Te verschijnen in: Prpic, K., Van der Weijden, I., Aseulova, N. (Eds.)
(Re)searching Scientific Careers. St. Petersburg: IHST/RAS & SSTNET/ESA, 2013.

ACADEMIC TALENT SELECTION IN GRANT REVIEW PANELS

Pleun van Arensbergen*, Inge van der Weijden**, Peter van den Besselaar***

* Science System Assessment, Rathenau Institute, The Netherlands

** CWTS, Leiden University, the Netherlands

***Department of Organization Science & Network Institute, VU University Amsterdam, The Netherlands

Abstract

Career grants are an important instrument for selecting and stimulating the next generation of leading researchers. Earlier research has mainly focused on the relation between past performance and success. In this study we investigate the evidence of talent and how the selection process takes place. More specifically, we investigate which quality dimensions (of the proposal, of the researcher and societal relevance) dominate, and how changes in weighing these criteria affect the talent selection. We also study which phases in the process (peer review, panel review, interview) are dominant in the evaluation process. Finally we look at the effect of the gender composition of the panel on the selection outcomes, an issue that has attracted quite some attention. Using a dataset of the scores of 897 career grant applications we found no clear 'boundaries of excellence', and only a few granted talents are identified as top talents based on outstanding reviews compared to the other applicants. Quite often, the scores applicants receive change after the interview, indicating the important role of that phase. The evaluation of talent can be considered to be contextual, as the rankings of applicants changed considerably during the procedure and reviewers used the evaluation scale in a relative way. Furthermore, talent was found to have different (low correlated) dimensions. Small changes of the weights of these dimensions do not influence the final outcomes much, but strong changes do. We also found that the external peer reviews hardly influence the decision-making. Finally, we found no gender bias in the decisions.

1. Introduction

Attracting and maintaining well-qualified staff is essential for organisations that want to improve their status and reputation. Therefore universities and research councils aim at selecting the most talented young researchers, using explicit and also often implicit criteria (Van den Besselaar & Leydesdorff, 2009). As the academic career opportunities are by far outnumbered by young researchers who hope to establish an academic career (Huisman, de Weert et al., 2002; Van Balen, 2010), there is a strong competition among researchers (De

Grande, De Boyser et al., 2010). Securing a personal career grant seems increasingly crucial for a successful academic career. Besides the necessary resources to conduct research, it provides recognition of ones talent by the scientific community. As both the quality of the research system and the careers of individual researchers depend on these selection processes, it is important to understand how they function.

Most research on grant allocation focuses on the outcomes, searching for predictors for success. The internal selection mechanism hardly has been studied, and we therefore do not know what happens during the selection process (Bornmann, Leydesdorff et al., 2010). Only few studies have been conducted on the individual steps of the selection process (e.g. Hodgson, 1995; Bornmann, Mutz et al., 2008). Bornmann et al (2008) applied a latent Markov model to grant peer review of doctoral and post-doctoral fellowships. Their model shows that the first stage of the selection procedure, the external reviewing, is of great importance for the final selection decisions. External reviews had to be positive for fellowship applicants to have a chance of being approved. However, Van den Besselaar & Leydesdorff (2009), using a different method, could not confirm this. And no correlation was found between the decision and the external review score within the top 50% of the applicants.

In this paper we study the process of selecting scientific talent through career grants. We will show how the selection proceeds through the various phases, how consistent these phases are with each other, and which phases and criteria are decisive for the final selection. We will also look at the differences between disciplinary domains and between the three grant schemes under study.

2. Theoretical background

Although ‘scientific excellence’ and ‘talent’ are commonly used (Addis & Brouns, 2004), the meaning of these concepts is contested (Hemlin, 1993). Much debated is e.g., whether talent is innate or acquired. Talent has been explained by innate factors (e.g. Gross, 1993; Baron-Cohen, 1998), but this research is often criticised as mainly anecdotal and retrospective (Ericsson, Roring et al., 2007). Talent is also conceived in terms of personality (and its genetic components), effecting scientific performance (e.g. Busse & Mansfield, 1984; Feist, 1998; Feist & Barron, 2003). However, others claim that people are not born to be a genius (Howe, Davidson et al., 1998), as excellence is mainly determined by environmental factors, including early experiences, training, preferences and opportunities. If that is the case, talent should not be considered as a quality in itself, but more as innate *potential*. Talent is a process that enhances training and with that performance. It involves domain-specific

expertise (Simonton, 2008). Consequently, it is difficult to decide who is a talented researcher and who is not.

Selection panel members review and discuss grant proposals or job applications and jointly identify the most excellent ones – often using peer review reports. This decision making process entails among other things reference to one's expertise, explanation of preferences, discussion between proponents and opponents, obedience (or not) to procedures and rules, and finally reaching agreement. To study this process of scientific reviewing and decision-making, different theoretical approaches can be used. A well-known approach which prescribes how scientists should behave according to the norms and values of science, the so called 'ethos of science', is the Mertonian sociology of science (Bornmann, 2008). One of these norms is *universalism*, which means that the judgement of knowledge claims should be based on scientific criteria only, without interference by personal or social backgrounds of the reviewed and reviewers (Merton (1973 [1942]). Applied to talent selection, access to scientific careers should be based on scholarly competence alone. In this context talent relates mainly to scientific excellence. However, Lamont (2009) describes this type of evaluation as a social, emotional and interaction process. In an observation study of grant review panels, she shows that scientific excellence does not mean the same to everyone. Panel members from different fields, with a variety of motivations, use different criteria. And even within fields, people define excellence in various ways. As excellence is not the same for everyone, but subject to discussion and (dis)agreement, one may consider talent to be 'socially constructed' (Smith, 2001). More generally, emerged by criticism on the Mertonian sociology of science, social constructivism poses that scientific knowledge and the judgement thereof is constructed through interpretations, negotiations, and accidental events (Knorr-Cetina, 1981). Cole (1992) used some elements of the constructivism approach to make a distinction between the research frontier and the core of scientific knowledge. The frontier consists of new work which is in the process of being evaluated by the community. The core involves a small number of contributions which are accepted by the community as important and true. In this respect, there is a low level of consensus on frontier knowledge and a high level of consensus on core knowledge.

Even within the Mertonian norms, grant applications (and job applicants) are not evaluated and selected separately, but in comparison to competing applications (Smith, 2001). Quality is socially and contextually defined from a specific point of reference that evolves during the evaluation process (Lamont, 2009). As a result of this *contextual ranking*, one may expect that the same grant application can be valued differently across panels, process phases, and time. This is exactly what Cole & Cole (1981) found in their study on the reviewing of applications for research grants from the National Science Foundation (NSF). After reviewing

all and selecting half of the applications, a second group of peers reviewed and ranked the same set again. The two rankings differed substantially. Several proposals that were rejected by the NSF would have been granted if the selection had been based on the second ranking. What then determines whether a proposal is evaluated to be more excellent than the other? How is talent selected within peer and panel review?

Engaging peers is essential, as they are best suited to review the work of 'colleagues' within their specialty (Eisenhart, 2002). However, peers are often close to the applicants, and this creates tension between peer expertise and impartiality (Eisenhart, 2002; Langfeldt & Kyvik, 2011). This relates to another tension: peer reviews ought to be neutral, but not scholarly neutral. Personal interests should be eliminated and the evaluation should be based on scholarly discretion. But where are the boundaries? A third tension exists between unanimity and divergence. Grant review panels are expected to reach a unanimous decision, but at the same time divergence is considered of great value. Divergent assessments lead to discussion and contribute to the dynamics of science (Langfeldt & Kyvik, 2011). As scientific excellence is not unambiguous, but defined by reviewers and panel members in their own way, grant allocation clearly is a dynamic process.

