Scientific analysis of data on proposals and the decision-making procedure of the FWF with particular focus on the programme "Stand-Alone Projects" in the years 2010-2019

Findings report

15 February 2020

Dr. Rüdiger Mutz,

Center for Higher Education and Science Studies,

University of Zurich

Prof. em. Dr. Hans-Dieter Daniel,

Professorship for Quantitative Research on Higher Education,

University of Zurich

TABLE OF CONTENTS

List of tables

List of figures

1. DATA

1.1 PRELIMINARY REMARKS

In the following, the structure of the data that is the basis of the scientific analysis of the application procedure will be described.

1. Because of missing values, the sample sizes in the statistical analyses may deviate from the sample sizes in the data structure.
2. Only the sessions of the Board of Trustees are listed in the data structure; proposals are occasionally also discussed in meetings of the Executive Board.

1.2 DATA STRUCTURE

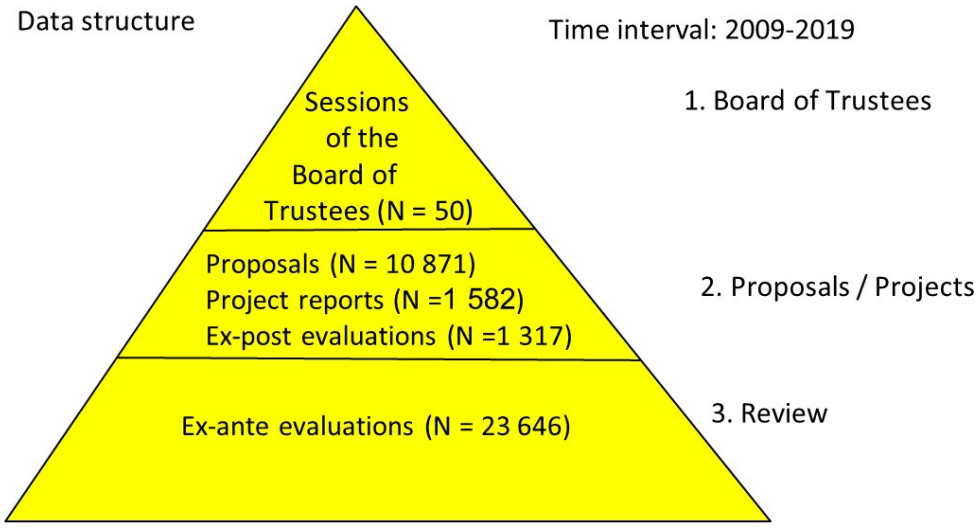

FIGURE 1: Data structure

2 HISTOGRAMS OF RATINGS OF REVIEWERS

2.1 PRELIMINARY REMARKS

- In the following, histograms of the distribution of reviewers' ratings are presented, in each case separately for the two grading scales (old grading scale till the year 2015, new grading scale from 2015 onwards).
- We distinguish between three types of scales: a) Raw scales: the old scale is a 100 point scale (1 = poor, 100 = excellent), the new scale is a 5-point ordinal scale (1 = excellent, 5 = poor). b) 100 point scale: To permit comparison between the scales, they are transformed to a 100 point scale (0 = excellent, 100 = poor). c) Scale of the National Institutes of Health: A scale in accordance with the National Institutes of Health (NIH, USA) is generated.
- Furthermore, a distinction is made between mean ratings of reviewers of a proposal and reviewers' single ratings per proposal.

2.2 METHODOLOGICAL BASIS OF THE NIH SCALE

The NIH scale is a percentile scale on which the individual proposals are ranked according to mean peer review ratings. The NIH scale is constructed as follows (https://www.niaid.nih.gov/grants-contracts/understand-paylines-percentiles):

"Step 1 – Following the discussion led by the primary reviewer, all reviewers rate the overall impact of an application, assigning a whole number from 1 to 9.

Step 2 – These scores are averaged, rounded mathematically to one decimal place, and multiplied by 10 to create the overall impact score, e.g., a 1.34 average yields a 13 overall impact score.

Step 3 – Percentiles are determined by matching an application's overall impact score against a table of relative rankings containing all scores of applications assigned to a study section during the three last review cycles.

Step 4 – NIH calculates percentiles using the following formula.

Percentile = 100 * (Rank - 0.5) / Total Number of Applications (The 0.5 percent is a standard mathematical procedure used for rounding.)"

2.3 POLICY-RELEVANT SUMMARY OF THE FINDINGS

The distribution of the mean ratings of reviewers both for the old and the new scale is skewed, i.e. not a normal distribution in the sense of a bell curve. On average, proposals are frequently rated "good", "very good" and "excellent". Overall, the variability of mean ratings of reviewers is higher for the new scale than for the old scale. This is expressed in the coefficient of variation (CV) (**Table 1**), which can be interpreted independently of the scale. The CV is 16.41 for the old scale (raw scale) and clearly higher, at 38.69, for the new scale. As regards reviewers' single ratings, the entire range of values of each scale is used by the reviewers. An alternative to the raw scale or the 100 point scale is the NIH percentile scale, which permits broad differentiation between proposals. Furthermore, the effect of the different scales (old, new) is reduced, as the NIH scale expresses percentiles or rankings of proposals with regard to mean ratings of reviewers.

2.4 FINDINGS

2.4.1 HISTOGRAMS OF MEAN RATINGS OF REVIEWERS OF A PROPOSAL – RAW SCALE

TABLE 1: Descriptive statistics of mean ratings of reviewers of a proposal

Grading scale	N	Mean	Median	SD	CV
Old scale					
Raw score	5,732	81.06	85.00	13.30	16.41
Percentage scale	5,732	19.92	16.00	13.28	66.67
NIH scale	5,732	52.65	53.00	27.90	53.00
New scale					
Raw score	4,749	2.11	2.00	0,816	38.64
Percentage scale	4,749	27.80	25.00	20.40	73.39
NIH scale	4,749	57.15	62.00	27.31	47.78

SD = standard deviation, CV = coefficient of variation

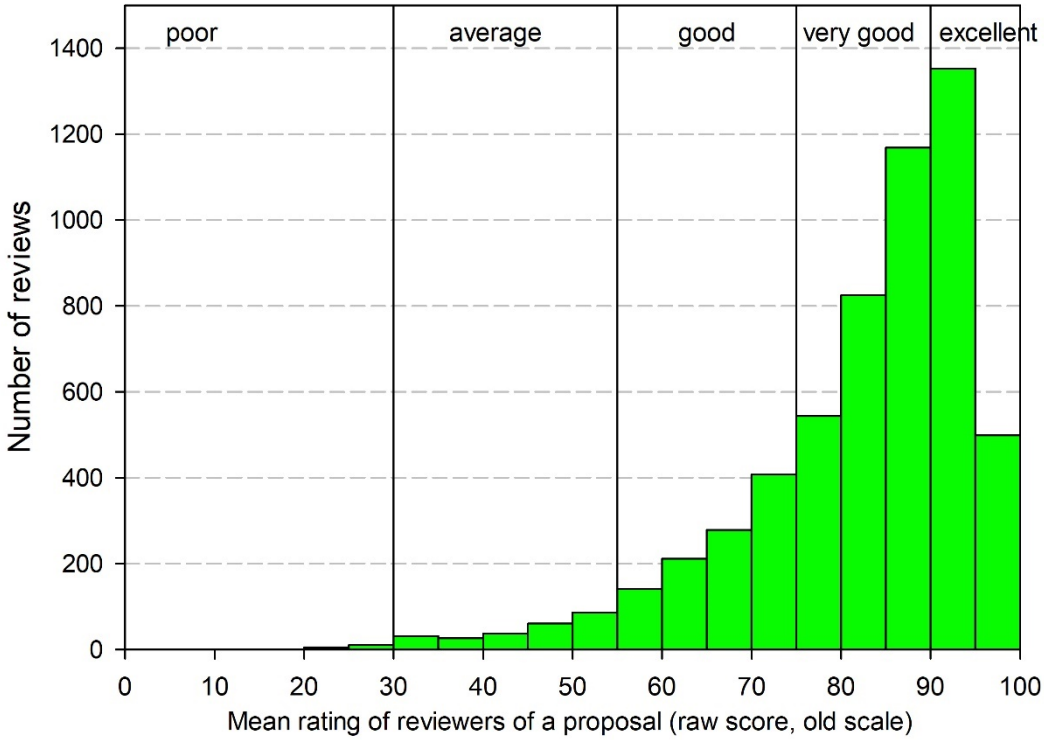

FIGURE 2: Histogram of mean raw ratings of reviewers of a proposal (old scale till the year 2015)

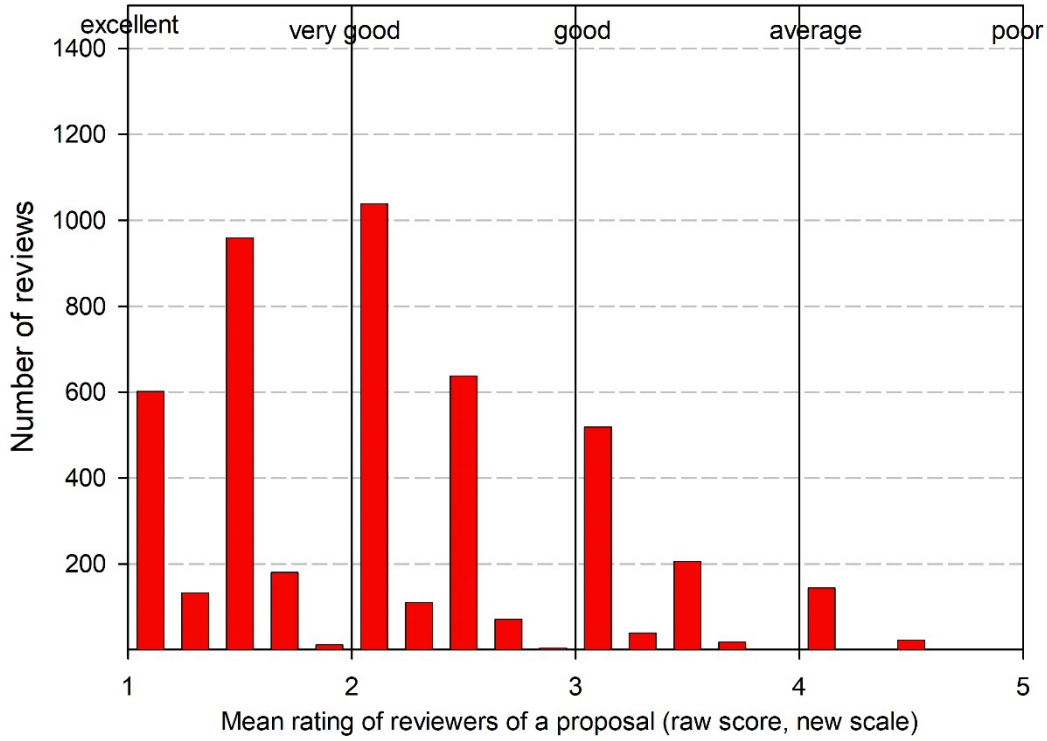

FIGURE 3: Histogram of mean raw ratings of reviewers of a proposal (new scale from 2015 onwards)

