Research Evaluation, 25(4), 2016, 416–426
doi: 10.1093/reseval/rvw002
Advance Access Publication Date: 1 April 2016
Article

OXFORD

Funding decision-making systems: An empirical comparison of continuous and dichotomous approaches based on psychometric theory

Rüdiger Mutz[1,*], Lutz Bornmann[2], Hans-Dieter Daniel[1,3]

[1]Professorship for Social Psychology and Research on Higher Education, ETH Zurich, Muehlegasse 21, 8001 Zurich, Switzerland, [2]Division for Science and Innovation Studies, Administrative Headquarters of the Max Planck Society, Hofgartenstr. 8, 80539 Munich, Germany and [3]Evaluation Office, University of Zurich, Muehlegasse 21, 8001 Zurich, Switzerland

*Corresponding author. Email: mutz@gess.ethz.ch

Abstract

Psychometrics questions the use of dichotomous decisions. For these reasons, De Los Reyes and Wang (2012) favour a continuous funding decision system, in which the funded percentage of a requested grant sum is coupled to the ratings that a proposal receives in the ex ante peer evaluation. In contrast to the 'winner takes all' philosophy in a dichotomous funding decision system, a continuous system takes the low reliability of peer review ratings into account. Funding decisions are mostly based on peer review rating systems that have rather low inter-rater reliability. This article aims to use psychometrics to simulate the two funding decision systems, to compare them to the funding decision system implemented by a real funding organization, and with this, to investigate for the first time the effects of measurement errors on funding decisions. We used peer review data from the Austrian Science Fund (FWF) (N = 8,496 proposals), which obviously implements a hybrid funding decision system. The approval rate at FWF is 44.5%; our findings show that the approval rate would be 32.1% using a purely dichotomous system and 58.4% using a continuous funding decision system. As the funded percentage of a proposal's requested grant sum increases with increasing mean ex ante peer evaluation of a proposal ($r = 0.23$), the FWF also shows elements of a continuous funding decision system. Relative to a continuous system, a dichotomous system reduces the approval probability of a proposal overall. This is even the case for high-quality proposals (approval probability ~0.70).

Key words: grant peer review; funding decision; psychometry; simulation

1. Introduction

Considering research funding at universities today, where university personnel resources and equipment are hardly adequate for larger research projects, external project funding is essential for research in many countries, even if Tatsioni, Vavva and Ionanidis (2010) came to a different conclusion in their analysis of funding resources related to the papers awarded the Nobel prize: 'A substantial portion of this exceptional work was unfunded' (Tatsioni, Vavva and Ionanidis 2010: p. 1335). However, funding decisions are mostly based on a peer review rating system that has rather low inter-rater reliability, as a large number of papers on journal peer review and peer review of grant proposals have shown (e.g., Cicchetti 1991; Marsh, Jayasinghe and Bond 2008; Marsh 2008; Bornmann, Mutz and Daniel 2010; Bornmann 2011; Mutz, Bornmann and Daniel 2012a).

In this article we examine empirically what an alternative approval decision procedure could look like that explicitly takes the unreliability in the peer review system into account. With this, we also want to take up the recently voiced demand that the effectiveness of different funding methods should be tested (Ioannidis 2011, 2012; Nicholson and Ioannidis 2012). Following the 'winner takes all' philosophy, current funding systems largely use the dichotomous approach to funding (i.e., small subgroup of total applications

receives funding, remainder receives no funding). An alternative to 'winner takes all' funding systems was sketched theoretically by De Los Reyes and Wang (2012). De Los Reyes and Wang (2012) utilized concepts from psychometric research and theory, which are used for the construction of psychological tests. They assumed that ex ante evaluations of proposals are based mostly on 'reviewers' subjective impressions of the scholarly merit of funding proposals' (De Los Reyes and Wang 2012: 298). Peers are asked to rate proposals subjectively using Likert-type rating scales, which—like the scales of a psychological test—can be examined psychometrically. This type of quantitative peer review system is used, for example, by the National Institutes of Health in the USA, the Australian Research Council, and the Austrian Science Fund (FWF). For each proposal, the funding organization has the individual referees' ratings on the uniform scale. De Los Reyes and Wang have proposed a continuously distributed funding system, in which the funded percentage of a requested grant sum is coupled to the ratings that a proposal receives in the ex ante peer evaluation. For the purpose of simplification, we use the term 'continuous funding decision' and the term 'dichotomous funding decision' instead of 'continuously distributed funding system' and 'dichotomous funding decision system' in the remainder.

De Los Reyes and Wang (2012) presented a purely theoretical analysis; their concept has not as yet been put into practice at a research funding organization, and it has not yet been modelled using data from a research funding organization. As we have access to an extensive set of data from a research funding agency, we undertook modelling the concept in an empirical study. In this study we carried out a psychometrically oriented reanalysis of data on the funding decision procedure used at the FWF. The FWF is Austria's central funding organization for basic research (Fischer and Reckling 2010; Sturn and Novak 2012). A first mention of De Los Reyes and Wang's (2012) psychometric concept in peer review research is found in Marsh, Jayasinghe and Bond (2008), but only here in this paper the concept is actually fleshed out.

The study focusses on the following questions:

1. *Impact*: What is the impact of the different funding systems, pure continuous (scenario I) and pure dichotomous funding decisions (scenario II) on the number of proposals receiving support? How are the decisions that result from these funding systems to be assessed, compared to actual practice at the FWF?
2. *Fairness*: Are women and men, younger and older researchers, and the different scientific disciplines treated fairly by the different funding systems?
3. *Reliability*: To what extent in the different funding systems do measurement errors (lack of reliability of the peer review system) have an effect on approval decisions?

The FWF has implemented a hybrid funding decision system (including dichotomous and continuous elements). Even though the referees rate the proposals on a rating scale, the FWF leaves the final approval decision (approval or rejection) to the FWF Board of Trustees. Beyond that, the amount of support granted and the originally proposed budget differ in over half of the proposals approved for funding by more than 10% (as is foreseen by the continuously distributed funding decision system). With the aid of the FWF data, it was now possible to simulate two extreme situations—strictly dichotomous funding decisions and continuous funding decisions. The hybrid funding decision procedure at FWF can be positioned between the two scenarios.

The article is organized as follows: The approach of De Los Reyes and Wang (2012) is described in the next section, followed by both the method part, which illustrates the data and the psychometric approach, and the result part, structured according to the three research questions of this study. The article ends with a summary and discussion of the results.

2. Continuously distributed funding system approach of De Los Reyes and Wang (2012)

De Los Reyes and Wang (2012) asked whether 'the process of funding research proposals' should 'be a dichotomous (i.e. "fund/not fund") or continuous (i.e. degrees of support beyond fund/not fund) decision-making process' (De Los Reyes and Wang 2012: 299). To answer this question, the researchers assumed that there is a theoretical construct like 'research fundability' that can be measured quantitatively relying on referees' ratings. Dichotomous decision making assumes a dichotomous distribution of research fundability 'in which the underlying theoretical question is, "Does the scholarly merit inherent in a research proposal support its fundability or not?"' (p. 299). In contrast, continuous decision making assumes, statistically, a continuous distribution of fundability based on the following question: 'Does a research proposal that achieves some minimum-degree threshold of scholarly merit inherently warrant some degree of fundability, commensurate to that merit?' (p. 299). De Los Reyes and Wang (2012) favour the latter alternative, as it has—from a psychometrical perspective—increased validity and could improve the quality of funding decisions. Fortin and Currie (2013) reached a similar conclusion in their study on the Natural Sciences and Engineering Research Council of Canada. They asked whether it is 'more effective to give large grants to a few elite researchers, or small grants to many researchers?' (Fortin and Currie 2013: 1). Their empirical findings showed that 'funding strategies that target diversity, rather than "excellence," are likely to prove to be more productive' (1).

