researchers that left the science system. Including this would be easy, using 'stopping to publish' as indicator for leaving the research system.



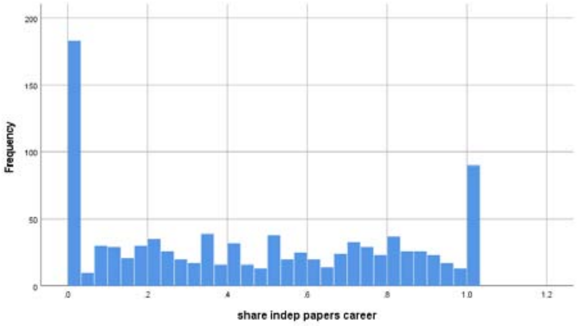
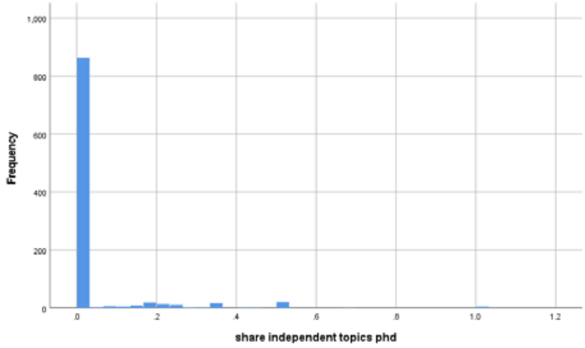


Figure 4: Independence PhD period, papers

Figure 5: Independence career, papers

The figures 6 and 7 show the distributions for the topics-based indicators. The figures show that topical independence is lower than paper-based independence, which is expected. This suggests that many of the papers not coauthored with supervisors still remain in the same topics. Finally, and also as expected, independence increases strongly over the career (Figure 7 versus Figure 6) but many researchers remain in the same topics as their supervisors worked in (which again includes those PhD students that left the science system after graduating).



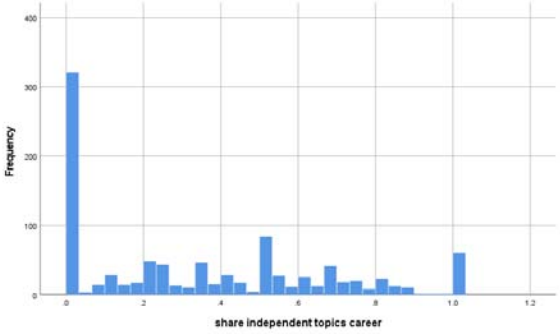


Figure 6: Independence PhD, topics

Figure 7: Independence career, topics

The case shows that using Scopus/Scival data provided us with a possibility to calculate the independence indicators[27] at a large scale. We do not enter here in the discussion whether the deployed *Scival topic clusters* (Scopus 2018) are accurate, as even when the topic classification is not perfect, it can be used as to calculate the topic overlap between the early career researcher and the PhD supervisor. And this avoids that we have to manually make topic maps for every individual PhD student and his/her supervisor(s), as was done in an earlier paper (Van den Besselaar & Sandstrom 2019).

However, the collection of the bibliometric data remains a time-consuming activity:

[27] Somewhat adapted compared to the original definitions (Van den Besselaar & Sandstrom 2019)

1. Collecting and cleaning the publication data of the researchers involved, although the Scopus author-ID proved to be helpful here.
2. Identifying the PhD-supervisor(s) of researchers for who one needs to calculate the independence indicator. However, one may assume that the supervisors are those with most co-authorship relations with the researchers during the PhD period. These can relatively easily be identified after having retrieved the publications of a researcher. By the way, even if the in that way identified supervisor(s) is/are not the former supervisor(s) (which in the Dutch context can be the case), they can be considered as the de facto supervisors.
3. Collecting and cleaning the publication data of the supervisors.

2.3.3 Conclusions

The two cases show that large scale calculating of independence indicators is possible, when using the Scopus data. Parts of the data collection remains labor intensive, such as finding PhD supervisors, but replacing those with the dominant coauthor may be a good solution. Finding the bibliometric data is also time consuming, but becomes easier with the increasing quality of the bibliometric datasets, and with the increased use and reliability of author-identifiers. For the cases within the GRANted project, the calculation of the independence indicators seems doable. However, for a reliable measurement, we may need a few more independence indicators that together might create a reliable scale.

2.4 Gender bias in awards and prizes, a case study[28]

There is evidence the gender bias plays a role in decisions who gets awards and prizes (Lincoln et al. 2012; Melnikoff & Valian 2019; Silver et al. 2018). Men also receive other signs of prestige more than women do, such as invitations for keynote speaker at scientific conferences (Casadevall & Handelsman 2014; Dumitra et al. 2019; Johnson et al. 2016; Klein et al. 2017; Sardelis & Drew 2016), for other invited talks (Nittrouer et al. 2017), or invitations for e.g., scientific advisory boards (Ding et al. 2012). Often these awards, prizes and signs of prestige follow success, but for GRANteD it is relevant whether awards also affect winning (prestigious) grants and support academic careers.

This asks for focusing on gender differences in early career awards, such as cum laude (with honors) awards for the PhD thesis. Such awards, especially when very selective, may impact decisions about grants and about academic positions. And the more selective a grant, the less gender balanced the outcomes generally are (Cruz-Castro & Sanz-Menéndez 2019). Therefore, the question becomes important whether the decisions about awards and prizes themselves are gender biased and have an indirect gendered effect on grant allocation.

We present here the results of a study on gender differences in awarding Cum Laude (CL) for doctoral dissertations in the Netherlands. Only some 4 percent of the PhD students receive CL, and men get that about twice as often than women do. This could be due to gender differences in the quality of the doctoral thesis, and if so, it would imply that men are overrepresented in the group of 5% best dissertations. Elsewhere we showed that the so-called productivity puzzle is mainly due to the fact that men are overrepresented in the top prolific researchers (Van den Besselaar & Sandstrom 2017), and this could be the same for the PhD students. However, the gender difference could also be based on biased decisions, as research on awards and prizes suggest. In this exploratory study, we try to answer the question whether the gender difference between men and women in receiving cum laude is based on quality differences or on biased decisions.

Measuring the quality of PhD theses

As the PhD thesis is assessed by committee members, and as there are no independent scores for the quality of a PhD dissertation, we need to develop a method to do this. As a proxy for the quality of dissertations we will use the papers published by the PhD student in the period of preparing the dissertation, as well as papers published in the three years after the year in which the student received the PhD degree. We assume that these papers are related to the dissertation work. Such a measure would contain error (it may include papers that are not PhD thesis related), but we assume also that this error is not related to relevant variables, especially not with gender. Then the question of the quality of the thesis becomes another question: what is the quality of the scientific output of the PhD students in the defined period? Using Scopus data and Scival indicators, we constructed three bibliometric variables to measure the quality of the thesis work: total impact, relative impact, and journal impact. The way this is done is explained in Section 2.1.1.

[28] This section is based on Annex 4 to this deliverable, which will be published as P van den Besselaar & C Mom (2023) Gender and Merit in Awarding Cum Laude for the PhD Thesis. In *Proceedings ISSI 2023*, Bloomington.

Of course, this does not work in an absolute way, and one would not expect correlation of 1 between the quality of the thesis and the number of highly cited papers coming out of the thesis work. However, one would expect that the *set of cum laude theses* would have higher *average* bibliometric scores than a set of very good, but not cum laude awarded dissertations. If the meritocratic system would work perfect, the quality of the PhD dissertations that received cum laude is higher than the quality of those that did not.

A next issue is the period covered by the measurement. The obvious period to take into account are the years of the PhD trajectory and a few years after the moment of awarding the PhD. But the committee may have not only assessed whether the thesis is excellent, but also whether the author is excellent and whether one can *expect (future) excellent performance*. This can be measured by the same performance indicators over a longer period after the dissertation was finished, and we use the period until 2018.

The case

The case is a large research-intensive university in the Netherlands, using the period between 2000 and 2018. In that period, 5732 doctorates were awarded by the VU, and the share of female PhDs increased from 39.5% in 2000 to 52.0% in 2018. In that period, 248 theses were granted cum laude, of which about 34% to women, implying that men have almost twice as high a probability to receive cum laude. This ratio is fairly constant over the period considered, and therefore we take the years together as one case. The year of the PhD is, however, used as a covariate. These PhD students were supervised by in total 3988 supervisors in different roles. Most of them (88.4%) were never involved in awarding cum laude, as one already would expect from the small number of CL awardees.

The basic data were provided by the university: Names of the PhD students, age, gender, the year of degree, faculty, cum laude or not, and the names of their supervisors. The data to measure the quality of the PhD thesis have been retrieved from Scopus. For a large number of supervisors, bibliometric data were retrieved for Scopus as well.

As we measure the quality of the thesis using citation-based indicators, we restrict the analysis to the period 2000-2014. Overall, this leads to a set of (only) 120 cum laude dissertations out of 3306 dissertations – which is 3,6%.

We only include in our analysis those research fields for which journal articles are the main research output. Consequently, we include all dissertations in the *sciences*, *mathematics* and *computer science*, in the *earth and life sciences*, in the *medical sciences* and *dentistry*, in *economics* and in *psychology* and *movement sciences*.

Methods

As publication, collaboration and citation behavior differs between the disciplines, we first *standardize* the performance indicators *by faculty* and in such a way that the mean of the performance indicators is 0 and the standard deviation is 1.

The data are analyzed using logistic regression (the dependent variable is binary: cum laude or no cum laude). The logistic regression is first done for men and women separately, in order to test whether the independent performance variables have a similar effect for women or not. If not, we have to include interaction terms between gender and the performance variables in the analysis.

Then we run a logistic regression also including gender as a variable. The odds ratio for gender gives a first insight in the effect of gender on getting cum laude. However, the odds ratio only gives the value of the dependent variable for the independent variables at value = 0. As logistic regression is non-linear, the gender differences may be dependent on the level of the independent (here: performance) variables (Mood 2009). In order to correct for this, we calculate the predicted probabilities (using the predicted margins in STATA) to display the probability to receive cum laude by gender for a variety of values of the performance variables.

Description of the population

Table 11 gives a break down by field, showing that the *stronger codified fields* (the STEM fields, and the mathematics-oriented fields like economics) have a low percentage of PhD theses with cum laude (between 1.6% and 3.3%). In these fields, the probability that men receive cum laude is twice as high (or more) as for women.

Table 11. Cum laude by gender and faculty (2000 - 2018)

Faculty	share	M-Total	W-Total	% CL	M-CL	W-CL	M-CL%	W-CL%	W/M
Earth and life Sciences	12.5%	379	338	2.0%	9	5	2.4%	1.5%	0.62
Natural Sciences	15.5%	645	243	4.3%	33	5	5.1%	2.1%	0.40
Dentistry	1.9%	56	53	2.8%	2	1	3.6%	1.9%	0.53
Medicine	36.0%	849	1211	3.3%	40	29	4.7%	2.4%	0.51
Business and Economics	6.7%	277	105	1.6%	6	0	2.2%	0.0%	0.00
Behavioral & movement sciences	9.8%	221	338	6.4%	21	15	9.5%	4.4%	0.47
Humanities	5.5%	174	143	7.9%	18	7	10.3%	4.9%	0.47
Social Sciences	5.1%	123	168	7.6%	11	11	8.9%	6.5%	0.73
Law	3.0%	81	88	8.3%	7	7	8.6%	8.0%	0.92
Theology and Religious Studies	4.1%	194	42	8.9%	18	3	9.3%	7.1%	0.77

M-total: total number male PhDs; W-total: total number female PhDs; %CL: percentage of PhDs with Cum Laude; M-CL: met with cum laude; F-CL: women with cum laude; M-CL%: percentage men with com laude; F-CL%: percentage women with cum laude; W/M = W-CL%/M-CL%

Within social sciences, law, and religion studies, the share of PhD students receiving cum laude is substantially higher (between 7.6% and 8.9%), and the differences between men and women are relatively small. Psychology and humanities have an ambiguous position. Psychology – which often presents itself as a STEM related field - belongs to that group when we look at the gender differences: men receive twice as often CL than women. However, psychology (and movement sciences) has a much higher percentage of PhD students that receive CL and are in this respect more similar to the other social sciences. The same holds for the humanities: with respect to the share of PhDs with cum laude, humanities are similar to the social sciences, but in terms of gender differences it belongs to the group with the STEM fields and economics (and psychology).

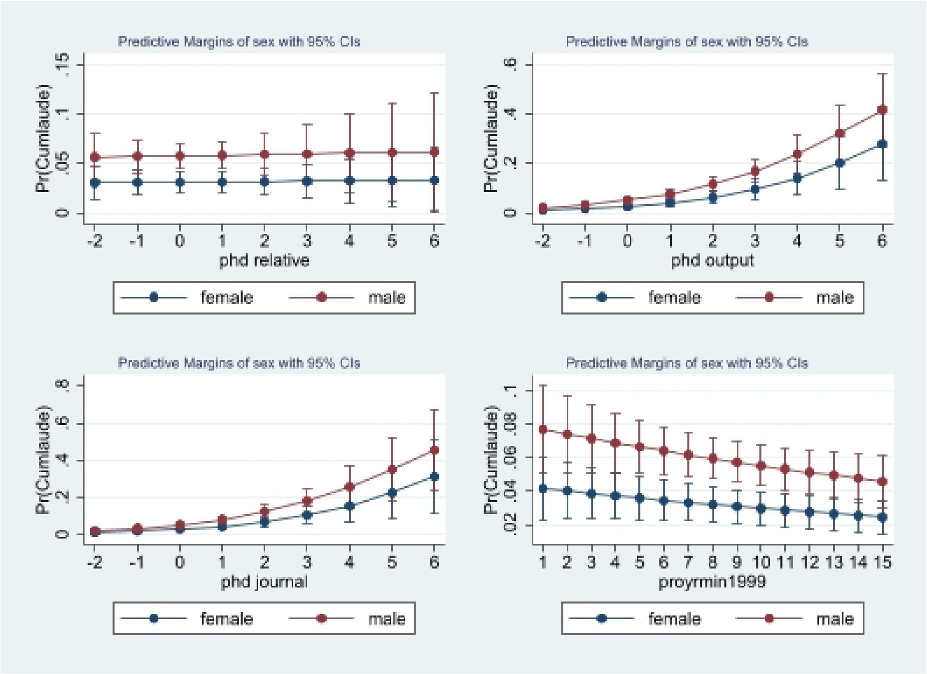
Gender bias?

We started with a logistic regression for men and women separately, and test whether the independent performance variables have the same effect on the probability to get cum laude for men and women. This is the case, and we therefore do not need to include interaction terms in the analysis. In order to test whether women have less chance to be awarded cum laude, we use a logistic regression with *yes/no Cum Laude* as dependent variable. As independent variables, we use the three impact variables and gender, and we add as control variables the faculty and the year of receiving the PhD. Table 12 shows the results. Two impact variables (total and journal impact) have a significant *positive* effect on the probability of obtaining cum laude, and the relative impact has *no effect*. Being *female* has a negative effect on the probability to get cum laude: the odds ratio for gender is 0.49 suggesting that men get twice as often CL than women after controlling for performance.

Table 12. Cum laude by gender, faculty and quality*

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
							Lower	Upper
Sex of PhD candidate(1)	-0.714	0.228	9.824	1	0.002	0.49	0.313	0.765
Faculty			33.673	5	0			
faculty(1)	-0.702	0.258	7.373	1	0.007	0.496	0.299	0.823
faculty(2)	-1.806	0.402	20.154	1	0	0.164	0.075	0.362
faculty(3)	0.004	0.268	0	1	0.988	1.004	0.594	1.697
faculty(4)	-1.677	0.495	11.462	1	0.001	0.187	0.071	0.494
faculty(5)	-1.125	0.764	2.164	1	0.141	0.325	0.073	1.453
PhD year	-0.041	0.024	2.967	1	0.085	0.96	0.916	1.006
relative impact (PhD period)	0.02	0.1	0.042	1	0.838	1.021	0.839	1.241
total impact (PhD period)	0.465	0.066	49.777	1	0	1.592	1.399	1.812
journal impact (PhD period)	0.474	0.093	26.051	1	0	1.606	1.339	1.926
Constant	-2.143	0.265	65.64	1	0	0.117		

* Logistic regression; all PhD students in the selected faculties; ** Indicators calculated over the t-3 ~ t+3 period; N=2579; Nagelkerke R Square: 0.177



Figures 8a-8d: Predictive margins of probability to receive cum laude by sex

The odds ratio for gender reported above is sensitive for non-linearity, and only measures the gender effect for independent variables = 0. Therefore, we calculate the predicted probability to receive cum laude for men and women *at the various levels of the independent variables*. This is done using STATA, and the results are in Figure 8a-8d.

For each level of the *relative impact*, men have a higher probability to get cum laude than women do, although the performance variable itself makes no difference (Fig. 8a). For the *total impact/output* (Fig 8b) and for the *journal impact* (Fig 8c) the analysis shows that at the lower levels of the independent variables there is no gender difference, but the higher the performance level, the larger the distance between men and women. One would expect that CL awardees are at the higher level of performance, so gender does play a role. At these high-performance levels, the 95% confidence intervals start to overlap, which is due to the low number of observations at those levels. These predictive margins support the gender difference found in the logistic regression. The last figure 8d shows that over the years, the probability to get cum laude declines, but the men-women ratio remains about constant: In each year men get it about twice as often as women do – even after controlling for performance. Of course, in further analysis many other relevant variables could be included in the analysis, such as the quality or reputation of the PhD supervisor, the social relation between PhD student and the supervisor, personality of the PhD student, etc., because the moderate value of the R^2 suggest that other variables not included in the model could also play a role in the explanation of the outcomes.

Conclusions: Is cum laude merit based?

The raw data showed that men get twice as often cum laude than women do. In the logistic regression analyses, we controlled for quality of the PhD thesis research, for the discipline, and for the year of obtaining the PhD degree. We consistently find that gender has a strong and for women negative effect on the probability to receive CL. In fact, the control variables included in the model do hardly change the strong effect of gender on the probability to receive a CL award. When comparing the predicted probabilities to get a cum laude, we find that for the low performing men and women the probability is equally low, but the higher the performance level the larger the gender difference: the highest performing women overall have a much lower probability to get cum laude then men do at that level of performance. As cum laude may have strong career effects, our result suggest that the current procedure may create unjustified differences.

Overall, the analysis shows that it is important to include (a) variable(s) representing awards and prizes in the analysis, as it is likely to represent a factor that (indirectly) contributes to gender differences and possibly to bias in grant allocation.

2.5 Gender differences in careers[29]

Not only are gender differences in grant allocation disputed, the same holds for gender differences in higher academic ranks. It is clear that in most countries (and especially in the Netherlands) the number of female (full) professors is much lower than the number of male (full) professors, and often this is directly characterized as *gender bias*, using metaphors like the leaky pipeline and the glass ceiling. However, several competing explanations have been put forward. In fact, all these mechanisms may play some role in creating gender differences sin career patterns.

- The cohort explanation: If 15 to 20 years ago few women entered the academic career through PhD programs in a certain field, we now see a low supply of qualified women and that translates into a low number of women appointed now as full professor in that field.
- Differences in career choices: men and women may have differential preferences as to pursue an academic career or not. This by the way, seems to differ between fields and countries.
- Women have more career interruptions then men due to pregnancy and child care, leading to a slower career and to less women reaching the highest academic positions. However, also here the evidence is not univocal.
- Differences between men and women when to start an academic career, also something that may differ between countries and time periods.
- Policies to increase the share of women in the academic staff may have a positive effect in reducing the gender differences that have emerged in the past.

Cohorts

Time plays an important role when analyzing careers. In the Netherlands, it takes on average seventeen years between being awarded a PhD and being appointed as a full professor, and - importantly - that would be the same for men and women (Rathenau Instituut 2019). This should be considered when assessing gender differences in academic careers. If seventeen years ago in a field only a few women were getting a PhD, then only a small pool of qualified female candidates exists *now* for full professor positions. This is the well-known issue of subsequent cohorts with a different demography: the cohort of 2000 has now reached the level and experience of entering into full professor positions, and the share of women among the fresh appointed full professors should be compared with the gender distribution in that cohort, and not with the cohort of PhD receivers now or with an average of a set of annual cohorts.

The cohort approach enables us also to investigate the leaky pipeline, whether it exists differently for men and women, and especially when the career pipeline leaks – in what career phase(s). The leaky pipeline suggests involuntary leaving, but leaving the system can also be based on individual career priorities.

In the analysis planned for D4.2, we aim to explain both (i) leaving the system and (ii) career steps (becoming associate and full professor).[30] Based on the literature review D1.1 (Cruz Castro, Sanz

[29] This is a rather extended version of an earlier Dutch paper that appeared on the political science blog *Stuk Rood Vlees*, August 2019. Part of the findings were presented during the *Gender Summit 17* in Amsterdam on October 3, 2019. https://gender-summit.com/142-gs17. This chapter is also a contribution to Work Package WP3. This section is based on Annex 5.

[30] Additionally, one can also ask the question whether the academic position influences grant success: Is securing a (prestigious) grant influenced by the academic rank such as associate professor or full professor.

Menendez, 2019) and the model developed in D2.1 (Van den Besselaar et al. 2020), the study will follow an *event history approach* (Allison 2014) using several independent variables like (1) grants – especially the career grants of the ERC and NWO[31] (ii) performance, (iii) gender, (iv) mobility, (v) level of independence, (vi) discipline, (vi) awards - in this case the Cum Laude award for the PhD thesis. The dataset is the same as used for sections 2.2, 2.3 and 2.4, consisting of some 5200 Dutch PhD recipients (2000-2018) in all fields. This work contributes also to WP3, as it enables the comparison between Sweden and the Netherlands, and therefore prevent WP3 from becoming too 'local'. Furthermore, in contrast to the Swedish case we do not compare grant winners with unsuccessful applicants but with potential applicants. In this section, we show some *preliminary descriptive results* for the case of the Netherlands.

The glass ceiling and the leaky pipeline[32]

An initial result is based on a subsample of about 250 PhD recipients from the medical school of a Dutch university covering a six years period (2000-2005). The sample is evenly distributed over men and women, and we have for all of them data covering a 17 years career period. We distinguished between those PhD receivers that remained in an academic position in a university, university hospital, or a research institute, and those that leave academia (or research in a broader sense). Everyone with publications up to 2018 is considered as still being within academia, which is not unreasonable for fields like biomedical research. Those that remain in academia, however, do not all have traditional academic ranks such as senior researcher, associate professor or full professor. Quite a few of those that publish are medical specialists in the first place. We then searched the web (and academic repositories such as Narcis[33]) for the careers of the PhD receivers: when did they get which position and how many years they remained in the position before getting promoted to a higher position. We also searched for the grants these researchers received at the Dutch NWO project database, the ERC ERAS-database.[34]

Firstly, about half of the PhD receivers leave academia, and this are more women (58%) than men (43%) – suggesting that the leaky pipeline as conventionally understood indeed exists. Using this, we find that women leave academia much more in the five years immediately after obtaining the PhD. After that women leave the academic system only slightly more than men do (see Figure 9).

[31] NWO is the Netherlands Research council.

[32] See Working paper 7 for more details. Some of the work reported here was done together with Davina van Perge in her master thesis project.

[33] Narcis is the Netherlands national research information system (www.narcis.nl)

[34] We may – if resources are sufficient – include the EC Cordis database, although this database is much less reliable at the individual level.

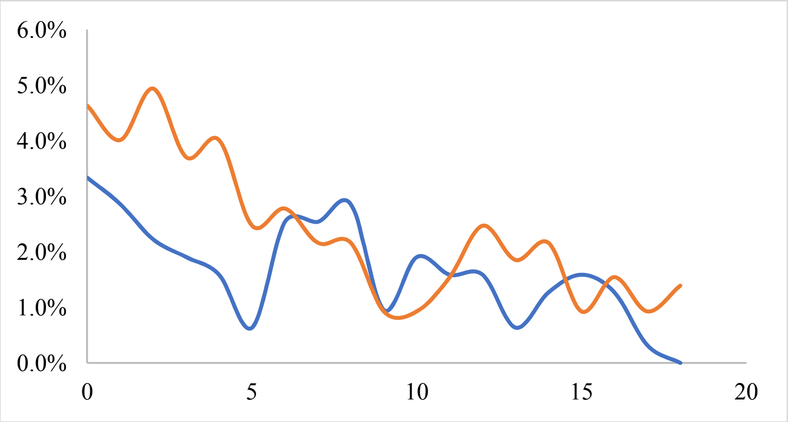


Figure 9. The share of men (blue) and women (red) leaving academia by years after the PhD

We found for almost all that remained within academia rather detailed data about academic positions. The following results emerged from this pilot study (Table 13): In our small sample, men have a slightly higher probability than women reach the higher positions of full professor and a much higher probability to become medical specialist, whereas women more often end (within the period under study) in the lower rank of senior researcher. These three positions cover most of the researchers in the sample. We also found that on average, women need more time to reach those positions than men do. Both results together suggest the existence of a glass ceiling, although for the full professor position much less than expected. But please note that this is based on a small sample only and only in the medical field.