2.4.2 HISTOGRAMS OF MEAN RATINGS OF REVIEWERS – 100 POINT SCALE

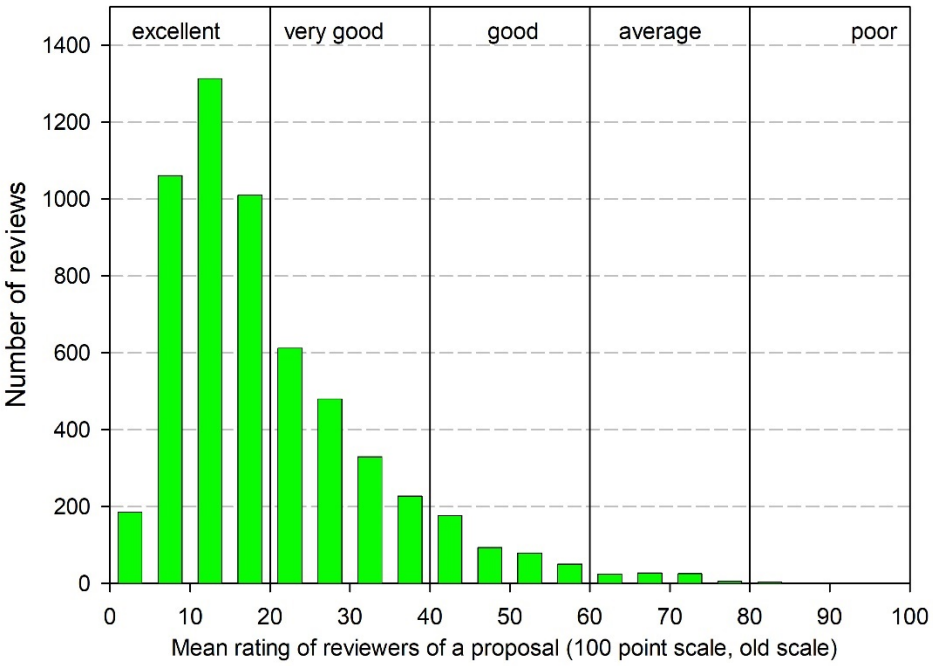

FIGURE 4: Histogram of mean ratings of reviewers of a proposal, transformed to a 100 point scale (old scale till the year 2015, scale labels were added in steps of 20 points)

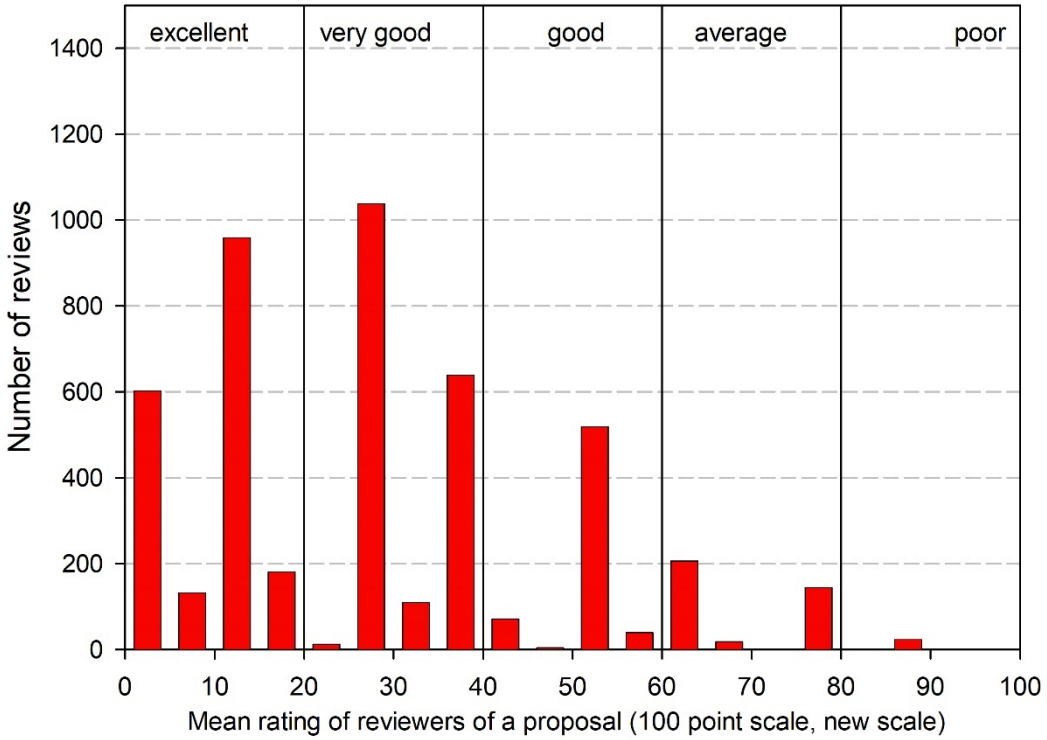

FIGURE 5: Histogram of mean ratings of reviewers of a proposal, transformed to a 100 point scale (new scale from 2015 onwards, scale labels were added in steps of 20 points)

2.4.3 HISTOGRAMS OF MEAN RATINGS OF REVIEWERS – NIH SCALE

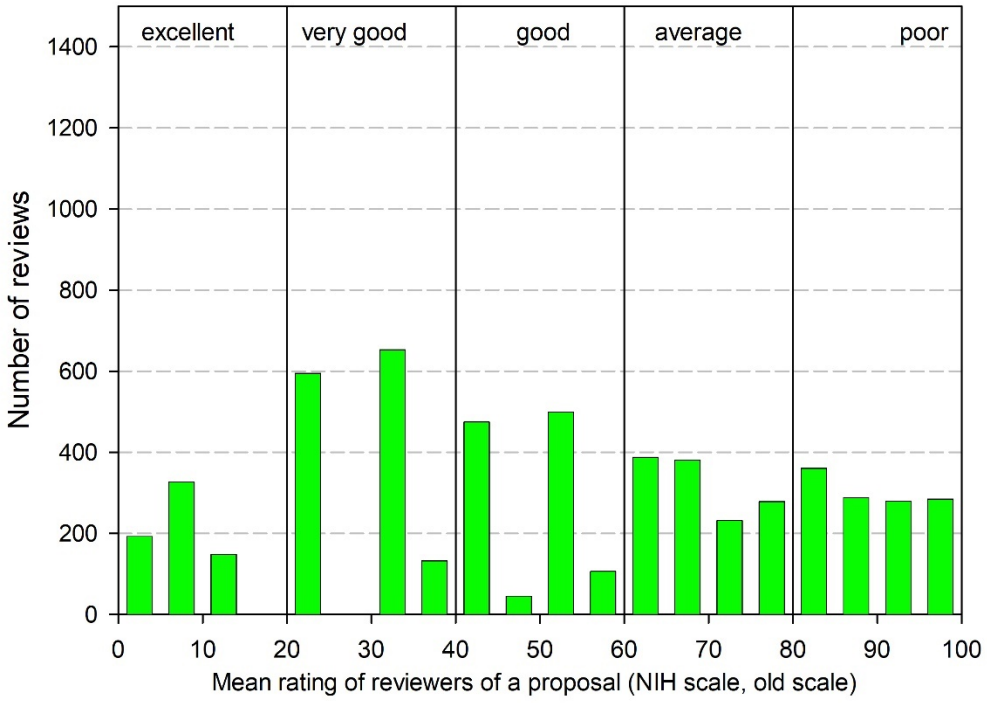

FIGURE 6: Histogram of mean ratings of reviewers of a proposal, transformed to the NIH scale (old scale till the year 2015, scale labels were added in steps of 20 points)

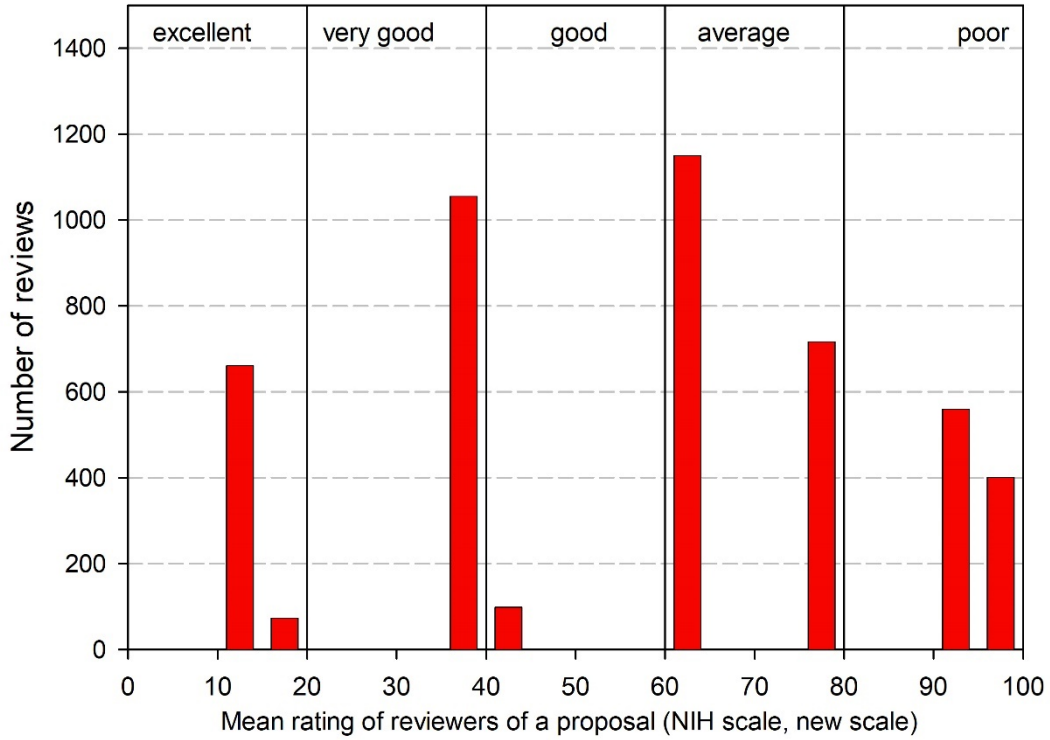

FIGURE 7: Histogram of mean ratings of reviewers of a proposal, transformed to the NIH scale (new scale from 2015 onwards, scale labels were added in steps of 20 points)