Earlier studies on selection of applications focused mainly on *past performance* of the applicant.[1] Melin and Danell (2006) compared the past performance of successful and just unsuccessful applicants to the Swedish Foundation for Strategic Research. As the mean number of publications only slightly differed between the two groups, the awarded applicants can hardly be considered to be more productive than the rejected applicants. A study of the past performance of grant applicants in the Netherlands did find the expected difference in track record between awarded and all rejected applicants (Van den Besselaar & Leydesdorff, 2007; Van den Besselaar & Leydesdorff, 2009). However, comparing the past performance in terms of publications and citations of the awardees with the most successful rejected applicants, the latter have a slightly better average past performance than the awarded applicants. A later study found the same for German career grants (Hornbostel, Bohmer et al., 2009) and for international career grants in molecular biology (Bornmann, Leydesdorff et al., 2010). In their classical study on reviews of grant applications at the NSF Cole et al (1981) found a weak correlation between past performance and granted funding, concluding that the allocation of grants seems to be determined about half by characteristics of the applicant and the proposal, and about half by chance. Other research showed academic rank (Cole, Cole et al., 1981), research field (Laudel, 2006), type of research (Porter & Rossini,

[1] For a more elaborate literature review of the process of grant reviewing and group decision-making: van Arensbergen et al. (*forthcoming*), Olbrecht and Bornmann (2010). For an elaborate review of peer review including the reviewing of scientific articles, see Bornmann (2011).

1985), and academic and departmental status (Cole, Cole et al., 1981; Bazeley, 1998; Jayasinghe, Marsh et al., 2003; Viner, Powell et al., 2004) (weakly) correlate with quality assessment of the application or applicant. Interestingly, there is hardly any literature on the predictive validity of peer review: do the selected applicants have a better ex post performance than the non-selected (Bornmann, 2011; Van den Besselaar, forthcoming).

The chance element reported by Cole et al (1981) can partly be ascribed to the subjective character of the reviewing process and the social construction of scientific quality. According to Lamont (2009) it is impossible to completely eliminate this subjectivity, given the nature of the processes. The outcomes of the review process therefore are affected by who is conducting the review and how the panel is composed (Langfeldt, 2001; Eisenhart, 2002; Langfeldt & Kyvik, 2011). Different mechanisms can be discerned. Firstly, panel members who are nominated by the applicants, give higher ratings (Marsh, Jayasinghe et al., 2008). Secondly, relations between reviewers and applicants influence the ratings. Researchers affiliated with reviewers received better reviews than those without this type of affiliation (Sandstrom & Hallsten, 2008). Thirdly, the way the review process is organized, influences the outcomes (Langfeldt 2001). Finally, the importance of the gender dimension is often debated. Given the low number of females in academic top positions, and consequently the lack of female reviewers (Wennerås & Wold, 1997), and the persistence of the so-called glass ceiling an empirical analysis is hard to come by. The available empirical evidence provides contradictory results. Broder (1993) examines the rating of proposals from the National Science Foundation and finds that female reviewers rate female-authored NSF proposals lower than do their male colleagues. The study of Zinovyeva and Bagues (2011) showed that gender composition of committees in Spanish universities strongly affects the chances of success of candidates applying to full professors positions, but has no effect on promotions to associate professors. De Paola and Scoppa (2011) did a similar study in an Italian university and showed that gender in the composition of evaluation committees does matter. In competitions in which the evaluators are exclusively males, female candidates are less likely to be promoted. However gender discrimination almost disappears when the candidates are judged in a panel of mixed gender.

3. Data, research questions and methods

3.1 The case

Our dataset consists of 1539 career grant applications. These involve personal grants for researchers in three different phases of their careers:

- The early career grant scheme (ECG) for researchers who got a PhD within the previous three years. The grant offers them the opportunity to develop their ideas further.
- The intermediate career grant scheme (ICG) for researchers who have completed their doctorates with a maximum of eight years and already spent some years conducting post-doctoral research. The grant allows them to develop their own innovative research line and to appoint one or more researchers to assist them.
- The advanced career grants scheme (ACG) for senior researchers with up to 15 years post-doctoral experience, and who have shown the ability to successfully develop their own innovative lines of research and to act as coaches for young researchers. The grant allows them to build their own research group.

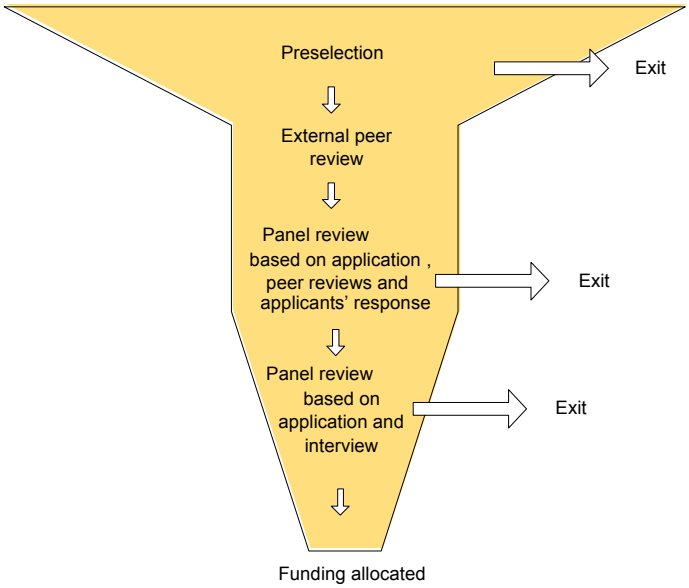

Figure 1. The general grant allocation procedure

Figure 1 briefly describes the selection procedure. If the number of applications in the ECG and ICG program is more than four times as high as the number of applications that can be awarded (as generally is the case), a pre-selection will take place – which resulted in our case into an overall rejection rate of about 40% of the applications, but with substantial differences between the fields. Because our dataset contains no further information on the criteria and assessments in the pre-selection, we do not include this phase in our study. In the ACG program, researchers first submit a pre-proposal. The selected applicants are

invited to submit a full more detailed proposal. Also the selection of pre-proposals is left out from our study, for the same reasons. This reduces the dataset to 897 applications.

Next the applications are sent to external referees, who are considered to be experts about the research of the applicant. The number of referees varies between two and six per proposal. The reviews and the applicants' rebuttal are sent to the review panel. Partly based on this input the panel evaluates every proposal on three criteria: quality of the researcher (QR), quality of the proposal[2] (QP), and research impact (RI)[3]. The score on research impact is only taken into account if it is better than the proposal score[4]. When this is the case (QP<RI), the final panel score is calculated as follows:

Total panel score = ½ QR + ¼ QP + ¼ RI

If the research impact is scored lower than the quality of the proposal (If QP>RI), the panel score is calculated as:

Total panel score = ½ QR + ½ QP

The total panel score leads to a ranking of the applications, which determines who proceeds to the next round: the interview, where the applicants present their proposal for the panel. Hereafter the panel again evaluates every interviewed applicant (N = 552) on the same three criteria, taking into account the information from the previous phases. To arrive at the final panel score, the same calculation rule is used as prior to the interview. The ranking of the final panel scores determines which applications will receive funding and which are rejected.

The research council consists of eight scientific divisions[5], which are aggregated into three domains[6]: 1) Social Sciences and Humanities (SSH), 2) Science, Technology and Engineering (STE), and 3) Life and Medical Sciences (LMS). In our analyses we will distinguish between these domains when relevant. Table 1 gives an overview of the number of applications per program and domain. As mentioned earlier we do not include the applications rejected in the pre-selection phase.

[2] More precisely this is the quality, innovative nature and academic impact of the proposed research.

[3] This is the intended societal, technological, economic, cultural or policy-related use of the knowledge to be developed over a period of 5–10 years.

[4] From 2012 the Research Impact score will always be included in the calculation of the total panel score.

[5] These are the following divisions: (1) earth and life sciences (ELS); (2) chemistry (CH); (3) mathematics, computer science and astronomy (MCA); (4) physics (PH); (5) technical sciences (TS); (6) medical sciences (MS); (7) social sciences (SS); (8) humanities (HU). About 7% of the applications are cross-divisional (CD).

[6] We aggregated the scientific dimensions to domain level as follows: SSH: social sciences and humanities; STE: chemistry, mathematics, computer sciences and astronomy, physics, and technical sciences; LMS: earth and life sciences, and medical sciences.

Our data include several attributes of

the applications and applicants: gender, the grant scheme, the scientific division and the domain of the application, the referee scores, the panel scores on the three criteria, and the decisions. Between a third (ACG) to a quarter (ICG) of the applications that made it through the pre-selection, received funding (table 1).