Table 13: Careers of men and women

N = 133; Male: 70; Female: 63	Men	Women
Share PhDs becoming full professors	23%	21%
Average nr years between PhD and becoming full professor.	9.6	11.2
Minimum # years	3	6
Maximum # years	17	19
Share MDs (medical doctor) becoming senior researcher	16%	32%
Average nr years between PhD and becoming senior researcher	9	10.5
Minimum # years	1	0
Maximum # years	20	26
Share MDs becoming medical specialist	43%	29%
Average nr years between MD and becoming medical specialist.	7.8	9.3
Minimum # years	2	4
Maximum # years	13	13

The glass ceiling indicator

The next analysis is based on statistics about PhD students in the Netherlands, and about full professors in the Netherlands. Table 14 below shows the increase in numbers of female full professors (2003-2015) and numbers of female PhDs seventeen years earlier (1984-1998), for different disciplines. Over the long period, the average growth in the proportion of female full professors in the non-medical fields is lagging only a little (3%) behind the proportion of female PhDs receivers seventeen years earlier. However, this proves to be a composition effect. It is mainly due to the technical sciences[35], where the percentage of female full professors grew 34% faster than the percentage of those who obtained their PhDs seventeen years earlier. In all other areas, the percentage of new female full professors lags behind the expectations that the average suggests. This applies very strongly to the agricultural sciences (39%), but also quite strongly to the social sciences (18%) and natural sciences (19%). The humanities show the smallest career disadvantage for women (7%). For the medical field we had data for a shorter recent period, but also these suggest that the number of appointed full professors lags considerably (35%) behind the expected value. The data show that women are catching up, because the number and proportion of female full professors is increasing everywhere. But the data also shows that the catching up goes slower than expected, that is slower than it should be given the supply of qualified candidates.

Table 14: Appointments of female full professors by the change in qualified candidates

	Humanities	Social sci. and Law	Science, Math, Comp sci	Technical fields	Agriculture, Veterin.	Total	Bio-medical	All UMCs	VUMC
Professors									
Appointments 2003-15	513	1507	508	581	86	3195	2016-2018		42
Of which women	137	326	65	67	12	607			13
% women	27 %	22 %	13 %	12 %	14 %	19 %		27 %**	31 %
Qualified candidates									
PhD awarded 1986-1998	2496	4722	5631	3836	1737	16145	1999-2001	2064	201
Of which women	717	1252	895	328	400	3224		840	96
% women	29 %	27 %	16 %	9 %	23 %	20 %		41 %	48 %
Expectation:	147	399	81	50	20	623			20
Difference with real Nr:	-10	-73	-16	17	-8				-7
Percentage:	-7 %	-18 %	-19 %	35 %	-39 %	-3 %		-34 %	-35 %

Data all fields apart from biomedical: Rathenau Instituut (professors), Statistics Netherland CBS (PhD awardees);
Data biomedical field: LNVH (professors), Statistics Netherland CBS (PhD awardees);
Data VUMC: Vrije Universiteit, VUMC. (VUMC = Medical center Vrije Universiteit)
Data all UMCs: Statistics Netherlands, own estimate (UMCs = aggregated for all University Medical Centers)
*: share appointed women (estimated)
Red: Negative gender gap for women; green: positive gender gap for women.

The following limitations may influence the results. (i) It takes on average for men and women the same number of (17) years between receiving the PhD degree and the being appointed as full professor (RI 2018, 2019). As many women give birth to children during that phase of the career, this finding is unexpected, and needs checking in the micro data. (ii) The data we use do not account for international mobility (RI 2018). However, if we assume that migration behavior is about equal for men and women, international mobility may not influence the outcomes of the analysis. (iii) Not all PhD-holders may prefer an academic career, and this may differ by discipline and, important for this analysis, with gender. After receiving their doctorate, men are one and a half times more likely to be employed *outside academia* than women (Sonneveld et al 2010; Van der Schoot et al 2012, 2020), suggesting that the analysis may even *underestimate* the gender gap.

[35] In the technical sciences, the share of female students is the lowest. This may explain why there so many female professors are appointed: The hope that this will attract more female students.

A useful measure of the strength of the glass ceiling (GC) would be the following, were we correct for possible gender differences in the academic job preference (AJP):

$$GC = \frac{\text{(\% women among newly appointed full professors in year T+17)}}{\text{(\% women among PhD receivers in year T)}} \times \frac{AJP(F)}{AJP(M)}$$

This indicator could be further extended with a factor correcting for the gender balance in migration, as the available pool of qualified women and men may be affected by different rates of incoming and outgoing migration.

Conclusion

In the above sections, we elaborated on the idea that studying gender differences and gender bias in careers requires a cohort approach which takes into account the time dimension. What we found is that the glass ceiling still exists in most fields, and that the leaky pipeline is leaking in the early career, and more for men than for women. The analysis suggests that the supply of qualified candidates cannot be the only factor explaining the low share of women among full professors in the Netherland. If that would be the case the gender gap in the full professor positions would be decreasing much faster. Our earlier research shows that other factors play a role (Van den Besselaar 2015; Van den Besselaar et al. 2015, 2016, 2017). From a policy perspective, it is therefore not enough to trust that changes in the labor market over time will do the job of creating gender equality in the occupational structure of universities and their medical centers.The role of grants in these processes has been investigated in a next step (Mom et al, 2022).

2.6 Concepts, categories, and measurement issues

The term “indicator” comes from the idea of “an index hand” or “pointer” (Merriam Webster, main definition), but an indicator is an indicator of *something.* That something is (or should be) always a *concept.*

Measurement in the social sciences has two potential instances (Cartwright & Runhardt 2014): “assigning a number to an individual unit” (e.g. citations to a country or institution) and “assigning the unit to a specific category” (e.g. characterizing a R&D project proposal as excellent or high quality).

The theory of measurement

It is important to distinguish concepts from categories. Concepts play a fundamental role in social life. By concepts we understand “the mental representations through which we identify and give meaning to our experiences” (Hannan et al 2019). The concepts are used by people to classify the entities, objects and situations that they encounter in their social life. Concepts can be regarded as expectations of what kind of properties an arbitrary object that we encounter will have.

By category we understand “a set of objects that have been recognized as fitting a concept” (Hannan et al 2019). Categorizing means assessing the objects (that could be individuals, objects, attributes, situations, etc.) in terms of the concepts, or the abstract mental representations of the world.

The traditional approach of crisp categorization has been challenged in the last decades with the empirical identification of “hybrids”. This has opened the way to the idea that “how an object is categorized depends on social and environmental factors.

For developing an indicator in our domain of interest, we need to consider at least three relevant attributes of the concepts: First, most concepts are socially constructed; second, concepts are not universally shared, they are contentious; and third, most concepts include moral and evaluative dimensions that change with the context.

As social constructs, concepts emerge, evolve and change, and to better understand them we need to consider the effects of social interaction, and, in particular, understand “how the concepts held by one individual evolve as a result of observing the everyday categorization behaviors of other individuals” (Hannan et al 2019), or how the emergence of some devices (e.g. rakings and other metrics or measurements systems (Espeland and Sauder 2016) shapes the categorization behavior of individuals and entities. All processes of conceptualization and categorization depend on social structures such as the “roles that individuals play in social settings or the social acceptance (legitimacy) that concepts have gained within a group” (Hannan et al 2019); what a research project was a hundred years ago is rather different to our present understanding of a research project (Serrano Velarde 2018).

Social construction also means the we should consider the extent to which meanings diverge or coalesce across individuals and groups. (Hannan et al 2019). Sharing concepts contributes to social order, but most concepts are neither universally agreed nor unique to each individual. Finally, many concepts of interest in our field are value-laden. For example, recent evidence (Cruz-Castro and Sanz-Menéndez 2021) shows consistent differences in academics regarding criteria to be used to assess merit in promotion.

In the domain in which we are moving we should consider sets of concepts that appear to be related and need to be considered together; to describe a set of concepts taken together in a space we use the term *"conceptual space*"; for instance: quality, value, worth, merit, talent, excellence, brilliance, are part of the same conceptual space.

Research Quality (RQ), is a "concept" that could be defined for different uses or purposes: for example, for "scientific use", in order to fit into a theory or to make predictions, or to serve "policy needs" or social "descriptive purposes". To clearly define a concept, the purpose needs to be explicit. Then, the concept and the proposed measures should be assessed against the objectives.

Concepts are often measured by indicators or indices. When we face concepts such as Research Quality, it is difficult to do much more than simply count up or list different indicators (as suggested by Aksnes et al 2019) or listing the criteria used for evaluating different objects (for example: Hug and Aeschbach 2020) and considering the context (Langfeldt et al 2020)

But as one refines more a concept and the more precise one tries to make it, the more one may lose some of the associations and original meaning, and comparability across uses may suffer. For instance, the link between citations as a bibliometric measurement of RQ is well recognized and it could be lost when we try to refine the concept. Increasing the number of dimensions and indicators means reducing their role in the definition, and then, establishing the relative weight of every dimension becomes the crucial issue to guarantee the validity of the measure.

Therefore, increasing the conditions of the concept to make more granular measures could reduce its precision. This is also the risk of approaches like the one suggested by Asknes et al (2019) based on dimensions' where the measurement procedure could combine different variables and issues (see below).

The establishment of indicators and the strategy of measurement should be subordinated to theory building, concept definition or to what others, in the context of statistical analysis, defined as "the variables in the model. In the case of research quality, we face two challenges: on the one hand, the possible conceptualizations of "quality" (of a research proposal) and, on the other hand, the application of the concepts to specific real environments (e.g. a research funding competition).

We are not aware of conceptual efforts to develop the conceptualization of quality of research proposal, but there is a significant stream of literature on the so called "predictive validity" of the grant decisions, comparing the rating of the evaluators (as indicators of the peer review) and the citations of the project outputs. (e.g. Cabezas-Clavijo et al 2013; Mom & Van den Besselaar 2021). In fact, such comparison is a way of assuming that the "rating" of the project as a peer review outcome could be a valid measure of quality (understood as worth according to the evaluation criteria stated in the call.)

In the area of assessing quality of research proposals we also face a problem related to the level of analysis, and to the low number of reviewers that usually assess a single project. Most peer review processes in competitive calls are implemented on the basis of evaluations of every project made by very few reviewers, and therefore the conclusions are generalized (at most) at the level of population of applicants (with the problem of the aggregation strategy). For allocations of funds, this could be considered a "fair" procedure, but for measuring quality it is somehow problematic.

Quality of the research proposal: dimensions and measurement challenges

A main difficulty in developing (quantitative) indicators to assess the quality of a research proposal lies in the complexity or even almost impossibility to approach such development disconnected from peer review. At present, there are no ready-made technologies for such an assessment, and this why the issue label as “predicted validity” of evaluation marks has attracted attention. In a peer review based evaluation of research proposals, several dimensions are assessed; the problem is moving from the many-fold concept of quality and its dimensions to the indicators.

We should not adopt a concept of quality based on necessary and sufficient conditions that the object must show to be considered an exemplar (typical of crisp categorization). Instead, we need a concept of research quality that could be the result of a probabilistic approach about how actors (reviewers) classify particular objects (research proposals) in relation to a semantic space defined by the attributes of the concept (quality), and in relation to one another in a conceptual space that represents the nearby concepts and categories.

Context and sequence are relevant here. In processes of evaluation within funding agencies, the way the process is organized is very important and the way in which applications are ordered, whether there is a first filter centered on the quality of the project or on the CV maybe determinant for the evaluation outcome. The fact that in most cases both the CV and the proposal are assessed simultaneously by the same reviewers is also of paramount importance. Also, the ways in which applications are compared individually, in pairs, in groups, etc. since (competitive) research funding implies comparative judgments of fit (among the different proposals) to the concept of quality that is being held.

A research proposal is not an article; it cannot be cited and, *a priori*, it cannot be assessed in terms of (past) impact. Developing indicators to assess the quality of a research project is very different from assessing the impact of a publication through, let’s say, citations. Research proposals include “promises” and “prospective” performance.

Additionally, research proposals should be understood in the context of the funding instrument. Although there may be some common structure, there is not a universal research proposal format. Some agencies and instruments allow for more or less open free style projects, whereas others are much more guided or structured. The quality of the proposal depends on the contextual definition of quality and the dimensions that are explicitly or implicitly included and embedded in the Call for applications.

Therefore, we should acknowledge that assessing the various dimensions of quality of a research proposal requires an analysis of the funding instrument, and that it is difficult to come up with “universal” indicators. A further acknowledgement is that establishing scientific quality involves peer review to more or less extent, with its advantages and shortcomings, including the spaces where bias can occur.

Many of the quality dimensions of a research proposal are not different from those one could identify in a publication; what is different and missing in a research project (but less so if it is evaluated jointly with the CV) are some important social and collective aspects of the science, namely, recognition, credit and visibility. The key and very complex issue is how to measure quality as a property not linked to performance, and also as a property not linked to the scores that the proposal eventually gets. What is similar between an article and a research proposal is that are both science within science and it is in those relational spaces where indicators can be developed.

The first step would be decomposing “quality” (solidity, creativity, value, validity, originality, innovativeness, feasibility, contribution, impact, interdisciplinarity, …) among many other components.

Aksnes et al. (2019) distinguish four dimensions of scientific quality: solidity and plausibility, originality and novelty, scientific value and societal impact. Their aim is different as they refer to those dimensions (and try to define them) to criticize the use of citations as “unique” measures of a quality of journal articles. Although their initial approximation is right, it is however incomplete as concepts are taken for granted; what is missing is a critical analysis of the concepts as categorization social devices, and of how actors locate the objects (in their case publications) in the space defined by those concepts. However, we do not believe these authors analyze the quality as an “empirical concept” but rather they define, according to non-explicit criteria or understandings, some normative dimensions that they believe (as experts) are related with the concept of quality.

Characterizing the concept of quality entails more than simply recognizing and understanding which properties (feature values) are more central. An important additional aspect of the representation of the concept of quality involves the distinction between better and worse instances of the concept, this is, the degree to which the instances (the proposals) fit the concept.

In any case, what is relevant here is that not all of those dimensions have the same score weight in the concept of quality, and they are neither equally prone to be measured through indicators. Although bellow we explain why this is the case in our view, we acknowledge that previous advances help us to present a more systematic approach.

Solidity and plausibility relate to whether the proposal is scientifically well founded, theoretically and methodologically, and likely to produce convincing results. These are probably the dimensions more dependent on peer review, as they are, in essence, a matter of expert judgement. Not even published articles are easy to evaluate on this dimension using the standard indicators.

Originality and novelty refer to whether the proposal poses new hypotheses, new methods, new models and new results. Here we slowly start to enter into a domain in which the proper use of repositories and qualitative semantic databases could be the basis of indicators, but up to now, this would be a very work intensive task. In this area, although mainly oriented to prevent fraud, a very powerful tool is already in place with the more and more extensive use of anti-plagiarism programs.

Scientific value includes aspects like the importance of the research question, the generalizability of results, the relevance to previous and future research and the opening of new avenues. These aspects are very different in relation to measurement. Whereas there is some consensus that the generalizability of results is strongly dependent on the use of (certain) methods and type of data, and the relevance to previous research is quite easy to assess referentially, on the contrary, the importance of a research question is much more subjective.

In both cases, there is a shared understanding that scientific value is related to the accumulation of substantive knowledge about the issue and technical knowledge about the validity and robustness of the methods.

Nevertheless, and acknowledging that check lists are not indicators, there are some common points usually included in the evaluation templates of reviewers which include:

-How well the research is designed (difficult to measure without peer review), and with big difference by fields in which there is consensus on methodological approaches and other.

-Whether it is free from obvious shortcomings, errors or flaws (difficult to measure without peer review)

-How original the research questions are (easier to measure trough bibliographic indicators).

-How much the research questions add to state of the art (possible to measure trough bibliographic and semantic indicators but only indirectly).

An additional complexity with the scientific value dimension is the difference between the quality of the proposal assessed against the formal criteria established in the call, and "expected quality" (excellence) of the (expected) results.

A further dimension of quality, as suggested by Aksnes et al. (2019) is the societal value, and this relates to the broader impact, which, depending on the field, is part of the of funding agencies' review criteria to award research grants. Although this dimension is often considered much harder to measure, in fact, whether the proposal is related to patents, whether it includes a diffusion/commercialization plan, if it has policy implications, or what is the scope of the interest (local or broader) are all aspects that can be verified by a more or less lay person. However, scientific knowledge is socially valuable per se, and social value is intrinsic to scientific knowledge. The issue here is that the social value of scientific research is often confused with its impact or with transferability, which are different concepts.

We have only mentioned some of the possible value dimensions of research quality, but, how valuable is the view of the concept of quality as a list of feature values? In fact, a conceptual schema has high dimensionality and complex connections between feature values; in other words, it is relational. To make the issue even more complicated, those relations can be positive or there can be trade-offs.

The development of metrics and indicators to evaluate the quality of a research proposal may be a question of time and technology; for various reasons related to the complex dynamics of authority sharing in the science system, it is unlikely that they will substitute peer review or administrative discretion in the agencies. However, it is possible that new advances based on cognitive sciences, in computational linguistics or in Bayesian inference speed up the process.

In the meantime, experimenting with evaluation procedures that filter the applications starting with the quality of proposals independently of the applicant CV could produce interesting results.

A research proposal is an object that could be treated as a "cultural object" (Griswold 1986) and in this case, quality is a value that results from the interaction between authors and readers (reviewers), that compose the audience who categorizes the object according to the concept. If we assume that the concept of quality (and its dimensions) is culturally and contextually constructed (including the cultures and contexts of RFOs and scientific fields) we should be very cautious with proposing general, universal and transferable ready-made quality indicators for research proposals. In that sense, a research proposal is a social process that involves the communication of science to a "restricted" public, and a dynamic of interaction between intentions of the authors and interpretations of those goals by the reviewers, a process that is affected also by the quality of the reviewers themselves (and their ability to understand the object).

Chapter 3: Conclusions and implications

The model developed in D2.1 constitutes the basis for the analyses in WP4 and the other WPs 3, 5, 6, and 7. The model consists of various parts, which together are needed for a full understanding of gender differences and gender bias in decision making about grants and careers:[36] (i) Explaining application behavior (WP4 and WP7); (ii) Explaining grand success (WP4), with input from WP5 (the policy context) and WP6 (processes at panel and organization level); (iii) Explaining career success and the role of grants (WP3 and WP4). The work reported here contributes to the different parts of the model.

Implications of the findings

(i) The measurement of past performance of the applicants can be improved by the approach outlined in *Chapter 2.1.1* Most studies using bibliometric data select some bibliometric indicators out of the many available for inclusion in the analyses. However, our experience is that the results of the analyses often depends on that choice, which makes the indicators not very reliable, and therefore also not valid. The approach suggested in this report, running an Exploratory Factor Analysis over the set of available indicators is from measurement theory an improvement. We used it for three large datasets and found the same three components in each of the cases, and these proved to be reliable scales with high Cronbach Alphas. The three scales have an obvious meaning: Total impact, relative impact and journal impact. We tested them also in other analysis, and especially total impact and journal impact work as expected. These indicators will be used in the different parts of the model.

(ii) *Chapter 2.1.2* (and Paper 1) propose a further classification of merit indicators, including several non-bibliometric indicators, with the main finding that a distinction needs to be made between *performance* indicators and *reputation* indicators. As performance and reputation may both have a gender dimension, this distinction is important for the analyses in the GRANteD project. Also this will be used in the different parts of the model.

(iii) *Chapter 2.2* studied whether team characteristics have an effect on e.g., performance of team members, and whether those effects do indicate gender differences. The preliminary study did not reveal clear patterns, but that does not preclude that team effects may play a role that can only be identified in more complex designs. In WP4 we will do some further tests.

(iv) The upscaling of the independence indicators worked well, as *Chapter 2.3* shows. Even though no significant influence of the independence indicator on funding decisions has been demonstrated so far on a small sample, it is generally seen as an important characteristic of creative ad productive scientists, and may play a role in several of the cases under study in the GRANteD project. The independence indicator is a highly relevant variable for all parts of the model developed in D2.1.

(v) *Chapter 2.4* showed that gender bias may be an important factor predicting why men more often than women receive awards that may play a role in (partly reputation based) grant decisions. The case under investigation showed that men twice as often as women receive a prestigious award for the PhD dissertation, and that this could not be

[36] Van den Besselaar et al 2020

'explained away' by performance variables. It is therefore important to collect information about awards and prizes that grant applicants may have received, as this may influence the grant decisions. For the GRANteD project this implies that the applicants survey needs items to collect such information.
Also the awards and prizes are relevant for all parts of the model. Do awards and prizes influence the probability to apply for a grant (or an academic position), and are grants and prizes taken into account by those that decide about grants and academic positions?

(vi) *Chapter 2.5* goes deeper into the issue of how to study academic careers, focusing on the importance of a cohort approach. We show that using such a cohort approach is fruitful and brings relevant insights in the glass ceiling and the leaky pipeline.
It needs to be applied in the third part of the model covering the effects of (bias in) grants on (bias in) academic careers.

(vii) Chapter 2.6 discusses theoretical and methodological aspects of the construction of variables (and indicators) than is often done. The arguments in this part of the deliverable are an important input for the further.

Further work

Firstly, the results of the work done for this deliverable will be applied in the various research activities, especially in WP4.3, where we test the models developed in WP2 (and D2.1) with existing datasets, and were possible with the newly collected data, specified in the Grant Agreement. Most indicators reported here will be used in the *second part of the model*, meant for *explaining the grant decision*: The independence indicator, awards, and the bibliometric variables with a distinction between performance and reputation. In order to do that, the existing datasets should be as much as possible be extended with other relevant variables to cover the model as good as possible, e.g., by adding bibliometric data and/or with CV data.

Secondly, for the *first part of the model (on application behavior)*, we have planned to analyze data from the applicant survey, but these do not include non-applicants. The latter can be done with the German scientists survey (both WP7). The new collected dataset (Paper 2) may be useful for this too, and if so, it would be a welcome addition as this dataset is not based on surveys, but on direct observation.

Thirdly, several results are also relevant for the third part of the model, meant for explaining career progress and the role of grant in that progress. Also here, the bibliometric indicators, the independence indicator and awards and prizes are relevant covariates in the planned event history analysis. The rich new dataset we produced makes a Dutch case study possible, using event history analysis, which can be compared with the Swedish case study (Deliverable D3.2). The sample for the study will be a set of more than 1,000 PhD holders, which got the PhD degree at the same university between 2000 and 2005, covering all disciplines. For this sample, we know the acquired (i) Dutch (NWO) and ERC grants (by year), (ii) gender, (iii) age, (iv) year of PhD, (v) bibliometric performance scores, (v) mobility, (vi) the (development of the) independence score, and (vii) whether they received an award for the PhD thesis. As dependent variables we have (i) when (if at all) they left the science system, (ii) when they were appointed to associate professor and (iii) when to full professor. Comparing the findings with the Swedish case may show differences, and if so the Dutch will help to find out what organizational and national differences may influence the different career dynamics.

Last but not least, more work is needed, in order to publish the relevant results of this deliverable. Several others are already published or under review. It depends on the progress of other tasks as to whether this additional work will be done in the timeframe of the GRANteD project. The SAB-reviewers suggested that some priority-setting is needed.

References

Aksnes, D. W., Langfeldt, L., & Wouters, P. (2019). *Citations, Citation Indicators, and Research Quality: An Overview of Basic Concepts and Theories*. SAGE Open, 9(1), 2158244019829575. https://doi.org/10.1177/2158244019829575

Angrist, J. D. (2014). The perils of peer effects. *Labour Economics*, 30, 98–108. https://doi.org/10.1016/j.labeco.2014.05.008

Bozeman & Youtee (2017) *The strength in numbers – the new science of team science*. Princeton and Oxford: Princeton University Press

Cabezas-Clavijo, Á., 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. https://doi.org/10.1371/journal.pone.0068258

Cartwright, N., & Runhardt, R. (2014). Measurement. In N. Cartwright & E. Montuschi (Eds.), *Philosophy of Social Science: A New Introduction* (pp. 265–287). Oxford University Press.