2.4.4 HISTOGRAMS OF REVIEWERS' SINGLE RATINGS – RAW SCALE

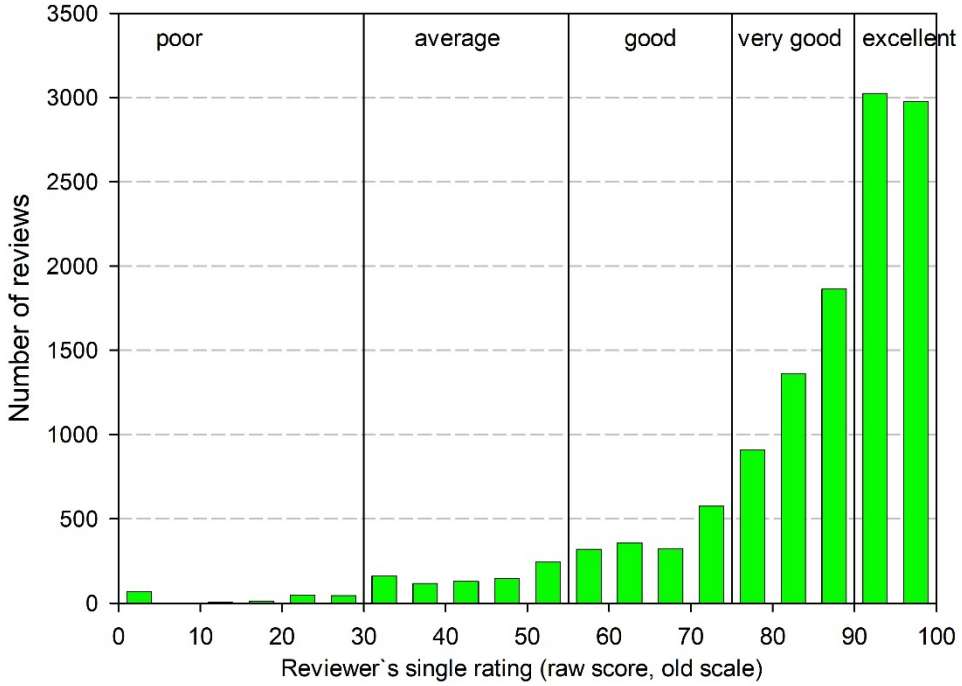

FIGURE 8: Histogram of reviewers' single raw ratings per proposal (old scale till the year 2015)

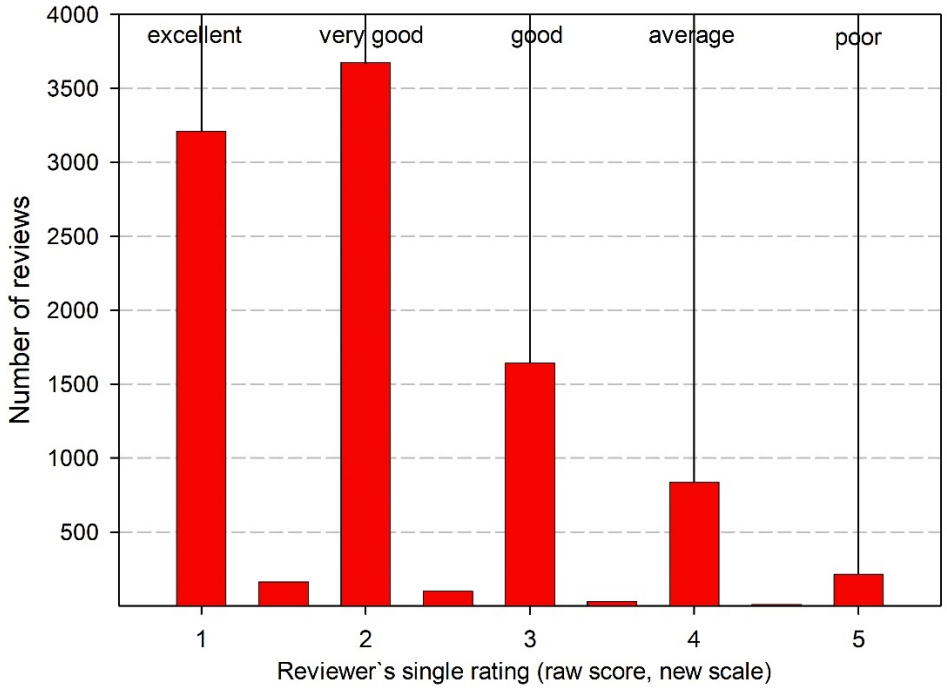

FIGURE 9: Histogram of reviewers' single raw ratings per proposal (new scale from 2015 onwards)

2.4.5 HISTOGRAMS OF REVIEWERS' SINGLE RATINGS – 100 POINT SCALE

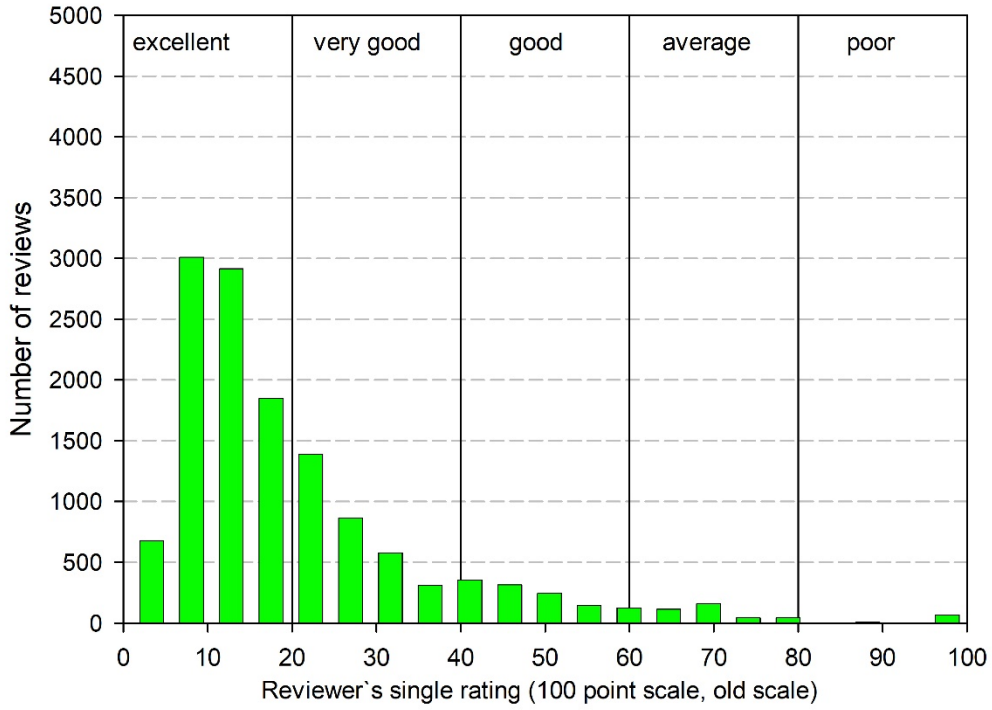

FIGURE 10: Histogram of reviewers' single ratings per proposal, transformed to a 100 point scale (old scale till the year 2015, scale labels were added in steps of 20 points)

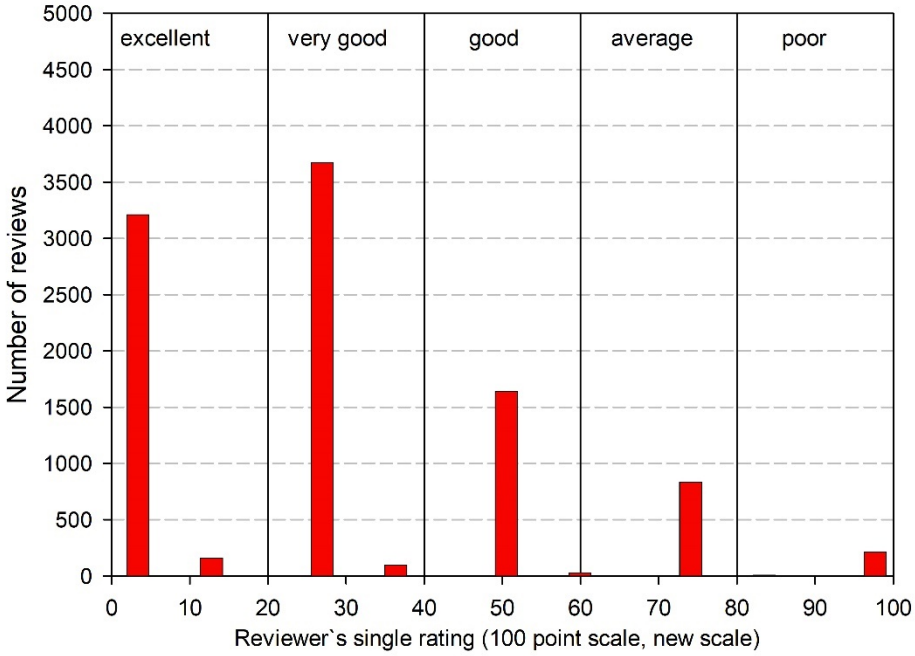

FIGURE 11: Histogram of reviewers' single ratings per proposal, transformed to a 100 point scale (new scale from 2015 onwards, scale labels were added in steps of 20 points)