De Los Reyes and Wang (2012) favour a 'continuously distributed funding system' for three main reasons: First, when funding decisions are made dichotomously, a quantitative variable is dichotomized. From the perspective of measurement theory, there is a loss of information connected with the transition from an interval scale to a dichotomous scale. Beyond that, dichotomization leads to a substantial reduction in reliability; here, De Los Reyes and Wang refer to a study by Markon, Chmielewski and Miller (2011). But especially pertinent here is the methodological literature on the subject of the 'dichotomizing of quantitative variables' (e.g., MacCallum et al. 2002; Magnano, Tannonia and Andrà 2015). According to McCallum et al. (2002), for example, the reliability for the reliability index $\rho_{xx(3)}$ of 0.80 prior to dichotomization is reduced to 0.60 after dichotomization. A reliability of 0.60 decreases to 0.40 after dichotomization. Secondly, De Los Reyes und Wang (2012) found that 'dichotomous decision-making results in wasted human capital' and further referred to Ioannidis (2011). As a result of the 'winner takes all' philosophy of dichotomous decision making, researchers are forced to devote a considerable amount of work time to writing research proposals that altogether represent more funds and projects than their research institutes could actually carry out. This is the practice in view of the fact that there is a low probability that any

one proposal will receive support. Thirdly, dichotomous decision making 'creates uncertainty as to whether funded projects might have addressed the same or highly similar research aims with less funding' (De Los Reyes and Wang 2012: 301). Researchers could be tempted to raise the budget requested in a proposal by the amount of the anticipated percentage reduction of the grant sum.

With the percentile-based method used in this study, the total grant funds of a research funding organization are distributed to considerably more research funding proposals than is the case with dichotomous funding decision making. Taking a fictitious example, De Los Reyes and Wang (2012) showed that in a dichotomous decision making 10% of the best-rated proposals receive the full proposed budget, but in a continuous decision system, the percentages increase to 20–26% of the best-rated proposals.

De Los Reyes and Wang (2012: 302) named the following disadvantages of a continuously distributed funding system: (1) After receiving considerably less funding than their originally proposed budgets, which is to be expected more frequently with continuous funding than with dichotomous funding, researchers would have to revise their projects; (2) The continuously distributed funding system would continue to contain a dichotomous decision component. It is necessary to identify a minimal threshold for approving a proposal for funding. As this is often a difficult undertaking, De Los Reyes and Wang 'encourage future research to identify optimal thresholds for reliably and validly distinguishing between clearly meritorious and not meritorious funding proposals' (2012: 303).

3. Data and methods

3.1 Data and variables

In this study we reanalysed data that were collected in the context of the usual review procedures at FWF (Mutz, Bornmann and Daniel 2012a, 2012b, 2013, 2014a, 2014b). The data pool contained all proposals for research projects called 'Stand-Alone Projects' that were submitted from 1999 to 2009 and make up 60% of all FWF grants (in addition to stand-alone projects, the FWF also categorizes proposals as 'Special Research Program', 'Awards and Prizes', and 'Transnational Funding Activities'). A total of N = 8,496 proposals, which were rated by external referees (on average by two to three referees for each proposal), with complete data for the variables used in this study were available. At the FWF it is the grant applicant's task to assign the proposal to the different scientific disciplines (maximal 4). They can assign the proposal to more than one discipline using percentages (e.g. 30% clinical medicine, 70% preclinical medicine). The discipline with the highest percentage is taken as the proposal's main discipline (Mutz, Bornmann and Daniel 2014b: 32).

Of the FWF grant applicants, 17.7% were women and 82.3% men. Their median age was 45 years (M = 46.7, SD = 9.77, MIN = 23, MAX = 87). For this study we divided the applicants into quartiles according to age (younger than age 39, age 39–44, age 45–54, and older than age 54). Overall, the FWF granted support to 44.5%, or 3,788, of the research proposals submitted.

The following five central variables were included in the empirical analysis:

- *EXANTE*: In an ex ante peer evaluation, referees rated the merit of each proposal on a continuous scale from 0 to 100 (from poor to excellent), which is assumed to be interval-scaled for measurement purposes. For the proposals in the data set there were a total of 23,977 individual referees' ratings. For the analysis, however, we used the mean ratings of a proposal as a proposal's mean score across all referees (called here EXANTE). The individual ratings of a referee and their variability were used in the reliability analysis.
- *Requested grant sum*: For each proposal the originally proposed budget in € was available to us.
- *Approved grant sum*: In addition to the requested grant sum, the actually allocated funding support in € was available to us.
- *Funded percentage*: For each approved proposal, the funded percentage of a proposal's requested grant sum was known. This is the proportion of the originally proposed budget that the actually allocated funding support represents.
- *Funding decision*: For each proposal, the funding decision (approved or not-approved for funding) was available.

3.2 Funding decision systems

In addition to the funding decision system implemented by the FWF, two further, contrasting funding decision systems (continuous, dichotomous) were simulated with the data.

It is characteristic of a *continuous funding system* that not the full originally proposed budget is approved but instead only a certain percentage thereof. This amount depends on the proposal's EXANTE: The higher EXANTE is, the higher the percentage of the requested grant sum that is approved (the highest allocated amount of funding support would be 100% of the total budget proposed in the grant application). To create an approval schema for this study, percentiles of EXANTE are used. The percentiles indicate what percentage of the sample of proposals has a score (i.e., referees' mean value in EXANTE) that is smaller than the score of a certain proposal (Bornmann et al. 2011: 864f). To create a continuous approval schema, in addition to the calculation of the percentiles, also proportions of requested grant sum need to be assigned to the percentiles. De Los Reyes and Wang (2012) chose the minimum and maximum as follows: The highest allocated amount of funding support would be 100% of the total budget proposed in the grant application, and the lowest would be 20% (the 'floor').

The percentiles were assigned to the proportions of requested grant sum as follows: Proposals with EXANTE placing them in the 95th percentile and higher are supported with 100% of the requested grant sum. Proposals that are in the 90th percentile and higher were supported with 90%. We implemented this schema as far as the floor, with 5th percentile intervals and dropping the percentage of the requested grant sum by 10% for each percentile interval. On the floor, proposals that are in the 43rd percentile and higher were supported with only 20% of the proposed budget. Proposals in the 1st–43rd percentile were not approved for any support. EXANTE scores in the 43rd percentile represented the minimum degree of merit that a proposal must meet for funding approval. Absolute lowest thresholds for approved grant sums, which certainly make sense for real funding organizations, were not defined in the simulation in this study. As the total funding budget available to the FWF we used the total of requested grant sums in the 3,788 approved proposals: 639.77 million €. With this upper limit it is ensured that the simulated funding decision systems remain within the financial limits of the FWF. The number of approved proposals under the continuous funding decision system was determined using the information on the percentiles, the assigned proportions, and the total funding budget.

A *dichotomous funding decision system* is based on an 'all-or-nothing' or 'winner takes all' decision rule. If a proposal is approved for funding, either 100% of the originally proposed budget is allocated as the amount of funding support, or nothing at all. We simulated this using the data available from the FWF as follows: All proposals were sorted in descending order by EXANTE scores, so that the highest rated proposal ranked the highest (1st place) and the proposal with the lowest score ranked the lowest (*N*th place, if *N* proposals). The requested grant sums were then cumulated across the ranking places. For example, the cumulated requested grant sum of the proposal ranking in third place is the sum of the requested grant sums of the proposals ranking in the first three places. If in the ranking sequence the cumulated requested grant sum of a proposal *i* exceeds the total funding budget of the FWF, then the proposal that is one place higher (*i*-1) is the last proposal to receive funding support. The ranking place number *i*-1 indicates the number of supported proposals. With the discrete funding decision system, the sum of the requested grant sums of the best proposals does not exceed the total funding budget of the FWF.