Casadevall A, Handelsman J (2014). The presence of female conveners correlates with a higher proportion of female speakers at scientific symposia. **mBio** 5 (1): e00846-13

Cruz-Castro, L and Sanz-Menendez, L (2019). *Grant Allocation Disparities from a Gender Perspective: Literature Review*. Synthesis Report. https://doi.org/10.13039/501100000780

Cruz-Castro, L., & Sanz-Menéndez, L. (2021). What should be rewarded? Gender and evaluation criteria for tenure and promotion. *Journal of Informetrics*, 15(3), 101196. https://doi.org/10.1016/j.joi.2021.101196

Ding WW, Murray F, Stuart TE (2012), From Bench to Board: Gender Differences in University Scientists' Participation in Corporate Scientific Advisory Boards. *Academy of Management Journal* 56, 5, 1443-1464

Dumitra TC, Trepanier M, Lee1 L, Fried GM, Mueller CL, Jones DB, Feldman LS (2019). Gender distribution of speakers on panels at the Society of American Gastrointestinal and Endoscopic Surgeons (SAGES) annual meeting, *Surgical Endoscopy*

Espeland, W. N., & Sauder, M. (2016). *Engines of Anxiety: Academic Rankings, Reputation, and Accountability*. Russell Sage Foundation.

Glänzel, W., Moed, H. F., Schmoch, U., & Thelwall, M. (Eds.). (2019). Springer Handbook of Science and Technology Indicators (Edición: 1st ed. 2019). Springer.

Hannan, M. T. (1971). Problems of Aggregation. In H. M. Blalock Jr. (Ed.), *Causal Models in the Social Sciences* (2nd Revised edition, pp. 403–436). Aldine.Atherton.

Hannan, M. T., Mens, G. L., Hsu, G., Kovács, B., Negro, G., Pólos, L., Pontikes, E., & Sharkey, A. J. (2019). *Concepts and Categories: Foundations for Sociological and Cultural Analysis*. Columbia University Press.

Hug, S. E., & Aeschbach, M. (2020). Criteria for assessing grant applications: A systematic review. *Palgrave Communications*, 6(1), 1–15. https://doi.org/10.1057/s41599-020-0412-9

Huijgen M (2019). *NRC checks "men rise more easily in science*" (In Dutch). NRC 25-2-2019

Johnson C, Smith PK, Wang C, (2016). Sage on the Stage: Women's Representation at an Academic Conference, *Personality and Social Psychology Bulletin*, November 2016;

Klein RS, et al. (2017). Speaking out about gender imbalance in invited speakers improves diversity, *Nature Immunology* 18, 5, 475– 478

Langfeldt, L., Nedeva, M., Sörlin, S., & Thomas, D. A. (2020). Co-existing Notions of Research Quality: A Framework to Study Context-specific Understandings of Good Research. *Minerva*, 58(1), 115–137. https://doi.org/10.1007/s11024-019-09385-2

Lincoln AE, Pincus S, Bandows Koster J, Leboy PS (2012). The Matilda Effect in science: Awards and prizes in the US, 1990s and 2000s, *Social Studies of Science* 42, 307-320

Melnikoff DE, Valian VV (2019). Gender Disparities in Awards to Neuroscience Researchers, *Archives of Scientific Psychology* 7, 4–11;

Mom C, van den Besselaar P, Möller T, Factors influencing the academic career – an event history analysis. *In N. Robinson-Garcia, D. Torres-Salinas, & W. Arroyo-Machado (Eds.), Proceedings 26th International Conference on Science and Technology Indicators, STI 2022*

Mood C (2009). Logistic regression: why we cannot do what we think we can do, and what we can do about it. *European Sociological Review* 26, 67-82

National Network of Female Professors LNVH (2018). *Monitor Female Professors 2018*. (In Dutch)

Nittrouer CL, Hebl MR, Ashburn-Nardo L, Trump-Steele RCE, Lane DM, Valian VV (2017) Gender disparities in colloquium speakers at top universities, *Proceedings of the National Academy of Sciences*, December 2017

RI (2018) Rathenau Instituut, *Fact sheet: Women in science* (In Dutch). January 19, 2018

RI (2019) Rathenau Instituut, *News: Share of new female professors rising faster than expected* (In Dutch). March 8, 2019

Sandström U & van den Besselaar P (2019). Performance of Research Teams: results from 107 European groups. *Proceedings 17th ISSI Conference*, pp 2240-2251 September 2-5, 2019, Sapienza University of Rome, Italy

Sandström U, Sandström E, Mom C, Van den Besselaar P, Möller T (2022), Two longitudinal papers on gender disparities in science careers. GRANreD deliverable D3.3.

Sardelis S, Drew JA (2016). Not "Pulling up the Ladder": Women Who Organize Conference Symposia Provide Greater Opportunities for Women to Speak at Conservation Conferences. *PLoS ONE* 11 (7): e0160015

Serrano Velarde, K. (2018). The Way We Ask for Money… The Emergence and Institutionalization of Grant Writing Practices in Academia. *Minerva*, 56(1), 85–107. https://doi.org/10.1007/s11024-018-9346-4

She Figures 2015 (2016) *She Figures 2015*, European Commission

She Figures 2021 (2021) *She Figures 2021, The path towards gender equality in research and innovation*, European Commission

Silver JK., Blauwet CA, Bhatnagar S, Slocum CS, Tenforde AS, Schneider JC, Zafonte RD, Goldstein R, Gallegos-Kearin V, Reilly JM, Mazwi NL (2018). Women Physicians Are Underrepresented in Recognition Awards from the Association of Academic Physiatrists. *American Journal of Physical Medicine & Rehabilitation* 97, 1, 34-40

Simpson, E.H. (1951), The Interpretation of Interaction in Contingency Tables, *Journal of the Royal Statistical Society*, Ser. B, 13, 238-241

Sonneveld HM, Yerkes M, van de Schoot, R (2010). *Trajectories and Labour Market Mobility. A survey of recent doctoral recipients at four universities in the Netherlands*. IVLOS, Netherlands Centre for Graduate and Research Schools, Utrecht

Statistics Netherlands CBS (2019). *PhDs by gender*. Download June 2019

Van Arensbergen P, van der Weijden I, van den Besselaar P (2014) Different views on scholarly talent – what are the talents we are looking for in science? *Research Evaluation* 23 273-284.

Van Arensbergen P, van der Weijden I, van den Besselaar P, Academic talent selection in grant review panels. In: Prpic, K., Van der Weijden, I., Aseulova, N. (Eds.) (2014). *(Re)searching Scientific Careers*. St. Petersburg: IHST/RAS

Van de Schoot R, Yerkes M, Sonneveld H (2010). *Dataset of the Netherlands Survey of Doctorate Recipients*

Van de Schoot R, Yerkes M, Sonneveld H (2012). The Employment Status of Doctoral Recipients: An Exploratory Study in the Netherlands. *International Journal of Doctoral Studies* 7, 331-348

Van de Wier M (2019). Interview with Marceline van Furth and Mpho Tutu (In Dutch). *Trouw*, 12 February 2019

Van den Besselaar P, Mom C, Cruz-Castro L, Sanz-Menendez L (eds.) (2020) *Identifying gender bias in grant allocation, and its causes and effects*. Deliverable D2.1, GRANteD project. https://www.granted-project.eu/wp-content/uploads/2021/04/GRANteD_D2.1.pdf

Van den Besselaar P, Mom C, Sandstrom U (2019) Recognition through performance and reputation. *Proceedings 17th ISSI Conference* 2019, pp 2065-2070, September 2-5, 2019, Sapienza University of Rome, Italy

Van den Besselaar P, Sandström U (2015). Early career grants, performance and careers; a study of predictive validity in grant decisions. *Journal of Informetrics* 9, 826-838

Van den Besselaar P, Sandström U (2016). Gender differences in research performance and in academic careers. *Scientometric*s 106, 143–162

Van den Besselaar P, Sandström U (2017). Vicious circles of gender bias, lower positions and lower impact: gender differences in scholarly productivity and impact. *PlosOne* 12, 8: e0183301

Van den Besselaar, P (2015). New figures: a glass ceiling on all levels of Dutch science (in Dutch). *StukRoodVlees Blog* 1-12-2015

van den Besselaar, P., & Sandström, U. (2019). Measuring researcher independence using bibliometric data: A proposal for a new performance indicator. *PLoS ONE*, 14(3), 1-20. [e0202712]. https://doi.org/10.1371/journal.pone.0202712

Van Furth M, Tutu M (2019). #IToo. *The Lancet* 393, issue 10171, p529

Wegner A (2019) *Result from the NCAPS 1st round in 2019* (Personal communication)

Wilgaard L, Schneider JW, Larsen B (2014). A review of the characteristics of 108 author-level bibliometric indicators. *Scientometrics* 101, 125-158

In: Proceedings 17th International Conference on Scientometrics & Informetrics ISSI 2019,
September 2-5, 2019, Sapienza University of Rome, Italy

Recognition through performance and reputation

Peter van den Besselaar[1], Ulf Sandström[2] & Charlie Mom[3]

[1] *p.a.a.vanden.besselaar@vu.nl*
Network Institute & Department of Organization< Science Vrije Universiteit Amsterdam (Netherlands)

[2] *ulf.sandstrom@indek.kth.se*
KTH Royal Inst Technol, 100 64 Stockholm (Sweden)

[3] *charlie@teresamom.com*
Amsterdam (Netherlands)

Abstract As the various disciplines have different forms of social and intellectual organization (Whitley 2000), scholars in various fields may depend less on their peers, and more on other audiences for recognition and funding. Following Merton (1973) we distinguish between *performance* and *reputation* for building up *recognition*. We show that there are indeed differences between the disciplines: in life sciences and social sciences, the reputation related indicators are dominant in predicting the score that grant applicants get from the panel, whereas in the natural sciences, the performance related indicators dominate the panel scores. Furthermore, when comparing within the life sciences the grantees with the best performing non-grantees, we show that the former score higher on the reputation indicators and the second score better on the performance variables, supporting the findings that in life sciences one probably gains recognition over reputation more than over individual performance. We suggest that this may not be optimal for the growth of knowledge.

Introduction

In the Credibility Cycle (CC) (Latour and Woolgar 1986), recognition is the step that follows publications and that precedes money – the resource that enables a new round of research – which then may result in publications and recognition. This is a somewhat traditional view on the research process, as recognition is not only based on publications.

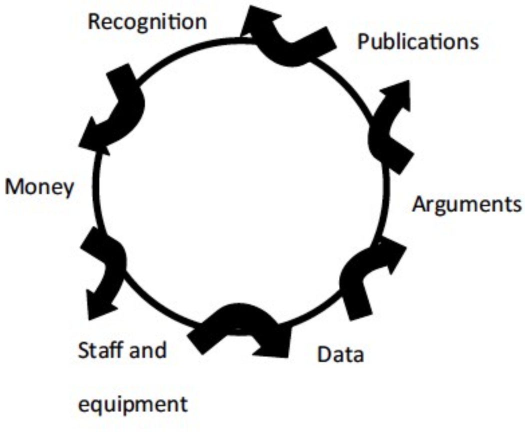


Figure 1: the Credibility Cycle

(Latour and Woolgar 1986)

As the various disciplines have different forms of social and intellectual organization (Whitley 2000), scholars in various fields may depend less on their peers, and more on other audiences for recognition and funding. For the CC this means that the phase “publications” should be combined with other sources of recognition, such as innovations (e.g., patents), policy reports, and contributions to societal problem solving and to the public debate. More recently, also outputs related to other tasks of scholars, including teaching, community service are claimed to contribute to recognition – the extent to which needs further investigation. Another more recent phenomenon that may modify the credibility cycle is that money (grants) is not anymore solely an effect of recognition and an input for research, but receiving a (prestigious) grant is more and more seen as a performance, and bringing directly additional recognition (Van Arensbergen et al 2014; Van den Besselaar et al 2018).

Publications have many properties that may help to increase recognition. It could be the number of publications (productivity), the number of highly cited papers, the number of citations received (impact), the size and quality of the co-author network, the international nature of the co-author network, the journal impact factor (reputation of the journal), and so on. Furthermore, when recognition comes into play in e.g., grant selection procedures, also other signs of recognition may play a role – which can be found in the CV of the applicant: reputation or performance of the organizations the applicant has worked and/or collaborated with or is going to work, the reputation of the PhD supervisor, and the amount of earlier acquired grants (see above). These contributing factors partly relate to the performance of the applicant and partly to the reputation, a distinction already made by Merton (1973), and also at the level of research organizations and universities (Paradeise and Thoenig 2013).

Research question

Based on the considerations above, we try to find out whether reputation, performance, or both constitute the recognition that leads to winning new grants.

Data

We use data on a large and prestigious European grant program. We have data for 3030 of the applicants, which is 95% of all applicants in that round. We do have data on personal characteristics, the subfield of the proposal, the scores the application got from one of the 25 panels, and the full CV of all applicants. Furthermore, we collected bibliometric data from the

Web of Science for (now 60%) of the applicants

For about 1800 of the applicants we downloaded the WoS records, and we checked for different name variants and for different (first) initials. The resulting records were processed by the BMX program (Sandström & Sandström 2009) and then first semi-automatically disambiguated. After that, a manual disambiguation was done. Using the BMX program, the scores for different performance variables were calculated – as mentioned in Table 1.

Table 1. Performance and reputation indicators.

Name	*Description*	*Type*
P-frac	Number of fractionally counted publications.	Performance
Citations(2y)	Field normalized citations, two years citation window.	Performance
Top5%	As above, but now the fields' top 5% cited papers.	Performance
Journal Impact	The field normalized average journal impact factor.	Reputation
Network quality	Median ranking (top 10%) score of linked organizations.	Reputation
Other Grants	The number of other obtained grants by the PI	Reputation

Several other variables could only be retrieved from the CVs of the applicants, and the various data processing steps were done using the SMS platform for data integration and enrichment.[37] The extracted organization names were linked to the Leiden Ranking, in order to determine the quality of the host institution (where the project will be done) in terms of the share of 10% highest cited papers. In the same way, the organizations the applicant has collaborated with or has worked are given a rank-score. We use the median of these scores as indicator for quality of the applicant's network. We manually extracted information about other grants of the applicant from the CVs.

For a smaller set (four life science panels) we have not for 60% but for all applicants the bibliometric data. This smaller set is used for one of the analyses, and for those we also have some additional variables as showed in table 2.

Table 2. Additional performance and reputation indicators

Name	*Description*	*Type*
Top10%	As above, but now the fields' top 10% cited papers.	Performance
Co-authors	Average number of co-authors	Reputation
International	Average number of international co-authors	Reputation
Host quality	The ranking (10% PP) of the host institution	Reputation

We now can differentiate between indicators for performance and indicators for reputation. Performance is measured by indicators that say something about the individual researcher's scholarly work, such as number of publications (fractional count), the number of top cited papers, and so on. Reputation indicators are more indirectly related to the work of the scholar, such as the Impact Factor of the journals one publishes in, the ranking of organizations in the network, and the earlier grants. The last column of Table 1 gives the classification.

Method

We use the variables mentioned above to predict the score the applicants get from the panels, using stepwise linear regression. As disciplinary differences are expected to influence what are considered important for accumulating recognition, we do the analysis by discipline. In this version of the paper, it is done at the level of meta-disciplines: (i) life sciences and medicine, (ii) physics and engineering, and (iii) social sciences and humanities. We report

[37] The SMS platform: www.sms.risis.eu.

here which variables play a role, and which not, but do not go into the numerical aspect of the regression outcome.

Then we look in more detail at the difference between the grantees and the group of best performing rejected applicants.

Findings 1: what predicts the panel score?

In Table 3, we show for the three different domains what variables are included in the stepwise regression outcomes, using the six variables from Table 1. We then compare the emphasis on reputation related aspects versus on the performance-based aspects. Please be aware that this is at the domain level, and that within the various domains the disciplines may differ. This is also the case within the social science and humanities, but as we use Web of Science bibliometric data, the scores for SSH are strongly dominated by economics and psychology. The two latter disciplines are strongly oriented at publishing in journal articles, whereas this is much less the case in the other SSG disciplines, such as literature, history, philosophy, anthropology, and sociology and political science.

What does Table 3 show? All the three reputation indicators are includedin predicting the scores of the life science applications, and only one of the performance-based indicators: top 5% cited papers. The overall score for *emphasis on individual performance* is low: 0.33. For the social sciences and humanities we find the same pattern, but with a different performance indicator: citations. Finally, physics and engineering show a very different pattern. All the three performance variables contribute to predicting the score, and two of the reputational variables. Interestingly, the *journal impact* indicator does not. The overall *emphasis on individual performance* is high: 1.5.

Table 3. Performance or reputation

Name	*Life sciences*	*Physics and engineering*	*Social sciences and humanities*	*Type**
P-frac		+		Perf
Citations(2y)		+	+	Perf
Top5%	+	+		Perf
Journal Impact	+		+	Rep
Network quality	+	+	+	Rep
Other Grants	+	+	+	Rep
Perf/Rep**	.33	1.50	.33	

Source: Van den Besselaar et al. (2016)

* Perf = performance indicator; Rep = reputation indicator;

** Perf/Rep = number of performance indicators divided by the number of reputation indicators

Findings 2: reputation and performance within the very good group

As showed elsewhere, the selection process of grant panels is not very strong, as the best rejected applicants are in average at least as good as the granted applicants (Bornman et al 2010), and the predictive validity is low (Van den Besselaar & Sandström 2015). For four panels where we have bibliometric data for all applicants, we compare the granted applicants with the set of best performing non-granted applicants. Table 4 presents the results.

Table 4. Granted versus best non-granted: performance versus reputation

Name	*Granted*	*better?*	*Best nongranted*	*F*	*Sign*
Performance					
P-fractional	1.9	=	1.9	0.006	0.940
Citations(2y)	3.05	=	2.88	0.360	0.550
Top10%	1.23	<	2.18	11.54	0.001
PModel*	17.3	<	22.5	2.101	0.150
Reputation					
Journal Impact	2.31	>	2.07	2.455	0.120
Network quality	195.5	=	196.9	0.005	0.946
Ranking host	0.14	=	0.13	0.518	0.473
# co-authors	6.76	=	6.52	0.225	0.636
# international co-authors	1.73	=	1.65	0.610	0.436
Other Grants	2.7	>	1.7	5.413	0.022

* The PModel indicator is explained in (Sandström, Sandström & Van den Besselaar 2019)

As the table shows, the granted applicants score (marginally) significant better on two of the reputation indicators, and equal on the other four than the best-performing non-granted. In contrast, the best-performing non-granted applicants score better than the granted applicants on two of the four performance indicators and equal on the two others.

Conclusions and discussion

The conclusion is that in the life sciences reputation is more important than performance for acquiring funding. For the social sciences and humanities, but mainly for economics and psychology, we observe the same pattern. On the other hand, in physics and engineering the pattern is different and there performance seems to be dominant over reputation.

As future work, we will repeat the analysis at a lower level of aggregation: for the individual disciplines, and we will relate the findings to other characteristics at the field level, among other with the occurrence of gender bias and nepotism. We suggest that fields focusing on performance may be less susceptible to bias and more strongly following Merton's CUDOS norms.

Acknowledgments

We acknowledge support of the ERC grant 610706 (the GendERC project) and EC-H2020 grant 824574 (the GRANteD project).

References

Bornman 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** 211-220

Latour B and Woolgar S (1986). *Laboratory life: The construction of scientific facts*, 2nd ed. London: Sage.

Merton, RK (1973). Recognition' and 'excellence': instructive ambiguities. In RK Merton, *The sociology of science: theoretical and empirical investigations* (pp. 419–437). Chicago: University of Chicago Press

Paradeise C, Thoenig J-C (2013). Academic Institutions in Search of Quality: Local Orders and Global Standards. *Organization Studies* **34** 189–218

Sandström U, Sandstrom, E & Van den Besselaar P (2019) The P-model: An indicator that accounts for field adjusted production as well as field normalized citation impact (in preparation)

Van Arensbergen P, van der Weijden I, van den Besselaar P (2014) Different views on scholarly talent – what are the talents we are looking for in science? *Research Evaluation* **23** 273-284.

Van den Besselaar P, Sandström U (2015). Early career grants, performance and careers; a study of predictive validity in grant decisions. *Journal of Informetrics* **9** 826-838

Van den Besselaar P, Schiffbaenker H, Mom C, Sandström U (2016) *Explaining grant application success: the effect of gender*. Deliverable 6.1 GendERC project.

Van den Besselaar P, Schiffbaenker H, Sandström U, Mom C (2018). Explaining gender bias in ERC grant selection. *STI 2018 Conference Proceedings*, 346-352

Whitley 2000 Whitley, Richard. 2000. The intellectual and social organization of the sciences, 2nd ed. Oxford: Oxford University Press.

In: Proceedings 17th International Conference on Scientometrics & Informetrics ISSI 2019,
September 2-5, 2019, Sapienza University of Rome, Italy

Performance of Research Teams: results from 107 European groups

Ulf Sandström[1], Peter van den Besselaar[2]

[1] *ulf.sandstrom@indek.kth.se*
KTH Stockholm (Sweden)
[2] *p.a.a.vanden.besselaar@vu.nl*
Vrije Universiteit, Amsterdam (the Netherlands)

Abstract[38]

This paper investigates what factors affect the performance of research teams. We combine survey data about the team with bibliometric data about the performance of the team. The analysis shows that teams with a few PIs perform better than single PI teams – of course controlling for team size. On the other hand, gender diversity does not have an effect on performance. The good news is that gender objectives can be realized, without any performance problem.

Introduction

A long discussion exists on the role of (gender) diversity on the performance of research teams. Teams gather resources and use them to publish papers. In that process of changing inputs into outputs, there are a number of factors that might have an impact on efficiency. In this paper, we include a considerable number of team characteristics, based on a survey among members of some 100 teams, and on bibliometric data covering those members. The resulting set of data is a rich source of analysis. We investigate whether gender diversity has a role, whether context factors like type of organization or national research systems have a role, whether team composition would change the output levels, and we look into team dynamics.

Team diversity is a complex question and it is not obvious how to do research concerning this topic. A useful definition is provided by Jackson et al. (2003): “We use the term diversity to refer to the distribution of personal attributes among interdependent members of a work unit.” Both general similarities are of importance for a group of people working in non-routines task but at the same time, there is a need for cognitive diversity in order to bring in different perspectives for solving the immediate problems. One example of this is gender diversity: this specific dimension (also of political interest) has been debated for quite some years. Some authors claim that gender diversity would lead to more efficient use of scarce resources as well as to increasing the scope and impact of research (e.g. Bear & Wolley 2011; Nielsen et al 2017). Research has been targeted at different levels: individuals, e.g. increasing the number of women researchers; on organizations, e.g. developing and implementing gender equality plans; on research projects, e.g., integrating the gender dimension in research to increase quality and relevance. However, as today’s research to a large extent is based on

[38] We acknowledge the Gedii project partners for their contributions, especially in collection of data and the construction of survey questions. The research was funded by the Gedii project (author 1) and the GRANteD project (author 1 and 2).

collaboration in teams, the effect of gender diversity on research can probably be expected to be part of the effect of team composition, team culture, and team dynamics.

One of our aims is to examine claims that gender diversity is beneficial to science and research, using data on gender diversity in research teams and on research performance at the team level. We focus on four aspects of research performance: field adjusted volume of publications (total production) as well as productivity per senior, and field normalized citation-scores (totals and per senior) as a marker of impact. In general, we take a variety of team characteristics into consideration: team composition, team dynamics, team culture, and a variety of dimensions related to gender diversity.

The paper is organized in the following way. *First*, we provide an overview of research on team research and small groups, a body of work that has examined how to make teams more effective. *Second*, we outline our methodological approach. We provide an account of how the bibliometric datamining and data analysis was conducted. Also, the variables used are described in more detail. *Third*, we discuss the results of our models including team characteristics and research performance. We conclude by a discussion of the significance of our findings. Last but not least we discuss the limitations of our work and suggest further research avenues.

Research teams and team science

In the analysis we identify three different types of teams: 1) the PI-led group with a number of post doc's and PhD's; 2) the research teams with about 2-4 seniors and assisting personnel including post docs and PhD's; 3) team science based on large-scale initiatives or consortia's of research teams (Stokols et al. 2008). In this paper we intend to focus on the two former types of research teams and will try to keep the third, "team science", questions aside as they have a slightly different focus: "Most science teams and larger groups are geographically dispersed, with members located across multiple universities or research institutions" (NRC 2015).