Table 1. Number of applications per scientific domain and funding program across the selection procedure

		ECG			***ICG***			***ACG***		
		1st review*	2nd review#	Granted	1st review	2nd review	Granted	1st review	2nd review	Granted
SSH	***N***	141	129	54	111	70	28	22	22	9
	%		91,5	38,3**		63,1	25,2		100,0	40,9
STE	***N***	151	70	40	124	65	33	34	34	12
	%		46,4	26,5		52,4	26,6		100,0	35,3
LMS	***N***	161	76	49	118	56	28	35	30	10
	%		47,2	30,4		47,5	23,7		85,7	28,6
Total	***N***	453	275	143	353	191	89	91	86	31
	%		60,7	31,6		54,1	25,2		94,5	34,1

*: external reviewers & 1st panel review; # 2nd panel review
**: If we include all applications, also those rejected in the pre-selection phase, the SSH success rate is lower than the two others. This is due to the very high rejection rate in the SSH pre-selection.

3.2 Research questions

The grant allocation procedure (figure 1) resembles a pipeline model. At the start there is a big pool of applicants, but as the procedure progresses the number of applicants decreases, with only a minority successfully reaching the end: receiving funding. In this study we aim to understand how applications pass the selection procedure and what determines which applications are eventually successful and which are expelled along the way. This should show how talents are identified or created by the selection process. We answer the following research questions:

1) *How evident is talent?*
   How strong are the correlations between the various reviewers' scores? The stronger they correlate, the more 'evident' talent is. Secondly, do scores vary strongly? Do the selected applicants have significantly higher scores than the non-selected? Thirdly, can a clear top be discerned, distinguishing top talents from the other talents?
2) *Is talent selection dependent on the procedure?*
   Do the rankings of applications in the different phases of the procedure correlate? Is the result stable, or does additional information in later phases result in strong fluctuations? And, are reviewers using the evaluation scales consistently through the procedure – do scores have a stable meaning?
3) *Which dimensions of talent can be distinguished?*

Do the three main criteria used by the panels represent different dimensions – or do they in fact measure the same? If they are different, are the rankings dependent on weighting the dimensions? And what does a change in weighting mean for the selection outcomes?

4) *Which phases of the process and which criteria eventually determine which applicants are considered to be talents*?
   A logistic regression analysis is used to identify which criteria and phases of the selection procedure have most influence on the final grant allocation decision.
5) *Is talent gender sensitive?*
   Does the gender composition of the panel influence the selection outcomes?

After answering these questions, we will discuss the implications of the findings for the system of selecting and granting research proposals.

3.3 Methods

Some of the following analyses are conducted on domain and program level, others on the complete dataset. In the latter case the data is standardized beforehand on the domain and program variables. This was done through calculating the z-scores at the level of programs and fields.

Agreement between reviewers is analysed by calculating the standard deviation and the maximum difference between review scores per application and by rank order correlation. We will rank the review scores per step of the selection process and compare these rankings to see if applicants were evaluated differently at various moments of the procedure. The use of the evaluation scale is analysed with Chi-square tests. Rank order correlations are calculated between the three evaluation criteria used by the panels. This will show whether talent has one or various dimensions. Finally, to identify the predictors for talent selection we conducted multiple logistic regression analysis.

4. Results

Evaluation practices differ between the scientific domains and the funding programs (for more details see Van Arensbergen & Van den Besselaar, 2012). Therefore we will distinguish between the three scientific domains and funding programs in our analyses.

4.1 The evidence of talent

The applications are refereed by external reviewers and (twice) by a panel. The number of external reviewers per proposal varies between two and six.[7] In general there are two reviewers for the ECG, three for the ICG and four for the ACG. The external reviewers assign scores from 1 (highest) to 6 (lowest). We calculated the difference between the maximum and minimum review score per proposal. As table 2 shows, the reviewers disagree least in the ECG scheme (M = 1,59; SD = 1,27) and most in the ACG scheme (M = 2,22; SD = 1,33). The level of disagreement differs significantly between the schemes, $F(2, 895) = 18{,}72$, $p < .001$, indicating that the further an applicant is in his career, the stronger the average disagreement about his quality.

Taking into account that the number of reviewers varies per grant scheme, we compare the average distribution of review scores per proposal (mean standard deviation, table 2). The standard deviation can range from 0 (if all reviewers totally agree) to 3.54 (when reviewers totally disagree). However, no significant difference was found between the programs. Although the maximum disagreement between reviewers increases with the career phases, the mean disagreement remains the same. The higher number of reviewers in the IGC and ACG scheme explains this: the more reviewers per proposal, the smaller the weight of reviews with extreme scores.

We repeated the analysis for each of the domains, to find out whether agreement on talent differs between the domains. Only in the ICG scheme the average disagreement (standard deviation) between reviewers significantly differs between the domains ($F(2,351) = 5.25$, $p < .01$). In the ECG and ACG schemes no significant differences were found. Finally, in all career phases the reviewers in the Social Sciences and Humanities seem to disagree stronger than in the other domains.

Table 2. Disagreement in evaluations by external referees per domain and funding program

	Early Career Grant		Intermediate Career Grant		Advanced Career Grant	
	Maximum disagreement*	Average Disagreement**	Maximum disagreement*	Average disagreement**	Maximum disagreement*	Average disagreement**
All	1,57	1,05	2,06	1,10	2,22	1,06
- SSH	1,60	1,13	2,25	1,21	2,75	1,28
- STE	1,68	1,09	1,76	0,95	2,03	0,95
- LMS	1,45	0,94	2,18	1,16	2,08	1,02

* Mean of *maximum difference between review scores per application*
** Mean of *standard deviation review scores per application*

The selection of interview candidates is done by a panel, taking into account the external reviews and the applicants' rebuttal. The correlation between the standardized external

[7] Note that the applications are sent to different external reviewers, so generally reviewers are involved in the evaluation of a single application.

review scores and the panel reviews is used to determine the extent to which evaluators in different phases of the procedure agree on the quality of applicants. In all domains the external reviews correlate moderately strong (ECG and ACG: $\tau = .53$, $p < .001$; ICG: $\tau = .52$, $p < .001$) with the first panel scores.[8] After the interview, the same panel evaluates the applicants again including the new information. The correlation between the panel scores prior to and after the interview is also moderately strong in the domains of STE and LMS ($\tau = .42$, $p < .001$) and strong in SSH ($\tau = .62$, $p<.001$).

The average scores are used to distinguish between the talented and the less talented applicants, but how strong to these scores discriminate? We ranked (for the complete set and per domain) all applications using the standardized average review score. As Figure 2 shows for the complete set, the distribution has no clear-cut off point, and a similar pattern exists at domain and program level. The dotted line indicates the de facto cut off point of applications selected for the next (interview) phase. However, this selection boundary does not follow from the scores, as the difference between success and just no success is very small. Similar patterns were found for the panel scores, where the difference between success and failure is very small too.

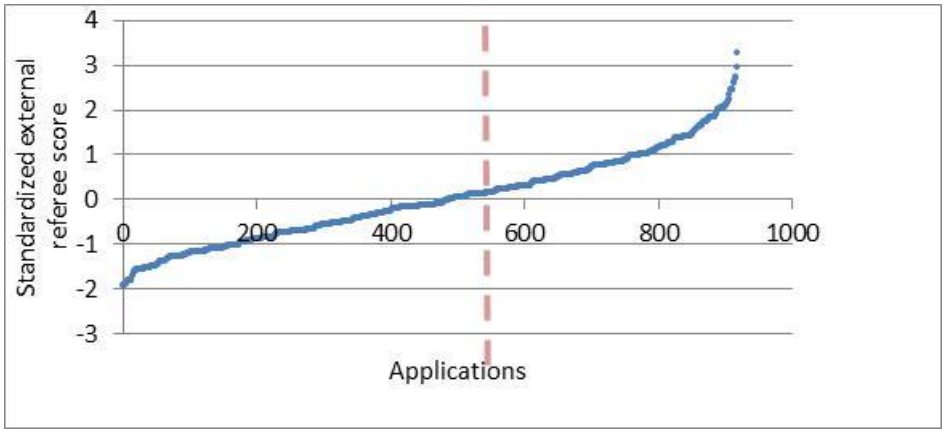

Figure 2. Standardized external referee scores for the complete set of applications

Concluding, no clear 'boundaries of excellence' could be identified between selected and not selected applicants. Moreover, the average scores in the three phases of the procedure only correlate moderately strong, and that may reflect considerable changes between the rankings. This issue will be addressed in the next section, after we have looked into the evidence of top talents.