3.3 Reliability analysis

In peer review, ratings can be expected to show systematic effects (e.g. judgment biases of the referees) and random fluctuations that were not taken into account above in the comments on simulation of the two funding decision systems. For example, some referees may judge a proposal too strictly, others more leniently (Siegelman 1991). Following Graves, Barnett, and Clark (2011), who without reference to a measurement theory simulated the effect of random fluctuations on funding decisions using a simple non-parametric bootstrap procedure, we took a measurement-theoretical approach here that should help determine the effect of measurement errors on funding decisions.

To model random fluctuations or random measurement error, classical test theory is particularly suitable for the following reasons. First, the approach proposed by De Los Reyes and Wang was defined within a psychometric framework. Secondly, psychometric concepts are well established in peer review research (e.g. Jayasinghe, Marsh and Bond 2003; Marsh and Bazeley 1999). Thirdly, the true-score concept represents central elements of the FWF peer review process. The FWF (and probably other funding organizations) uses the average of the peer referees' ratings as an empirical base for funding decisions. Classical test theory is a prominent concept in psychometrics, or the methodology of psychological testing. Test theory assumes that an observed score is made up of two elements: a true score, τ, and a random error term, ε (Gulliksen 1950; Novick 1966; Lord and Novick 2008):

$$x = \tau + \varepsilon \tag{1}$$

With regard to grant peer review, the following can be used as the observed score x: the individual rating, x_{ji}, of a referee *i* of a proposal *j*, or the mean peer review rating of a proposal *j* across all referees *i* of that proposal, x_j. As individual referees' ratings in funding decision procedures are not of interest for the research questions in this study, we used the overall referees' mean rating of a proposal (EXANTE) as the observed score, which can be divided into two elements as follows:

$$x_j = \tau_j + \varepsilon_j \tag{2}$$

The true score, τ_j, represents the measurement-error-corrected EXANTE of a proposal *j*, the so-called expected value $E(x_j)$, and should not be necessarily equated with the mean of x_{ji} of a proposal *j* across all referees *i* of a proposal. As stated above, the true score is only corrected for random fluctuations, represented by ε_j but not for any systematic biases. However, as the nonparametric bootstrap procedure used by Graves, Barnett, and Clark (2011) starts with the individual referees' ratings, it certainly overestimates the variability of ratings and thus the effect of random fluctuations on funding decisions.

The equation above (Equation 2) does not permit separate estimation of the two components τ_j and ε_j. Therefore, repeated measurements of the observed scores are necessary. One option is to have a second set of referees rating the same proposal again. Another possibility for repeated measurements is to divide the referees' rating of a proposal into two groups. The mean ratings of the second group serve as the repeated measurement of the mean ratings of the first group. In a third option, which we follow here, the inter-rater reliability, in particular the intra-class correlation, is estimated at the level of the individual referees' ratings in the sense that the single referee's rating of a proposal provide for the replicates (Jayasinghe, Marsh and Bond 2003: 287). To complete the measurement error theory, the following additional assumptions (Equations 5–6) are made, whereby *i* indicates the replicate (e.g. referee, different referee set, time point):

$$E(\varepsilon_{ji}|\tau_j) = E(\varepsilon_{ji}) = 0 \tag{3}$$

$$\sigma(\varepsilon_{ji},\ \tau_j) = 0 \tag{4}$$

$$\sigma(\varepsilon_{j'i},\ \varepsilon_{ji}) = 0 (j' \neq j) \tag{5}$$

$$\sigma(\varepsilon_{ji'},\ \varepsilon_{ji}) = 0 (i' \neq i) \tag{6}$$

$$\sigma^2(x_j) = \sigma^2(\tau_j) + \sigma^2(\varepsilon_j) \tag{7}$$

The conditional expected value given the true score, and the expected value in general of the error component, is zero (Equation 3). The true score and the error component are not correlated (Equation 4). Neither the error component between replicates of referees' mean ratings within a proposal (Equation 5), nor the error component between different proposals (Equation 6) are correlated. Consequently, the variance of the observed scores, x_j, can be decomposed into true-score variance, $\sigma^2(\tau_j)$, and error variance, $\sigma^2(\varepsilon_j)$ (Equation 7). The assumptions will be violated, if, for instance, the same referee reviews different proposals (Equation 6). Possible violations of assumptions of that kind were not further considered in this study. The model can also be represented as a multilevel or mixed-effects model, where the random effects represent the true scores (Kamata, Bauer and Miyazaki 2008: 346f; Jin and Mesbah 2014).

The reliability of the observed scores, x_j, is the ratio of true score variance and total variance: $r_{tt} = \sigma^2(\tau_j)/\sigma^2(x_j)$, which varies between 0 (no reliability) and 1 (perfect reliability). With respect to referees' ratings, the following intra-class correlation for the mean rating, ρ_M, provides for the reliability measure, r_{tt}, using the Spearman-Brown prediction formula (Bartko 1976: 763f; Jayasinghe, Marsh and Bond 2003: 287):

$$\rho_M = N\rho/(1 + (N - 1)\rho), \tag{8}$$

where ρ is the single-rater reliability or intra-class correlation of individual referees' ratings (Spearman-Brown calculation), and N is the (average) number of referees rating a proposal.

The standard error of measurement, SEM, which is an index of the variation of the random error, is $SEM = \sigma(x_j)\sqrt{(1-r_{tt})}$

(Harvill 1991: 34), where $\sigma^2(x_j)$ is the variance of referees' mean ratings across all proposals. Given the true score of a proposal, the observed referees' mean rating of that proposal varies within a 95% confidence interval of $\tau_j \pm 1.96$ SEM (Dudek 1979).

However, for the research questions in this study, the opposite case is of interest: Starting with the known observed score (the referees' mean rating of a proposal), a score band around an observed score is calculated to obtain the unknown true score. Based on Gulliksen (1950), Harvill (1991: 37) provided the following formula for the score band (especially, the 95% confidence intervals of the assumed true score):

$$\left[M + r_{tt}\left(x_j - M\right)\right] \pm 1.96\,\sigma\left(x_j\right)\left(\sqrt{}(1 - r_{tt})\right)\left(\sqrt{}r_{tt}\right), \qquad (9)$$

where the term in bracket, $[M + r_{tt}\ (x_j - M)]$, is the estimate of the *expected true score* of proposal *j*, and the term, $\sigma(x_j)(\sqrt{}(1\text{-}r_{tt}))\ (\sqrt{}r_{tt})$, is the SEM with respect to proposal *j*. M is the overall average of the referees' mean rating of a proposal. If reliability is perfect ($r_{tt} = 1.0$), the true score (term in bracket) will be equal to the observed score, x_j. If the reliability is far from perfect ($r_{tt} < 1.0$), the observed score of a proposal will be shrunk to the mean score M to obtain the expected true score. Furr and Bacharach (2008: 153) speak of a regression to the mean effect. The lower the reliability, the higher the shrinkage. In the extreme case of no reliability ($r_{tt} = 0$), the expected true score equals the overall mean score M. Therefore, the referees' mean rating of a proposal will equal its true score only if the reliability is perfect. Eventually, the true scores are nothing but Empirical Bayes Estimates, as they are estimated in multilevel modelling to obtain expected values for the groups means on the aggregate level (Mutz and Daniel 2007; Hox 2010: 30). In addition, the confidence interval of the true scores for a given observed score of a proposal increases as the reliability approximates 0.5 (maximum).