Why do researchers team up in more or less stable groups? An increased need for channels for knowledge flow between scientists is the main cause why collaboration is becoming more important over time since decades ago (Adams et al. 2005). In that situation there are two different *strategies* that can be applied for research teams: First, to keep the group small (PI-led, one senior) and collaborate on papers nationally and internationally. Second, another strategy would be to join forces with a small number of qualified seniors who hold complementary competencies/skills and at the same time similar views and cognitive capabilities. To find these attractive components in long-distance collaborations can be costly, and, of course, to find these colleagues for co-location (same city) might include search costs and need for mobility of staff.

Many factors of this type have an influence on team performance, as a recent review of team science (Hall et al. 2018) shows, but there is a low level of precision; and not at all clarified how to operationalize the categories. De Saá-Pérez et al. (2017) points at two different paradigms in a research group or team research: on the one hand similar-attraction and on the other and the cognitive resource diversity (c.f. Hurley 2005). The former accentuates that similarities within the group might promote mutuality and cohesion within the group and that diversity would spur conflict and tensions. The latter paradigm, cognitive diversity, underscores that there should be unique combinations of cognitive resources brought together by team senior members. As often is the case there is need for a balanced approach, which here would translate to a combination of similarity and diversity but in different respects (ibid.). The similarity could be related to bio-demographic dimensions and diversity to the cognitive task-related issues, but there are many other possible combinations.

Vague concepts like the ones mentioned here are often of the type that they hardly can be operationalized, as they include parts that can be interpreted from several perspectives. This has spurred contradictory findings (van Knippenberg & Schippers 2007). We propose for the "balanced version" of diversity to use the concept of inwardness ("internal network density") which is the ratio between the number paper fractions from the own team in relation to all fractions within the total paper network. With this concept, we cannot distinguish between similarity and/or diversity but we can build an indicator which is sensitive to the capability of the group to contribute to larger parts of papers without the help from outside. Probably, this demands some of the features from the similarity side, to ease the collaboration, but it also makes it necessary to understand in the team how different competencies contribute to the cognitive tasks the team is working on. Hence, the more the groups have of inside competence and capability the more the team produces by itself, but that is dependent on a similar vision and a similar overall background of understanding within the specialty that is under investigation, as well as a willingness to collaborate with others outside of the team.

Connected to this inwardness is another important factor: seniority which gives the basis for richer research experience, better cognitive resources, and levels of prestige; all factors instrumental for the production of knowledge. Seniority might also be needed for even more advanced levels of research productivity (Hinnant et al. 2012).

There is quite convincing evidence that team size has an effect, mostly in the form a 'critical mass' thresholds and taking the form of an inverted U-curve-pattern: productivity seems to rise with increasing size of team up to about six or eight persons, above which there is usually little or no extra gain per researcher (Von Tunzelman et al. 2003; Verbree et al. 2015; De Saá-Pérez et al. 2017)

Several other bio-demographic variables affect research performance and they almost inevitably intervene to complicate or twist any simple relationship between size and efficiency. The age structure, of both individuals and institutions, is often found to be relevant to research performance. (Von Tunzelman et al. 2003; Verbree et al. 2015). Career phases also matter. Although smaller groups produce less output about a more restricted range of research topics, they are easier to manage— especially for less-experienced group leaders—and the coordination costs to organize scholarly communication and collaboration among group members are much lower (Carayol and Matt 2004, 2006; Heinze et al. 2009).

These team composition factors are complemented by factors consisting of the *context* for team activities. That could be the larger organization type (university, public research institute, company etc.), but also the fact that research groups are embedded in a national science system which affects important aspects of research and differ between national systems, as do governance and regulations which in turn influence the performance levels (Bonaccorsi and Daraio 2005).

Team dynamics concerns time-dependent factors like average team tenure (time in the team) reflecting staff mobility, the share of temporary contracts, and age of the personnel. These are quite straightforward variables and it is easy to see the influence of time for developing a common team culture and a similar cognitive understanding for how and with what different members can contribute to the team production. Kozlowski and Ilgen (2006) show that social interaction, sharing of perspectives, and collective sense-making might have a considerable effect on team performances. We introduce a number of variables for *team culture* in order to test whether these are important for team performance, and we keep in mind that these factors also coincide with aspects of team composition.

One topic in the literature is the role of gender diversity on team functioning and performance. On the whole, there is no consensus yet from this body of work on whether gender diversity has a positive effect on group performance or not. Nielsen et al. (2017)

describe several experimental research items on team science pointing towards a positive productive team mechanism in problem-solving due to diversity. However, as Campbell et al. (2013) make clear, there are several and quite differing results when it comes to actual performance. In fact, many results find a negative effect on performance (Bowers et al 2000; Stewart et al. 2006; Webber et al. 2001).

Where a positive relationship is found, this is attributed to gender diversity allowing for more complex tasks to be solved, for more collective intelligence to support the team, or for more social sensitivity. Recent studies based on topic modeling and text-mining indicates that gender diverse groups seem to have a higher propensity to search outside of the box (see Nielsen et al. 2017 for refs). Likewise, Joshi (2014) has done compelling research showing that recognition of team member competence differs in male-dominated compared to groups dominated by the other sex. These results and reasoning are transposed and used in policy documents. For example, the European Commission states that "*European research still suffers from a considerable loss and inefficient use of highly skilled women*" (e.g. European Commission 2012 page 12) that hinders both the quality and relevance of research. Calls are made for the potential of women in the scientific workforce to be taken on-board and set in motion, as a strategy to make the most of hitherto unused competent researchers.

Furthermore, it remains an issue how to measure the performance of teams. Here we focus on the scholarly output and measure this using bibliometric data – which is at the team level an accepted way of performance measurement (Hinnant et al. 2012). However, what indicators should be used is still open for discussion: here, we introduce an innovative method for this which takes care of differences in means of production between areas (Koski et al. 2016; Sandström & Wold 2015).

From all the factors that may influence team performance, we use in our view the most important subset: we hypothesize that team performance is influenced by some contextual variables, by team composition, and team culture, by team dynamics, and by gender diversity in teams. Several variables we use are new, such as the overall indicator for gender diversity, but also the variable "coverage of needed skills", i.e. inwardness. The latter refers to how complete the team covers skills needed, and could be seen as measuring the level of independence from the environment. The variable "team dynamics" measures the stability of team membership. Not included in our model are the processes leading to the establishment of teams. However, we use "objective" data based on author address collected from the bibliometric files per individual/team in order to categorize the team as co-located (in the same city) on the one hand and nationally or internationally dispersed. This together with information on cohesion (coverage of skills needed for article production) we would guess that set of data we use constitute a more trustful and comprehensive data set than survey data only on meetings and communication within the team.

This leads to the following model:

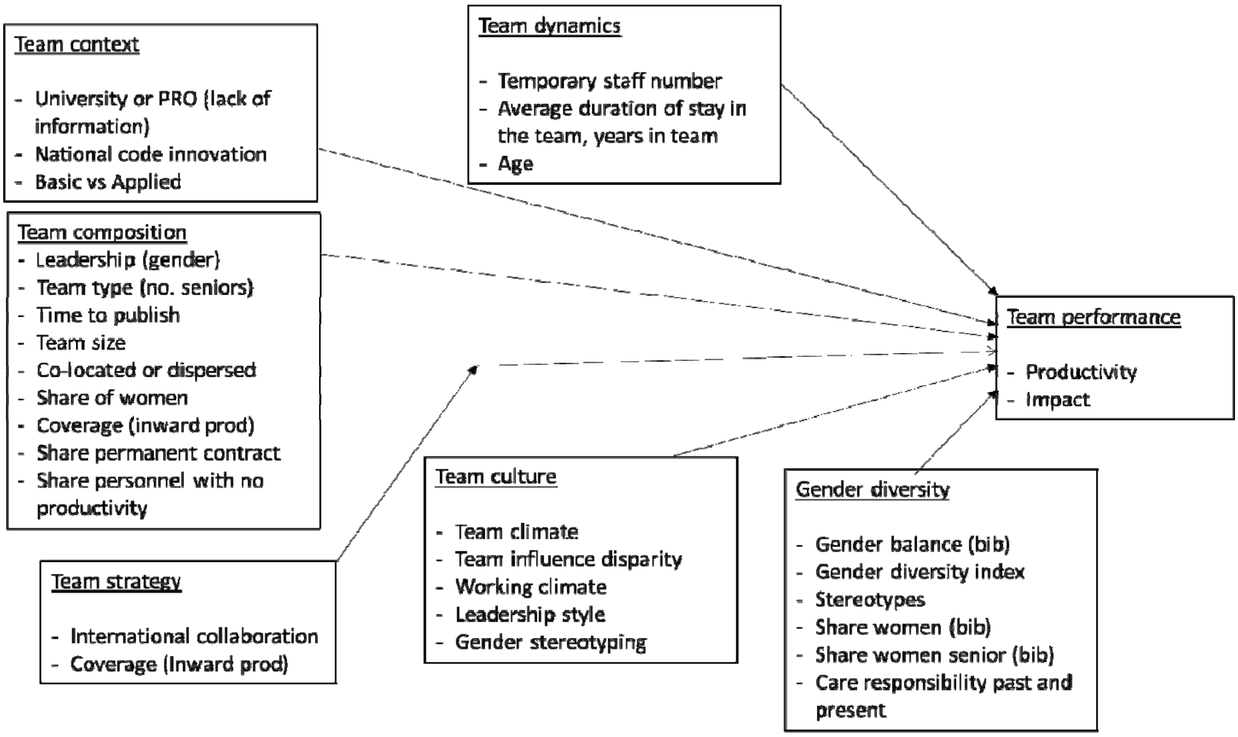


Figure 1: The model with its different components.

Methods and data

The initial sampling strategy was to focus on Transportation Studies and Biomedical Engineering. We retrieved from the Web of Science (WoS) all authors who have at least three publications in the period from 2011 until 2016 within the two subject categories. Based on this information all their publications in WoS were retrieved in order to achieve information about the e-mail address for potential research group leaders. These procedures were the major part of the recruitment process, and this takes us immediately outside of the two starting subfields, actually all fields are activated in this study. The survey is based on responses from seventeen EC countries. Altogether, 159 teams participated in the survey, representing 1,357 individuals, but due to applied rules for selection, there are 107 research teams representing 1,272 individuals in the final round for analysis. Due to missing values, 94 teams were used for the analysis.

The sample of research teams which participated in the surveys is based on self-selection, with oversampling of transportation and medical engineering, and of highly productive researchers. So, the sample consists of research-intensive productive teams. There were four contacts with the teams: First to ask them to participate, and to provide a list of team members. Second a survey among all the members of the teams, with a response rate of 77%. The survey included scales measuring gender stereotyping, diversity climate (Settles et al. 2006), team influence (Curşeu and Sari 2015), team climate (Anderson and West 1998), and team leadership (Berger et al. 2011). In addition, items regarding team communication (Pinto and Pinto 1990), care responsibilities, mentoring, working conditions, science communication activities, as well as acquired funding, were included. Thirdly, the bibliometric data were collected and sent to the contact person for validation. Fourthly, a short survey was sent to the team leader asking for information regarding the team such as its founding year, co-location of members, working methodology, size, and gender composition.

This study is, to our knowledge, one of the few that have taken the opportunity to combine survey data with advanced bibliometric performance indicators. Our design is similar to Verbree et al. (2015), but that study uses the research leader as an informant for the group which may affect the quality of the measurement.

Variables in the model

We have chosen to use four *dependent variables* in the analysis: 1) *Production* which covers the last 5 years of publications from the whole team no matter where their publications were produced. Five years to facilitate for a comparison between groups no matter when they were started or when members joined the team; 2) *Productivity* which is in this case FAP (Koski et al. 2016) divided by the Number of bibliometric seniors; 3) *Impact* is the influence over subsequent literature based on FAP figures but with Percentile Model applied (for further explanation see Sandström & Wold 2015), and. 4) *Impact per senior (bib).* More on the calculation of the respective indicators are given in a later section (see the headline: Dependent variables – the performance indicators). In this paper, team role has not been taken into consideration. This implies that everyone that was assigned membership at the moment of initial recruitment (which includes administrative personnel, master students etc.) affect the averages for all team level data from the survey. The following *independent variables* were used:

Team composition is measured through various variables. 1) *Team type* is based on number of senior researchers with a bibliometric profile over at least five years. We distinguish two types, namely (i) single leader teams; (ii) teams with 2-4 leaders. 2) *Coverage of skills* needed also named *inwardness* and measured as the share of publication fractions within the group related to all publications (bibliometric variable). 3) *Mean time available to publish*/patent (based on the survey). 4) *Team size* (provided in initial contact). 5) The percentage of women in the team (average of the scores of the surveys). 6) The share of senior team members.

Team culture has several dimensions, measured through the team member survey: 1) *Team climate*. 2) *Team influence (*disparity). 3) *Working environment climate*. 4) *Leadership style*. 5) *Gender stereotyping*.

Team dynamics is hard to establish using cross-sectional data but we use a few proxies. 1) Number of staff with a *temporary* contract. 2) Average *duration of stay* in the team. 3) Average *age* of team members.

Team context may affect performance, such as the 1) the *national context* – here measured as a dichotomous variable separating high performing and less high performing science countries. 2) The type of *organization* they are embedded in (University versus Public Research Organizations). 3) The level of *co-location* with three values: co-located, nationally dispersed, and internationally dispersed. 4) Applied vs. basic research, based on the categorization of ScienceMetrix (http://science-metrix.com/?q=en/classification).

Gender diversity is measured in several ways. 1) Firstly gender balance in team membership, using the Blau index (Blau 1977). 2) Secondly, a composite indicator was developed: the Gender Diversity Index (GDI), covering gender processes along seven diversity aspects: age, education, care responsibilities, marital status, team tenure, seniority and type of contract (Humbert & Günther 2018).

Method

Because of overdispersion of the dependent variables, we use from generalized linear models the negative binomial version. Multi-collinearity can represent a significant source of error in modeling. An indication for this is the strong change in regression coefficients when including new variables in the model – which is not the case. We then use the independent variables block by block (see Figure 1). After having done this, we keep from each block the (marginally) significant variables (p <0.20) for the final model. See Table 4 for the results.

Results

Team production: Production increases with teams that include 2-4 seniors that have a bibliometric profile covering several years of the period. Having a single PI-led team is clearly negative for production. This can be understood in terms of complementary skills in the team which is accentuated by the inward factor. Hence, in most cases, we find that higher levels of inwardness, greater coverage of needed competencies in the team makes it easier to publish papers, and not being dependent on colleagues from outside. Being able to do more inside the team increases paper production to higher levels. Overall production becomes better.

Regarding the variables **Gender Diversity Index** and Gender Balance, it turns out that both variables have a weak or no significant effect. Other factors are stronger and more important for the production figures. Among them, we should not forget the time devoted to publishing or patenting activities. As reported by the respondents this has a significant effect on production. Of course, this is a trivial result but might also be understood as a validation of the survey instrument as such. That production becomes higher with more senior members is likewise trivial, here it serves as confirmation of the overall design.

Team productivity: Productivity per senior is a more relevant indicator of performance. We have worked from the hypotheses that seniority is a crucial factor and that there is a need for seniors with a consecutive production of papers in Web of Science, i.e. in international scientific journals. However, the implication is not that more junior personnel also contribute in the same way to productivity. As shown in Table 4 this, team size is significantly negative, thereby indicating that the theoretical prediction seems to be corroborated. Having more personnel in the learning process might take down per senior productivity.

Team type is for productivity a significant *and* positive variable. Teams with two, three and four seniors have a significantly higher output per person. These groups might have more efficient team communication and a better understanding of the common objectives which translates into productivity. Again, team coverage of needed competencies (inwardness) is significant and contributes strongly to productivity. Unfortunately, we do not have communication patterns between seniors as a question in the survey, and therefore we use inward as an indicator for several processes of that type.

Surprisingly, *team culture* and *team dynamics* seem unrelated to productivity. Another slightly negative influence is the percentage of women in the team. It is not a significant factor but this result goes well together with a long series of research and this productivity gap can actually be traced back to the 1930s according to Cole and Zuckerman (1984). Since the 1930's it seems to be the case that women produce two-thirds only of what men produce in general (p.221). These results were confirmed by Xie & Schauman (1998) for the period 1960's to the 1980's. Recently Van den Besselaar & Sandström (2017), based on a Swedish dataset, and showed that these patterns were still at hand also after the millennium: "women are vastly underrepresented in the group of most productive researchers". Looking into the details we find that 80% of the highly productive individuals in the sample (having more than FAP points five times over the normal) are male researchers.

Table 4. What influences research performance?

		Production	Productivity	Impact	Impact/senior
		Beta	*Beta*	*Beta*	*Beta*
(Intercept)		7,200***	6,047***	8,660***	8,014***
Org type-	Univ	0,16 (ns)			
	PRO	0,17 (ns)			
	Other	0#			
Nation	high	0,22 ns			0,471**
	low	0#			0#
Team Type 1		-1,1***	-0,9***	-1,09***	
Type 2		0#	0#	0#	
Geography	Co-located	0,1 (ns)			
	Distributed	0#			
Orientation-	Applied				
	Basic				
Team size			-0,04**		
Senior(bib)					
Time to Pub		0,38***	0,352**	0,612***	0,378**
Inward Pub		4,05***	2,711***	2,822***	1,393
Women share			-0,05 ϕ	0,082 ϕ	-0,595 ϕ
PowerInfluenceDisparity					
Team Climate					
Leadership Style					
Working Climate					
Gender Balance		0,18 (ns)	0,086 (ns)	0,63 ns	
Gender Diversity Index		0,04 (ns)	0,408 (ns)		
Stereotypes			0,367 (ns)	0,101 ns	
Temporary staff		0,03 ϕ		0,02 ϕ	0,046 ϕ
Yrs in team		0,00 ϕ	0,015**	0,001 ϕ	
Age personnel					
Experience in area(yrs)					
Observations		86	86	86	86
Model fit(score/df)		0,923/61	0,946/88	1,096/71	1,289/87
AIC		1558	1505	1825	2040

Variables not included in this table are found (marginally) significant in the analysis per analytical dimension. Plus all gender diversity-related variables (also when not at all significant in the analysis per dimension).

# When dummy variables are applied redundant variables are set to zero.

Significance levels ns = non-significant, ϕ < .20, *<.10, **<.5, ***<.01%

Note: Based on dataset Müller et al. (2019).

Team Impact: The so-called percentile model applied here takes citation impact into account. This indicator expresses the total production and impact in one score. The pattern that evolves from Table 4 is very much the same as the one based on the Productivity indicator. However, there are differences and there are two factors that should be stressed in this context: First, team size is not significantly related to impact, and clearly the female share of the team is as a factor unrelated and not significant. The latter result is also in line with results from the literature. Van den Besselaar & Sandström (2017) showed that women are not among the high profiles of impact, but in the regions where women have the same number of papers the share of top papers are about the same as for men, i.e. no difference in impact and there are a several of investigations that have given evidence on this point.

Impact per senior: When impact is investigated as impact per senior (bib) a slightly different result emerges. The model is statistically weaker, and fewer variables are significant.

Findings show that team type no longer influences the average number of impact points, but overall the same pattern emerges again concerning what factors that have an effect on citation performance. Female share is negatively affecting the performance, inward is only marginally operative and time for publication is strongly and positively significant. Interestingly, the variable "Nation" now comes back as a strong factor; highly innovative countries – according to the EC indicators - seem to produce more influential research which is taken up and discussed in subsequent work by colleagues. In that respect, the Nordic countries and north-western Europe seems to perform better than southern countries. Here we should remind ourselves about the self-selection effects for the representability of the sample. This is important as several of the groups from the second round happen to be high performing groups.

Conclusions and discussion

The overarching research question for this paper is whether gender diversity makes a difference in the shaping of research teams performance? To answer that question three types of data were used: 1) a list of members of research teams, 2) survey data from these groups (77% response rate); 3), bibliometric data covering all (100%) members listed as participating in the teamwork.

We developed a model and consisting of different variables that could be hypothesized to affect performance: we divided the variables into different blocks representing a multitude of aspects of team research that we have found to be of interest based on literature studies. The model includes team context, team composition, team culture, team dynamics, and gender diversity. The latter has been a general point of departure for the project in the context of which this paper has been written. However, *the gender variables seem to be without significant effect*, other variables are more important for the dependent variables which cover production, productivity, and impact. This holds for the GEDII composite gender diversity indicator – which may due to the problematic nature of composite indicators – but also for other gender-related variables.[39] One gender relevant variable, well known from earlier research, negatively affects productivity in research teams: the share of women in the team.

What is important for creating successful research teams? Our findings indicate firstly that teams with a few experienced researchers with seniority tend to have greater production of papers than teams of single PIs. Such combinations are crucial for producing high-level output. Secondly, we found that team composition matters. When the senior team members complement each other in terms of skills and resources (and there are components of similarities, e.g. in the understanding of objectives, of how to handle research questions), the effect on performance is positive. Basically, this is what we find with the coverage or inward factor.

The sample did not include many groups with an interdisciplinary composition of competencies (measured as publication activity in different ESI fields), but many teams do have variations of specialization, so-called "small IDR". For example, in a team focusing on materials for dentistry, one of the senior members had a main specialization in materials science, but a second one in dentistry, whereas another senior member had most publication in dentistry, but material science as the second specialization. Probably this can be interpreted in the following way: the constellation is built on a common understanding of a research problem and competencies are gathered around this problem, some with a more emphasized grounding in the first ESI field and some with a concentration in the second ESI field. They

[39] We tested this (not reported here).

can do a lot of papers together; they have a mutual understanding and they share a vision (Heinze et al. 2009). Thirdly, also other trivial characteristics matter, e.g. time for writing publications should not be forgotten.

The next question is what factors are marginally or not important for production and productivity? From earlier research, we expected that gender diversity wouldn't be a game changer, but we wanted to know whether it would be possible to find something new using an innovative composite indicator. This is not corroborated. Are there other indications of change? We have a high number of female group leaders (34 out of 107), so that may be on its own already indicating a change in teams: what is sometimes called "transformational leadership" (Jeong & Choi 2015). The Gedii project developed a composite indicator which covers many dimensions, not only gender diversity but also functional and educational diversity. We might have to wait for more detailed studies but the results from the analysis are that these factors are of less importance both for the production as well as for the productivity. So, gender diversity does not result in higher performance, but on the other hand, it also does not imply lower performance, despite significant performance differences at the individual level between female and male researchers.[40] For impact, influence over subsequent research, we find that basic structures are still more important than diversity measures.

Also, the type of organization does not seem to matter. Independently of whether the team is at universities or public research organizations (institutes) they seem to be able to produce the same level of output per senior as long as there is room for an orientation towards research activity. Whether the team is co-located or dispersed over a country or internationally is of low importance for scholarly performance. Another finding is that the orientation towards basic research or applied research is not at all important for production and productivity. Not even the factors in the team dynamics block, such as age, number of temporary staff and team tenure, seem to be of determining performance. Finally, team culture variables did not have a significant effect.

In all, we conclude with the following: For productivity and citation impact there is one thing that is crucial and that is to construct research teams that have the capacity to combine and use different competencies in a creative way.

Limitations and future research

1) The survey is a cross-sectional analysis at a certain moment in time (2017), but the publication and citation analysis build on more "historical" data from 2012 up until 2016, for those we know were in the team in 2017. So, we perform a cross-sectional analysis, where one would prefer longitudinal data: do changes in team characteristics precede changes in performance. On the other hand, performance and team characteristics are expected to be rather stable during a short period, as is considered here. The same holds for the bibliometric performance, as good groups more easily attract good people than not so good groups. Furthermore, the performance data build on a fairly long period of time as we need to get rid of the year-to-year fluctuations on an individual level. That also makes that quite a high proportion of the team members do have bibliometric contributions. Finally, half of the sample would qualify as Top10 % best researchers in Sweden, i.e. the best 4,700 researchers out of 47,000 in total. That the less good groups would change considerably in the next few years is possible but definitely not expected.

[40] If gender diversity does not influence team performance, it may influence within team performance differences. For example, in more gender balanced teams, performance differences between male and female researchers may be smaller than in other teams.

2) We do not control for which phase of development the research group is going through at the moment of data gathering (2017). Any type of team was considered as interesting and we did not investigate when the group was actually started and how their funding and financing had developed over time. This is a factor that eventually could supersede many of the factors controlled for in this paper. Entrepreneurial research leaders with new funding arrangements available create by themselves an atmosphere of creativity and productivity. So in future studies, these aspects might be taken into account.