3 INTER-RATER RELIABILITY OF RATINGS OF REVIEWERS

3.1 PRELIMINARY REMARKS

- The scales used are hybrid scales; strictly speaking, they are neither continuous nor categorical. Accordingly, though the old 100 point scale is continuous, it has, additionally, categorical ratings (e.g. “excellent”, “very good”). The new scale used from 2015 onwards is a categorical/ranking scale (1 = excellent, 2 = very good ...), but has intermediate levels, such as 1.5. To permit comparison between the scales and time periods, on the one hand, a scale was generated that transforms all ratings of reviewers to a 100 point scale (1 = excellent, 100 = poor). On the other hand, a ranking scale variable was generated with 5 categories (1 = excellent, 5 = poor), and for the new scale, intermediate levels such as, for example, 1.5 were randomly assigned to one or the other level.
- The question of how much the ratings of reviewers agree is linked to the question of the scale of measurement. Accordingly, measures of inter-rater agreement (IA) for categorical variables and ranking scale variables can be distinguished from measures of inter-rater reliability (IRR) for continuous scales. IA measures serve to assess absolute agreement (concordance) of reviewers in rating a proposal. The more identical proposals are rated in the same category (e.g. “excellent”) by different reviewers, the higher the IA coefficient. Measures of inter-rater reliability are not focused on absolute agreement, but instead on the question of whether or not reviewers’ ratings are capable, on average, of distinguishing excellent and very good proposals from not so good and poor ones. Generally, this presupposes continuous ratings of reviewers.
- We distinguish between different measures of inter-rater agreement and inter-rater reliability (R packages “stats”, “psych”, “irr”).

  <u>Measures of inter-rater agreement:</u>

  - *Percentage agreement*: Percentage of absolute agreement, not adjusted for chance, in ratings of proposals [0, 1].
  - *Overall Cohen’s kappa (kappa for short)*: Chance-adjusted measure of agreement between two reviews in ratings of applications [0, 1]. The higher the coefficient, the greater the agreement. The kappa coefficient effectively indicates the share of proposals with beyond-chance agreement. Only the diagonal of a cross table is taken into account.
  - *Category-specific kappa*: Chance-adjusted measure of agreement regarding one rating category (e.g. “excellent”) as compared to all other categories [0, 1].
  - *Weighted kappa*: While Cohen’s kappa indicates mean absolute agreement independently of rating category, weighted kappa gives more weight to agreement in adjacent rating categories (e.g. “excellent”, “very good”) than to agreement in non-adjacent rating categories (e.g. “excellent”, “poor”). All cells of a cross table are taken into account, not just the diagonal. Weighted kappa requires ranking scale variables. Weighted kappa (categorical data) is situated between Cohen’s kappa and intraclass correlation (continuous data).

Measures of inter-rater reliability:

- *Intraclass correlation of individual ratings (ICC1)*: ICC1 serves to examine to what extent reviewers' single ratings are capable of distinguishing between proposals regarding their quality, i.e. distinguishing excellent and very good proposals from not so good and poor ones. The higher ICC1, the higher the distinction capacity.
- *Intraclass correlation of aggregated ratings (ICC2):* ICC2 serves to examine to what extent mean ratings of reviewers of a proposal are capable of distinguishing between proposals regarding their quality, i.e. of distinguishing excellent and very good proposals from not so good and poor ones. The higher ICC2, the higher the distinction capacity. It is assumed that averaging of reviewers' ratings per proposal eliminates chance variation and increases reliability.

- Determinants of inter-rater reliability: There are four different influencing factors on ratings of reviewers that can be distinguished statistically: 1) The *quality of the proposal*: The higher the quality of the proposal, the higher the ratings of reviewers. 2) The *applicant*: Applicants have often submitted several proposals in the past. It can be assumed that these persons differ from others in terms of the mean quality of proposals (e.g. excellent vs. average applicants). 3) The *reviewer*: Reviewers who have repeatedly reviewed proposals may differ in terms of strictness or generosity; accordingly, their ratings may tend to be stricter or more generous on average. 4) *Random factors*: Proposals are complex and may cause reviewers to give different ratings of the same proposal (e.g. methodology, content).

- Limitations: Strong divergences in the frequencies of categories have an influence on the amount of the kappa coefficient. The more divergent the frequencies of the rating categories, the greater the likelihood of underestimating the kappa coefficient. Such divergences can be observed in the present data. Intraclass correlation requires a broad variability of ratings, which is the case for the present data, as shown in the histograms.
- Overview: In the following, we will first present a policy-relevant summary of the findings, and then report the findings on inter-rater agreement and inter-rater reliability, including coefficients and cross tables.

3.2 POLICY-RELEVANT SUMMARY OF THE FINDINGS

Overall, actual agreement of ratings of reviewers of a proposal is very low, with Cohen's kappa ~.07, and this holds true regardless of the grading scale or the discipline. Agreement between reviewers is somewhat higher if the rating categories in the secondary diagonals (e.g. "excellent" and "very good") are also rated as agreement (weighted kappa Ķ~.18). Category-specific kappa is highest for the combination of the ratings "excellent" and "excellent" (~.15). Weighted kappa, at .18, is similarly high as the intraclass correlation for single ratings (ρ = .23).

If the mean value of reviewers' ratings of a proposal is implicitly or explicitly relied on for to evaluate and decide on a proposal , ICC2 is relevant for mean ratings of reviewers of a proposal. ICC2 is ρ = .43 for an average of k = 2.4 reviewers per proposal (Jayasinghe, Marsh, & Bond, 2003). This value is slightly below the value (ICC2=.48) reported in the previous study

conducted by the authors in 2012 (Mutz, Bornmann, & Daniel), but still in the range of reliabilities as reported by Ersoheva, Martinkova, & Lee (2021, p. 3) in an overview of various studies on inter-rater reliability for research funding organisations.

Less than one quarter of systematic variance attributable to both the applicant and the quality of the proposal can be attributed to the quality of the proposal (around 7.2% of total variance). The effect of the reviewer (e.g. strictness) accounts for a share of around one third. There are no major differences between disciplines.

3.3 FINDINGS

3.3.1 INTER-RATER AGREEMENT

3.3.1.1 COEFFICIENTS

TABLE 2: Kappa between two reviewers (below the diagonal) and weighted kappa (above the diagonal)

	Reviewer 1	Reviewer 2	Reviewer 3
Reviewer 1	1.00	0.18	0.17
Reviewer 2	0.07	1.00	0.18
Reviewer 3	0.08	0.09	1.00

TABLE 3: Overall kappa coefficients for the first two and first three reviewers of a proposal, each with 95% confidence intervals in brackets

Number of reviewers	Percentage agreement	Cohen's kappa	Weighted Cohen's kappa
2	0.36 [0.35, 0.37]	0.07 [0.05, 0.08]	0.18 [0.07, 0.30]
3	0.36 [0.35, 0.37]	0.08[+]	0.18[+]

[+]Confidence intervals could not be estimated

TABLE 4: Overall Kappa coefficients for the first two reviewers of a proposal, separated for the old scale (till 2015) and the new scale (from 2015 onwards), with 95% confidence intervals in brackets

Number of reviewers	Percentage agreement	Cohen's kappa	Weighted Cohen's kappa
Old scale	0.37 [0.35, 0.38]	0.06 [0.04, 0.07]	0.17 [-0.03, 0.36]
New scale	0.35 [0.33, 0.36]	0.07 [0.05, 0.09]	0.19 [0.06, 0.33]

TABLE 5: Overall Kappa coefficients for the first two reviewers of a proposal, separated for the main disciplines (WD1), with 95% confidence intervals in brackets

Disciplines	Percentage agreement	Cohen's kappa	Weighted Cohen's kappa
Natural Sciences	0.38 [0.36, 0.39]	0.07 [0.05, 0.08]	0.19 [-0.01, 0.38]
Technical Sciences	0.36 [0.32, 0.41]	0.07 [0.01, 0.13]	0.20 [-0.33, 0.73]
Human Medicine and Health Sciences	0.31 [0.29, 0.34]	0.02 [-0.01, 0.05]	0.13 [-0.20, 0.46]

Agricultural Sciences and Veterinary Sciences	0.34 [0.25, 0.43]	0.07 [-0.04, 0.17]	0.29 [-0.46, 1.00]
Social Sciences	0.32 [0.29, 0.35]	0.06 [0.02, 0.10]	0.13 [-0.15, 0.41]
Humanities	0.37 [0.35, 0.40]	0.08 [0.04, 0.11]	0.16 [-0.08, 0.40]

3.3.1.2 CROSS TABLES

3.3.1.2.1 TOTAL

TABLE 6: Cross table of categorical ratings for the combination of the first and second reviewer of a proposal (N= 10,141 reviews, 730 missing values)

		Second reviewer					
	Count *Row per cent* **Kappa**	Excellent	Very good	Good	Average	Poor-rejected	Total
First reviewer	Excellent	1,169 *36.7* **0.15**	1,360 *42.7*	417 *13.1*	188 *5.9*	52 *1.6*	3,186 31.4
	Very good	1,199 *26.9*	2,006 *45.0* **0.07**	795 *17.9*	372 *8.4*	82 *1.8*	4,454 43.9
	Good	333 *21.5*	643 *41.6*	345 *22.3* **0.05**	175 *11.3*	51 *3.3*	1,547 15.3
	Average	108 *15.6*	276 *39.9*	181 *26.2*	88 *12.7* 0.9 **0.03**	39 *5.6*	692 6.8
	Poor-rejected	48 *18.3*	101 *38.6*	56 *21.46*	34 *12.9*	23 *8.8* **0.06**	262 2.6
	Total	2,857 28.2	4,386 43.3	1,794 17.7	857 8.5	247 2.4	10,141 100

Note: kappa = 0.07 [0.05, 0.08], weighted kappa = 0.18 [0.07, 0.30]

3.3.1.2.2 SEPARATED FOR GRADING SCALES

TABLE 7: Cross table of categorical ratings for the combination of the first and second reviewer of a proposal for the old scale till 2015 (N= 5,571 reviews, 161 missing values)

	Second reviewer					
Count Row per cent **Kappa**	Excellent	Very good	Good	Average	Poor-rejected	Total
Excellent	516 *33.3* **0.12**	713 *46.6*	200 *12.9*	90 *5.8*	29 *1.9*	1,548 27.8