The parameters in Equation (9) could be estimated based on the FWF data, so that for each proposal the expected true score and the 95% confidence interval of the true score could be calculated. As the analysis showed, the overall average of the referees' mean ratings, M, amounted to 81 (100 point grading scale). The standard deviation of the referees' mean ratings, $\sigma(x_j)$, was 12.47. According to our study on the reliability of the FWF peer review system (Mutz, Bornmann and Daniel 2012a), the intra-class correlation for the individual ratings amounted to $\rho = .259$. Regarding the average number of reviews of 2.82 referees per proposal, the intra-class correlation ρ_M for the mean ratings (Equation (8)) could be calculated ($\rho_M = .495$) and also the reliability coefficient, r_{tt}. For example, if the mean rating of a proposal was 90, the expected true score, τ_j, of this proposal amounts to $[M + r_{tt}\ (x_j - M)] = (81 + 0.495 \times (90 - 81)) = 85.5$. The assumed true scores vary within a 95% confidence interval of $\pm 1.96\ \sigma(x_j)(\sqrt{}(1\text{-}r_{tt}))\ (\sqrt{}r_{tt}) = 1.96 \times 12.47 \times \sqrt{}((1 - 0.495) \times (0.495)) = \pm 12.22$ units around the expected true score [73.3, 97.7].

Based on the psychometric approach with true score and error component outlined above, a simulation could be carried out that explicitly took into account the reliability of the peer review system. The analysis was carried out at the level of the calculated true scores. There were three steps to the simulation:

1. For the referees' mean rating (EXANTE) of each proposal, the corresponding expected true score and SEM were calculated (see Equation (9)).
2. With the help of a random number generator and under the assumption of a normally distributed error component, 10 replicates of true scores for each proposal were obtained, which were distributed according to the expected true score and the SEM of a single proposal. Simulated true scores above 100 (below 0) were set to 100 (0).
3. The two funding decision systems (continuous, dichotomous) were applied to each set of replicates. For each proposal, the proportion of replicates that received funding was calculated. The simulation was done not only for different funding decision systems but also for the different number of referees that was reported in the data for each proposal (1 or single rater reliability, 2, 3, 4, 5, 6, and more referees). Additionally, a simulation was done for all proposals under the condition of perfect reliability ($r_{tt} = 1.0$) of EXANTE.

All analyses were done using the statistical software product SAS 9.4 (SAS Institute Inc. 2014).

Due to non-normally distributed data, in this study we chose the Spearman rank correlation as the correlation coefficient.

To test whether there were any differences between scientific disciplines in the overall results, two-dimensional χ^2-tests were used. With these analyses, the fairness of funding decisions with regard to disciplines can be analysed only under the aspect of potential biases (as only the EXANTE peer evaluation is used). Comprehensive analysis of fairness questions is possible only if, in addition, project success, or an ex post evaluation, is included in the analysis; an approach of that kind for the examination of peer review processes regarding the FWF as an example is found in Mutz, Bornmann and Daniel (2014a).

4. Results

4.1 Funded percentage of grant proposals for different funding systems

Fig. 1 shows the results of the simulation of the two funding systems based on EXANTE, or referees' mean ratings. In the simulation we examined the funded percentage of a proposal's requested grant sum in relation to the percentile rank of a proposal in the EXANTE. The higher the percentile rank (shown at left) was, the higher the EXANTE was (meaning, the higher the referees' mean rating of the merit of the proposal was).

In the actual (real-life) funding system used by the FWF, 44.5% of proposals (N = 3,788) were approved for funding, whereas in the dichotomous funding system (red bars in Fig. 1) only 32.1% of all proposals (N = 2,726) were approved for funding. This latter percentage corresponds to a threshold value for EXANTE of 88.3 (percentile rank of 32.1%). With the introduction of a purely dichotomous funding system, the approval rate would decrease by 12.4%—that is, approximately 28% fewer proposals would be approved for funding than in the existing funding system at the FWF.

When we simulate a continuous funding system using the FWF data, 58.4% of the proposals (N = 4,960) were approved for funding (blue bars in Fig. 1). The approval rate thus increased by 13.9% compared to the approval rate with the existing funding system at the FWF, which means 30.9% more proposals approved for funding than in the FWF's hybrid funding system. The approval rate corresponded to a threshold for EXANTE of 81.6. This value was somewhat higher than the mean of EXANTE. Due to a slightly skewed distribution, the mean value (M = 81) did not equal the median of EXANTE (median = 84).

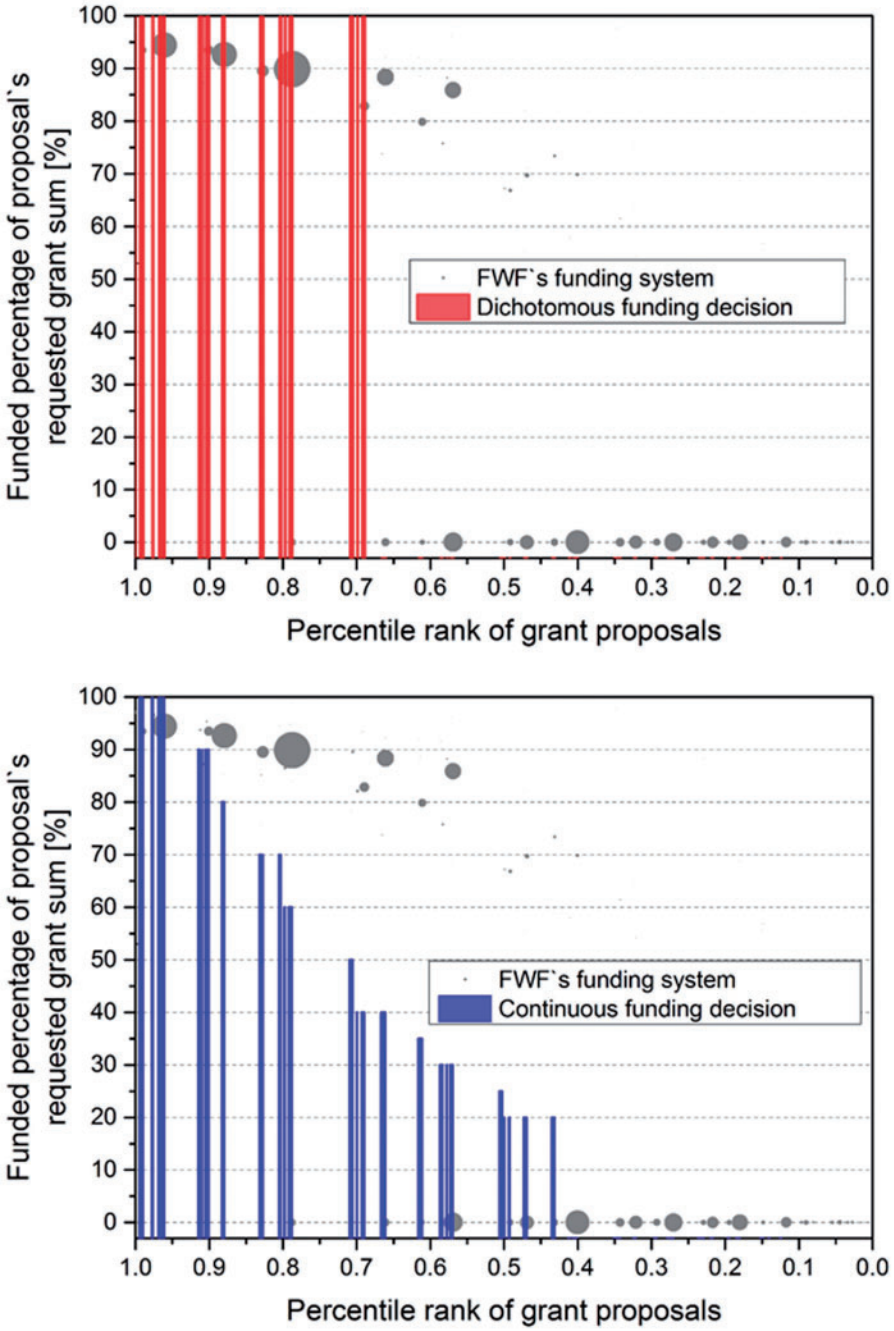

Figure 1. Funded percentage of proposal's requested grant sum in dependence on the percentile ranks (EXANTE) for different funding systems (red = dichotomous, blue = continuous). For the hybrid FWF funding system (see the grey bubbles), the median of funded percentages was calculated for each percentile. The size of a bubble represents the number of proposals with the same percentile and median value.