References

Adams JD, Black, GC Clemmons JR; Stephan PE (2005). Scientific teams and institutional collaborations: Evidence from US universities, 1981-1999. *Research Policy* **34** (3): 259-285.

Anderson, N., & West, M. A. (1998). Measuring climate for work group innovation: development and validation of the team climate inventory. *Journal of Organizational Behavio*, **19**, 235–258.

Bear JB & Wolley AW (2011). The Role of Gender in Team Collaboration and Performance. *Interdisciplinary Sciece Reviews* **36** (2): 146–153.

Bell, ST; Brown, SG; Colaneri, A & Outland, N (2018). Team Composition and the ABCs of Teamwork. *American Psychologist* **73** (4): 49–362 http://dx.doi.org/10.1037/amp0000305.

Berger, R., Romeo, M., Guardia, J., Yepes, M., & Soria, M. A. (2012). Psychometric Properties of the Spanish Human System Audit Short-Scale of Transformational Leadership. *Spanish Journal of Psychology* **15** (1) 367–376.

Blau PM (1977). *Inequality and Heterogeneity*. New York: Free Press.

Bonaccorsi A & Daraio C (2005). Exploring size and agglomeration effects on public research productivity. *Scientometrics* **63** (1): 87–120.

Bowers CA, Pharmer JA, & Salas E (2000). When member homogeneity is needed in work teams: A meta-analysis. *Small Group Research* **31** (3): 305-3–27.

Campbell, L. G., Mehtani, S., Dozier, M. E., & Rinehart, J. (2013). Gender-heterogeneous working groups produce higher quality science. *PloS One* **8** (10), e79147.

Carayol N & Matt M (2004). Does research organization influence academic production? Laboratory level evidence from a large European university. *Research Policy* **33**(8):1081-1102.

Cohen SG & Bailey DE (1997). What makes teams work. Group Effectiveness Research from the Shop FIoor to the Executive Suite. *Journal of Management* **23** No. 3.239-290.

Cole JR & Zuckerman H (1984). The productivity puzzle: persistence and change in patterns of publication of men and women scientist. *Advances in Motivation and Achievement* **2**, 217-258.

Curşeu, P. L., & Sari, K. (2015). The effects of gender variety and power disparity on group cognitive complexity in collaborative learning groups. *Interactive Learning Environments* **23** (4), 425–436.

De Saá-Pérez P, Díaz-Díaz NL, Aguiar-Díaz I, Ballestreros-Rodríquez JL (2015). How diversity contributes to academic research teams performance. *R&D Management* **47**, 2: 165-179.

European Commission (2013). *Research and Innovation Performance in EU Member States and Associated Countries: innovation union progress at country level*. <ec.europa.eu>.

European Commission (2012). *Communication from the Commission to the European Parliament, the Council and The European Economic and Social Committee and the Committee of the Regions: A reinforced European Research Area Partnership for Excellence and Growth* (European Commission, Brussels).

Heinze T, Shapira P, Rogers JD & Senker J (2009). Organizational and Institutional Influences on Creativity in Scientific Research. *Research Policy* **38** (4):610-623.

Hinnant CC, Stvilia B, Wu SH, Worrall A, Burnett G, Burnett K, Kazmer MM, Marty PF (2012). Author-team diversity and the impact of scientific publications: Evidence from physics research at a national science lab. *Library and Information Science Research* **34** (2012) 249–257.

Humbert AL & Günther E (2018). *Measuring gender diversity in research teams: methodological foundations of the Gender Diversity Index*. Deliverable D3.2 <www.gedii.eu>.

Hurley SK (2005). The Compositional Impact of Team Diversity on Performance: Theoretical Considerations. *Human Resource Development Review* **4**, 2 June: 219-245.

Jackson SE, Joshi A & Erhardt NL (2003). Recent research on team and organizational diversity: SWOT analysis and implications. *Journal of Management* **29**: 801–-830.

Joshi A (2014). By whom and when is women's expertise recognized? The interactive effects of gender and education in science and engineering teams. *Administrative Science Quarterly* **59** (2), 202-239.

Jeong S & Choi JY (2014). Collaborative research for academic knowledge creation: How team characteristics, motivation, and processes influence research impact. *Science and Public Policy* **42**: 460–473

Koski T, Sandström E & Sandström U (2016). Towards field-adjusted production: Estimating research productivity from a zero-truncated distribution. *Journal of Informetrics* **10** (4): 1143–1152.

Kozlowski SWJ. (2012). Groups and teams in organizations: Studying the multilevel dynamics of emergence. In A.B. Hollingshead and M.S. Poole (Eds.), *Research Methods for Studying Groups and Teams: A Guide to Approaches, Tools, and Technologies* (pp. 260–283). New York: Routledge.

Müller J, Busolt U, Callerstig A-C, Guenther EA, Humbert AL, Klatt S, Sandström U (2019). *GEDII Survey on Research Teams Dataset.* https://doi.org/10.5281/zenodo.2545196.

Nielsen MW Alegria S, Börjeson L, Etzkowitz H, Falk-Krzesinski HJ, Joshi A, Leahey E, Smith-Doerr L, Woolley AW, & Schiebinger L (2017). Gender diversity leads to better science. *PNAS* **114** (8):1740-1742. (opinion paper).

NRC (2015). Committee on the Science of Team Science; Board on Behavioral, Cognitive, and Sensory Sciences; Division of Behavioral and Social Sciences and Education; National Research Council; Cooke NJ, Hilton ML, editors. Washington (DC): National Academies Press (US); 2015 Jul 15.

Sandström U & Sandström E. (2009). The Field Factor: towards a metric for Academic Institutions. *Research Evaluation* **18** (3) 243–250.

Sandström U & Wold A. (2015). Centres of Excellence: Reward for gender or top-level research? In Björkman, & Fjaestad (Eds.), *Thinking ahead: Research, funding and the future*. Stockholm: Makadam Publ. (pp. 69–89).

Settles, I. H., Cortina, L. M., Malley, J., & Stewart, A. J. (2006). The climate for women in academic science: The good, the bad, and the changeable. *Psychology of Women Quarterly* **30** (1), 47–58.

Stewart GL (2006). A meta-analytic review of relationships between team design features and team performance. *Journal of Management* **32** (1): 29–54.

Stokols D, Hall KL, Taylor BK & Moser RP (2008). The science of team science: overview of the field and introduction to the supplement. *American Journal of Preventive Medicine* **35** (2 Suppl), S77–89.

Van den Besselaar P & Sandström U (2017). Vicious circles of gender bias, lower positions, and lower performance: Gender differences in scholarly productivity and impact. *PLoS ONE* **12**(8): e0183301. https://doi.org/10.1371/journal.pone.0183301.

Van Knippenberg D & Schippers MC (2007). Work group diversity. *Annual Review of Psychology* **58**, 515–541.

Verbree M, Horlings E, Groenewegen P, Van der Weijden I, Van den Besselaar P (2015). Organizational factors influencing scholarly performance: a multivariate study of biomedical research groups. *Scientometrics* **102**, 25-49.

Webber SS & Donahue LM (2001). Impact of highly and less job-related diversity on work group cohesion and performance: A meta-analysis. *Journal of Management* **27**: 141–-162.

Von Tunzelmann N Ranga M, Martin B, Geuna A (2003). *The Effects of Size on Research Performance: A SPRU Review*. SPRU Report, Sussex, U.K.

Wuchty S, Jones BF, Uzzi B (2007). The increasing dominance of teams in production of knowledge. *Science* **316** (5827):1036–1039.

Xie Y & Shauman KA (1998). Sex Differences in Research Productivity: New Evidence about an Old Puzzle. *American Sociological Review* **63** (6): 847-870.

STI2022
GRANADA | SEP 7+8+9

26th International Conference on Science and Technology Indicators
"From Global Indicators to Local Applications"
#STI2022GRX

Full paper

STI 2022 Conference Proceedings

Proceedings of the 26th International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Proceeding Editors
Nicolas Robinson-Garcia
Daniel Torres-Salinas
Wenceslao Arroyo-Machado

Citation: Möller, T., van den Besselaar, P., & Mom, C. (2022). What is researcher independence and how can it be measured?. In N. Robinson-Garcia, D. Torres-Salinas, & W. Arroyo-Machado (Eds.), *26th International Conference on Science and Technology Indicators*, STI 2022 (sti2292). https://doi.org/10.5281/zenodo.6975312

Collection: https://zenodo.org/communities/sti2022grx/

26th International Conference on Science and Technology Indicators | STI 2022
"From Global Indicators to Local Applications"
7-9 September 2022 | Granada, Spain
#STI22GRX

What is researcher independence and how can it be measured?[41]

Torger Möller[*], Peter van den Besselaar[**] and Charlie Mom[***]

[*] *moeller@dzhw.eu*
German Centre for Higher Education Research and Science Studies (DZHW), Schützenstrasse 6a, Berlin, 10117, (Germany)

[**] *p.a.a.vanden.besselaar@vu.nl; peter@vandenbesselaar.net*
Faculty of Social Sciences, Vrije Universiteit Amsterdam, Amsterdam, (Netherlands)

[***] *charlie@teresamom.com*
TMC Research, Middenweg 203, Amsterdam, 1098 AN, (Netherlands)

Introduction

„Independence‟ is a core criterion in selecting grant or job applicants. In this paper we develop – after a theoretical introduction – a few indicators for independence that can be calculated also for large datasets. After having done so, we apply the indicators to two cases of prestigious early career grants. We end with ideas for follow-up work to further develop the indicators.

Independence

For many selection procedures in science, such as the prestigious ERC starting grant and the German Emmy Noether program, independence is mentioned as a review criterion. But what does the concept of independence mean? We propose that independence relates with negative freedom (*freedom from*) and with positive freedom (*freedom to*), a distinction developed in the philosophical discourse on freedom of action.[42] The negative freedom is the *freedom from* influences, especially from other actors (personal or organizational) or constraints. This freedom is "negative" because it is undetermined what a person does with the *freedom from* and how he or she then acts concretely. This is where the term positive freedom comes in, as the freedom of a person to act (in a self-determined direction). Negative freedom can be understood as a precondition for positive freedom, as *freedom from* enables *freedom to*. The realized particular action makes other potential options no longer possible. In this respect, actions limit the scope of further actions. These *path dependencies* are not perceived by the self-determined actors as a restriction of their scope of action, but as a realization of their *freedom to* go their own way. Positive freedom in this sense does not mean to keep all paths open and to live in a free-floating unboundedness. Martin Heidegger, for example, distinguishes in this respect between a "freedom as unbound, *freedom from*

[41] This work was supported by grant nr. 824574 of the EC: The GRANteD project (2019-2023).

[42] The difference between freedom of will and freedom of action will not be discussed here. However, the concept of negative and positive freedom can also be applied to freedom of will.

(negative freedom)" and a "freedom as being bound to, libertas determinationis, *freedom to* (positive freedom)" (Heidegger, 1971, p. 106).

We apply the above distinction to the concept of researcher independence. In the following we distinguish between *independence from* other actors or constraints, e.g. the doctoral supervisor or the university in which one works, and an *independence to* develop one's own research agenda and to undertake that research.

In this sense, researcher independence means both breaking ties (*independence from*) and entering into new ties (*independence to*). This point is important to emphasize as independence should not be interpreted as the lowest number of ties. Being an independent researcher in a world of team science does not mean doing research alone, but having the *freedom to* choose the collaborations that are most appropriate for one's particular research agenda. Independence should not be confused with disconnectedness.

In the academic literature and in practical guidelines, various definitions of researcher independence can be found that more or less comprise the above distinction without making it explicit. For example, the Research Excellence Framework Guidance on Submission defines an independent researcher as someone who "undertakes self-directed research, rather than carrying out another individual"s research programme" (REF Guidance, 2019, p. 118). The definition implies that *independence from* "carrying out another individual"s research programme" is a prerequisite for *independence to* "undertake self-directed research".

Documents about academic career development refer to a process of *becoming independent* instead of *being independent*. A guidebook from a research-oriented university states: "Research degrees are training in independent research" (Code of Practice, cited by Guccione 2020). It could be summarized that from entering a university as a student, through the positions of a doctoral student, postdoc, senior scientist, and assistant/associate/full professor, the independence evolves and increases. We assume that successful academics show an increasing *independence from* and *independence to* at each stage, although independence above a certain level may not necessary always better (Gläser, et al., 2021). We agree with this view, but consider the distinction between *independence from* and *independence to* is useful to answer the question at what career stage what degree of independence is appropriate.

Summarizing, the distinction between negative and positive freedom (*freedom from* and *freedom to*) borrowed from philosophy can be profitably applied to the question of researcher independence. It focuses on the process of becoming independent, instead of being independent, which can be analyzed in terms of researchers' career development over time, in which a co-evolution of *independence from* and *independence to* seems to be most conducive. Furthermore, defining independence in this way shows that researcher independence is not contradictory to collaborative research and team science.

Earlier work

Elsewhere, we have developed the concept of independence in relation to the PhD supervisors, as this is the initial dependent collaboration relation that early career researchers engage in (Van den Besselaar & Sandström 2019). After graduating, the postdoc period is then seen as the period to start developing research independence, which should result in new collaborations and new research topics. We translated that concept in terms of indicators and tested those on a small sample of researchers. The following indicators were proposed:

1. Becoming independent means that the role of the former supervisors should become less prominent in the collaboration network of the early career researcher, which can be measured through the *eigenvector centrality* of the researcher and the *clustering coefficient* of the supervisor(s). This is in terms of the previous section: becoming "independent from" the supervisor's network and "independent to" create an own network.
2. Becoming independent also means the role of former *supervisor(s) as coauthor* declines – which can be measured by dividing the number of the publications of the researcher without supervisors as coauthor by all publications of the researcher. In terms of the previous section: becoming "independent from" the supervisor and "independent to" publish.
3. Finally, an independent researcher is expected to enter in *new research topics*, different from those of the environment where the PhD research was obtained. To calculate this, we retrieved all publications of the supervisor(s) and the researcher from WoS, did some topic mapping of the set of publications, and then calculated the share of independent topics. This is in terms of the previous section: becoming "independent from" the supervisor‟s themes and "independent to" develop an own research agenda.

A disadvantage is that the calculation of indicators 1 and 3 is time-consuming, and therefore we propose to use slightly different indicators.

Large scale indicators: concepts, data and methods

Independent publications: We produced a script that splits the publications of a young researcher into two groups: those with and those without any of the supervisors as co-author. The query basically looks in the author-name or Scopus authors-ID-field which contains the name or the IDs of all the co-authors. If one of the names or IDs belongs to one of the supervisors, this paper is counted as „dependent‟. In case none of the supervisor IDs is found in the mentioned field, the paper is counted as „independent‟. One can do this for full papers and for fractional paper counts, but that seems to make no difference. And it can be done for different periods in the career, in order to be able to investigate the development of independence over time. Basically, this leads to the formula:

Independence = (Number independent publications)/(All publications) (1)

Independent research topics: A second script was developed to find all research topics in which the former PhD student has published without any of the supervisors as co-author. Here we use Scopus Scival data, and one has the choice between about 40 fields of research, some 1,500 topic clusters, and about 96,000 topics[43]. We decided to do this analysis at the level of *topic clusters*, as topics are too fine-grained (as we would have more topics than papers in the analysis).

Technically, this means that we need to identify the set of *topic clusters* in which the former PhD student has been publishing, and then identify the subset of those topic clusters in which one or more of the supervisors have been publishing (with or without co-authoring with the former PhD student). The remaining subset (so the topic clusters in which the supervisors are not visible) can be counted as the independent topic clusters. However, one could argue that in cases where the former PhD student first published in a topic cluster without the supervisor as co-author, that those topic clusters should also be counted as independent. We indeed include both sets of topic clusters in the *independent topic cluster set*. We did not use a threshold for the period between the former PhD student entering a topic cluster and a supervisor entering that topic cluster. Again, this can be done for different periods in the career. The indicator would be:

[43] Source: https://www.elsevier.com/solutions/scival/features/topic-prominence-in-science

Independence = (Number of independent topics) / (all topics) (2)

We applied the derived indicators on two cases, both of early career researchers during the PhD period and the early postdoc period.

Calculating the indicators: an example from the Netherlands

We use a dataset of 899 researchers who received their PhD from the same Dutch university between 2000 and 2014, in fields where journal articles are the main output medium. Bibliometric data were also collected for 812 supervisors, this set of PhD students had on average 2.05 supervisors (sd = 1.24). We use Scopus/Scival data, because every record (paper) contains a field with *all author-identifiers*.[44] Another advantage of Scival is the *very detailed research topic classification*. The data were manually collected and cleaned.

We produced a script that splits the publications of a researcher into two groups: those with and those without any of the supervisors as coauthor. We then calculated the independence indicator for the PhD period[45], and for the whole career (up to 2018). The latter is important, as one may expect an optimal career pattern where the researcher is very close to (and dependent on) the supervisors during the PhD period, as that may lead to the strongest learning effect, and then becoming independent in the later part of the early career. Becoming independent too early may be disadvantageous. We also use fractionally counted papers.

A second script was developed to find all research topics in which the former PhD student has published without any of the supervisors as co-author. Also here we use Scival data on 1,500 topic clusters.

Table 1 shows that average independence in terms of *published papers* almost doubles from the PhD period to the later career, from 0.27 to 0.47. Independence in terms of *research topics* was very low during the PhD period (0.035), which is expected as the PhD period is about learning to do research and then one would expect a PhD student being embedded in the research lines of the supervisors. Topical independence increases strongly in the later career.

Table 1. Share of the independent papers and topics (the PhD period and career)

	Period	Mean	Std. Dev.	Min	Max	N
Papers (fractional)	PhD period	0.2750	0.3296	0	1	925
Papers (fractional)	Career	0.4574	0.3511	0	1	925
Topics	PhD period	0.0345	0.1186	0	1	925
Topics	Career	0.3352	0.3191	0	1	925

Table 2 shows, the correlation between the two career-indicators is moderate[6] (.577), but for the PhD period, the correlation between two indicators is rather weak (.266), suggesting that independence develops in two stages: first own papers, then in the later career own topics. „Own" here does not mean single authored, but independent from the supervisors.

[44] Several PhD students and supervisors have more than one ID, which we merged.

[45] Defined as the three years before until the three years after the year of obtaining the PhD degree. [6] See e.g.: https://www.statstutor.ac.uk/resources/uploaded/pearsons.pdf (Evans 1996)

Table 2. Correlations between the independent indicators

	Career: topics	PhD period: Papers (frac)	PhD period topics
Career: papers (frac)	.577*	.692*	.170*
Career: topics		. .417*	.248*
PhD period: papers (frac)			.266*

* Correlation is significant at the 0.01 level (2-tailed). N=925

Figures 1 and 2 show the frequency distributions of the independence indicators. (i) Quite a large share of the PhD students has an independence score of 0. (ii) The relative high frequency of the score 1 is due to the PhD students with only one publication, without supervisors as co-authors. (iii) Over time, early career researchers tend to become more independent, and the number of researchers that score 0 declines considerably, and all other values increase. However, quite a few remain rather dependent. This latter group may contain researchers that left the science system. Such an analysis could be done easily, using „stopping to publish" as indicator for leaving the research system.

Figure 1. Independence PhD period (papers)

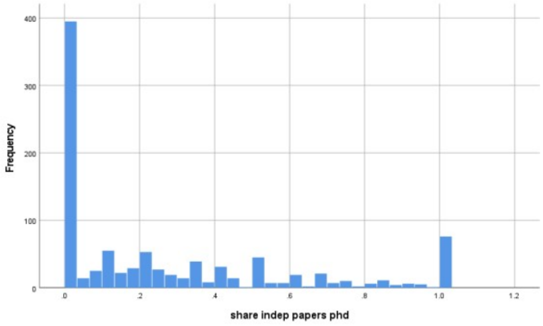
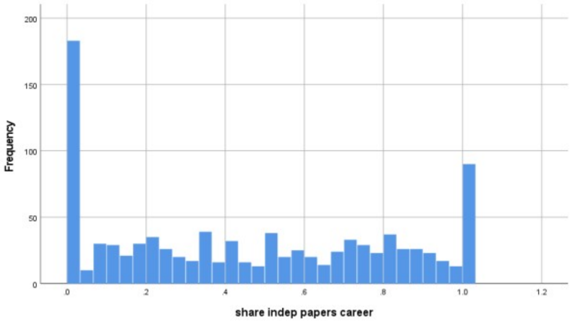


Figure 2. Independence career (papers)



Figures 3 and 4 show the distributions for the topics-based indicators. The figures show that topical independence is overall lower than paper-based independence, which is expected. This suggests that many of the papers not co-authored with supervisors still remain in the same topic clusters. Finally, and also as expected, independence increases strongly over the career (Figure 4 versus Figure 3) but many researchers remain in the same topics as their supervisors worked (which again includes those PhD students that left the science system after graduating).

Figure 3. Independence PhD (topics)

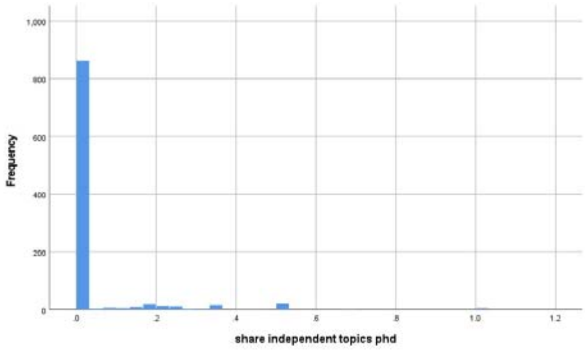
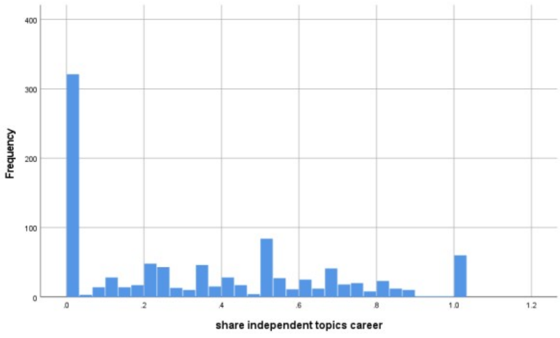


Figure 4. Independence career (topics)



This shows that_using Scopus/Scival enables the calculation of the independence indicators at a large scale. We do not enter here in the discussion whether the deployed *Scival topic clusters* (Scopus 2018) are accurate, as even when the topic classification is not perfect, it can be used as to calculate the field overlap between the early career researcher and the PhD supervisor. This avoids that we have to manually produce field maps for every individual PhD student and his/her supervisor(s), as was done in the earlier paper (Van den Besselaar & Sandstrom 2019). In the next section we apply the indicators on a second case, to explore whether independence explains partly grant success.

Applying the indicators, an example from Germany

Researcher independence is a central review criterion for the Emmy Noether Program of the German Research Foundation (DFG). The program is open to academics from all fields who have completed their doctorate with an outstanding result and submitted an excellent research proposal within four years after the date of dissertation. They should have published demanding articles in internationally renowned journals and have already proven their scientific independence and substantial international research experience in the postdoctoral period (Böhmer, Hornbostel & Meuser, 2008, p. 18). Our data cover funding decisions between 2000 and 2006 in biology, chemistry and physics, and consist of 328 Emmy Noether (predominantly male) applicants (table 5).

The success rate in the period under consideration is relatively high (56.7%, table 5), which is probably mainly due to the high formal entry requirements. After easing the rules, the success rate dropped to the current 20% (DFG 2021). In the basic population (table 6), the average age of the applicants was 32.6 years. The doctorate was awarded at an average age of 29.1 years. The applicants submitted their proposal 3.5 years after receiving their doctorate. Women are on average half a year younger at the time of their doctorate and half a year older at the time of their application.

Table 5. All applicants (population)

N=328	male	female	sum
granted	47.0%	9.8%	56.7%
not granted	32.6%	10.7%	43.3%
sum	79.6%	20.4%	

Table 6. Used sample (supervisors known)

N=107	Male	female	sum
granted	65.4%	7.5%	72.9%
not granted	23.4%	3.7%	27.1%
sum	88.8%	11.2%	

The doctoral theses were searched via the German National Library. Then the names of the supervisors were investigated, through (where available) (i) the metadata from the German National Library, (ii) names of supervisors from the online publications of the doctoral thesis or (iii) from an internet search of the applicants. The first two methods yielded hardly any hits, and also the internet search hardly found names of the supervisors, and if so, usually only that of the first supervisor. This leads to a selection bias, since persons who left the scientific system could not be found on the internet or did not have names of their supervisors on their homepages. All data required for the analysis could only be found for 107 researchers. Successful male applicants were overrepresented in this sample (table 6).