[8] Since the data set is characterized by a large number of tied ranks, we use Kendall's tau instead of Spearman's rho.

Top talents

Figure 2 showed no clear delineation of talent, but more gradual differences in talent assessment. Experienced reviewers often claim to "easily identify the real top, there are always a few top talents who stand out from the beginning" (Van Arensbergen, Van der Weijden et al., forthcoming). To test this claim, we looked at the average total review scores per panel in order to identify the top talents. We determined i) the number of positive outliers (= exceptionally high scores) in the evaluation round prior to and after the interview; ii) the distance between the outliers and the best of the gross evaluation scores; iii) the number of stable outliers (the same outliers in both evaluation rounds).

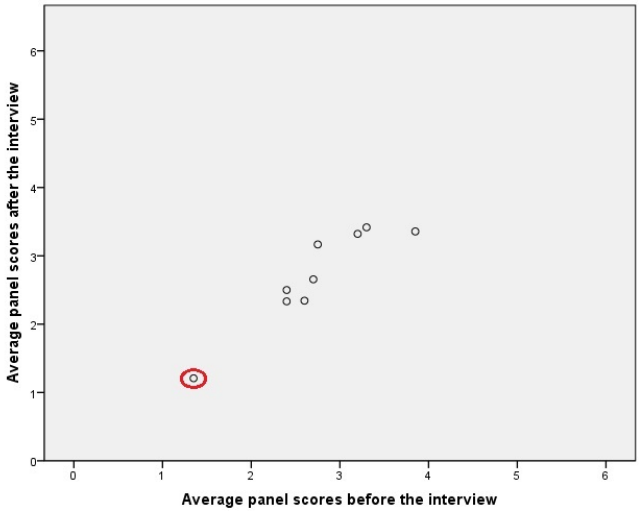

Figure 3a. Average panel scores before and after interview in panel which clearly identified a top talent in both review rounds

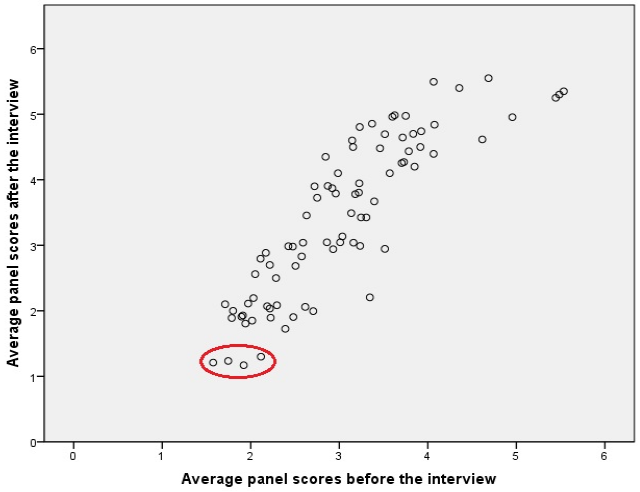

Figure 3b. Average panel scores before and after interview in panel which identified top talents only after the interview

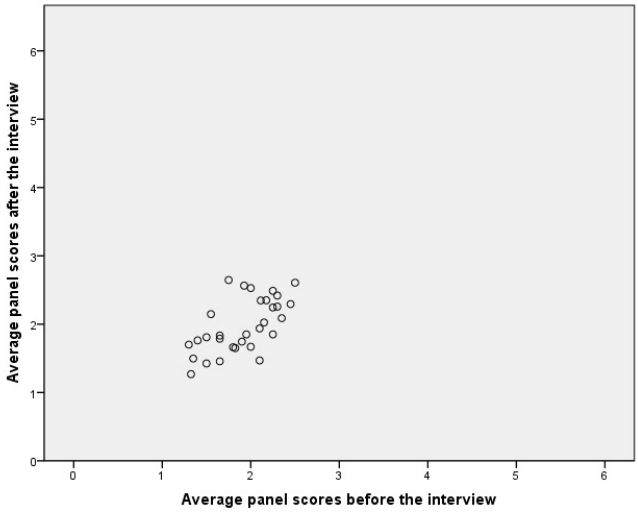

Figure 3c. Average panel scores before and after interview in panel which identified no top talents

Figure 3a is an example of a panel that clearly identified a top talent both before and after the interview. Figure 3b shows that a clear top was identified only after the interview. Looking at the x-axis, the four applicants eventually identified to be the top talents did not stand out in the eyes of the panel members before the interview. An example of a case in which no top is recognizable, but all applicants being close together is depicted in figure 3c. In general we found that a clear top was identified more often after the interview than before (table 3), making figure 3b most representative for the 27 panels. In more than half of the panels no applicants stood out from the rest before the interview, while after the interview twenty of the panels identified a top. This top predominantly consists of one person, with a maximum of four. For example, seven panels identified one top talent in the first selection phase, whereas two panels identified four top talents.

Also after the interview the distance between the (lowest in the) top and the (highest in the) middle group is on average a little larger (0.51, SD = 0.19) than before the interview (0.48, SD = 0.18). Panel members use an evaluation scale from 1 to 6. These average distances of 0.48 and 0.51 clearly differentiate a top from the large middle area, where there is much overlap and most applications are very close to each other in terms of their review scores (see figures 2 and 3).

Table 3. Number of panels (n = 27) which identified top talents before and after the interview, and which identified the same top talents in both selection phases

Number of identified top talents	Before interview	After interview	Both before and after interview
0	14	7	17
1	7	8	7
2	3	3	1
3	1	6	2
4	2	3	0

When we look at the stability of the top, we found that only in a few cases the same applicants were identified as top talents both before and after the interview. In 17 out of 27 panels, none of the applicants was identified as a top talent in both the evaluation rounds. In seven panels we discerned one stable top talent. In total of the 53 applicants who were in the top at some point of the evaluation process, 15 belonged to the top in both rounds and can be considered to be stable top talents. But the far majority of selected applicants (210 out of 263) was never scored as exceptional talent.

4.2 Effects of the procedure

The selection procedure includes three evaluation phases in which new information is added which may influence the resulting assessment. Figures 4 and 5 show how applications are evaluated differently at different moments of the procedure, based on the standardized review scores. Right of the diagonal in figure 4 are the applications that had a better (= lower) first panel score than external review score. On the left side are the applications that had a better external review score. Clearly, the scores and the relative position of applications changes during the procedure. If external (peer) review scores would have been leading, the set of applicants invited to the interview would have been rather different. Since both evaluations are based on about the same information, this implies that talent evaluation depends on the way it is organized – it is 'contextual'.

In figure 5, the panel reviews before and after the interview are compared, with right from the diagonal those applications that score lower (= better) after the interview than before, whereas left of the diagonal the opposite is the case. Panels adjust their assessments after the interview, and quite some applicants' score rather different after the interview compared with before.