Regarding the hybrid FWF funding system, Fig. 1 made three things clear (see grey bubbles): First, at around the 55th percentile rank, there was a sudden change in the funded percentage, the proportion of the originally proposed budget that the actually allocated funding support represents, from about 80% of the originally proposed budget to 0% (no funding at all). This sudden change reflected the dichotomous element of the hybrid funding system used by the FWF. Secondly, however, with decreasing percentile rank, also the funded percentage of proposal's requested grant sum decreased slightly (see grey bubbles). This decrease reflected the continuous element of the funding system: With decreasing percentile rank, a smaller proportion of the requested grant sum was approved. Thirdly, the data showed clearly that the FWF's funding decisions are not based only on the EXANTE (referees' ratings) but that the FWF Board of Trustees (bearing in mind the EXANTE) made the decisions. For example, some proposals were not approved for funding even though they had a comparatively high percentile rank of higher than 60 (see grey bubbles with

Table 1. Descriptive statistics, shown separately for approved and not-approved proposals, and the correlations for the approved proposals (Spearman rank correlation)

Variable	Proposal approved?	Descriptive statistics					Spearman correlation			
		N	M	SD	Min	Max	A	B	C	D
A. EXANTE	No	4,708	73.91	12.09	0	97.50				
	Yes	3,788	89.85	5.14	0	100.00	1.0			
	Total	8,496	81.02	12.47	0	100.00				
B. Requested grant sum (k€)	No	4,708	230.48	109.24	4.36	806.60				
	Yes	3,788	226.43	106.31	4.80	812.46	.06	1.0		
	Total	8,496	228.67	107.95	4.36	812.46				
C. Funded percentage (%)	Yes	3,788	77.64	27.78	0	158.22	.23	−.26	1.0	
D. Approved grant sum (k€)	Yes	3,788	182.84	90.38	3.27	625.26	.21	.88	.16	1.0

Table 2. Testing for differences in approval rates for selected covariates, shown separately for the three different funding decision systems (N = 8,496 proposals)

Covariate	FWF			Dichotomous			Continuous		
	χ^2	*df*	Cramer's V	χ^2	*df*	Cramer's V	χ^2	*df*	Cramer's V
Applicant's gender	4.34*	1	−0.02	2.95	1	−0.02	5.37*	1	−0.03
Applicant's age	13.22*	3	0.04	37.50*	3	0.07	20.39*	3	0.05
Discipline	314.9*	21	0.19	522.7*	21	0.25	491.3*	21	0.24

*$P < 0.05$.

0% funded percentage). There were also proposals with percentile ranks lower than 70 that were approved for funding (see grey bubbles with a funded percentage greater than 0%). In addition to the FWF Board of Trustees, differing approval rates in the scientific disciplines could be responsible for the deviations from the EXANTE as well.

Accurate information on the correlations between EXANTE, requested grant sum, percentage of funding, and approved grant sum gave the descriptive statistics for the four variables, separately for approved and not-approved proposals (Table 1). Whereas the requested grant sum of funded proposals did not differ much from the not-approved projects, there were very definite differences between funded and rejected proposals in the EXANTE (73.91 versus 89.85). Some of the values of the funded percentage were conspicuous; they ranged from a minimum of 0 to a maximum of 158.22. These are certainly isolated cases, but we did not eliminate them, as they are real values from a funding agency.

In the next step, we analysed the correlations (Spearman rank correlation) between EXANTE, requested grant sum, funding percentage, and approved grant sum (Table 1). With a continuous funding system, the EXANTE should correlate with the funded percentage of the proposal's requested grant sum. The higher the referees' mean rating of a proposal, the higher the funded percentage had to be. This association was actually found in the data, but the correlation was low ($r = 0.23$). This relationship is higher ($r = 0.34$), if the correlation is statistically controlled by requested grant sum or if the requested grant sum is the same for all proposals. With a continuous funding system, besides the funding percentage no other of the variables mentioned above should correlate with the funded percentage. But this could not be confirmed by the data. There was a slightly negative correlation ($r = -0.26$) between requested grant sum and funded percentage: The higher the requested grant sum, the less it was cut back in the allocated support. Apparently, the requested grant sums were cut back in the allocated support not only depending on EXANTE but also depending on the amount of funding originally requested in the proposal.

All in all, the results show that the FWF had a *hybrid funding decision system*, in which the element of a dichotomous funding system (decisions by a Board of Trustees) dominated, but that the FWF also had elements of a continuous funding system.

4.2 Testing for fairness of the different funding systems

If ratings and decisions are very different for certain groups (e.g. male and female grant applicants), this is a first indication of violations of fairness in the system. For this reason we checked for such indications in the three funding decision systems (FWF, dichotomous, continuous). As potential bias variables we included applicant's gender, applicant's age, and scientific discipline in the analysis, which could lead to violations of fairness (Table 2). For the cross-tabulated relationships between approval decision (approved or not approved for funding) and the three categorical covariates, two-dimensional χ^2 tests were calculated. As effect size measure we used Cramer's V, which varies from 0 (no relationship or difference) to 1 (perfect relationship).

The results were statistically significant for applicant's gender and also applicant's age, but given the huge sample size and a very small Cramer's V, these differences are negligible. But for scientific discipline larger differences were found for all funding decision systems, and they were somewhat higher for the dichotomous and continuous funding decision systems than for the hybrid funding system implemented by the FWF (Cramer's V, χ^2).

More information about possible violations of fairness in differing handling of the grant applications in the disciplines was revealed

Table 3. Approval rates and correlations of EXANTE/requested grant sum with funded percentage, shown separately for scientific disciplines and funding decision systems

Discipline	N_{prop}	FWF		Dichotomous		Continuous		Corr	
		$Rate_{Ap}$	z_{res}	$Rate_{Ap}$	z_{res}	$Rate_{Ap}$	z_{res}	r_{exant}	r_{grant}
Mathematics	317	0.66	5.78*	0.53	6.47*	0.80	5.07*	0.18	−0.23
Computer sciences	324	0.38	−1.79	0.21	−3.43*	0.51	−1.76	0.17	−0.30
Physics, mechanics, astronomy	745	0.55	4.22*	0.40	3.94*	0.71	4.37*	0.21	−0.29
Chemistry	574	0.45	0.19	0.30	−0.90	0.63	1.52	0.24	−0.14
Biology	1,079	0.48	1.68	0.32	−0.17	0.62	1.72	0.19	−0.18
Botany	204	0.45	0.00	0.32	−0.06	0.56	−0.38	0.37	−0.10
Zoology	256	0.45	−0.01	0.36	0.98	0.58	−0.04	0.31	−0.20
Geosciences	451	0.46	0.56	0.29	−1.06	0.63	1.21	0.02	−0.27
Other natural sciences	87	0.36	−1.25	0.26	−0.93	0.53	−0.67	0.32	−0.28
Technical sciences	449	0.39	−1.92	0.24	−3.17*	0.50	−2.35*	0.13	−0.37
Preclinical medicine	1,245	0.37	−3.87*	0.20	−7.28*	0.48	−4.59*	0.34	−0.11
Clinical medicine	488	0.30	−4.92*	0.19	−5.16*	0.37	−6.21*	0.30	−0.34
Agriculture, forestry, veterinary medicine	121	0.36	−1.49	0.26	−1.26	0.52	−0.91	0.18	−0.25
Social sciences	373	0.27	−5.22*	0.18	−4.91*	0.37	−5.40*	0.15	−0.34
Psychology	99	0.31	−1.98*	0.25	−1.20	0.49	−1.16	−0.02	−0.14
Jurisprudence	89	0.39	−0.74	0.31	−0.10	0.51	−0.97	0.02	−0.14
Economics/business	157	0.35	−1.79	0.26	−1.32	0.45	−2.26	0.36	−0.15
Philosophy/theology	191	0.44	−0.13	0.35	0.73	0.57	−0.24	0.04	−0.30
Historical sciences	559	0.57	4.23*	0.53	8.71*	0.75	5.13*	0.18	−0.24
Linguistic/literature	345	0.54	2.51*	0.49	5.64*	0.68	2.30*	0.14	−0.36
Arts	199	0.55	2.26*	0.56	5.90*	0.78	3.60*	0.05	−0.29
Other area of humanities	144	0.58	2.47*	0.53	4.53*	0.67	1.41	0.01	−0.48
Total	8,496	0.45		0.32		0.58			