For those 107 applicants all publications were retrieved from Scopus by using the Scopus author IDs. When researchers have several IDs (Aman 2018; Kawashima & Tomizawa 2015), we merged the IDs. In the following analyses, the first (collaboration) *independence* indicator was applied (see above, both full and fractional count). The value of the independence indicator ranges from 0 (no independent publication) to 1 (all publications without the supervisor). The indicator is calculated for three periods: (i) publications up to and including the year of the PhD thesis, (ii) publications up to and including the year of application, and (iii) publications up to four years after application.

The researcher *independence* is lowest during the doctorate and increases step by step thereafter (fig 5). This seems to fulfil a central assumption of the independence model, namely that researchers become more and more independent over time. Furthermore, full and fractional counts hardly differ (Pearson‟s correlation: 0.95).

A logistic regression model was tested with grant success as the dependent variable and independence, gender and subject as independent variables. The findings suggest that independence has no effect on early career grant decisions (Table 6) even if it is a formal review criterion and *independence from* the supervisor increases over time (Figure 5).

Figure 5. Independence indicator for publications in three time periods

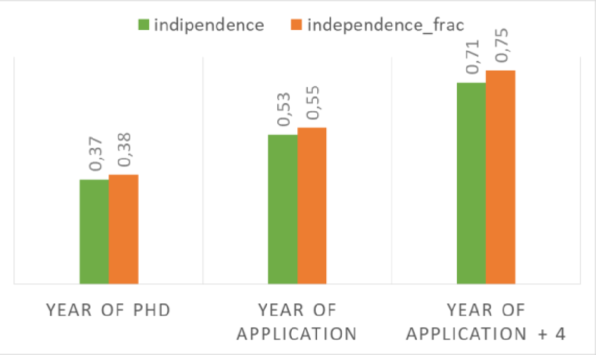


Table 6. Grant success by independence*, gender and subject** of the applicants

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.6014	0.5845	2.740	0.00615
Independence (full count)	-0.7082	0.7309	-0.969	0.33260
Gender female	-0.3082	0.6675	-0.462	0.64423
Subject Chemistry	-0.2865	0.5601	-0.512	0.60893
Subject Physics	-0.2627	0.5522	-0.476	0.63425

Logistic regression; *: full count & publications until the year of application;
**: Biology, Chemistry, Physics.

The case illustrates how the independence indicator can be used, although the above findings (table 6) are not yet final. In a follow up analysis, we will include other variables in the model that may influence the selection process such as past performance. Another issue that need to be solved is that we do not have all information about supervisors for all applicants, which leads to an overestimation of independence. This also affects the sample size and may lead to a selection bias.

Conclusions

The paper shows that large scale calculating of independence indicators is possible, using Scopus data. The main data-issue is identifying the PhD supervisors, which is not always easy. One solution might be to take the dominant co-author of the PhD period as the (informal) supervisor. The dominant co-author can be identified quite easily, and it is not unlikely that this is the formal supervisor.

With respect to the independence model and the independence indicator(s), we showed (Figure 5) that the *independence from* the supervisor (or other senior researchers) indicator is already a good indication of the way the independence of junior researchers advances. Other indicators could be developed, to measure different dimensions of independence: *(in)dependence from* research infrastructures / labs, and *independence to* enter other infrastructures or to create an own lab. Also the *independence to* create own collaborations through international mobility could be measured. Finally, for a reliable measurement, we may need an underlying variable that is based on such larger set of independence related indicators.

Acknowledgement
The work for this paper was supported by grant nr. 824574 of the EC: The GRANteD project (2019-2023).

References

Aman, V. (2018). Does the Scopus Author ID Suffice to Track Scientific International Mobility? A Case Study Based on Leibniz Laureates. *Scientometrics, 117*(2), 705–720.

Böhmer, S., Hornbostel, S., & Meuser, M. (Eds.) (2008). *Postdocs in Deutschland: Evaluation des Emmy Noether-Programms.* Bonn: iFQ.

Gläser, J., Ash, M. G., Bunstorf, G., Hopf, D., Hubenschmid, L., Janßen, M., Laudel, G., Schimank, U., Stoll, M., Wilholt, T., Zechlin, L. & Lieb, K. (2021). *The independence of research - a review of disciplinary perspectives and outline of interdisciplinary prospects* [Preprint].

Guccione, K. (2020, 12 November). Can We Define Research Independence? Can We Teach It? *A Community Blog, on Doctoral Supervision Relationships and Pedagogies*.

Heidegger, M. (1971). Schellings Abhandlung über das Wesen der menschlichen Freiheit (1809). Niemeyer.

Kawashima, H., & Tomizawa, H. (2015). Accuracy Evaluation of Scopus Author ID Based on the Largest Funding Database in Japan. *Scientometrics, 103*(3), 1061–1071.

van den Besselaar, P., & Sandström, U. (2019). Measuring Researcher Independence Using Bibliometric Data: A Proposal for a New Performance Indicator. *PloS ONE, 14*(3). https://doi.org/10.1371/journal.pone.0202712

Gender and Merit in Awarding Cum Laude for the PhD Thesis[46]

Peter van den Besselaar[1] and Charlie Mom[2]

[1] *peter@vandenbesselaar.net*
Prof. Em., Vrije Universiteit Amsterdam, (Netherlands)

[2] *charlie@teresamom.com*
TMC Research, Middenweg 203, 1098 AN Amsterdam (Netherlands)

Abstract

In the Dutch academic system, PhD theses can be awarded with 'Cum Laude', and only some 5% of all PhD graduates receive this selective award. In this paper, we investigate whether there is gender bias in awarding cum laude, using data from a major Dutch research university. We measure the quality of the PhD thesis using bibliometric data. A main result is that the set of PhD theses receiving cum laude do on average not have a higher quality than the best theses not getting cum laude. A second main result is that, after controlling for the quality of the PhD thesis, women have a substantially lower probability to be granted cum laude. These results strongly suggest that the distribution of awards may contribute to gender bias. These findings lead to a strong doubt about the adequacy of the procedures leading to awarding cum laude for the PhD thesis.

Introduction

Within the Netherlands higher education system, awarding a PhD thesis with the qualification 'cum laude' (with honors; abbreviated as CL) does not happen often. It may differ between universities and faculties, but the overall 4.3% of our case seems a good estimate. It has been shown that men are awarded *cum laude* for their PhD thesis more frequently than women are, and this seems to be the case in all Dutch universities (De Bruin 2018). This is not unimportant, as cum laude awards are of considerable help in the early academic career (Brouns et al. 2004). This may even have increased in the more recent period (Reschke et al. 2018).

In this report we try to answer the question is whether this observed gender difference is a form of gender bias. Is the difference between men and women because the peer reviewers adequately pick out the best dissertations and that there happen to be more male than female authors among those? Or is the decision biased, and in this case especially gender biased, and what mechanisms might cause such bias? As gender differences in *cum laude* do not fit with the generally accepted policies within Dutch higher education to promote gender equality, an explanation for the existing disparities is needed.

We analyze data for one of the Dutch universities in order to contribute to the understanding of this phenomenon. Is the gender difference in *cum laude* an effect of gender bias, or do peer and panel reviews correctly reflect the quality differences of the dissertations?

There are hardly studies available on gender bias in cum laude awards. But there are studies on other types of awards and prizes, all suggesting the existence of gender bias (Lincoln et al. 2012; Melnikoff & Valian 2019; Silver et al. 2018). Furthermore, men also receive other signs of prestige more than women do, such as invitations for keynote speaker at scientific conferences (Casadevall & Handelsman 2014; Dumitra et al. 2019; Johnson et al. 2016; Klein

[46] Forthcoming in Proceedings of the ISSI 2023 conference, Bloomington, July 2023.

et al. 2017; Sardelis & Drew 2016), for other invited talks (Nittrouer et al. 2017), or invitations for e.g., scientific advisory boards (Ding et al. 2020?)

Apart from that, the general peer review literature has shown the many uncertainties and biases in peer and panel review processes (Chubin & Hackett 1990; Cole et al. 1977, 1982; Cole 1991; Lamont 2010; Van Arensbergen et al. 2014), making it unlikely that the committee-based evaluation of PhD theses would be completely unbiased.

Gender differences versus gender bias

If there are strong gender differences in getting cum laude for the PhD thesis, the question is whether we have to classify those differences as gender bias. Bias implies a deviation of what would be correct and merit-based decisions (Merton 1973). Therefore, the selection of a frame of reference for measuring merit is crucial. One perspective is *distributional equality*, and that would imply that the share of female PhD students that receive a cum laude should be equal to the share of male PhD students that get cum laude. Here we use a widely accepted view that science is a *merit-based system* and that implies that cum laude awards should go to the best PhD students with the best doctoral dissertations. Bias can be defined as deviation from this norm.

This of course does not immediately solve the problem, as one may have different opinions how to decide what are the best dissertations and who are the best PhD students. E.g., should the most talented PhD students or the best performing ones receive the cum laude award? And how does one objectively measure who are the best PhD students or what are the best theses? We suggest that cum laude is meant as a recognition of the quality of the thesis itself, and we assume that the higher the quality, the more highly cited papers result from the thesis. Of course, this does not work in an absolute way, and one would not expect correlation of 1 between the quality of the thesis and the number of highly cited papers coming out of the thesis work. However, one would expect that the *set of cum laude theses* would have higher *average* bibliometric scores than a set of very good, but not cum laude awarded dissertations.

If we measure the quality of the PhD research by measuring the impact of the related output, the question can be answered whether the observed *gender differences* in receiving cum laude are a form of *gender bias*. In an earlier study, we showed that there are persisting differences between men and women researchers in terms of output and impact, and relevant for this study, these gender differences appear to occur specifically in the top of the performance distribution (Fig 1): Among the group of high productive and high impact researchers, male researchers are strongly overrepresented (Van den Besselaar & Sandström 2017).

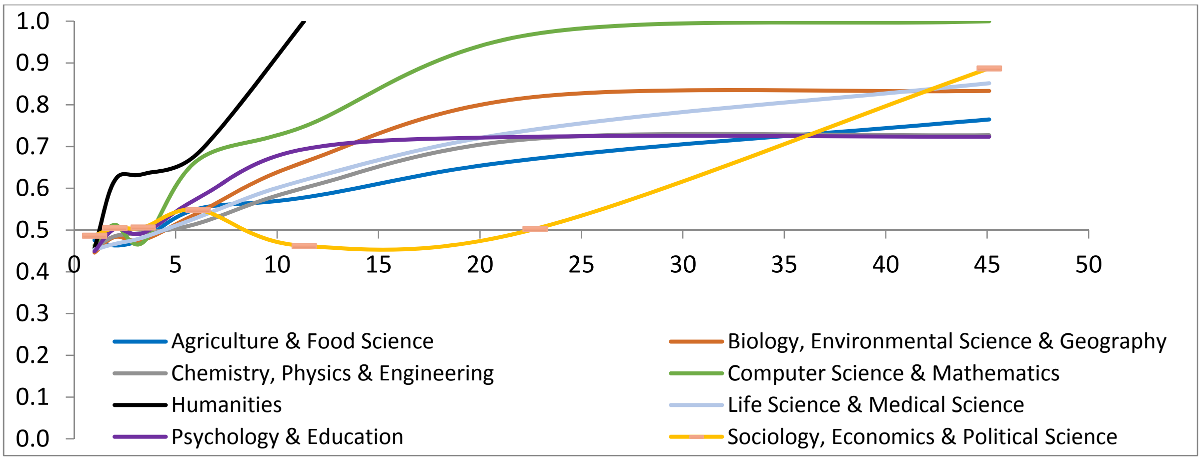


Figure 1: Share of men (Y-axis) by productivity level (X-axis: full counts)
(Note that the number of researchers declines fast with the increase of productivity)

The figure shows only publication-counts, but elsewhere we showed that high numbers of publications lead to high numbers of highly cited papers (Sandström & Van den Besselaar 2016). This finding may be relevant for the current study, as the cum laude recipients are expected to belong to that top segment. This means that the fact that women less often receive cum laude may not be bias, but the effect male researchers being over-represented in the top-performers segment. We will test that in this paper.

The case

The case is a large research-intensive university in the Netherlands, in the top 100 in the Leiden Ranking (2019) – based on the share of top 10% cited papers[47]. We cover in this study the period between 2000 and 2018. In that period, 5732 doctorates were awarded by the university, with a substantial annual growth: In 2000 it where 177 doctorates, increasing to 433 in 2018. The share of female PhDs increased from 39.5% in 2000 to 52.0% in 2018, with an average of 48%. In that period, 248 theses were granted cum laude, of which about 34% to women. As table 1 shows, men have almost twice as high a probability to receive cum laude.

Table 1. Cum laude by gender*

	Male	Female	Total	F/M
Cum laude	165	83	248	0.50
No cum laude	2834	2646	5481	0.93
Total	2999	2729	5728	0.91
Share cum laude (%)	5.5	3.0	4.3	0.55

* Source: Vrije Universiteit Amsterdam, 2000-2018, 1 case has missing gender

If we distinguish between disciplines, quite some differences become visible. First of all, the numbers of PhDs per discipline differ strongly in this case. Some 36% of all PhD degrees are from medicine plus 2% from dentistry, and 28% from the two science faculties. The remaining 34% are divided among the other disciplines, with behavioral and movement sciences (with a share of about 10%) as largest of the smaller fields. Table 2 gives the details.

[47] The university has 14.1% of its publications within the top 10% category.

Table 2. Faculty by cum laude (2000 - 2018)

faculty	# PhDs	Share of PhDs	# CL	Share of CL	Faculty share
Theology and Religious Studies	236	4.1%	21	8.5%	8.9%
Social Sciences	291	5.1%	22	8.9%	7.6%
Humanities	317	5.5%	25	10.1%	7.9%
Law	169	2.9%	14	5.6%	8.3%
Natural Sciences	888	15.5%	38	15.3%	4.3%
Earth and life Sciences	717	12.5%	14	5.6%	2.0%
Behavioral & movement sciences	559	9.8%	36	14.5%	6.4%
Business and Economics	383	6.7%	6	2.4%	1.6%
Dentistry	109	1.9%	3	1.2%	2.8%
Medicine	2060	36.0%	69	27.8%	3.3%
Total	5729		248		

This set of 5729 PhD students were supervised by in total 3988 supervisors in different roles. Most of them (88.4%) were never involved in awarding cum laude, as one already would expect from the small number of CL awardees (Table 3). Table 3 shows that the actual cum laude chance of a PhD student in group of 'cum laude giving supervisors' is 8.4% (248 out of 2733). This leads to an obvious question: is there a *quality difference* between the group of PhDs that were supervised by belong to the 'gave cum laude' set of supervisors and those that belong to the 'never gave cum laude' set?[48] If the answer would be no, it would suggest that there is an inequality coming from a difference in attitude between supervisors: those that are inclined to propose cum laude versus those that seem not inclined to propose cum laude. We tested that, but that is not the case.

Another question is whether those supervisors with PhD students that received CL perform better themselves than those supervisors without CL-awardees. Finally, it would be interesting to study individual difference between supervisors, as some have a quite high percentage of PhD students with com laude. These questions related to the differences between supervisors are not addressed in this version of the paper.

Table 3. Supervisors who supervised a PhD student that received CL (all years and faculties)

Supervisors who have at least	# supervisors	# cum laude PhD students	# non cum laude PhD students
one PhD student with cum laude)	462	248	2733
share of total	11.6%	4.3 %	47.7 %

Data and methods

The basic data were provided by the university: Names of the PhD students by year of degree, faculty, cum laude or not, over the period 2000-2018. We also received for each PhD student the names of their supervisors. The data to measure the quality of the PhD thesis have been retrieved from Scopus. For a large number of supervisors, bibliometric data were retrieved for Scopus as well.

[48] The decision to award cum laude is reserved for the PhD committee, but the supervisor generally is the one who proposes this. For readability, we use a formally incorrect phrase. Where we write "supervisors awarding cum laude" one should read "supervisors who had at least one PhD student that was awarded cum laude by the PhD committee". And where we write "Supervisors who never gave cum laude", one should read: "Supervisors who have no PhD student that was awarded cum laude".

Measuring the quality of a PhD thesis

If the meritocratic system would work perfect, the quality of the PhD dissertations that received cum laude is higher than the quality of those that did not. As there is no independent scoring of the quality of a PhD dissertation, we need to develop a method to do this. As a proxy for the quality of dissertations we will use the papers published by the PhD student in the period of preparing the dissertation, as well as papers published in the three years after the year in which the student received the PhD. We assume that these papers are related to the dissertation work. Then the question of the quality of the thesis becomes another question: what is the quality of the scientific output of the PhD students in the defined period? The bibliometric toolbox Scival of the Scopus database provides a series of indicators *at the paper level* (Elsevier, no date). We aggregated those Scival indicators to the author level, using full and fractional counts, and absolute and relative (to the total oeuvre of the author) values. The non-field normalized and the field normalized indicators were both included. The resulting list of indicators is:

1.	P	Number of full papers
2.	Pfrac	Number of papers fractionally counted by number of authors
3.	C	Total number of citations
4.	Cfrac	Fractional share of total citations
5.	C/P	Citations per publication
6.	P10%	Number of papers in the set of 10% highest cited papers
7.	P10%FN	Number of papers in the set of 10% highest cited papers, field normalized
8.	PP10%	Share of top 10% cited papers in the oeuvre
9.	PP10%FN	Share of top 10% cited papers in the oeuvre, field normalized
10.	P10%frac	Number of papers in the set of 10 % highest cited papers, fractionally
11.	PP10%FNfrac	Share of top 10% cited papers cited papers, field normalized, fractionally
12.	Average FWCI	Average field weighted citation impact
13.	Sum FWCI	Sum field weighted citation impact
14.	SNIP	Source (journal) normalized impact per paper
15.	SJR	SCimago Journal Rank ('impact factor')

Publication behavior and citation behavior differ between fields, and therefore the indicators are often field normalized. And as most papers are co-authored, and co-author behavior also differs between fields, many indicators correct for the number of co-authors ('fractional counting'). The indicators measure different aspects of quality and do so in different ways. For example, P, Pfrac, P10%, and P10%frac each measure some aspect of *productivity*. P simply measures full output, Pfrac corrects for the number of authors, P10% only takes the top cited papers into account, and P10%frac does this while correcting for the number of co-authors. The last two indicators also measure *impact*, as is also done by other citation-based indicators. Finally, some indicators do not measure the impact of the work of the authors, but the impact of the journals the papers are published in. These indicators are probably measuring reputation, more than impact or quality.

A next issue is the period covered by the measurement. The obvious period to take into account are the years of the PhD trajectory and a few years after the moment of awarding the PhD. But the committee may also have assessed the *potential* of the researchers, and may have based the cum laude decision also on *expected (future) excellence*. This can be measured by the performance indicators over a longer period after the dissertation was finished.

As we measure the quality of the thesis using citation-based indicators, we restrict the analysis to the period 2000-2014. The more recent years cannot be included, as we lack a sufficient long period for measuring the citations: the bibliometric data were collected in late 2019. Overall, this leads to a set of (only) 120 cum laude dissertations out of 3306 dissertations – 3,6%.

We only include in our analysis those research fields for which journal articles are the main research output. This excludes from this study most of the social sciences, the humanities, law and theology. Fortunately, those are also the fields with a lower gender difference in awarding cum laude (Table 7). Psychology and economics are included, as in those fields journal articles are dominant. Consequently, we include all dissertations in the *sciences*, *mathematics* and *computer science*, in the *earth and life sciences*, in the *medical sciences* and *dentistry*, in *economics* and in *psychology* and *movement sciences*. Due to the large number of medical dissertations, we decided to include all the medical cum laude dissertations and 50% of the dissertations that did not receive cum laude.[49]

Data

The names and other information about PhD students (such as gender and whether they were granted cum laude) was provided by the university. For 2592 researchers from our sample of 3306 we collected the publication and citation data using Scopus (summer 2019), which then were controlled and cleaned, and then used to retrieve the mentioned indicators covering the selected periods.

For selecting the supervisors, we followed the following approach.

- We made a list of all first, second and third supervisors, and first, second and third co-supervisors. From that we made a list of unique supervisors.
- For all supervisors with 3 and more PhD students the bibliometric data were collected.
- Then we checked whether we had for all PhD students at least one supervisor with bibliometric data. For those we hadn't, we selected one of the other supervisors of that PhD student and collected the bibliometric data also for these supervisors. In this way we succeeded to find for almost every PhD student bibliometric data for at least one of his/her supervisors. Where we had these indicators for more than one of the supervisors, we used the one with the highest performance in the statistical analyses.

As table 6 shows, we collected and cleaned for 3306 researchers their bibliometric data, which was a very time-consuming activity.

Table 6. Available bibliometric data

	PhD students	Supervisors
Collected bibliometric data (Scopus)	2592	812
No bibliometric data	714*	3176

* All of these are in medicine.

[49] We use a stratified sample that represents the distribution over years and over gender in the total set. Within each category, the 50% was selected randomly. The very labor-intensive part of the study was the collection of performance data from (in this case) Scopus.

Methods

As publication, collaboration and citation behavior differs between the disciplines, we first *standardize* the performance indicators *by faculty* and in such a way that the mean of the performance indicators is 0 and the standard deviation is 1.

As explained in the previous section, the bibliometric indicators all measure in some way quality, and therefore we consider the various indicators *as items*, together measuring various underlying dimensions of the construct *quality*. A principal component analysis (oblique rotation; component scores saved with regression) resulted in three dimensions, together explaining almost 80% of the variance.[50] The resulting variables represent three distinct quality dimensions[51]:

- Dimension 1: *Relative impact* of the oeuvre of a researcher: *share* of top cited papers in the total oeuvre of the researcher, the average FWCI, and citations per publication:
- Dimension 2: *Total impact (and output)* of the researcher, in fractional count, consisting of the following indicators:[52]
- Dimension 3: *Journal impact*:

The *total impact* dimension includes the indicators measuring total output, total impact, and the sum of highly cited papers, suggesting that higher quantity produces also a higher quantity of top papers (Sandström & Van den Besselaar 2016). We tested whether the indicators also form reliable scales, and that is the case. The Cronbach Alpha for the three indicators are 0.911 (5 items), 0.861 (3 items), and 0.851 (2 items) respectively, and cannot be improved by leaving out one of the items[53]. Concluding, we have a quality concept with the three mentioned dimensions (total, relative, and journal impact) and the scales used to measure those are reliable.

The data are analyzed using logistic regression (the dependent variable is binary: cum laude or no cum laude). The logistic regression is first done for men and women separately, in order to test whether the independent performance variables have a similar effect for women or not. If not, we have to include interaction terms between gender and the performance variables in the analysis.

Then we run a logistic regression also including gender as a variable. The odds ratio for gender gives a first insight in the effect of gender (bias) on getting cum laude. However, the odds ratio only gives the value of the dependent variable for the independent variables at value = 0. As logistic regression is non-linear, the gender differences may be dependent on the level of the independent (here: performance) variables (Mood 2009). In order to correct for this, we calculate the predicted probabilities (using the predicted margins in STATA) to display the probability to receive cum laude by gender for a variety of values of the performance variables.

After having done this, we did a few controls, that is we restrict the sample of PhD receivers to those with at least one supervisor that has at least one PhD student with cum laude, and run

[50] The principal component analyses were done on both the whole career data and the PhD period data.

[51] The items belonging to each component are explained above.

[52] We do not use the full-count based indicators, but when including those, they load together with the fractional-count based publication and citation indicators (here component 2).

[53] For the career data the Cronbach Alpha's are .911 .929 and .849 for component 1,2 and 3 respectively.

the same analyses – this because only a small subgroup of supervisors has supervised a PhD student that was awarded cum laude. By restricting the analysis to this set of supervisors, we may get a better picture of what variables affect the probability to get cum laude. Secondly, we restrict the sample to the cum laude awardees plus the best performing non-cum laude PhD receivers, to assess whether CL was indeed awarded to the best PhD students.

Findings

Who receives the CL-award? Some basic findings.
As reported above, overall about 4.3% of the PhD recipients were awarded cum laude, and men twice as often as women. Did this change over time? In Figure 2, we show a breakdown over the period 2000-2018. The figure shows for each year the share of female PhD students that received cum laude divided by the share of male PhD students who received cum laude, as well as the opposite: male divided by female. In 2001, only men (14 in total) received cum laude, and no women. Given these strong fluctuations, we show in Figure 2 the raw data, as well as a 5-years moving average.