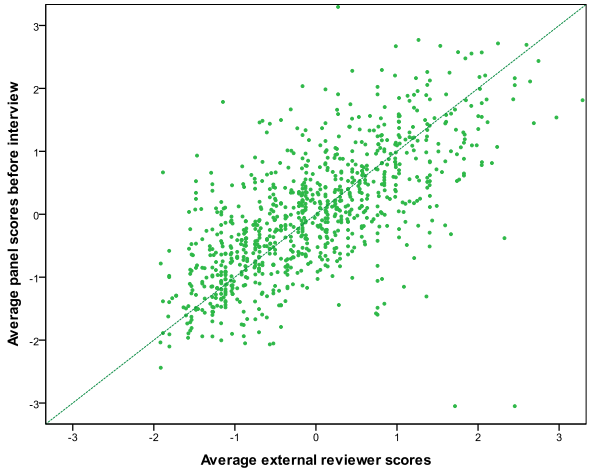

Figure 4. 1st panel review by external referee score

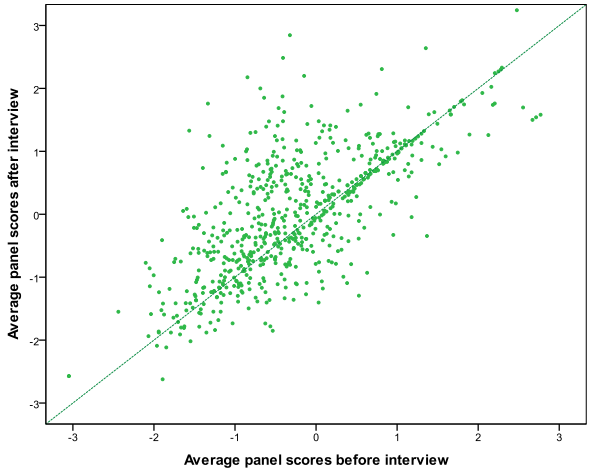

Figure 5. 2nd panel review by 1st panel review

This implies that if grant allocation had been based on the evaluation scores before the interview, the outcome would have been different. How strong is this effect? To answer that question, we compare the rankings of applications between the three evaluation moments, showing the importance of the various phases of the selection process.[9] We found that 48 (17%) of the interview candidates would not have been invited for the interview if the external referee scores had been paramount. According to the procedure, the panel score is decisive. However, there were 24 rejected applicants with a higher total panel score than the selected applicants. This means that 9% of the successful applicants was not selected because they were among the highest total panel scores. The panel thus has in fact additional autonomy in decision-making.

Grant allocation is the final step in the selection procedure. If the grant allocation had been based entirely on the evaluation by the external referees, 26% of the applicants would not have been allocated a grant. If interviews would not have been part of the procedure – and this is the case in many funding schemes – and the first panel reviews would have determined the grant allocation, 22% would have been allocated to currently unsuccessful applicants. These findings imply that the interview considerably changes the assessment of talent.[10] As the procedure prescribes, the eventual allocation decision largely corresponds to the final panel score, only 2% of the granted applicants had a lower panel score than the best rejected applicants.

Differentiating between the funding programs and scientific domains, differences were found between domain-program combinations, but no pattern could be identified (for more details see Van Arensbergen & Van den Besselaar, 2012).

What do the scores represent?

After showing how the *perception of talent* did change, we will now study changes in the *use of the evaluation scale* (as distinct from the evaluation of the applications). The six point scale ranges from excellent (1), very good (2), and very good / good (3), to good (4), fair (5) and poor (6), clearly an 'absolute scale'. The panel members assign a score between 1 and 6 to each application on three criteria (quality researcher, quality proposal and research impact). Table 4 shows the mean scores and standard deviations for two typical evaluation panels, before and after the decision about which applicants are invited for an interview.

[9] In some divisions and in the ACG all applicants were invited for the interview; these are excluded from this part of the analysis.

[10] In a follow-up study we investigate the dynamics, the criteria (implicitly) applied, and the effects of the interview (Van Arensbergen, Van der Weijden et al., forthcoming).

Table 4. Use of evaluation scale

Case	program & domain		1st panel review all applications			1st panel review selected applications			2nd panel review selected applications		
			researcher	proposal	total	researcher	proposal	total	researcher	proposal	total
1	ECG-STE	N	34	34	34	18	18	18	18	18	18
		Mean	2.89	3.43	3.10	2.28	2.87	2.55	2.56	3.23	2.84
		SD	.84	.78	.73	.54	.45	.47	.83	.95	.84
2	ICG-SSH	N	34	34	34	34	34	34	34	34	34
		Mean	1.69	2.18	1.91	1.69	2.18	1.91	1.79	2.20	1.98
		SD	.38	.43	.35	.38	.43	.35	.38	.49	.39

In case 1, about 50% highest scoring applications were selected. As expected, the means for all applicants (1st review, all applications) are lower than the means for selected applicants only (1st review selected applicants).[11] The standard deviation of the whole set of applicants is larger than for the selected only – indicating an expected smaller variation among the selected applicants. However, average and standard deviation of the scores *after* the interview (2nd panel score) are equal to the values for all applicants in the 1st review, suggesting that the panel again has applied *the whole scale*: some of the applications scoring very good and excellent in the first round are now only fair or even poor. In this case, the scale is used in a *relative* way, and not as an *absolute* one. In case 2 no selection took place, as all applicants were interviewed. The interview did influence individual scores, but the average and the standard deviation before and after the interview remain about the same. No changes in the use of the scale seem to have occurred here.

Comparing the 14 'selective' panels with the 12 'non-selective' panels (in table 5) shows a significant correlation between the change of context (selection between the phases or not) and the change of the use of the scale (relative or absolute scale). Consequently, the assessment of talent depends on the context, on the procedure: e.g., an interview, as showed in the previous section, and the number of competitors, as showed in this section.

Table 5. Changing use of the scores by changing context (n = 26)

		reduction of nr applicants after 1st panel evaluation			
		yes[a]		no	
decrease average score*	no	4	(28.6%)	10	(83.3%)
	yes[b]	10	(71.4%)	2	(16.7%)
increase standard deviation**	no	4	(28.6%)	8	(66.7%)
	yes[b]	10	(71.4%)	4	(33.3%)
Total		14	(100%)	12	(100%)

[a] yes = changing context.
[b] yes =using the score values in a relative way.
* X^2=7.797, p=0.005; ** X^2=3.773, p=0.05.

4.3 The dimensions of talent

[11] Please note that also here lower scores correspond with higher numbers.

Earlier we showed that the external reviews correlated moderately strong with the panel reviews. Distinguishing between the three criteria used by the panel shows that this moderate correlation is mainly due to a relative weak correlation between external reviews and the panel scores for research impact, $\tau = .22$, $p<.001$ (SSH); $\tau = .29$, $p<.001$ (STE); $\tau = .36$, $p<.001$ (LMS). In the LMS domain however, the external referee scores correlate even weaker with the panel scores for the researcher, $\tau = .32$, $p<.001$. The external reviews are strongest related to the panel scores for the proposal, $\tau = .55$, $p<.001$ (SSH); $\tau = .55$ $p<.001$ (STE); $\tau = .64$, $p<.001$ (LMS).

Table 6. Correlations between the standardized panel review scores for the three criteria per domain

		SSH		STE		LMS	
		QR	QP	QR	QP	QR	QP
Before	QP	.50*		.50*		.44*	
interview	RI	.33*	.41*	.38*	.49*	.37*	.47*
After	QP	.57*		.56*		.59*	
interview	RI	.37*	.49*	.31*	.44*	.41*	.47*

QR = quality researcher; QP = quality proposal; RI = Research impact
* $p<.001$

The three criteria are found to correlate moderately with each other (table 6). Research impact correlates weakest to the quality of the researcher in all domains both before and after the interview, ranging from $\tau = .31$ to .41. The correlation between quality of the proposal and quality of the researcher increased after the interview in all domains, strongest in LMS, from $\tau = .44$ before the interview to $\tau = .59$ after the interview.

This suggests that the three criteria represent different dimensions. The total score of the panel (as calculated with the formulas from the method section) therefore depends on the weights attributed to the different dimensions. This may change with the decision making context. In 2012 a change in the weighting of the research impact score was implemented in the review procedures. From now on, research impact accounts for 20% of the total panel score, and the quality of the researcher and the proposal both for 40%. We applied this new procedure to our dataset to explore how this would affect the selection outcomes.

The issue that comes up, is to what extent the changing of weights influences the selection procedure: would other applicants have been selected if the three criteria are weighted differently? To answer that question we did some simulations, in which we change the weights. Two analyses can be done. (i) A rank order correlation between the different simulated scores informs us about the impact of the scores. The lower the rank order correlation, the more effect the weighting has on the resulting order of applicants. This, by the way does not imply that changing the weight would also influence the decisions, as the altered rank order may be within the set of successful and within the set of unsuccessful

applicants. Therefore (ii) one should check whether the changed order would move applicants from below the success threshold to a place above the threshold and vice versa.

(i) Does changing weights imply changes in the rank order?

We simulated the outcomes using five different sets of weights, as shown in table 7. We check it here for the first decision whether an applicant is invited or rejected for the interview. For each of the sets, we calculated the score of the applicant, and this leads to five rank orders. Using Spearman's Rho (see table 8).