Note: $Rate_{Ap}$ = approval rate; z_{res} = standardized residuals of observed approval to the total approval rate; r_{exant} = correlation of EXANTE with funded percentage; r_{grant} = correlation of requested grant sum with funded percentage (Spearman rank correlation).

*$P < 0.05$.

by the approval rates broken down by discipline, separately for the three funding decision systems (Table 3). For a simplified interpretation of the approval rates, standardized residuals, z_{res}, between the approval rate per discipline and the overall approval rate for the specific funding decision system (total) were calculated. Values higher than 1.96 (97.5th percentile) indicated statistically significant deviations from the overall approval rate. Because of the standardization, these differences can be compared not only within a funding decision system but also between the systems.

The results showed that for all disciplines the approval rates with the continuous funding decision system were higher and the approval rates with the dichotomous system were mostly lower than the approval rates with the FWF funding system (there was one exception: arts). The approval rates with the FWF funding system ranged from 0.27 (social sciences) to 0.66 (mathematics), and the approval rates with the dichotomous system ranged from 0.18 (social sciences) to 0.56 (arts). In the simulation of the continuous funding decision system, the approval rate ranged from 0.37 (clinical medicine) to 0.80 (mathematics).

The disciplines profiting the most from a dichotomous system were mainly arts and humanities (e.g. historical sciences, linguistics/literature, arts) and mathematics, which showed distinctly higher approval rates than the average approval across all disciplines (which was 0.32). Although—in absolute terms—the rates were lower than the rates with the FWF funding system, the relative deviations from the average approval rate were higher (z_{res}). Disciplines with comparatively poor approval rates with this system were preclinical medicine, clinical medicine, social sciences, and technical sciences, whose approval rates were distinctly lower than the average approval rate of 0.32.

All in all, the results showed that in absolute terms in a continuous funding system more proposals are approved (over all disciplines), but in relative terms not all disciplines profit from this (not profiting are mainly preclinical medicine, clinical medicine, and social sciences).

In addition, we also examined correlations between the EXANTE or the requested grant sum and the funded percentage of proposal's requested grant sum (r_{exant}, r_{grant}), separately for disciplines. Systematic differences were found, but they were small. Whereas r_{exant} ranged from −0.02 (psychology) to 0.37 (botany), the correlations of the requested grant sum were all negative and ranged from −0.48 (other humanities) to −0.10 (botany). With a continuous funding decision system, we could expect to find markedly positive correlations between the funded percentage of proposal's requested grant sum and EXANTE as well as rather negative correlations between the funded percentage of proposal's requested grant sum and the requested grant sum. A continuous decision system of that kind was found the most for botany, zoology, preclinical medicine, and economics/business, with correlations r_{exant} over 0.30 and r_{grant} below −0.10.

Overall, the results showed that the various scientific disciplines profited very differently from the different funding decision systems. There were winners and losers. The calculation of the correlation coefficients showed that there were elements of a continuous funding decision system in the hybrid FWF system. These elements were

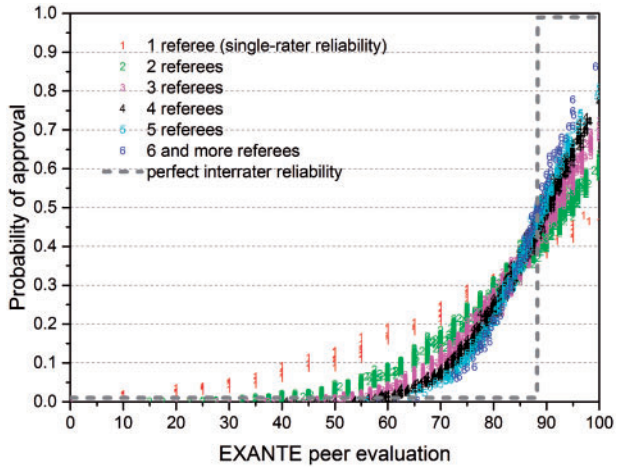

Figure 2. Dichotomous funding decision system.

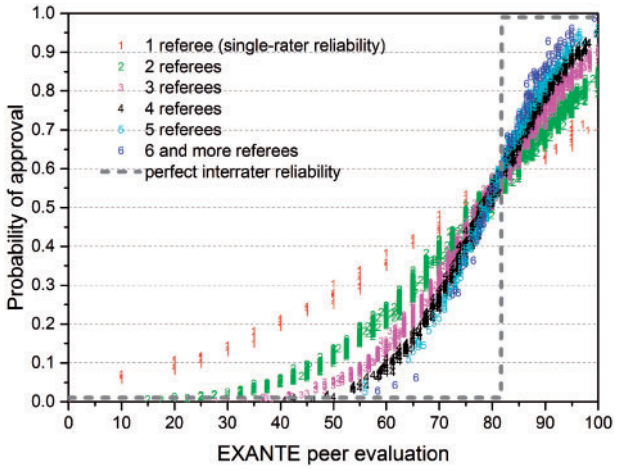

Figure 3. Continuous funding decision system.

more or less pronounced depending on the discipline. The results indicated that factors such as the amount of the requested grant sum play an important role in the FWF funding system.

4.3 Impact of measurement error on approval rates for different funding decision systems

The analyses reported so far were on the observed EXANTE—that is, the referees' mean rating of a proposal. However, many studies have pointed out the low reliability of peer review ratings. But mostly, the studies drew no conclusions from this concerning the statistical analysis of peer review data (e.g. correlation, regression analysis) (Mutz, Bornmann and Daniel 2014a). For this reason, in this study we simulated 10 measurement-error corrected true scores based on the referees' mean rating of a proposal (EXANTE)—the observed scores. We applied the two funding decision systems to each of the 10 sets of true-score replicates of a proposal to calculate the probability of approval (approved/not-approved) for each proposal (Figs 2 and 3). As we mentioned above, true score and observed score are the same only in the case of perfect reliability. In all other cases, the true score of a proposal varies around an expected true score within a band. As compared to the referees' mean rating, the expected true score shrinks towards the overall mean value.

At first glance, the curves for the two funding decision systems looked similar (Figs 2 and 3). Actually, we would expect to find a clear cut-off between the proposals approved and those not approved for funding, as the two funding decision systems plan for. In fact, however, we found s-shaped curves like those found for growth trajectories. With increasing EXANTE the probability of approval increased exponentially up to an inflection point. After the inflection point the curve continues to rise but with a continually decreasing climbing rate. With an increasing number of referees, the reliability of EXANTE and thus its discrimination power increased. The curves in the figures were therefore steeper. The separation of approved and not approved became more distinct. With perfect reliability there was a clear threshold in the EXANTE. If the EXANTE of a proposal surpassed this threshold, the proposal was approved for funding; if not, it was not approved for funding.