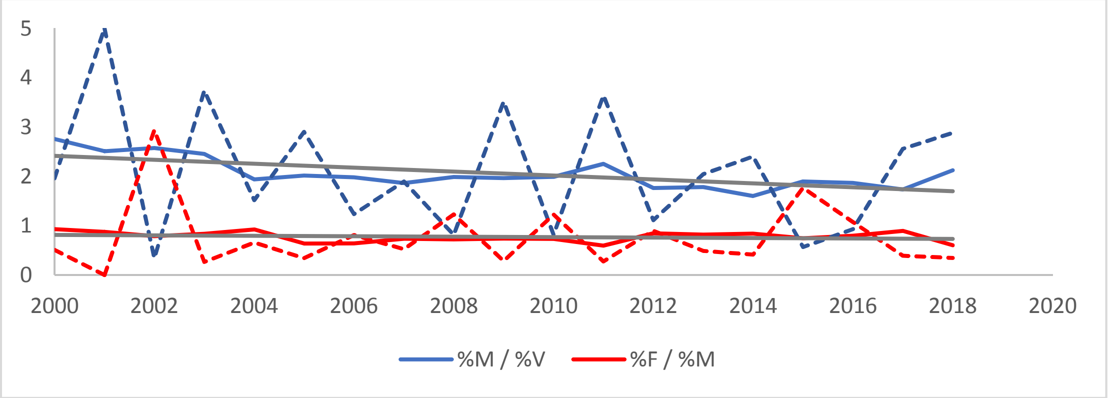


Figure 2. Cum laude by gender over time

The linear regression line suggests convergence, but this is mainly due to the extreme 2001 score – when deleting that year, no change is visible on the longer run. This makes our study easier: we can take the whole period as one data set – which is an advantage as the cum laude numbers are rather low.

Table 7 gives a break down by field, showing that the *stronger codified fields* (the STEM fields, and the mathematics-oriented fields like economics) have a low percentage of PhD theses with cum laude (between 1.6% and 3.3%). In these fields, the probability that men receive cum laude is twice as high (or more) as for women.

Within social sciences, law, and religion studies, the share of PhD students receiving cum laude is substantially higher (between 7.6% and 8.9%), and the differences between men and women are relatively small. Psychology and humanities have an ambiguous position. Psychology – which often presents itself as a STEM related field - belongs to that group when we look at the gender differences. However, behavioral (psychology) and movement sciences have a much higher percentage of PhD students that receive cum laude. These are more similar to the other social sciences. The same holds for the humanities: with respect to the share of PhDs with cum laude, humanities are similar to the social sciences, but in terms of gender differences it belongs to the group with the STEM fields and economics (and psychology).

Table 7. Cum laude by gender and faculty (2000 - 2018)

Faculty	share	M-Total	W-Total	% CL	M-CL	W-CL	M-CL%	W-CL%	W/M
Earth and life Sciences	12.5%	379	338	2.0%	9	5	2.4%	1.5%	0.62
Natural Sciences	15.5%	645	243	4.3%	33	5	5.1%	2.1%	0.40
Dentistry	1.9%	56	53	2.8%	2	1	3.6%	1.9%	0.53
Medicine	36.0%	849	1211	3.3%	40	29	4.7%	2.4%	0.51
Business and Economics	6.7%	277	105	1.6%	6	0	2.2%	0.0%	0.00
Behavioral & movement sciences	9.8%	221	338	6.4%	21	15	9.5%	4.4%	0.47
Humanities	5.5%	174	143	7.9%	18	7	10.3%	4.9%	0.47
Social Sciences	5.1%	123	168	7.6%	11	11	8.9%	6.5%	0.73
Law	3.0%	81	88	8.3%	7	7	8.6%	8.0%	0.92
Theology and Religious Studies	4.1%	194	42	8.9%	18	3	9.3%	7.1%	0.77

M-total: total number male PhDs; W-total: total number female PhDs; %CL: percentage of PhDs with Cum Laude; M-CL: met with cum laude; F-CL: women with cum laude; M-CL%: percentage men with com laude; F-CL%: percentage women with cum laude; W/M = W-CL%/M-CL%

Who awards cum laude: basic findings

Not all supervisors have supervised a PhD student that received cum laude; in fact, most of the supervisors (88.4 %) have not as Table 8a shows. Of all students that had at least one supervisor that had at least once awarded cum laude, the share of female PhD students is 48.9%), Of all PhD students that had only supervisors who were never involved in awarding cum laude, the share of female PhD students is 46.29%. Therefore, the low share of women among cum laude awardees is not a distributional effect: that would only be so if those supervisors that award cum laude would have a very low number of female PhD students – which is not the case.

Table 8a. Who awards cum laude?

	# sup	% sup	# CL	# PhD	% CL
has awarded cum laude	462	11.60%	248	2981	8.3%
has not awarded cum laude	3526	88.40%	0	4884	0.0%

* Source: Vrije Universiteit Amsterdam, 2000-2018, 1 case has missing gender. Note that there are non cum laude PhDs who have supervisors from both groups, as a consequence the bottom number of PhDs contains some of the same PhDs as the top number.

Table 8b. Who receives cum laude?

	# PhD	% PhDs	# CL	% CL	# F PhD	# MPhD	% F PhD
PhDs with CL granting supervisor	2981	52.0%	248	8.3%	1457	1523	48.9%
PhDs without CL granting supervisor	2748	48.0%	0	0	1272	1476	46.3%

* Source: Vrije Universiteit Amsterdam, 2000-2018, 1 case has missing gender

Furthermore, those supervisors that did award cum laude, do that in various frequencies. Some have high percentages cum laude receivers as, whereas others have lower percentages – but we found that those supervisors who do award a cum laude, do that on average for about 8.3 of his/her PhD students – much higher than the overall average. Of course, when you have only one or two PhD students, and one of them gets a cum laude, the percentages are 100% and 50%, but also quite a few of the supervisors with many PhD student have a high share of cum laude receivers. The other way around, there are also supervisors with many PhD students that never awarded cum laude. This leads to the following question: Do those supervisors that never awarded cum laude have lower quality PhD students, or do they have other reasons for this? We will address this issue later in the report.

Summarizing, awarding CL is very unevenly distributed among supervisors, and the different success rates for male and female PhD students is not explained by the gender distribution of the PhD students among those two groups of supervisors.

Gender bias?
We started with a logistic regression for men and women separately, and test whether the independent performance variables have the same effect on the probability to get cum laude for men and women. This is the case, and we therefore do not need to include interaction terms in the analysis. In order to test whether women have less chance to be awarded cum laude, we use a logistic regression with *yes/no Cum Laude* as dependent variable. As independent variables, we use the three impact variables and gender, and we add as control variables the faculty and the year of receiving the PhD. Table 9 shows the results. Two impact variables (absolute and journal impact) have a significant *positive* effect on the probability of obtaining cum laude, and the relative impact has *no effect*. Being *female* has a negative effect on the probability to get cum laude: the odds ratio for gender is 0.49 suggesting that men get twice as often CL than women after controlling for performance.

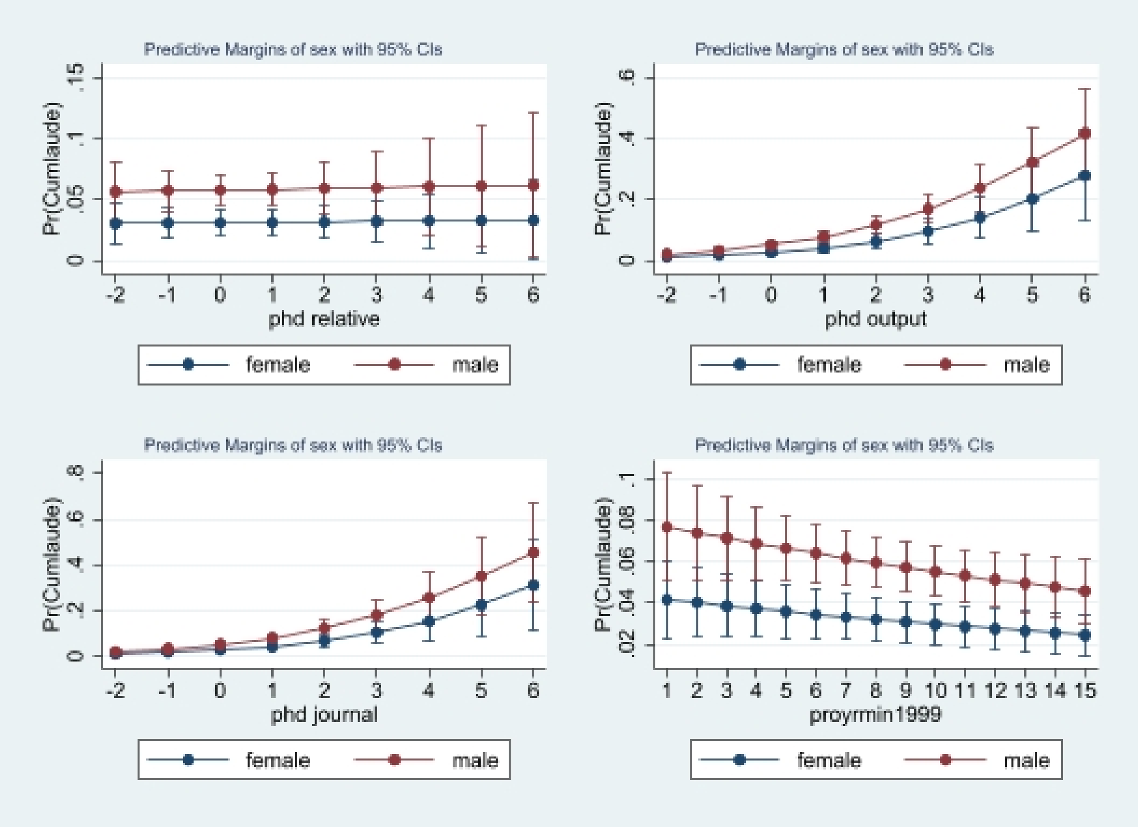
Table 9. Cum laude by gender, faculty and quality*

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
							Lower	Upper
Sex of PhD candidate(1)	-0.714	0.228	9.824	1	0.002	0.49	0.313	0.765
faculty			33.673	5	0			
faculty(1)	-0.702	0.258	7.373	1	0.007	0.496	0.299	0.823
faculty(2)	-1.806	0.402	20.154	1	0	0.164	0.075	0.362
faculty(3)	0.004	0.268	0	1	0.988	1.004	0.594	1.697
faculty(4)	-1.677	0.495	11.462	1	0.001	0.187	0.071	0.494
faculty(5)	-1.125	0.764	2.164	1	0.141	0.325	0.073	1.453
PhD year	-0.041	0.024	2.967	1	0.085	0.96	0.916	1.006
relative impact (PhD period)	0.02	0.1	0.042	1	0.838	1.021	0.839	1.241
total impact (PhD period)	0.465	0.066	49.777	1	0	1.592	1.399	1.812
journal impact (PhD period)	0.474	0.093	26.051	1	0	1.606	1.339	1.926
Constant	-2.143	0.265	65.64	1	0	0.117		

* Logistic regression; all PhD students in the selected faculties; ** Indicators calculated over the t-3 ~ t+3 period; N=2579; Nagelkerke R Square: 0.177

The odds ratio for gender reported above is sensitive for non-linearity, and only measures the gender effect for independent variables = 0. Therefore, we calculate the predicted probability to receive cum laude for men and women *at the various levels of the independent variables*. This is done using STATA, and the results are in Figure 2a-2d.

For each level of the *relative impact*, men have a higher probability to get cum laude than women do, although the performance variable itself makes no difference. For the *absolute impact/output* and for the *journal impact*, the analysis shows that at the lower levels of the independent variables there is no gender difference, but the higher the performance level, the larger the distance between men and women. One would expect that CL awardees are at the higher level of performance, so ender does play a role. At these high-performance levels, the 95% confidence intervals start to overlap, which is due to the low number of observations at those levels. These predictive margins support the gender difference found in the logistic regression. The last figure 3d shows that over the years, the probability to get cum laude declines, but the men-women ratio remains about constant: In each year men get it about twice as often as women do – even after controlling for performance.



Figures 3a-3d: Predictive margins of probability to receive cum laude by sex

Cum laude recipients versus the best non-recipients

As we controlled for the three quality variables, the result suggests that there should be enough high performing women that would – based on their performance – qualify for cum laude. Therefore, it is useful to have a closer look into the set of researchers with high high-quality dissertations. That group is defined in the following way. We select for each of the disciplines and for the two quality dimensions (*total impact/output* and *journal impact*)[54] the best scoring non-cum laude researchers. The threshold is chosen in such way that we include about 10% of all researchers in a field in the 'top sample'. Then we aggregate these two 'top samples' with the CL-recipients into one sample consisting of 519 researchers. However, there one case with a bibliometric data error, resulting a final set of 518.

We then deploy the same logistic regression for this smaller top group. Within the top group, the relative impact variable has no significant effect, but the two other quality variables (absolute impact and journal impact) have a negative effect: within the top, higher performance results in a lower probability to get cum laude – suggesting a flawed selection process (table 10).

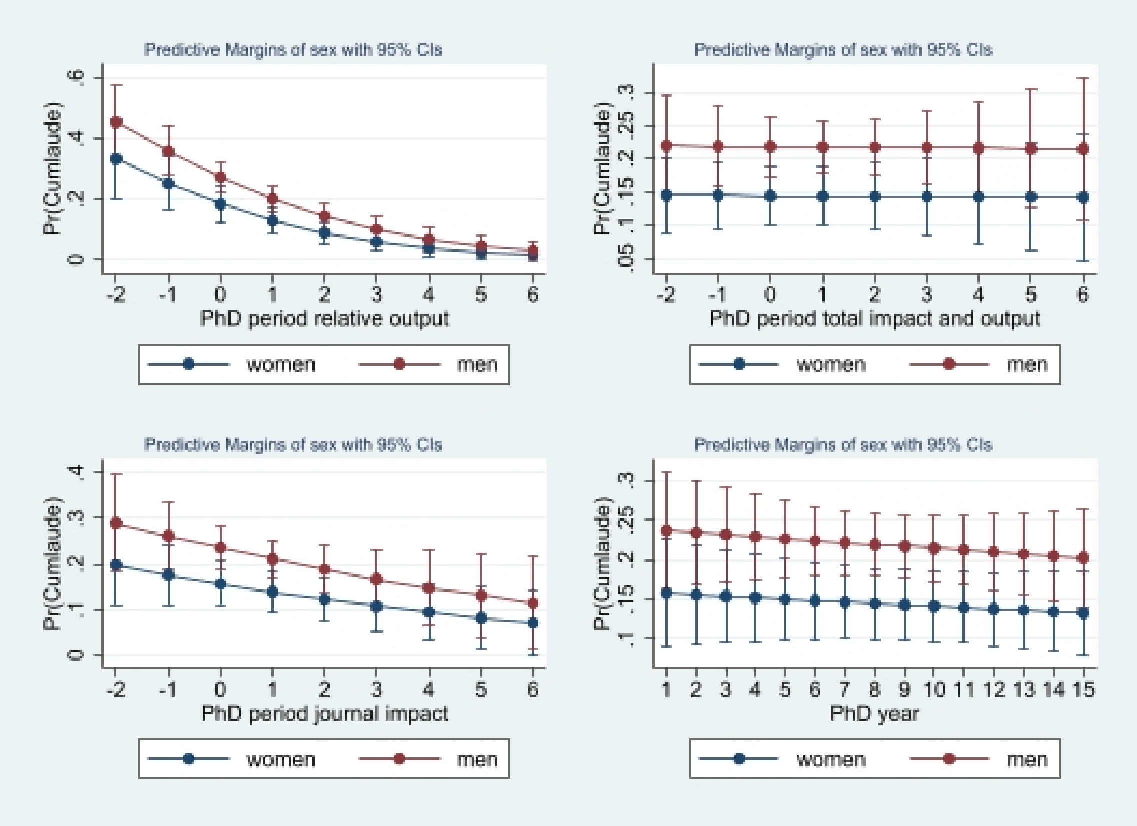
[54] We exclude the relative impact, as that variable did not have any effect on getting cum laude (see table 9)

Within the top group, being female lowers the probability to get CL considerably: the odds ratio is 0.582, not much different from the result of the previous analysis. The predictive margins show again that with a similar score on the independent variables, men have a higher probability than women to be awarded cum laude.

Table 10. Cum laude by gender, faculty and quality*

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
							Lower	Upper
Sex of PhD candidate	-0.542	0.258	4.41	1	0.036	0.582	0.351	0.965
faculty			38.106	5	0			
faculty(1)	0.418	0.296	1.989	1	0.158	1.519	0.85	2.714
faculty(2)	-1.056	0.411	6.611	1	0.01	0.348	0.156	0.778
faculty(3)	1.553	0.357	18.945	1	0	4.726	2.348	9.51
faculty(4)	-0.876	0.513	2.918	1	0.088	0.416	0.152	1.138
faculty(5)	-0.481	0.812	0.351	1	0.553	0.618	0.126	3.035
PhD year minus 1999	-0.007	0.027	0.06	1	0.807	0.993	0.941	1.048
PhD period relative output	-0.124	0.113	1.191	1	0.275	0.884	0.708	1.103
PhD period total impact and output	-0.150	0.081	3.417	1	0.065	0.861	0.734	1.009
PhD period journal impact	-0.364	0.113	10.37	1	0.001	0.695	0.557	0.867
Constant	-0.538	0.308	3.056	1	0.08	0.584		

* Logistic regression; all PhD students in the top segment; ** Indicators calculated over the t-3 ~ t+3 period; N=518; Nagelkerke R Square: 0.173



Figures 4a-4d: Predictive margins of probability to receive cum laude by sex, 3 factor top group.

Conclusions

The raw data showed that men get twice as often cum laude than women do. In the logistic regression, we controlled for quality of the PhD thesis, for the faculty of the PhD student, and (where needed) for the year of obtaining the PhD degree. We find that gender has again a strong and for women negative effect on the probability to receive CL. In fact, the control variables do hardly change the strong effect of gender on the probability to receive a CL award.

Secondly, when comparing the predicted probabilities to get a cum laude, we find that for the low performing men and women the probability is equally low, but the higher the performance level the larger the gender difference: the highest performing women overall have a much lower probability to get cum laude then men do.

Thirdly, we find that cum laude PhDs do not outperform the best performing PhDs who did not get cum laude. On the contrary, the cum laude receivers are in some impact indicators outperformed by the best performing non-cum laude PhDs – especially those that were supervised with non-cum laude supervisors. This indicates that some of the best PhD students may not have received CL because awarding CL was not an option for their supervisors.

Fourthly, the data show that several of the CL-recipients have a very low (Scopus covered) output and impact – in the PhD period but also over the career until 2018. In the fields under study here, this is rather remarkable.

As cum laude may have strong career effects, our result suggest that the current procedure may create unjustified differences.

Limitations

This study has a few limitations that should be mentioned. Firstly, one may dispute the way the quality of the PhD theses is measured using bibliometric indicators. However, as no other independent assessments of the theses exist, this is the best option for the moment. We also would defend the approach: an excellent PhD thesis can be expected to lead (in the fields we covered here) to impactful papers. And although a lot or randomness may affect the citations received at the individual level, at the group level one would expect that CL-dissertations would lead to higher scores on the impact variables we deploy than the non-CL-dissertations. A second issue that should be mentioned is the selection of the top groups: Our top includes about 24% of the total population. One may find this too high, and it would be reasonable to use a more selective definition. However, if we would make the top group smaller, then the difference with the cum laude receivers will even increase, as the more restricted group of non-CL top performers will have even higher impact scores. Then the cum laude receivers will clearly underperform compared to the others in the top. Thirdly, the explained variance (Nagelkerke pseudo R^2) found in the various regression analyses are rather low, suggesting that other factors than gender and impact may be behind the cum laude decision.

Further work

This last issue leads to further work. We would expect that three factors not included in this report may partly explain the cum laude decision: (1) the personal relation between the supervisor(s) and the PhD student, (2) the personality of the PhD student, and (3) differences between the supervisors.

As supervisors have to select who they would like propose for cum laude, the quality of the personal relations between the PhD student and the supervisor may be important. We are currently analyzing that using the acknowledgement texts that are included in most of the PhD theses. Can we derive from those text indicators for the personal relation between the supervisor and the PhD student? One would expect as language is an expression of relationships, opinions and emotions (Pennebaker et al. 2001; Tauzik & Pennebaker 2010).

Another factor is the personality of the PhD student: can we distinguish cum laude receivers from others in terms of personal characteristics? In order to test this, we have distributed a questionnaire to measure personality dimensions of the PhD students, and the analysis we are doing now should give a first indication whether the CL award may be more awarding the person(ality) than the quality and impact of the science. By the way both variables (social relation with the supervisor and personality) may have a gender dimension.

Finally, some characteristics of the supervisors may play a role, such as the scientific quality and the attitude towards the cum laude award. Also these we intend to add in a follow-up paper.

References

Brouns M, Bosman R, van Lamoen I (2004). *Een kwestie van kwaliteit. Loopbanen van cum laude gepromoveerde vrouwen en mannen.* Groningen: Rijksuniversiteit Groningen.

De Bruin E (2018). Helft van de promovendi is vrouw. Maar cum laude krijgen ze zelden. *NRC.* October 19, 2018

De Bruin E (2019). Universiteiten bestrijden m/v-verschil bij promoties. *NRC.* June 14, 2019

Casadevall A, Handelsman J (2014). The presence of female conveners correlates with a higher proportion of female speakers at scientific symposia. *mBio* **5** (1): e00846-13

Chubin DE, Hackett EJ (1990). *Peerless Science: Peer Review and U.S. Science Policy*. SUNY Press: Albany.

Cole, S (1992). *Making Science: Between Nature and Society*. Harvard University Press: Cambridge.

Cole S, Rubin, L, Cole JR (1977). Peer review and the Support of Science. *Scientific American* **237** (4): 34-41

Cole S, Cole JR, Simon GA (1981). Chance and consensus in peer review. *Science* **214** (4523) 881–886

Ding WW, Murray F, Stuart TE (2012), From Bench to Board: Gender Differences in University Scientists' Participation in Corporate Scientific Advisory Boards. *Academy of Management Journal* **56**, 5, 1443-1464

Dumitra TC, Trepanier M, Lee1 L, Fried GM, Mueller CL, Jones DB, Feldman LS (2019). Gender distribution of speakers on panels at the Society of American Gastrointestinal and Endoscopic Surgeons (SAGES) annual meeting, *Surgical Endoscopy*

Elsevier Research Intelligence (without year) *Research metrics guidebook.*

Johnson C, Smith PK, Wang C, (2016). Sage on the Stage: Women's Representation at an Academic Conference, *Personality and Social Psychology Bulletin*, November 2016;

Klein RS, et al. (2017). Speaking out about gender imbalance in invited speakers improves diversity, *Nature Immunology* **18**, 5, 475– 478

Lamont, (2010). *How professors think; inside the curious world of academic judgment*. Harvard University Press: Cambridge.

Lincoln AE, Pincus S, Bandows Koster J, Leboy PS (2012). The Matilda Effect in science: Awards and prizes in the US, 1990s and 2000s, *Social Studies of Science* **42**, 307-320

Melnikoff DE, Valian VV (2019). Gender Disparities in Awards to Neuroscience Researchers, *Archives of Scientific Psychology* **7**, 4–11;

Merton RK (1973). The normative structure of science. In: Merton RK, *The sociology of science*. U of Chicago press, 267-278

Mood C (2009). Logistic regression: why we cannot do what we think we can do, and what we can do about it. *European Sociological Review* **26,** 67-82

Nittrouer CL, Hebl MR, Ashburn-Nardo L, Trump-Steele RCE, Lane DM, Valian VV (2017) Gender disparities in colloquium speakers at top universities, *Proceedings of the National Academy of Sciences,* December 2017

Pennebaker et al. (2001). *Linguistic inquiry and word count LIWC 2001*. Mahway, NY: Lawrence Erlbaum.