Table 7. Used weights for the three criteria

Weights:	**1**	**2**	**3**	**4**	**5**
Researcher	0,5	0,5	0,33	0,4	0,4
Proposal	0,5	0,5	0,33	0,4	0,2
Societal impact	0	+	0,33	0,2	0,4

+: If 'societal impact' scores higher than proposal, a new value for 'proposal' is calculated as the mean of the old value of 'proposal' and the value of 'societal impact'

Using these weights, we found for the interview selection that the rank order correlations are rather high. Within almost all instrument/field combinations, Rho remained between 0.83 and 0.97 (table 8, left part). The lowest correlations (between 0.62 and 0.80) were all between weights set 1 (where societal impact would not be taken into account) and weights set 5 (where societal impact would be strongly taken into account). If it is taken into account, the exact weight may not be very important for the rank order of the applications, as the correlation remains in all cases above 0.83. For the granting decision, we find a similar pattern (table 8, right part).

Table 8. Simulations: average correlations between rank orders based on five weights for each funding program and field*

	decisions before the interview			**decisions after the interview**		
	ECG	ICG	ACG	ECG	ICG	ACG
ELS	0,93	**	0,90	0.89	**	0.97
CH	0,91	0,87	0,93	0.94	0.82	0.97
MCA	0,90	0,92	0,90	0.82	0.84	0.88
CD	0,94	0,90	0,97	0.95	0.90	0.83
HU	**	**	0,88	**	**	0.99
SS	0,84	0,88	0,84	0.83	0.92	0.93
PH	0,87	0,93	0,97	0.96	0.98	0.92
TS	0,88	0,88	0,96	0.89	0.89	0.99
MS	0,83	0,88	0,88	0.86	0.92	0.86

* We use here the more detailed division in fields (see notes 5 and 6)
** Societal impact scores not available

(ii) What would this mean in terms of the decisions and success rates?

We checked that for the final decisions in the ECG and ICG programs. Table 9 shows the findings.

Table 9. Scenario 5* versus scenario 2**: Number of different grantees

	ECG		ICG	
	different grantees	%	different grantees	%
ELS	1	5.6	-	-
CH	1	10.0	4	57.1
MCA	3	33.3	3	50.0
CD	1	11.1	2	28.6
SS	3	10.3	0	0.0
PH	0	0.0	0	0.0
TS	1	8.3	1	5.9
MS	3	11.1	1	5.0
Total	13	10.4	11	14.4

* Impact with heavy weight
** Reality (until 2012)

The table shows that the selection of grantees does depend on the selected weights. Scenario 5 would have changed the grant allocation between 10.4% (ECG) and 14.4% (ICG), and this is of course important for the involved applicants. Furthermore, the table shows that there is large variety between the fields, as in some fields the percentage of different grantees under scenario 5 would be more than 50%. Independently of whether this would have an effect on the science system, the analysis suggests that what counts as talent, indeed is context dependent.

4.4 Predictors for talent selection

The first decision is when panels select and reject applications for the interview round, based on the external reviews, the applicants' responses to these reviews, and the panels' own scoring on three criteria. In order to determine which of these variables best predict whether an application will be selected for the interview, we conducted a stepwise logistic regression analysis, including the average external referee score and the three panel scores[12].

The model with only the external reviews predicts in 69.1% of the cases correctly who goes through to the interview, slightly above the random correct prediction of 61.5%. In the full model, only the panel scores for the quality of the proposal and the researcher's quality are

[12] As the following results show, the stepwise method eliminates two variables since they do not contribute significantly to the model.

included, whereas the other variables are excluded (table 10). These two variables predict in 77.3% of the cases correctly whether a researcher was invited for the interview or not.

Table 10. Logistic regression of the selection of interview candidates

	B (SE)	X^2 (df)	Nagelkerke R^2	% correct
Constant	-0.61* (0.10)			
Quality Researcher	0.71* (0.13)			
Quality Proposal	1.36* (0.15)			
Model		283,96* (2)	.48	77,3
Not included				
External Reviews	*0.23 (0,16)*			
Research Impact	*0.15 (0.14)*			

*p < .001

After the interviews, the panel again scores the applications on the three criteria. A logistic regression analysis was done to predict the allocation decisions from the external referee scores and the three panel scores (table 11). Again, external referee scores and the research impact scores do not contribute significantly to the prediction. The panel scores for the proposal and for the researcher result into a correct classification in 83.1% of the cases. The model with only the external reviews predicts in 65.2% of the cases correctly who receives funding, slightly above the random correct prediction of 52,3%.

Table 11. Logistic regression analysis to predict grant allocation decisions

	B (SE)	X^2 (df)	Nagelkerke R^2	% correct
Constant	0.46* (0.15)			
Quality Researcher	1.40* (0.23)			
Quality Proposal	1.80* (0.23)			
Model		294.97* (2)	.65	83,1
Not included				
External Reviews	*0.21 (0.18)*			
Research Impact	*0.08 (0.19)*			

*p < .001

Distinguishing between the three funding programs, in short we found that for early career researchers to a greater extent other factors are taken into account in the decision-making. And the *de facto* weights of both included criteria are found to differ between the funding programs. For the early career researchers the evaluation of the proposal and the researcher almost evenly determine the final selection decision, whereas for the intermediate and advanced career researchers the quality of the proposal is more important than the quality of the researcher (for more details see Van Arensbergen & Van den Besselaar, 2012).

4.5 Is talent gendered?

As suggested in the literature, panel composition is often found to influence decision-making: decisions of panels with no or only a few female members are found to be gender biased. As councils increasingly claim to support female applications, it is interesting to investigate whether this effect still exists. Do 'male dominated' panels still exist, and if so, do these panels decide more often in favor of male than of female applicants? If no gender bias would exist, then one would expect that the percentage of granted application within the set of female applicants is similar to the percentage of granted applications in the set of male applicants.

This is under the assumption that the male and female applicants and applications are in average of equal quality. One way of tentatively testing this is by comparing the referee scores for female and male applicants. These are individually given by external reviewers – before the proposals enter the decision making process. We found that the mean score of male applicants is slightly higher (9%) than the average score of female applicants. In most fields, this difference is not statistically significant (if we may consider the data as a random sample), and as far as the differences are significant, it is in the more advanced career schemes. For the early career grants (ECG), differences are small(er) and never significant. The latter is in line with the findings about disappearing gendered performance difference in the younger generations of researchers (Van Arensbergen, Van der Weijden et al., 2012). We therefore assume that the – comparable – peer review scores are hardly gender biased - if at all (Marsh, Bornmann et al., 2009).

We analyze here the relation between gender composition of panels and the final selection decision. One may do the same for the interview decision. Figure 6 shows gender bias by the number of women in the panel. As the figure shows, there are still panels with no or only one female member. However, one cannot conclude that these panels show a gender bias against female applicants. In the lower range of female panel membership, we actually find a large variation in the bias variable. If there is a pattern, it seems actually more often in favor of female applicants. Panels with larger numbers of female members consistently seem to have no gender bias in the decisions.