First reviewer	Very good	659 *24.3*	1,301 *48.2* **0.04**	468 *17.4*	214 *7.9*	56 *2.1*	2,698 48.4
	Good	151 *18.7*	384 *47.5*	174 *21.5* **0.04**	75 *9.3*	24 *3.0*	808 14.5
	Average	51 *14.3*	145 *40.5*	98 *27.4*	38 *10.6* **0.02**	26 *7.3*	358 6.4
	Poor-rejected	28 *17.6*	63 *39.6*	36 *22.6*	21 *13.2*	11 *6.9* **0.04**	159 2.9
	Total	1,405 25.2	2,606 46.8	976 17.5	438 7.9	146 2.6	5,571 100.0

Note: kappa = 0.06 [0.04, 0.07], weighted kappa = 0.17 [-0.03, 0.36]

TABLE 8: Cross table of categorical ratings for the combination of the first and second reviewer of a proposal for the new scale from 2015 till 2020 (N= 4,565 reviews, 184 missing values)

		Second reviewer					
	Count *Row per cent* **Kappa**	Excellent	Very good	Good	Average	Poor-rejected	Total
First reviewer	Excellent	653 *39.9* **0.14**	647 *39.5*	217 *13.3*	98 *6.0*	23 *1.4*	1,638 35.9
	Very good	540 30.8	705 *40.2* **0.03**	327 *18.6*	158 *9.00*	26 *1.5*	1,756 38.5
	Good	182 *24.6*	259 *35.1*	171 *23.1* **0.05**	100 *13.5*	27 *3.7*	739 16.2
	Average	57 *17.1*	131 *39.2*	83 *24.9*	50 *15.0* **0.04**	13 *3.9*	334 7.3
	Poor-rejected	20 *20.4*	38 *38.8*	20 *20.4*	13 *13.3*	7 *7.1* **0.05**	98 2.2
	Total	1,452 31.8	1,780 39.0	818 17.9	419 9.2	96 2.1	4,565 100.0

Note: kappa = 0.07 [0.05, 0.09], weighted kappa = 0.19 [0.06, 0.33]

3.3.1.2.3 SEPARATED FOR DISCIPLINES

TABLE 9: Cross table of categorical ratings for the combination of the first and second reviewer of a proposal for Natural Sciences (N= 5,394 reviews, 284 missing values)

		Second reviewer					
	Count *Row per cent* **Kappa**	Excellent	Very good	Good	Average	Poor-rejected	Total
First reviewer	Excellent	660 *37.5* **0.14**	802 *45.5*	204 *11.6*	84 *4.8*	12 *0.7*	1,762 32.7
	Very good	694 *12.9*	1,152 *46.9* **0.05**	412 *16.8*	166 *6.8*	30 *1.2*	2,454 45.5
	Good	164 *21.4*	336 *43.9*	174 *22.7* **0.06**	75 *9.8*	17 *2.2*	766 14.2
	Average	48 *15.4*	137 *43.9*	79 *25.3*	35 *11.2* **0.04**	13 *4.2*	312 5.8
	Poor-rejected	16 *16.0*	47 *47.0*	20 *20.0*	13 *13.0*	4 *4.0* **0.03**	100 1.9
	Total	1,582 29.3	2,474 45.9	889 16.5	373 6.9	76 1.4	5,394 100

Note: kappa = 0.07 [0.05, 0.08], weighted kappa = 0.19 [-0.01, 0.38]

TABLE 10: Cross table of categorical ratings for the combination of the first and second reviewer of a proposal for Technical Sciences (N= 480 reviews, 50 missing values)

		Second reviewer					
	Count *Row per cent* **Kappa**	Excellent	Very good	Good	Average	Poor-rejected	Total
First reviewer	Excellent	46 *35.4* **0.15**	61 *46.9*	10 7.7	8 6.2	5 3.9	130 27.1
	Very good	57 26.5	103 47.9 **0.08**	40 18.6	13 6.1	2 0.9	215 44.8

First reviewer	Good	16 19.3	39 47.0	15 18.1 **0.04**	10 12.1	3 3.6	83 17.3
	Average	6 15.8	15 39.5	5 13.2	9 23.7 **0.13**	3 7.9	38 7.9
	Poor-rejected	2 14.3	6 42.9	4 28.6	1 7.1	1 7.1 **0.04**	14 2.9
	Total	127 26.5	224 46.7	74 15.4	41 8.5	14 2.9	480 100.0

Note: kappa = 0.07 [0.01, 0.13], weighted kappa = 0.20 [-0.33, 0.73]

TABLE 11: Cross table of categorical ratings for the combination of the first and second reviewer of a proposal for Human Medicine and Health Sciences (N= 1,599 reviews, 156 missing values)

		Second reviewer					
	Count *Row per cent* **Kappa**	Excellent	Very good	Good	Average	Poor-rejected	Total
First reviewer	Excellent	77 21.4 **0.06**	174 48.3	69 19.2	35 9.7	5 1.4	360 22.5
	Very good	144 19.1	328 43.5 **0.08**	173 22.9	95 12.6	14 1.9	754 47.2
	Good	54 17.9	114 37.8	77 25.5 **0.02**	49 16.2	8 2.7	302 18.9
	Average	17 12.2	51 36.7	43 30.9	18 13.0 **-0.01**	10 7.2	139 8.7
	Poor-rejected	4 9.1	16 36.7	17 38.6	6 13.6	1 2.3 **-0.00**	44 2.8
	Total	296 18.5	683 42.7	379 23.7	203 12.7	38 2.4	1,599 100.0

Note: kappa = 0.02 [-0.01, 0.05], weighted kappa = 0.13 [-0.20, 0.46]

TABLE 12: Cross table of categorical ratings for the combination of the first and second reviewer of a proposal for Agricultural Sciences and Veterinary Medicine (N= 124 reviews, 16 missing values)

		Second reviewer					
	Count *Row per cent* **Kappa**	Excellent	Very good	Good	Average	Poor-rejected	Total
First reviewer	Excellent	9 *29.0* **0.19**	15 *48.4*	5 *16.1*	2 *6.5*	0 *0.0*	31 25.0
	Very good	8 *15.1*	28 *52.8* **0.12**	11 *20.8*	5 *9.4*	1 *1.9*	53 42.7
	Good	4 *16.0*	10 *40.0*	4 *16.0* **-0.05**	6 *24.0*	1 *4.0*	25 20.2
	Average	1 *14.3*	2 *28.6*	2 *28.6*	1 *14.3* **0.00**	1 *14.3*	7 5.7
	Poor-rejected	0 *0.0*	3 *37.5*	2 *25.0*	3 *37.5*	0 *0.0* **-0.03**	8 6.5
	Total	22 17.7	58 46.8	24 19.4	17 13.7	3 2.4	124 100.0

Note: kappa = 0.07 [-0.04, 0.17], weighted kappa = 0.29 [-0.46, 1.00]

TABLE 13: Cross table of categorical ratings for the combination of the first and second reviewer of a proposal for Social Sciences (N= 992 reviews, 89 missing values)

		Second reviewer					
	Count *Row per cent* **Kappa**	Excellent	Very good	Good	Average	Poor-rejected	Total
First reviewer	Excellent	76 *31.5* **0.08**	87 *36.1*	43 *17.8*	24 *10.0*	11 *4.6*	241 24.3
	Very good	88 *21.2*	172 *41.4* **0.08**	81 *19.5*	58 *13.9*	17 *4.1*	416 41.9

First reviewer	Good	35 *19.2*	77 *42.3*	43 *23.6* **0.05**	17 *9.3*	10 *5.5*	182 18.4
	Average	15 *13.8*	43 *39.5*	28 *25.7*	16 *14.7* **0.02**	7 *6.4*	109 11.0
	Poor-rejected	9 *20.5*	12 *27.3*	7 *15.9*	9 *20.5*	7 *15.9* **0.10**	44 4.4
	Total	223 22.5	391 39.4	202 3.5	124 12.5	52 5.2	992 100.0

Note: kappa = 0.06 [0.02, 0.10], weighted kappa = 0.13 [-0.15, 0.41]

TABLE 14: Cross table of categorical ratings for the combination of the first and second reviewer of a proposal for Humanities (N= 1,552 reviews, 135 missing values)

		Second reviewer					
	Count *Row per cent* **Kappa**	Excellent	Very good	Good	Average	Poor-rejected	Total
First reviewer	Excellent	301 45.5 **0.15**	221 33.4	86 13.0	35 5.3	19 2.9	662 42.7
	Very good	208 37.0	223 39.7 **0.10**	78 13.9	35 6.2	18 3.2	562 36.2
	Good	60 31.8	67 35.5	32 16.9 **0.03**	18 9.5	12 6.4	189 12.2
	Average	21 24.1	28 32.2	24 27.6	9 10.3 **0.03**	5 5.8	87 5.6
	Poor-rejected	17 32.7	17 32.7	6 11.5	2 3.9	10 19.2 **0.13**	52 3.4
	Total	607 39.1	556 35.8	226 14.6	99 6.4	64 4.1	1,552 100.0

Note: kappa = 0.08 [0.04, 0.11], weighted kappa = 0.16 [-0.08, 0.40]

3.3.2 INTER-RATER RELIABILITY

TABLE 1 Overview of reported IRR in selected studies on grant peer review

Study	Proposals	Restricted range	IRR methods	IRR_1	IRR_n (n)
Cicchetti (1991)	150 NSF	No restriction	ANOVA	0.18–0.37	0.48[c] (4.24)–0.68[c] (3.69)
Jayasinghe et al. (2001)	2331 ARC	Top 78%	ANOVA and HLM	0.15[a]	0.44 (4.3)
Jayasinghe et al. (2003)	2331 ARC	Top 78%	ANOVA and HLM	0.17[b]	0.46 (4.2)
Carpenter et al. (2015)	260 AIBS	No restriction	ANOVA	0.14–0.41	0.25–0.58[c] (2.00)
Mutz et al. (2012)	8329 FWF	No restriction	HLM	0.26	0.50 (2.82)
Pier et al. (2017)	25 funded NIH	≈ top 18%	Krippendorf alpha	0.08	0.22[c] (3.00)
Pier et al. (2018)	25 funded NIH	≈ top 18%	HLM	0.00	0.00[c] (2-4)

Notes: All reported IRR estimates are for reviewer scores uninformed by panel discussion. IRR_1 denotes single-rater IRR, IRR_n denotes multiple-rater IRR based on n average number of raters.