Reschke BP, Azoulay P, Stuart TE, (2018). Status Spillovers: The Effect of Statusconferring Prizes on the Allocation of Attention, *Administrative Science Quarterly* **63**, 819–847

Sandström U & van den Besselaar P (2016). Quantity and/or Quality? The importance of publishing many papers. *PLoS ONE* **11** (11): e0166149

Sardelis S, Drew JA (2016). Not “Pulling up the Ladder”: Women Who Organize Conference Symposia Provide Greater Opportunities for Women to Speak at Conservation Conferences. *PLoS ONE* **11** (7): e0160015

Semin & Fiedler (1988). The cognitive functions of linguistic categories in describing persons: social cognition and language. *Journal of personality and social psychology* **54**, 5, 558-568

Silver JK., Blauwet CA, Bhatnagar S, Slocum CS, Tenforde AS, Schneider JC, Zafonte RD, Goldstein R, Gallegos-Kearin V, Reilly JM, Mazwi NL (2018). Women Physicians Are Underrepresented in Recognition Awards from the Association of Academic Physiatrists. *American Journal of Physical Medicine & Rehabilitation* **97**, 1, 34-40

Tauzik & Pennebaker (2010). Psychological meaning of words. LIWC and computerized text analysis methods. *Journal of language and social psychology* **29,** 24-54

Van Arensbergen P, Van der Weijden I, Van den Besselaar P (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 den Besselaar P, Sandström U (2015). Early career grants, performance and careers; a study of predictive validity in grant decisions. *Journal of Informetrics* **9**, 4, 826-838

Van den Besselaar P, Sandström U (2016). Gender differences in research performance and in academic careers. *Scientometrics* **106,** 1,143–162

Van den Besselaar P, Sandström U (2017). Vicious circles of gender bias, lower positions and lower impact: gender differences in scholarly productivity and impact. *PLS ONE* **12** (8): e0183301.

Van den Besselaar P, Sandström U, Schiffbaenker H (2018). Using linguistic analysis of peer review reports to study panel processes. *Scientometrics* **117**, 1, 313-329

The glass ceiling, full professorship and availability of qualified candidates[55,56]

Peter van den Besselaar

Vrije Universiteit Amsterdam, Department of Organization Sciences & TMC Research Amsterdam

p.a.a.vanden.besselaar@vu.nl

Abstract

The claim that a glass ceiling exists within academia is generally based on the fact that the higher the academic rank, the lower the share of women within that rank: The share of women among PhD students is higher than the share of women among postdocs, and that share is higher than the share of women among assistant professors. The share of women among associate professors is even lower, and the share of women among full professors is the lowest. However, within this line of argument it is neglected that academic careers take time, and that therefore a cohort approach should be adopted. In this presentation, we show the results of such cohort approach for the case of the Netherlands. The findings suggest that except for the engineering sciences, the glass ceiling still exists.

Introduction

As the share of women in the higher academic positions is still low in many countries, the long-standing controversy continues about whether a glass ceiling exists in science. This also holds for the Netherlands, where the issue of the glass ceiling is not only discussed within the science community, but also within the political and the public debate. A driving force in this debate is the Netherlands Network of Female Professors[57], which since 2006 publishes the *Women Professors Monitor*.[58] The report focuses on the gender gap in science, among others with respect to full professor positions (LNHV 2018). The share of female full professors goes up, but in 2018 it arrived just above 20%. The report includes the *Glass Ceiling Index*, with has declined in from 2.0 in 2000 to 1.4 in 2017. This index compares the share of women among associate professors with the share of women among full professors. This suggest that the step from associate professor to full professor has become easier, but it remains difficult to make that last academic career step. There seems to be consensus about the too low level of women among full professors in the Netherlands, and even the Minister of Education and Science shares that view as she stated in her opening of the Gender Summit 2019 in Amsterdam.

[55] Paper presented at the *Gender Summit 17*, Amsterdam, October 3, 2019 (https://gender-summit.com/past-summits/gs17-eu/about-gs17)

A shorter Dutch version of this article appeared on the political science blog *Stuk Rood Vlees*, August 2019: https://stukroodvlees.nl/is-het-glazen-plafond-in-de-wetenschap-plots-verdwenen-dat-lijkt-me-sterk/.

[56] The paper is an outcome of the GRANteD project. This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 824574.

[57] https://www.lnvh.nl/

[58] https://www.lnvh.nl/monitor2019/

Also the newspapers are involved in the debate on the glass ceiling. Early 2019, Marceline Van Furth, professor of pediatrics at the Vrije Universiteit Amsterdam, stated in the Dutch newspaper *Trouw* that "men have an easier and faster career in science than women". The interview followed on the Lancet special issue on women in medical science (Van Furth & Tutu 2019). Another Dutch newspaper *NRC* responded to this interview. The *Fact Check* column of the *NRC* newspaper apparently wondered what a professor of pediatrics knows about the dynamics of academic careers, and confronted Van Furth's statement with a *Factsheet* published by the *Rathenau Institute*.[59] The journalist concluded that an important variable is missing in Van Furth's argument and in the figures presented by the LNVH Women Professors Monitor 2018: the time factor. Taking that factor into account, the Rathenau Institute claimed that the share of women among the newly appointed full professors is not lower but even higher than the expected level, implying that no glass ceiling exists for women in science.

In this paper, we reanalyze the data and show that a correct – disaggregated – analysis of the data shows a different result: the glass ceiling has not disappeared in the Netherlands, on the contrary. Then we discuss how to measure the ‘glass ceiling’.

Is there a pool of qualified candidates?

Time plays an important role when analyzing careers. According to the calculations of the Rathenau Institute (RI), it takes on average nineteen years between being awarded a PhD and being appointed as a full professor, and - importantly - that would be the same for men and women (Rathenau Instituut 2018). This of course should be taken into account when assessing gender differences in academic careers. If nineteen years ago in a field only a few women were getting a PhD, then only a small pool of qualified female candidates exists *now* for full professor positions. This is the well-known issue of subsequent cohorts with a different demography: the cohort of 2000 has now reached the level and experience of entering into full professor positions, and the share of women among the fresh appointed full professors should be compared with the gender distribution in that cohort, and not with the cohort of PhD receivers now or with an average of a set of annual cohorts.

The RI *factsheet* claims that the percentage of women among newly appointed full professors is higher than the percentage of women among PhDs nineteen years earlier. This would have been a stable phenomenon since at least 2004 (RI 2018). The good news is that these findings suggests that there is no glass ceiling or gender bias against women – and that the opposite could be the case.

In a later publication, the RI corrects the analysis: recalculating they found that the career from the PhD examination to a full professor appointment lasts on average seventeen years and not nineteen years (RI 2019), but they did not discuss the effects of this on the earlier conclusions. If we use the same data, but a 17-years time lag instead of a 19-years time lag, the share of new female professors does not rise faster than expected, but at about the expected rate. This can be seen in Figure 1, where the (blue) graph representing increase of the share of women among full professors and the (red) graph representing the increase of the share of female PdD recipients 17 years earlier are superimposed. Figure 1 suggest that from a cohort perspective the Dutch academic career system is gender neutral.

[59] The Rathenau Institute is the Netherlands Institute for science system assessment, and it produces reports on the state of the Netherlands science system and science policy advice. https://www.rathenau.nl/en

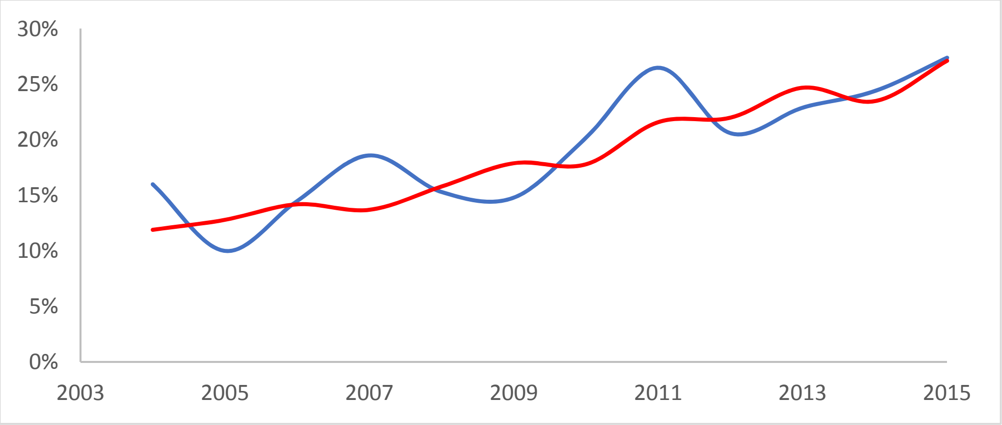


Figure 1: Appointment of women as professors and the labor market

Red line: Expected share of women among appointed full professors = Share of women among PhD grantees 17 years earlier.
Blue line: Real share of women among appointed full professors.
Source: Rathenau Instituut, Factsheet Women in science (RI 2018). The underlying microdata were not made available.

This would lead to the conclusion that gender differences in full professor positions have little to do with slower careers or with glass ceilings. It is a cohort effect: What looks like a glass ceiling may simply be the consequence of the time that careers takes. What looks like a lasting gender gap, it is easy to explain the lasting gender gap: If there is are less female than male PhD recipients at a certain moment in a specific field, then there will be less qualified female than qualified male candidates for full professor positions many years later. In the meantime, the number of female PhDs has been growing rapidly, and if the findings of Figure 1 are correct, we will see the effects of that on the share of female full professors in the future.

In order to find out whether there is gender bias in academic careers, one needs to follow the careers of men and women in a cohort: Disappear women more than men from the cohort (the ‘leaky pipeline’), does the career of women goes slower than the career of men, and do less women than men reach the highest positions (the ‘glass ceiling’). We will use this *cohort* approach to reanalyze the data, but before that a few caveats need to be discussed, and each of those deserves further investigation:

1. Firstly, according to a report of the Rathenau Instituut (RI 2018), it takes on average for men and women the same number of years between receiving the PhD degree and the being appointed as full professor. As many women give birth to children during that phase of the career, this finding of equal average periods is unexpected. Many women during this period spend much more time than men on child care, and therefore have less time for working on their academic career than men do. So one would expect that women on average need a (somewhat) longer period to become full professor than men do. Further inspection of the underlying microdata would be needed to clarify this.

2. Secondly, a caveat is that the data do not account for international mobility (RI 2018). However, if we assume that the number of male scientists entering the Netherlands from abroad equals the number that leaves the Netherlands, and that this also would hold for female scientists, international mobility may not influence the outcomes of the analysis.

3. Thirdly, the most important caveat is that the pool of PhD-holders is not the same as the pool of qualified candidates, as not all PhD-holders may prefer an academic career, and this

may differ by discipline and gender. For example, German data suggest that many medical PhD-holders aim at a career in health care and not one in research, whereas most PhD-holders in humanities aim for an academic research (Wegner 2019). For this paper, it is important whether these preferences differ by gender. After receiving their doctorate, men are one and a half times more likely to be employed *outside academia* than women (Sonneveld et al 2010; Van der Schoot et al 2012), and the disaggregated data suggest that this is also the case within the medical domain (Van der Schoot et al 2010). If this finding can be generalized, the analysis may even underestimate the gender gap, as frame of reference for assessing the share of women in current full professor appointments is not the share of women among PhD receivers seventeen years ago, but the much (150%) higher share of women in the group *PhD receivers that aim for an academic career*.

Fact checking

There are at least two problems with *'factsheet'* (RI 2018) that argues that no glass ceiling exists in academia in the Netherlands. Firstly, it only covers part of the science system, because the *university medical centers* are not included. This is rather problematic because (bio)medical research accounts for a large proportion of the total scientific research, of all PhD students, and of all professors: almost 40%. Above that, the proportion of female PhD students within the biomedical domain is also very high: since 2000 this has been above 40%, and since 2007 even above 50%.

Secondly, the data suggest that for the non-medical fields, the percentage of female professors appointed seems to more or less follow the percentage of women that got a PhD seventeen years earlier (see Figure 1). Regardless of whether this will continue in the future, the pattern found is the sum of developments within the different disciplines, some of which have higher percentages of

female PhDs and professors and others have lower ones. If the cohort effect described above (the supply of qualified candidates) were to exist, it should also be visible at individual discipline level.

The biomedical domain

Data about the University Medial Centers (UMCs) are even more difficult to get than data about the universities. The above-mentioned report (LNHV 2018) has some data about the changing share of women among medical full professors, but for analyzing the cohort effect one needs data about the annually new appointed full professors by gender. I was able to collect data about full professor appointments for one of the seven UMCs: the *Vrije Universiteit Medical Center* in Amsterdam (VUMC). This is a good case, because the VUMC has a higher overall percentage of female professors (figure 2, green dashed line – medium dots) than the other UMCs (figure 2, blue dashed line).

Because it the numbers of PhDs per year are small, and the numbers of annually appointed full professors are even smaller, I use a 5-years moving average to remove the annual fluctuations.

In the 2016-2018 period, 42 new full professors were appointed at the VUMC, of whom 13 were women. That is 31% (Figure 2, green straight line).[60] This then has to be compared with

[60] In 2016 and 2018, four women versus ten men were appointed as full professor. In 2017 (the *Westerdijk year*) additional funds were made available by Dutch government to appoint more female full professors. At the VUMC, five women and nine men were appointed as full professor. If this has meaning -

the percentage of women among those who obtained their PhDs seventeen years earlier (Figure 2, green dotted line). What was the percentage of women among those who obtained a PhD at the VUMC in 1999-2001, seventeen years before the 2016-2018 period? Figure 2 (green dotted line) shows the pattern, and we see that in the 1999-2001 period about half (48%) of those who obtained a PhD in biomedical fields were female. If the pattern presented in the Rathenau Institute report (RI 2018, 2019) was to be found here, it would have been about that percentage of 48% of women among the newly appointed professors in 2018. But that is not the case: It is much lower, only 0.62 of the expected value. We can conclude that the gender gap is rather large and far too few female full professors were appointed at the VUMC in the recent period. Table 1 (right column) summarizes the findings.

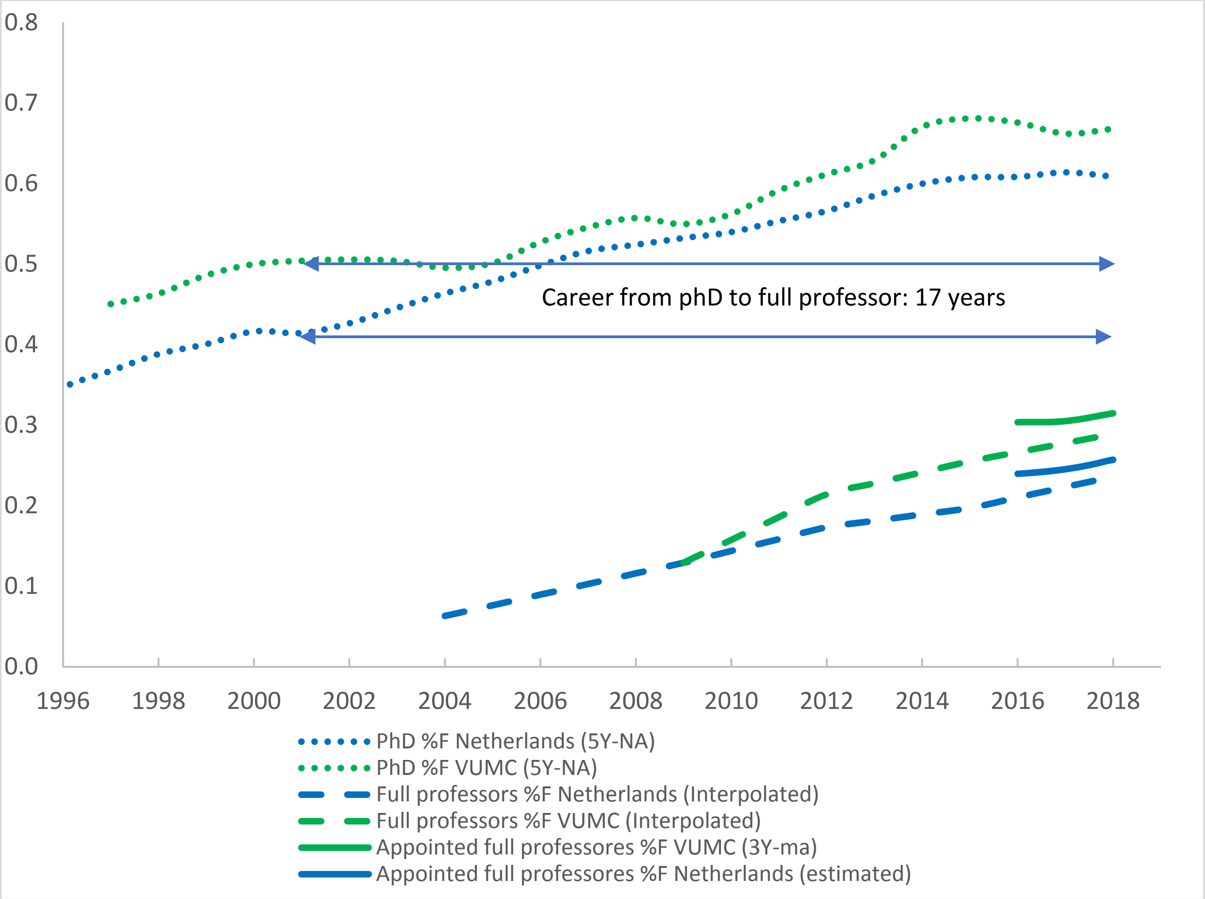


Figure 2: Developments of gender differences within the biomedical domain

Data on PhD students (1997-2018) and appointments (2016-2018): VUMC and Statistics Netherlands CBS
Data on percentage of female professors at the VU and at all UMCs: LNVH Monitor, various years.
Graph is interpolated. We use five and three year moving averages to eliminate short term fluctuations form the graphs.

Of course, this is a too restricted approach, as not all VUMC-PhDs stay at the VUMC, and not all professors at the VUMC are recruited from the own pool of PhD recipients. Therefore, we try to generalize this to a national pattern for the medical domain. Figure 2 shows that the

the numbers are low - then the percentage of appointments of women (without additional incentives) is around 29% and the gender gap is slightly larger.

percentage of women among the professor appointments at the VUMC is a few percentage points above the percentage of female professors at the VUMC (Figure 2, straight green line versus dashed green line), and that leads to the (moderate) growth of the proportion of female full professors at the VUMC. Because the recent growth curves for the share of female full professors at VUMC and for the nationwide development are parallel (blue and green dashed lines in Figure 2) we can safely assume that the proportion of new appointed women is also a few percentage points above the proportion of women for the entire medical domain.

The difference between the Netherlands and the VUMC with regard to appointments of women to full professor is approximately equal to the difference between the Netherlands and the VUMC in share of women among PhD recipients. That would imply that within the medical domain the gender gap nationwide is roughly the same as that at the VUMC (Table 1). In any case, it appears that within the medical domain, the number of women in full professor positions is increasing much slower than expected based on the availability of qualified candidates. So, the pattern found by the RI (figure 1) is not found in the biomedical domain.

Tabel 1: Appointments of female full professors by the change in qualified candidates

	Humanities	Social sci. and Law	Science, Math, Comp sci	Technical fields	Agriculture, Vetrin.	Total	Bio-medical	All UMCs	VUMC
Professors Appointments 2003-15	513	1507	508	581	86	3195	2016-2018		42
Of which women	137	326	65	67	12	607			13
% women	27 %	22 %	13 %	12 %	14 %	19 %		27 %*	31 %
Qualified candidates PhD awarded 1986-1998	2496	4722	5631	3836	1737	16145	1999-2001	2064	201
Of which women	717	1252	895	328	400	3224		840	96
% women	29 %	27 %	16 %	9 %	23 %	20 %		41 %	48 %
Expectation:	147	399	81	50	20	623			20
Difference with real Nr:	-10	-73	-16	17	-8				-7
Percentage:	-7 %	-18 %	-19 %	35 %	-39 %	-3 %		-34 %	-35 %

Data all fields apart from biomedical: Rathenau Instituut (professors), Statistics Netherland CBS (PhD awardees);
Data biomedical Netherlands: LNVH (professors), Statistics Netherland CBS (PhD awardees);
Data VUMC: Vrije Universiteit, VUMC. (VUMC = Medical center Vrije Universiteit)
Data all UMCs: Statistics Netherlands, own estimate (UMCs = aggregated for all University Medical Centers)
*: share appointed women (estimated)
Red: Negative gender gap for women; green: positive gender gap for women.

Averages are misleading

Do we find the average pattern in the other areas? That appears not to be the case, as the data published by the RI show (RI 2018, 2019). Table 1 below shows the increase in numbers of female full professors (2003-2015) and numbers of female PhDs seventeen years earlier (1984-1998).

Over the long period, the average growth in the proportion of female full professors is lagging only a little (3%) behind the proportion of female PhDs receivers seventeen years earlier. However, this proves to be a composition effect. It is mainly due to the technical sciences, where the percentage of female full professors grew 34% faster than the percentage of those who obtained their PhDs seventeen years earlier. In all other areas, the percentage of new female full professors lags behind the expectations that the average suggests. This applies very strongly to the agricultural sciences (39%) and the medical sciences (35%), but also quite strongly to the social sciences (18%) and natural sciences (19%). The humanities show the smallest career disadvantage for women (7%).[#]

The conclusion of the Rathenau Instituut is therefore wrong: in most areas, not more but fewer women have been appointed as full professors than could be expected. Averages are misleading, something we know as the Simpson paradox (Simpson 1951).

The gender gap decreases but remains substantial

The data show that women are catching up, because the number and proportion of female full professors is increasing everywhere. But the data also shows that the catching up goes slower than expected, that is slower than it should be given the supply of qualified candidates.

The earlier mentioned caveats such as gender differences in international migration of scientists and different job preferences of men and women may influence the results. However, the gender differences in job preferences seems to point at reasonably strong *underestimation* of gender inequality. The last word about the glass ceiling has not been said yet.

Concluding, the supply of qualified candidates influences the catching up process, but our analysis suggests that this cannot be the only factor. If that would be the case the gender gap in the full professor positions would be decreasing much faster. Our earlier research shows that other factors play a role (Van den Besselaar 2015; Van den Besselaar et al. 2015, 2016, 2017). From a policy perspective, it is therefore not enough to trust that changes in the labor market over time will do the job of creating gender equality in the occupational structure of universities and their medical centers.

Literature

Huijgen M (2019). NRC checks “men rise more easily in science” (In Dutch). *NRC* 25-2-2019

National Network of Female Professors LNVH (2018). *Monitor Female Professors 2018.* (In Dutch)

RI (2018) Rathenau Instituut, *Fact sheet: Women in science* (In Dutch). January 19, 2018.

RI (2019) Rathenau Instituut, *News: Share of new female professors rising faster than expected* (In Dutch). March 8, 2019.

Simpson, E.H. (1951), The Interpretation of Interaction in Contingency Tables, *Journal of the Royal Statistical Society*, Ser. B, 13, 238-241

Sonneveld HM, Yerkes M, van de Schoot, R (2010). *Trajectories and Labour Market Mobility. A survey of recent doctoral recipients at four universities in the Netherlands.* IVLOS, Netherlands Centre for Graduate and Research Schools, Utrecht

Statistics Netherlands CBS (2019). *PhDs by gender*. Download June 2019.

Van Furth M, Tutu M (2019). #IToo. *The Lancet* **393**, issue 10171, p529.

Van de Wier M (2019). Interview with Marceline van Furth and Mpho Tutu (In Dutch). *Trouw*, 12 February 2019

Van den Besselaar, P (2015). New figures: a glass ceiling on all levels of Dutch science (in Dutch). *StukRoodVlees Blog* 1-12-2015

Van den Besselaar P, Sandström U (2015). Early career grants, performance and careers; a study of predictive validity in grant decisions. *Journal of Informetrics* **9**, 826-838

Van den Besselaar P, Sandström U (2016). Gender differences in research performance and in academic careers. *Scientometrics* **106**, 143–162

Van den Besselaar P, Sandström U (2017). Vicious circles of gender bias, lower positions and lower impact: gender differences in scholarly productivity and impact. *PlosOne* **12**, 8: e0183301.

Van de Schoot R, Yerkes M, Sonneveld H (2010). Dataset of the Netherlands Survey of Doctorate Recipients.

Van de Schoot R, Yerkes M, Sonneveld H (2012). The Employment Status of Doctoral Recipients: An Exploratory Study in the Netherlands. *International Journal of Doctoral Studies* **7**, 331-348

Wegner A (2019) Result from the NCAPS 1st round in 2019 (*Personal communication*).

[#] **Annex**

The picture does not change much if we take the nineteen-year time lag between receiving the PhD degree and the appointment to full professor instead of a seventeen-year time lag. Even then the percentage of women named falls short of expectations in most areas, but the gender effect is less strong. Furthermore, the score for the humanities changes from slightly negative to slightly positive. The next table shows the differences between the two different time lags.

Domain	time lag of 17 years	time lag of 19 years
Humanities	-7%	3%
Social sciences & Law	-18%	-9%
Science, Math, Computer science	-19%	-6%
Technical fields	35%	71%
Agriculture	-39%	-33%
Biomedical	-34%	-32%