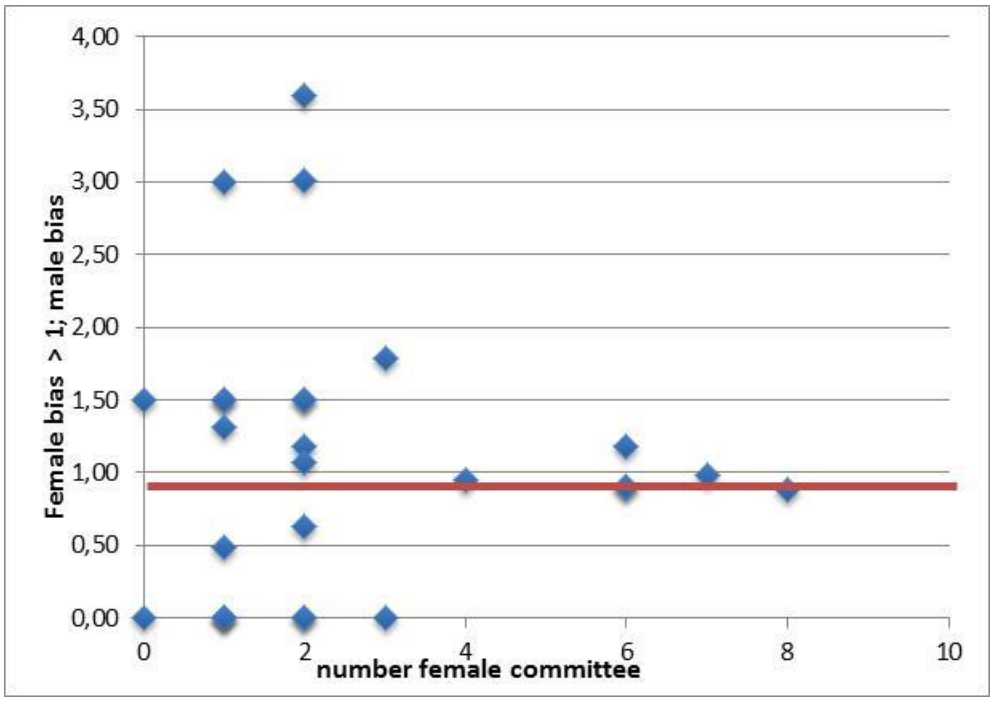

Figure 6. Gender bias by number of female panel members

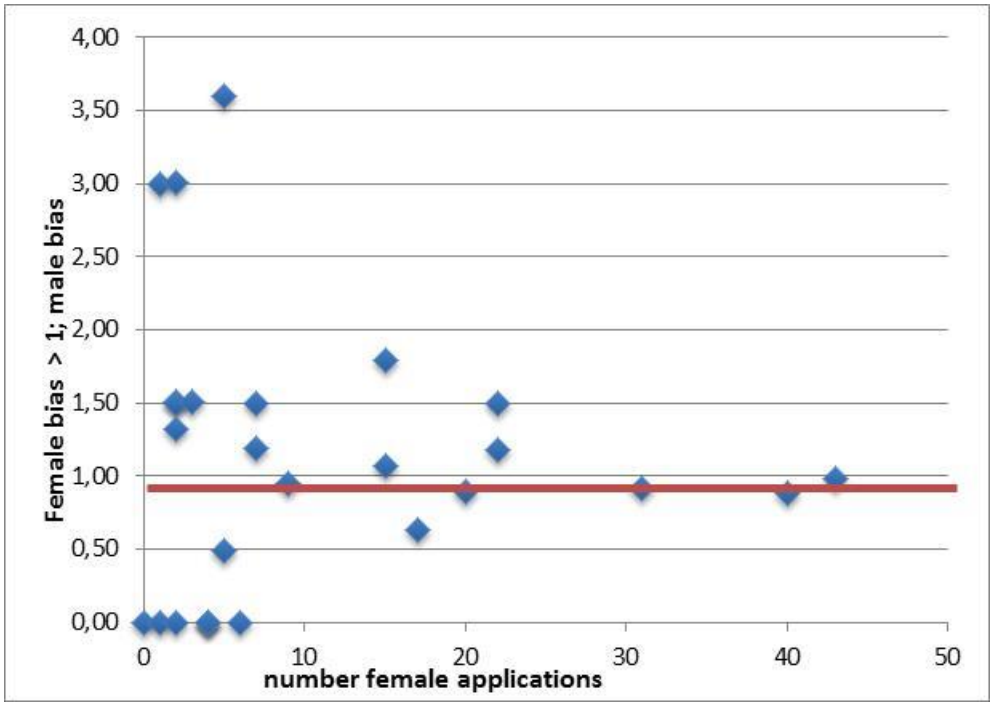

Figure 7. Gender bias by number of female applicants

Why this difference occurs needs further investigation. However, one factor may be whether a field has many or only a few female applicants. In the latter case, the success rate of women is heavily influenced by a single decision. Indeed, as figure 7 shows, in the fields with few female applicants, the spread in success rate is large, whereas this is not the case in fields with many female applicants. And, one may expect that fields with only a few female

applicants also have rather male dominated panels – as these fields may simply lack female researchers to occupy panels. A study of Van den Brink (2009), suggests a gender bias in promotion decisions is due to the composition of panels. However, we cannot conform this, as our data suggest no correlation between the number or percentage of women in a panel and the gender bias in the results.

5. Conclusions and discussion

First of all, the moderate correlations between the criteria indicate that talent has different dimensions. This implies that the weight of the criteria may strongly influence the selection process. E.g. the weight of research impact is very low in the case we studied, but the current tendency to include expected societal impact more strongly in the evaluation of proposals, potentially leads to the selection of other types of applicants as "the most excellent". However, our simulations suggest that this may only happen if the weight of the societal relevance criterion is more substantial than currently implemented.

Secondly, the scores change considerably between the phases. Some applicants, top ranked by the external referees, are not even invited for an interview by the panels. And these same panels regularly change their evaluation of applicants radically after the interview. A clear top can more often be distinguished after the interview than before, however the actual number of identified top talents is relatively low. The interview seems decisive, but how this works needs further investigation. Does the interview bring new information, leading to a different evaluation? In that case the procedure does influence the outcome considerably, which can of course be intended and desirable. Shouldn't then the many existing procedures without interviews be abandoned?[13] Or is it because other aspects of talent (such as communicative skills) and several cognitive, motivational and social processes (Lamont, 2009) play a role during the interview, as well as various psychological factors (Hemlin, 2009)?

Thirdly, the role of the external peer review in the quality assessment seems modest (Langfeldt, Stensaker et al., 2010). Using only external review scores as predictor, the percentage correctly predicted applications is only slightly higher than random (65,2% versus 52,3%), much lower than for the two other predictors (83.1%). Combined with the moderately (but not very) high correlation ($\tau = .52$) between reviewers scores and panel scores, this suggests that the panel takes the review scores into account, but not very strong.[14] Further

[13] Interestingly, the very prestigious ERC advance grants do not include an interview with the applicants.

[14] This is in line with the findings by Hodgson (1995), and contrasts with the findings of Bornmann

study is needed, to reveal whether and how panel members value and use the peer review reports.

Fourthly, reviewers disagree, and the further a researcher is in his/her career, the more reviewers disagree. In line with earlier studies, consensus about quality is lower in the social sciences and humanities than in sciences, technical sciences and life sciences (Cicchetti, 1991; Simonton, 2006). Panels and external reviewers also do not draw a clear line between talented and less-talented researchers, as for the middle group very small differences in scores eventually decide who receives a grant and who does not. As the funding decisions are of great importance for the careers of (especially) young researchers, career success becomes partly a question of luck.

Finally, the composition of the panel does not seem to result into a gender bias in the decisions. This suggests that councils' policies to stimulate female participation in science, seems effective – at least at the level of their panels. Under these conditions, gender bias in outcomes seems to be related to the low number of female candidates in some fields.

Summarizing, our findings clearly indicate the contextuality of evaluation and decision-making. For improving transparency, quality and legitimacy of grant allocation practices, it would therefore be important to uncover more deeply the details of the *de facto* (implicit and explicit) applied criteria. As the selection procedure influences the evaluation of scientific talent, we suggest using a variety of procedures, instead of standardizing. The interview was found to have an important impact on the evaluation of the applicants. If communicative skills and self-confidence are decisive in this phase of the procedure, the selection outcomes will be biased towards these qualities, when all procedures would include interviews. Since no evident pool of talents could be identified based on the various scores, and as differences between granted and eventually rejected applications were small, a variety of procedures may result into the selection of a variety of talent.

References

Addis, E. & M. Brouns, Eds. (2004). *Gender and excellence in the making*. Bruxelles.

Baron-Cohen, S. (1998). Superiority on the embedded figures test in autism and in normal males: Evidence of an "innate talent"? *Behavioral and Brain Sciences,* 21(3): 408-+.

Bazeley, P. (1998). Peer review and panel decisions in the assessment of australian research council project grant applicant: What counts in a highly competitive context?. *Higher Education*, 35: 435-452.

---

et al. (2008).

Bornmann, L. (2008). Scientific peer review. An analysis of the peer review process from the perspective of sociology of science theories. *Human Architecture: Journal of Sociology of Self-Knowledge*, 6(2): 23-38.