[a]Proposal quality for humanities, social sciences, and STEM disciplines.

[b]Proposal quality for STEM disciplines only.

[c]Calculation based on Spearman-Brown formula $IRR_n = \frac{n * IRR_1}{1 + (n-1) * IRR_1}$.

FIGURE 12: Overview of reported inter-rater reliability coefficients (Ersoheva, Martinkova, & Lee, 2021, p. 3)

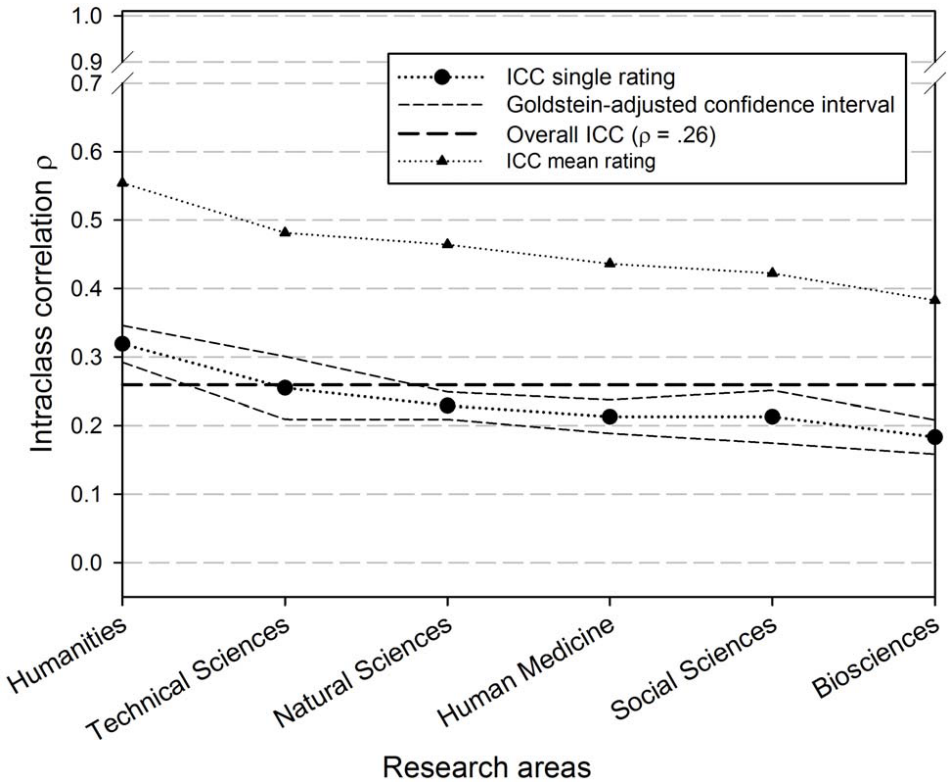

FIGURE 13: Intraclass correlations for single and mean ratings, separated for different disciplines for the years **1998-2008** (Mutz, Bornmann, & Daniel, 2012)

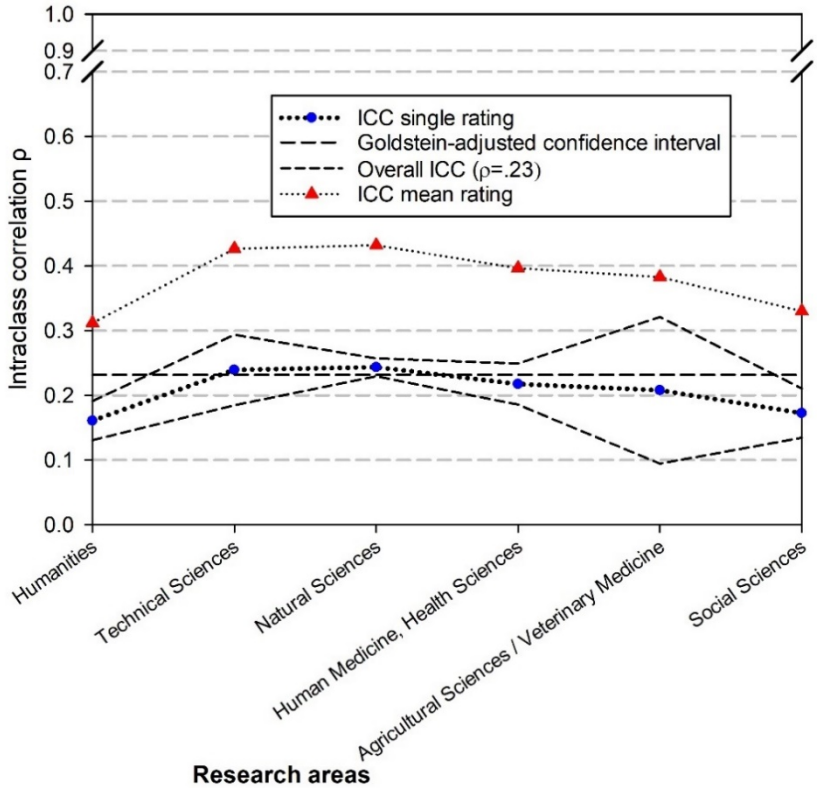

FIGURE 14: Intraclass correlations for single and mean ratings, separated for different disciplines for the years 2010-2019

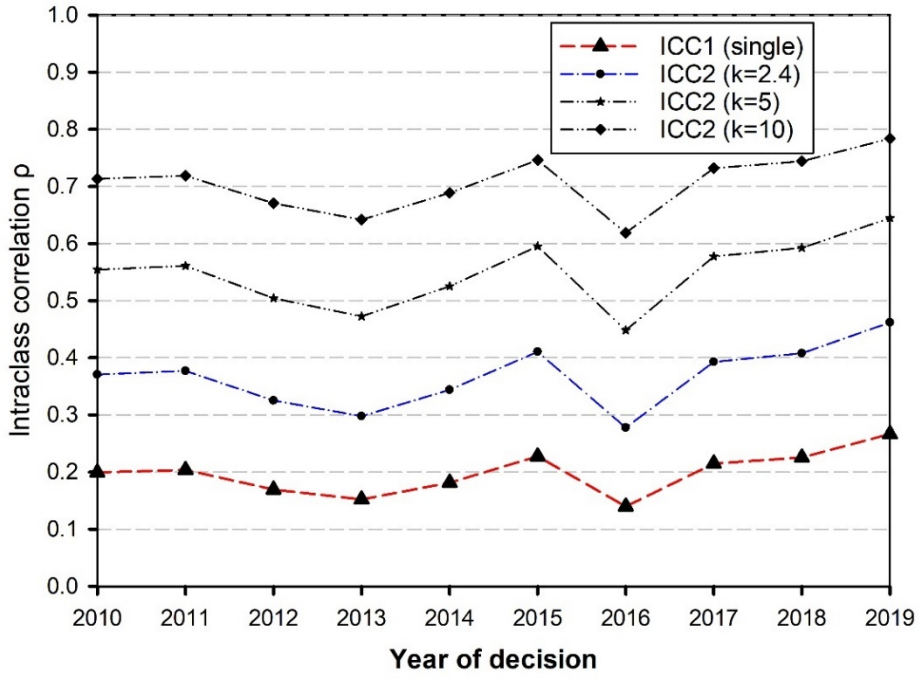

FIGURE 15: Intraclass correlations for single and mean ratings, separated for different years of funding decisions (2010-2019)

3.3.3 Determinants of reviewers' single ratings

Table 15: Number of applicants with more than one proposal and number of reviewers with more than one review, separated for the main disciplines

Discipline	Number of applicants with more than one proposal	Number of reviewers with more than one review
Biology and Medicine	913 (62.62%)	1,414 (20.09%)
Natural and Technical Sciences	926 (55.58%)	1,269 (17.88%)
Social Sciences and Humanities	717 (51.58%)	726 (14.33%)
Total	2,556 (56.62%)	3,409 (17.75%)

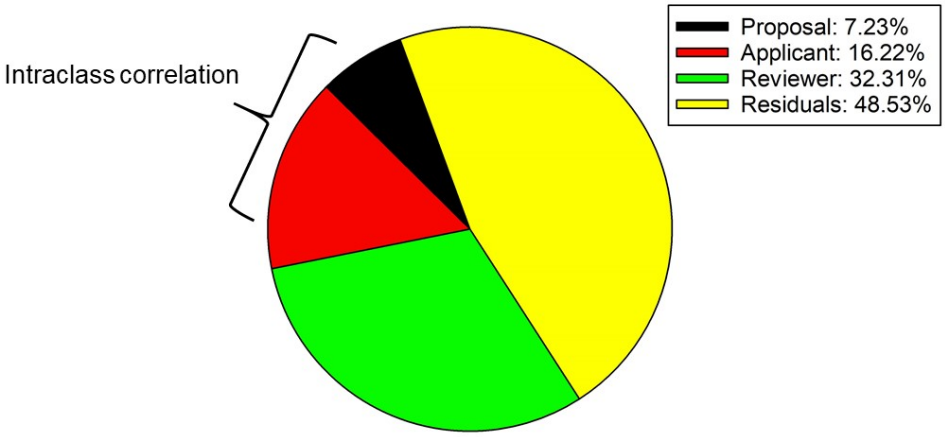

Figure 16: Determinants of reviewers' single scores for all reviewers' ratings of a proposal

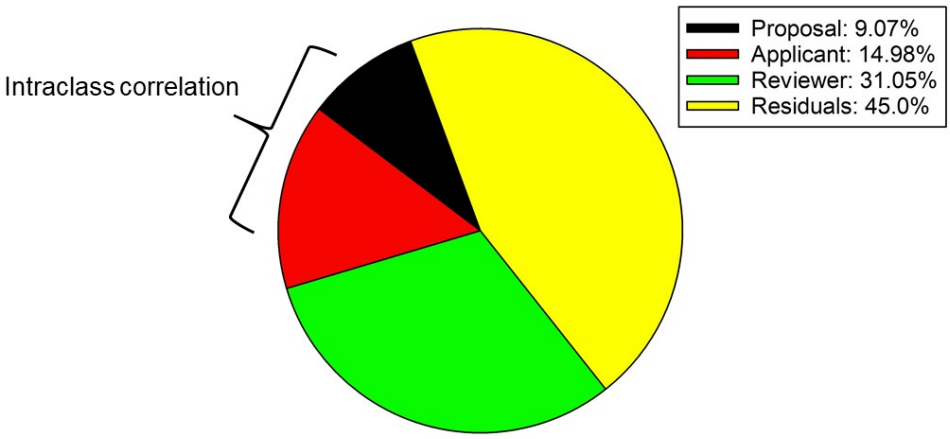

FIGURE 17: Determinants of reviewers' single scores for all reviewers' ratings of a proposal for the main discipline "Biology and Medicine" (FWF)