However, a closer look at the figures revealed considerable differences between the continuous and dichotomous funding decision systems. The level of the approval probability was clearly lower for the dichotomous system than for the continuous funding decision system. With a dichotomous system, proposals had a lower probability of being funded than with the continuous funding decision system. With a dichotomous system, even proposals with a very good EXANTE rating had a maximal probability of 0.70/0.80 to be approved for funding. If a proposal was rated by two referees, the approval probability for proposals receiving an EXANTE of 100 was only 0.60. The lacking reliability of peer review ratings had a considerable effect on approval decisions. The situation was found to be different with continuous funding decisions, however. Here, the approval probability for very good EXANTE ratings reached 0.80/0.90. If in this simulation a proposal was rated by two referees, the approval probability for proposals receiving a very good EXANTE of 100 grade points was 0.80. As in the simulation of the continuous system the approval probability increased overall as compared to the simulation of the dichotomous system, proposals that received relatively poor EXANTE still had a certain chance (0.20/0.30) of being approved for funding.

A continuous funding system increased the possible number of false positives in a peer review process. False positives were proposals that were supported by a funding agency but later showed no project success (meaning, for example, no papers on the project were published, or publications on the project had low citation rates). With a dichotomous system, there was also a chance that poorly rated proposals would be approved for funding, but it was considerably lower. With a continuous funding decision system, a proposal with an EXANTE of 70 (rated by two referees) had a chance of approximately 0.30 to be approved for funding; with a dichotomous funding decision system, the approval probability dropped to only 0.18.

5. Discussion

According to De Los Reyes and Wang's (2012) considerations based on psychometric research and theory, dichotomous funding decision systems, which are implemented by many funding agencies, lead to

unreliable funding decisions. As an alternative approach, they therefore recommend a continuous funding decision system, which they consider to have advantages in other aspects as well (higher approval rate, considering reliability issues,...).

Following De Los Reyes and Wang's (2012) considerations, in this study we simulated the two funding decision systems using real data from the FWF and taking measurement errors into account. We compared the results of the simulations of the two funding systems with the results of the real (hybrid) decision procedure implemented by the FWF. With this, for the first time in grant peer review research the effects of measurement errors on funding decisions were investigated based on a measurement theory (psychometrics).

This study did not aim primarily to prove De Los Reyes and Wang's (2012) hypotheses empirically. The evidence follows more or less analytically from the existing psychometric concepts. Instead, our goal was to work out the psychometric approach for peer review research, to estimate the concrete effects of these two systems on the number of approved proposals, to check the fairness of the systems, and to estimate how low reliability affects funding decisions.

The analyses yielded the following results, which are structured in accordance with our research questions:

- *Question 1: Impact of different funding decision systems*: As expected, considerable differences in the number of approved proposals result from the different funding decision systems: At the FWF, 44.5%, or 3,788, of the research proposals submitted were approved for funding. With a dichotomous funding decision system, the number of approved proposals decreases to 2,726 proposals (approval rate: 32.1%); the approval rate thus decreases by 12.4 percentage points as compared to the approval rate at the FWF. This means that 28% fewer proposals were approved than with the hybrid FWF system. With a continuous funding decision system, the approval rate increases by 13.9 percentage points to 58.4%; the number of approved proposals increases by 30.9% to a total of 4,960 proposals. The analyses show clearly that the FWF funding system is (as expected) a hybrid system. Elements of a dichotomous funding decision system predominate, whereby there is a corrective factor: The funding decisions are made by a Board of Trustees. The board's decisions have an effect mainly in the middle area of the EXANTE scale. But the FWF funding system also contains elements of a continuous funding decision system. The higher the EXANTE score is, the less the requested grant sum is cut in the actual support approved ($r = 0.23$). The coefficient is higher if the association is statistically controlled by requested grant sum ($r = 0.34$).
- *Question 2: Fairness of the funding decision systems*: We checked the fairness of the funding systems with regard to potential biases concerning applicant's gender, applicant's age, and scientific discipline. For age and gender there are no effects worth mentioning that could indicate potential biases. However, with regard to disciplines, there are clear differences: The disciplines that profit the most from a dichotomous system are arts and humanities (e.g. historical sciences, linguistics/literature, arts) and mathematics, which show clearly higher approval rates than the average approval rate of 0.32. Although in absolute terms more proposals are approved for funding in a continuous funding system, and this is the case for all disciplines. In relative terms the disciplines that profit relatively the least from this system are preclinical medicine, clinical medicine, and social sciences. All in all, a discipline-specific adjustment of the funding system would therefore be necessary.
- *Question 3: Impact of measurement errors*: A decisive argument against the implementation of a dichotomous funding decision system set out by De Los Reyes and Wang (2012) is that dichotomous funding decisions are less reliable. For this reason, we repeated all analyses based on measurement-corrected true scores. We found that through a dichotomous funding decision system the approval probability of a proposal decreases. Beyond that, due to the lack of reliability of the peer review system, even proposals estimated to have very good merit (EXANTE > 90) have an average approval probability of only 0.70. This is not really a problem with a continuous system, where the approval probability of proposals with a very good EXANTE after all reaches 0.80/0.90. The reduced approval probability in the two systems occurs due to the fact that with not-perfect reliability, for the calculation of the true score, the observed score (EXANTE) shrinks towards the mean.

  The overall higher approval probability of a proposal in a continuous funding system applies also to proposals that receive an only average to below EXANTE from the referees. These proposals, too, have a higher probability of approval. As a result, the proportion of false positives in a peer review procedure increases, and proposals are approved for funding that later turn out to be less than successful research projects.

Finally, we would like to point out some limitations of this study:

- The two funding decision systems investigated in this study were simulations based on more or less arbitrary determinations. However, the simulations were based on real data and are not purely Monte Carlo studies. The advantage of simulations is that it is not necessary to take into account decision makers, which at many funding agencies make the final funding decisions.
- We can only speak of violations of fairness as the result of an empirical study if in addition to the potential bias variables we also include performance measurements (e.g., the research output of a proposal) in the analysis. This analysis (potential bias versus real bias) has already been carried out for the FWF (Mutz, Bornmann and Daniel 2014a). Unfortunately, it was not possible to consider performance measurements in the analyses of the systems simulated here.

Given the limitations of this study, we would like to draw the following conclusions and offer some recommendations for the design of funding systems:

- In the face of the low reliability of the peer review system as it exists at many funding organizations, a purely dichotomous funding decision system is problematic. In a system of this kind, there is a risk (which should not be underestimated) that proposals rated to have very good merit will not be approved for funding. All in all, in this system, very good proposals have a lower chance of being approved.
- A continuous funding decision system has psychometric advantages, but with this system also poorly rated proposals have a chance (which should not be overlooked) of being approved for funding. This system implies an increased risk of supporting research projects that will show no success later on.

- Regarding the FWF, we would like to suggest that it would be easier for the FWF Board of Trustees if only those proposals within (mid-range) percentile ranks from 0.55 to 0.75 (EXANTE) were discussed and decided. For the other ranges of EXANTE, a modified continuous decision system could be implemented that is adjusted to the various main disciplines. The absolute lowest percentile rank threshold for approving funding could be 0.40 (EXANTE). For the percentile intervals from 0.40 to 0.55 and further from 0.75 to 1.00, a proportional decreasing funding percentage schema could be developed.

In conclusion, we would like to point out that it is not our objective in this study to explicitly vote for one or the other funding decision system. Our goal was to present empirical results on the possible systems that can serve as an empirical basis for discussion on alternative funding procedures. We hope that in the future especially aspects of the lack of reliability of peer review ratings will be taken into account in the examination of funding organizations.