Bornmann, L. (2011). Scientific peer review. *Annual Review of Information Science and Technology*, 45: 199-245.

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

Bornmann, L., R. Mutz & H. D. Daniel (2008). Latent markov modeling applied to grant peer review. *Journal of Informetrics*, 2(3): 217-228.

Broder, I. E. (1993). Review of nsf economic proposals: Gender and institutional patterns. *American Economic Review*, 83(4): 964-970.

Busse, T. V. & R. S. Mansfield (1984). Selected personality-traits and achievement in male scientists. *Journal of Psychology,* 116(1): 117-131.

Cicchetti, D. V. (1991). The reliability of peer-review for manuscript and grant submissions - a cross-disciplinary investigation. *Behavioral and Brain Sciences*, 14(1): 119-134.

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

Cole, S., J. R. Cole & G. A. Simon (1981). Chance and consensus in peer review. *Science,* 214: 881-886.

De Grande, H., K. De Boyser & R. Van Rossem (2010). Carrièrepaden van doctoraathouders in belgië. Loopbaanpatronen naar wetenschapsgebied. U. Gent.

De Paola, M. & V. Scoppa (2011). Gender discrimination and evaluators gender: Evidence from the italian academy. Italy, Department of Economics and Statistics. University of Calabria. Working paper No. 06-2011.

Eisenhart, M. (2002). The paradox of peer review: Admitting too much or allowing too little? *Research in Science Education*, 32(2): 241-255.

Ericsson, K. A., R. W. Roring & K. Nandagopal (2007). Giftedness and evidence for reproducibly superior performance: An account based on the expert performance framework. *High Ability Studies*, 18(1): 3-56.

Feist, G. J. (1998). A meta-analysis of personality in scientific and artistic creativity. *Personality and Social Psychology Review*, 2(4): 290-309.

Feist, G. J. & F. X. Barron (2003). Predicting creativity from early to late adulthood: Intellect, potential, and personality. *Journal of Research in Personality,* 37(2): 62-88.

Gross, M. U. M. (1993). *Nurturing the talents of exceptionally gifted individuals. International handbook of research and development of giftedness and talent*. K. A. Heller, F. J. Mönks and A. H. Passow, Routledge: 473-490.

Hemlin, S. (1993). Scientific quality in the eyes of the scientist - a questionnaire study. *Scientometrics*, 27(1): 3-18.

Hemlin, S. (2009). Peer review agreement or peer review disagreement. Which is better? *Journal of Psychology of Science and Technology,* 2(1): 5-12.

Hodgson, C. (1995). Evaluation of cardiovascular grant-in-aid applications by peer-review - influence of internal and external reviewers and committees. *Canadian Journal of Cardiology,* 11(10): 864-868.

Hornbostel, S., S. Bohmer, B. Klingsporn, J. Neufeld & M. von Ins (2009). Funding of young scientist and scientific excellence. *Scientometrics,* 79(1): 171-190.

Howe, M. J. A., J. W. Davidson & J. A. Sloboda (1998). Innate talents: Reality or myth? *Behavioral and Brain Sciences*, 21(3): 399-+.

Huisman, J., E. de Weert & J. Bartelse (2002). Academic careers from a european perspective - the declining desirability of the faculty position. *Journal of Higher Education,* 73(1): 141-+.

Jayasinghe, U. W., H. W. Marsh & N. Bond (2003). A multilevel cross-classified modelling approach to peer review of grant proposals: The effects of assessor and researcher attributes on assessor ratings. *Journal of the Royal Statistical Society Series a-Statistics in Society*, 166: 279-300.

Knorr-Cetina, K. (1981). *The manufacture of knowledge: An essay on the constructivist and contextual nature of science*. Oxford, UK, Pergamon Press.

Lamont, M. (2009). *How professors think. Inside the curious world of academic judgment*. Cambridge London, Harvard University Press.

Langfeldt, L. (2001). The decision-making constraints and processes of grant peer review, and their effects on the review outcome. *Social Studies of Science*, 31(6): 820-841.

Langfeldt, L. & S. Kyvik (2011). Researchers as evaluators: Tasks, tensions and politics. *Higher Education*, 62: 199–212.

Langfeldt, L., B. Stensaker, L. Harvey, J. Huisman & D. F. Westerheijden (2010). The role of peer review in norwegian quality assurance: Potential consequences for excellence and diversity. *Higher Education,* 59(4): 391-405.

Laudel, G. (2006). The 'quality myth': Promoting and hindering conditions for acquiring research funds. *Higher Education,* 52: 375-403.

Marsh, H. W., L. Bornmann, R. D. Mutz & A. O'Mara (2009). Gender effects in the peer reviews of grant proposals: A comprehensive meta-analysis comparing traditional and multilevel approaches. *Review of Educational Research,* 79: 1290-1326.

Marsh, H. W., U. W. Jayasinghe & N. W. Bond (2008). Improving the peer-review process for grant applications. Reliability, validity, bias, and generalizability. *American Psychologist,* 63(3): 160-168.

Melin, G. & R. Danell (2006)."The top eight percent: Development of approved and rejected applicants for a prestigious grant in sweden. *Science and Public Policy,* 33(10): 702–712.

Merton, R. K. (1973 [1942]). *The normative structure of science. The sociology of science. R. K. Merton.* Chicago, The University of Chicago Press: 267-278.

Olbrecht, M. & L. Bornmann (2010). Panel peer review of grant applications: What do we know from research in social psychology on judgment and decision-making in groups? *Research Evaluation*, 19(4): 293-304.

Porter, A. L. & F. A. Rossini (1985). Peer review of interdisciplinary research proposals. *Science, Technology & Human Values*, 10(3): 33-38.

Sandstrom, U. & M. Hallsten (2008). Persistent nepotism in peer-review. *Scientometrics*, 74(2): 175-189.

Simonton, D. K. (2006). Scientific status of disciplines, individuals, and ideas: Empirical analyses of the potential impact of theory. *Review of General Psychology,* 10(2): 98-112.

Simonton, D. K. (2008). Scientific talent, training, and performance: Intellect, personality, and genetic endowment. *Review of General Psychology,* 12(1): 28-46.

Smith, S. R. (2001). The social construction of talent: A defence of justice as reciprocity. *Journal of Political Philosophy,* 9(1): 19-37.

Van Arensbergen, P. & P. Van den Besselaar (2012). The selection of scientific talent in the allocation of research grants. *Higher Education Policy,* 25: 381 – 405.

Van Arensbergen, P., I. van der Weijden & P. van den Besselaar (2012), Gender differences in scientific productivity, a persisting phenonemon? *Scientometrics*, 93, 857-868

Van Arensbergen, P., I. Van der Weijden & P. Van den Besselaar (*forthcoming*). The selection of talent as a group process.

Van Arensbergen, P., I. Van der Weijden & P. Van den Besselaar (forthcoming). The notion of talent: What are the talents they are looking for in academia?

Van Balen, B. (2010). *Op het juiste moment op de juiste plaats. Waarom wetenschappelijk talent een wetenschappelijke carrière volgt*. Den Haag, Rathenau Instituut.

Van den Besselaar, P. (forthcoming). More competition, better science? Predictive validity of competitive grant selection.

Van den Besselaar, P. & L. Leydesdorff (2007). *Past performance as predictor of successful grant applications.* Den Haag, Rathenau Instituut

Van den Besselaar, P. & L. Leydesdorff (2009). Past performance, peer review, and project selection: A case study in the social and behavioral sciences. *Research Evaluation,* 18(4): 273-288.

Van den Brink, M. (2009). *Behind the scenes of science: Gender in the recruitment and selection of professors in the Netherlands*. Nijmegen, Radboud Universiteit.

Viner, N., P. Powell & R. Green (2004). Institutionalized biases in the award of research grants: A preliminary analysis revisiting the principle of accumulative advantage. *Research Policy,* 33: 443-454.

Wennerås, C. & A. Wold (1997). Nepotism and sexism in peer-review. *Nature*, 387: 341-343.

Zinovyeva, N. & M. Bagues (2011). *Does gender matter for academic promotion? Evidence from a randomized natural experiment?* IZA Discussion Paper No. 5537.