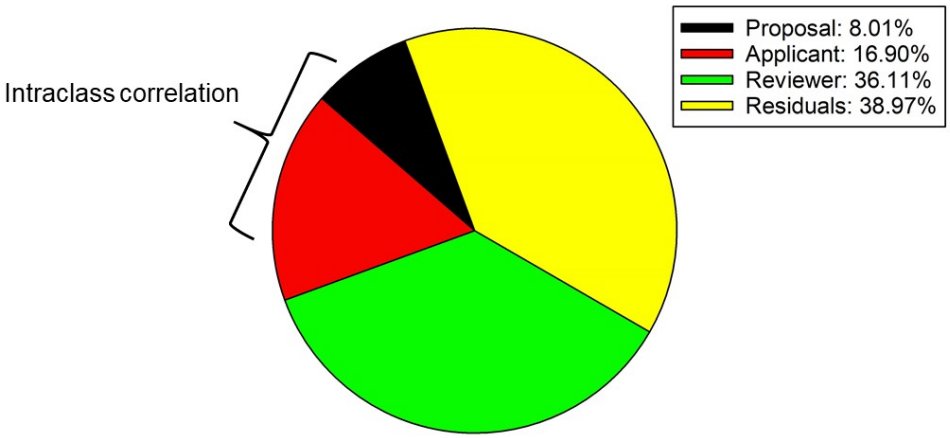

FIGURE 18: Determinants of reviewers' single scores for all reviewers' ratings of a proposal for the main discipline "Natural Sciences and Technical Sciences" (FWF)

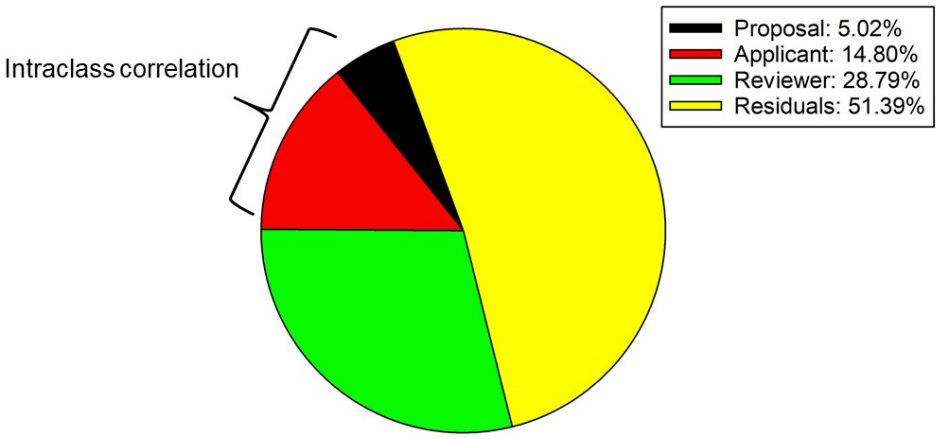

FIGURE 19: Determinants of reviewers' single scores for all reviewers' ratings of a proposal for the main discipline "Social Sciences and Humanities" (FWF)

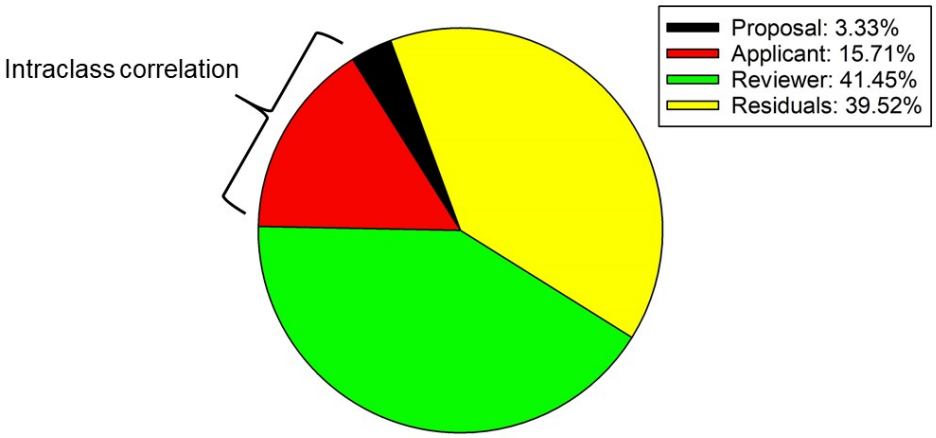

FIGURE 20: Determinants of reviewers' single scores for all reviewers' ratings of a proposal for the old scale (till 2015) (FWF)

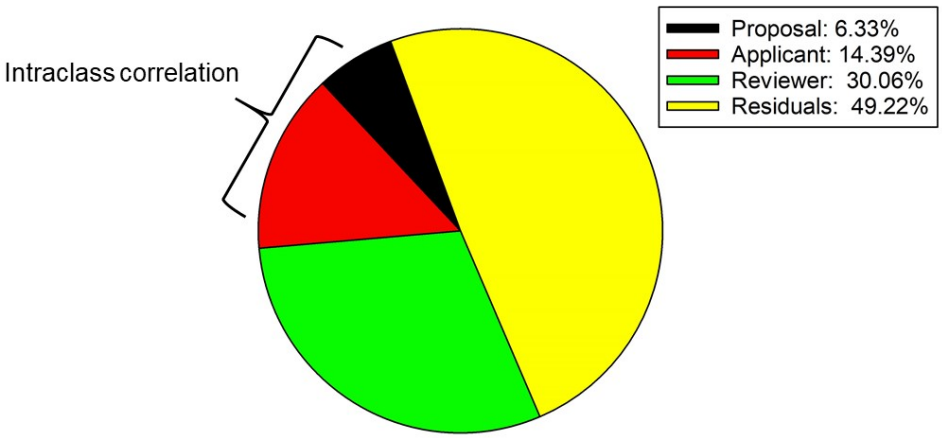

FIGURE 21: Determinants of reviewers' single scores for all reviewers' ratings of a proposal for the new scale (from 2015 onwards) (FWF)

4 BIAS AND FAIRNESS OF THE REVIEW AND DECISION-MAKING PROCEDURE – MEDIATION ANALYSIS

4.1 PRELIMINARY REMARKS

- We speak of "bias" if a property of a proposal or of an applicant has an influence on the ratings of reviewers or the funding decision, but that property is irrelevant to the review and decision-making procedure (e.g. age, gender, institution, country), i.e. it has nothing to do with the quality of the proposal (Lee, Sugimoto, Zhang, & Cronin, 2013). If there is no effect for a property, then there is also no bias, and the review and decision-making procedure is fair with regard to that property.

- A distinction must be made between *potential* bias and *actual* bias. For instance, differences in the ex-ante evaluation of proposals submitted by universities of different sizes (personnel, funds) may actually also reflect differences in quality and later project performance (ex-post evaluation). Such a case would not constitutea real bias.

- The following are potential bias variables: *gender, age of the principal investigator, funding decision in the first vs. the last quarter of a year, University of Vienna, TU Vienna, Medical University of Vienna, University of Innsbruck, University of Graz, University of Natural Resources, and a random variable to check for potential distortions in the statistical procedure (no bias).*

- Although it is problematic to assume causality of effects per se, the concept of causal inference from the statistics of experiments can nonetheless serve to assess the evidence for specific conclusions.

- In a causal mediation model, ratings of reviewers can be understood as mediators of the influence of bias variables on a funding decision (**FIGURE 22**). Accordingly, reviewers deliver assessments of a proposal that may be distorted by properties of the proposal. In turn, the ratings of reviewers and their potential distortions have an influence on funding decisions. These indirect effects must be distinguished from direct effects. This means that properties of a proposal may have a direct influence, without mediation by the peer review system, on the funding decision (FWF Office, presentation of the proposal to the Board of Trustees, conditions of the sessions of the Board of Trustees, ...).

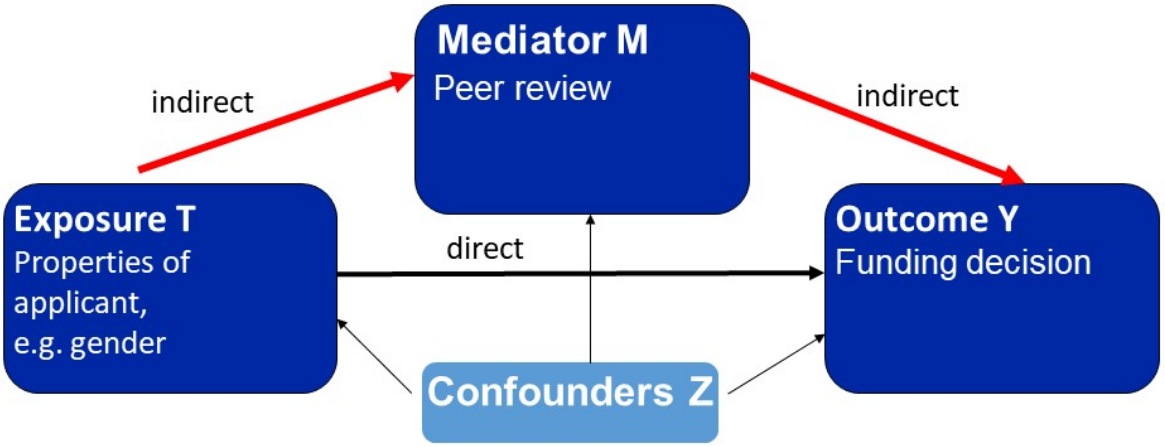

FIGURE 22: Mediation model of peer review

- *Odds ratio:* To show the effects on the funding decision, odds are used, i.e. the ratio of the probability that a proposal will be approved to the probability that a proposal will not be approved (p/(1-p)). Accordingly, odds of 1:3 mean that 1 out of 4 proposals will be approved, i.e. that the probability of a proposal being approved is p = 1/4. The odds ratio (OR) puts the odds of two groups, e.g. men and women, in relation to each other:

$$\mathrm{OR}=\frac{\frac{p_{Female}}{(1-p_{Female})}}{\frac{p_{Male}}{(1-p_{Male})}}$$