References

Bartko, J. (1976), 'On Various Intraclass Correlation Reliability Coefficients', *Psychological Bulletin*, 83/5: 762–65.

Bornmann, L. (2011), 'Scientific Peer Review', *Annual Review of Information Science and Technology*, 45: 199–245.

Bornmann, L., Mutz, R. and Daniel, H.-D. (2010), 'A Reliability-generalization Study of Journal Peer Reviews: A Multilevel Meta-analysis of Inter-Rater Reliability and its Determinants', *PLos One*, 5/12: e14331.

Bornmann, L., Mutz, R., Marx, W., Schier, H. and Daniel, H.-D. (2011), 'A Multilevel Modelling Approach to Investigating the Predictive Validity of Editorial Decisions: Do the Editors of a High Profile Journal Select Manuscripts That Are Highly Cited After Publication?', *Journal of the Royal Statistical Society Series a-Statistics in Society*, 174: 857–879.

Cicchetti, D. V. (1991), 'The Reliability of Peer Review for Manuscript and Grant Submissions: A Cross-disciplinary Investigation', *Behavioral and Brain Sciences*, 14: 119–86.

De Los Reyes, A. and Wang, M. (2012), 'Applying Psychometric Theory and Research to Developing a Continuously Distributed Approach to Making Research Funding Decisions', *Review of General Psychology*, 16/3: 298–304.

Dudek, F. J. (1979), 'The Continuing Misinterpretation of the Standard Error of Measurement', *Psychological Bulletin*, 86/2: 335–7.

Fischer, C. and Reckling, F. (2010), *Factors Influencing Approval Probability in FWF Decision-Making Procedures*, Vienna, Austria: Austrian Science Fund.

Fortin, J. M. and Currie, D. J. (2013), 'Big Science vs. Little Science: How Scientific Impact Scales with Funding', *PLos One*, 8/6: e65263.

Furr, R. M. and Bacharach, V. R. (2008), *Psychometrics - An introduction*, London: Sage.

Graves, N., Barnett, A. G. and Clarke, P. (2011), 'Funding Grant Proposals for Scientific Research: Retrospective Analysis of Scores by Members of Grant Review Panel', *British Medical Journal*, 343: d4797.

Gulliksen, H. (1950), *Theory of Mental Tests*, New York: John Wiley.

Harvill, L. M. (1991), 'Standard Error of Measurement', *Educational Measurement: Issues and Practice*, 10/2: 33–41.

Hox, J. (2010), *Multilevel Analysis - Techniques and Applications*, New York: Routledge.

Ioannidis, J. P. A. (2011), 'Fund People not Projects', *Nature*, 477/7366: 529–31.

Ioannidis, J. P. A. (2012), 'Research Needs Grants, Funding and Money - Missing Something?', *European Journal of Clinical Investigation*, 42/4: 349–51.

Jayasinghe, U. W., Marsh, H. W. and Bond, N. (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.

Jin, Z. and Mesbah, M. (2014), 'Unidimensionality, Agreement and Concordance Probability', in V. Couallier, L. Gerville-Reache, C. Huber-Carol, N. Limnios, and M. Mesbah (eds.) *Statistical Models and Methods for Reliability and Survival Analysis*, pp. 3–19. Hoboken: Wiley.

Kamata, A., Bauer, D. J. and Miyazaki, Y. (2008), 'Multilevel Measurement Modeling', in A. A. O'Connell and D. B. McCoach (eds.) *Multilevel Modeling of Educational Data*, pp. 345–88. Charlotte, NC: Information Age Publishing.

Lord, F. M. and Novick, M. R. (2008), *Statistical Theoy of Mental Test Scores*, New York: Addison-Wesley.

MacCallum, R., Zhang, S., Preacher, K. J. and Rucker, D. D. (2002), 'On the Practice of Dichotomization of Quantitative Variables', *Psychological Methods*, 7/1: 19–40.

Magnano, G., Tannonia, C. and Andrà, C. (2015), 'A Priori Reliability of Tests with Cut Score', *Psychometrika*, 80/1: 44–64.

Markon, K. E., Chmielewski, M. and Miller, C. J. (2011), 'The Reliability and Validity of Discrete and Continuous Measures of Psychopathology: A Quantitative Review', *Psychological Bulletin*, 137/5: 856–79.

Marsh, H. W. (2008), 'Improving the Peer Review for Grant Applications', *American Psychologist*, 63/3: 160–68.

Marsh, H. W. and Bazeley, P. (1999), 'Multiple Evaluations of Grant Proposals by Independent Assessors: Confirmatory Factor Analysis Evaluations of Reliability, Validity, and Structure', *Multivariate Behavioral Research*, 34/1: 1–30.

Marsh, H. W., Jayasinghe, U. W. and Bond, N. W. (2008), 'Improving the Peer-Review Process for Grant Applications - Reliability, Validity, Bias, and Generalizability', *American Psychologist*, 63/3: 160–8.

Mutz, R. and Daniel, H.-D. (2007), 'Development of a Ranking Procedure by Mixed Rasch Model and Multilevel Analysis - Psychology as an Example', *Diagnostica*, 53/1: 3–16.

Mutz, R., Bornmann, L. and Daniel, H.-D. (2012a), 'Heterogeneity of Inter-Rater Reliabilities of Grant Peer Reviews and its Determinants: A General Estimating Equations Approach', *Plos One*, 7/10: e48509.

Mutz, R., Bornmann, L. and Daniel, H.-D. (2012b), 'Does Gender Matter in Grant Peer Review? An Empirical Investigation using the Example of the Austrian Science Fund', *Zeitschrift für Psychologie*, 220/2: 121–9.

Mutz, R., Bornmann, L. and Daniel, H.-D. (2013), 'Types of Research Output Profiles: A Multilevel Latent Class Analysis of the Austrian Science Fund's Final Project Report Data', *Research Evaluation*, 22/2: 118–33.

Mutz, R., Bornmann, L. and Daniel, H.-D. (2014a), 'Testing for the Fairness and Predictive Validity of Research Funding Decisions: A Multilevel Multiple Imputation for Missing Data Approach Using Ex-Ante and Ex-Post Peer Evaluation Data From The Austrian Science Fund', *Journal of the Association for Information Science and Technology*, 66/11, 2321–2339.

Mutz, R., Bornmann, L. and Daniel, H.-D. (2014b), 'Cross-Disciplinary Research: What Configurations of Fields of Science are Found in Grant Proposals Today?', *Research Evaluation*, 24/1: 30–6.

Nicholson, J. M. and Ioannidis, J. P. A. (2012), 'Conform and be Funded', *Nature*, 492/7427: 34–6.

Novick, M. R. (1966), 'The Axioms and Principal Results of Classical Test Theory', *Journal of Mathematical Psychology*, 3: 1–18.

SAS Institute Inc. (2014), *SAS/STAT®13.2 User's Guide*, Cary, NC: SAS Institute Inc.

Siegelman, S. S. (1991), 'Assassing and Zealots - Variations in Peer Review - Special Report', *Radiology*, 178/3: 637–42.

Sturn, D. and Novak, R. (2012), 'Was koennen Forschungsevaluationen durch den FWF zur Entwicklung von Hochschulen beitragen? [What Can FWF's Research Evaluations Contribute to the Development of Institutions of Higher Education?]", in Oesterreichische Qualitaetssicherungsagentur (ed.), *Braucht Forschung Qualitaetsmanagement?*, pp. 59–70. Vienna, Austria: AQA.

Tatsioni, A., Vavva, E. and Ioannidis, J. P. A. (2010), 'Sources of Funding for Nobel Prize-Winning Work: Public or Private?', *Faseb Journal*, 24/5, 1335–9.