If there is no difference between the approval rate for women and the approval rate for men ($p_{Male} = p_{Female}$), then the odds ratio OR equals 1.0. There is no bias. If OR is greater than 1, women have higher odds of their proposals being approved than men; if OR is smaller than 1, they have lower odds. An OR of 2 means that the odds that a proposal will be approved are twice as high for women as the odds for men, e.g. 1:1 (women) to 1:2 (men). If the approval rate for proposals submitted by men is p_{Male} = 0.35 and higher by 5 percentage points than the approval rate for women (p_{Female} = 0.30), then OR equals ~0.80 for women. If the approval rate for proposals submitted by men is higher by 10 percentage points than the approval rate for proposals submitted by women (p_{Female} = 0.30), then OR equals 0.64. Conversely, if the approval rate for proposals submitted by men is lower by 10 percentage points (p_{Male} = 0.20) than the approval rate for proposals submitted by women (p_{Female} = 0.30), then OR equals 1.71. Odds ratios are proportions and also depend on the absolute figures of the probabilities involved (OR(p_{Male} = 0.40, p_{Female} = 0.30) = 1.55 ≠ OR(p_{Male} = 0.30,

p_{Female} = 0.20) = 1.71). *OR is statistically significant if it shows statistically significant deviation from 1.00, i.e. if 1.00 is not in the 95% confidence interval. The tests were not adjusted for multiple testing (control for familywise Type I error rate).*

- *Confounders:* Since funding decisions may depend on a variety of factors, isolation of a bias variable without consideration of other variables may lead to false conclusions. For instance, there may be differences between disciplines with regard to the level of ratings of reviewers as well as to the gender ratio. Potential differences in approval rates would not necessarily reflect gender-specific differences, but differences between the disciplines' specific female ratios and levels of ratings of reviewers. Therefore, the following variables were taken into account as confounders in the mediation model (**FIGURE 22**) to control or adjust for such effects: *age, gender, affiliation with the University of Vienna, Austrian citizenship of the applicant, number of proposals, grading scale, requested funding amount, processing time from submission to funding decision, year of funding decision (linear time trend), Biology and Medicine, Natural Sciences and Technical Sciences, Social Sciences and Humanities (FWF disciplines)*. If a variable, e.g. gender, was used as a bias variable, it was deleted from the list of confounders.
- *Effects:* For representing the "causal" effects, three types of effects on the funding decision are distinguished, each of which can be represented as an OR: a) *natural indirect effects* (NIE) and b) *natural direct effects* (NDE), which when multiplied yield c) *total effects* (TE). The results are shown for the case where no adjustment was made, and for the case where an adjustment was admitted, as well as a potential interaction of the bias variable with the mediator. If there is such an interaction, then the same mean ratings of proposals are associated with different odds of approval for different bias groups.

4.2 METHODOLOGICAL DETAILS

- Mediation analysis is a branch of causal inference which has been strongly influenced by Don Rubin in terms of statistics (Rubin Causal Model). A central aspect of the RCM is the potential outcomes framework, developed from randomised control trials. For statements on causality to be made, a person (or unit) would have to be examined for an effect (e.g. pain perception) both with treatment (e.g. headache pill) and in the control condition (no pill). Unfortunately, that is basically impossible. There are two potential outcomes for a person, of which only one can become reality. What is possible, however, is to observe the mean values of groups and to assume that individuals are assigned to the groups on a purely random basis (randomisation), or to use propensity score matching if randomisation is not possible.
- Based on this fundamental consideration, Rubin developed the concept of direct and indirect effects. In our case, the experimental variation, e.g. gender, is ultimately a label. An indirect effect of gender on a funding decision, mediated by the peer review system, can be distinguished from a direct effect on the funding decision, which includes everything apart from the peer review (e.g. FWF Office, session of the Board of Trustees, ...). The direct effect is the effect of gender on the funding decision for each mean peer review score observed per proposal. The indirect effect is the effect on the funding decision if proposals submitted by men (= 0) were rated like proposals submitted by women (= 1) ("what if", potential outcomes framework), regardless of

whether the proposals were actually submitted by women or by men. Ultimately, a total of three regression analyses were performed, which were combined to yield a result: X->Y, X->Z, Z->Y , where X = gender, Z = peer review, Y = funding decision. The problem is that there is no randomised control trial, i.e. women and men also differ in terms of other co-variables that have an influence on funding decisions, e.g. discipline (confounding). However, it is partly possible to control statistically for the influence of these variables. Moreover, there may be an interaction of the mediator "reviewers' mean score" with the bias variable. If comparable mean peer review scores for men and women lead to different funding decisions for women and men, then there is an interaction X*Z. The R package medflex was used for the analysis (https://cran.r-project.org/web/packages/medflex/ vignettes/ medflex.pdf). Since it was of course impossible to control for all factors, in particular for the actual quality of a proposal, caution should be exercised when interpreting the findings.

4.3 POLICY-RELEVANT SUMMARY OF THE FINDINGS

Overall, no *total bias effect* on funding decisions was found for the variables "gender", "age at submission" and "first vs. last quarter of a year" (**TABLE 16**). The FWF's review and decision-making procedure is fair overall with regard to these variables. However, a small, statistically significant gender effect was found in the review procedure. Taking into account only the peer review in the funding decision, the odds of approval are lower by 0.85 (adjusted: 0.90) for proposals submitted by women than for proposals submitted by men. Since OR is 1.0 or slightly higher in processes not subject to peer review (FWF Office, sessions of the Board of Trustees), these processes compensate for the distorting effect of the peer review system. Conversely, we found with regard to age that based on peer review alone, the odds of a proposal being approved are slightly higher for applicants older than 41 years, by a statistically significant 1.12 (adjusted: 1.07). This distorting influence is also compensated for by processes other than peer review with slightly lower odds for older applicants, resulting in the procedure being fair overall (total effect).

The situation is different when it comes to the selected universities (**TABLE 17**). In each case, proposals from a university are compared to proposals from all other universities.

With the exception of the Medical University of Vienna, the processes other than peer review (FWF Office, sessions of the Board of Trustees) are fair, with OR at 1.0. However, there is a slight bias effect for the University of Vienna and the University of Natural Resources. While the odds of approval are overall higher by 1.31 (adjusted: 1.29) for proposals from the University of Vienna, the odds of approval are lower by 0.69 (adjusted: 0.68) for proposals from the University of Natural Resources. For both universities, the higher and lower odds, respectively, of proposals being approved are attributable to the peer review system. For the Medical University of Vienna, the odds of proposals being approved drop to 0.85 (adjusted: 0.90) as compared to all other universities. This drop is compensated for by processes not subject to peer review (OR = 1.16, adjusted), meaning that overall, the funding decision is fair, at least in the adjusted case. For the other universities under consideration, the review and decision-making procedure is fair.

4.4 FINDINGS

4.4.1 APPROVAL RATES

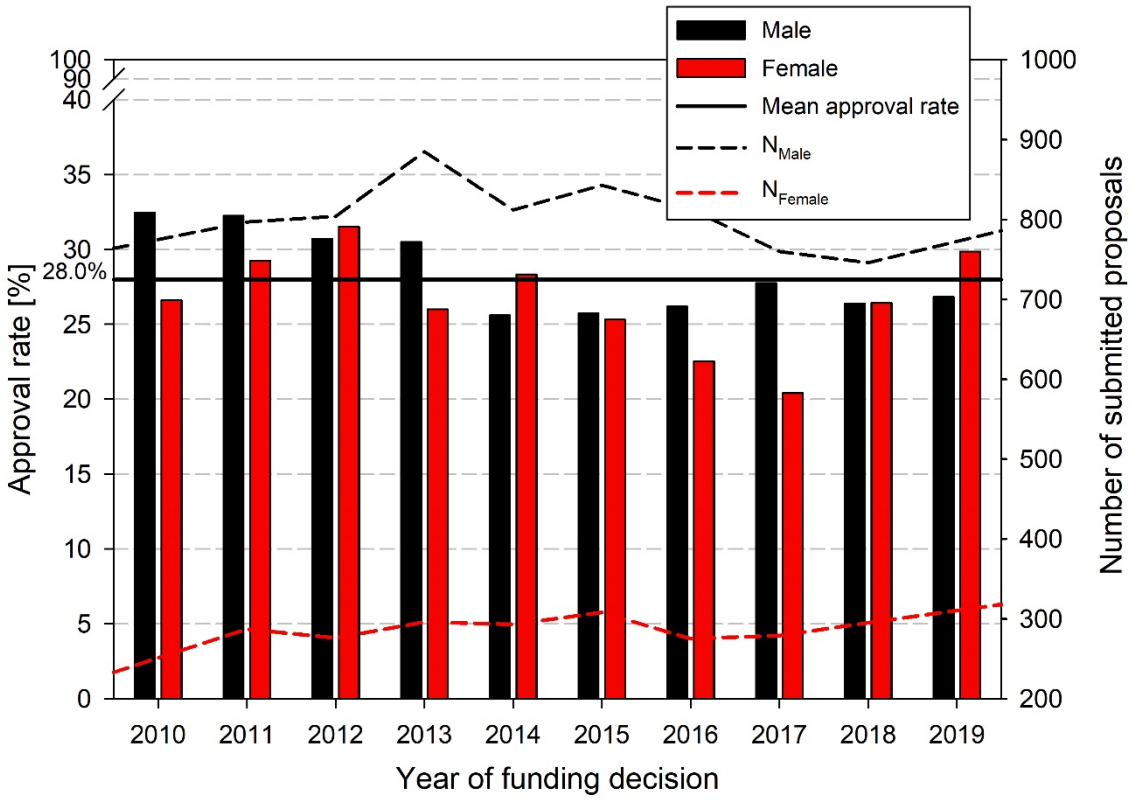

FIGURE 23: Approval rates separated for different years of funding decisions

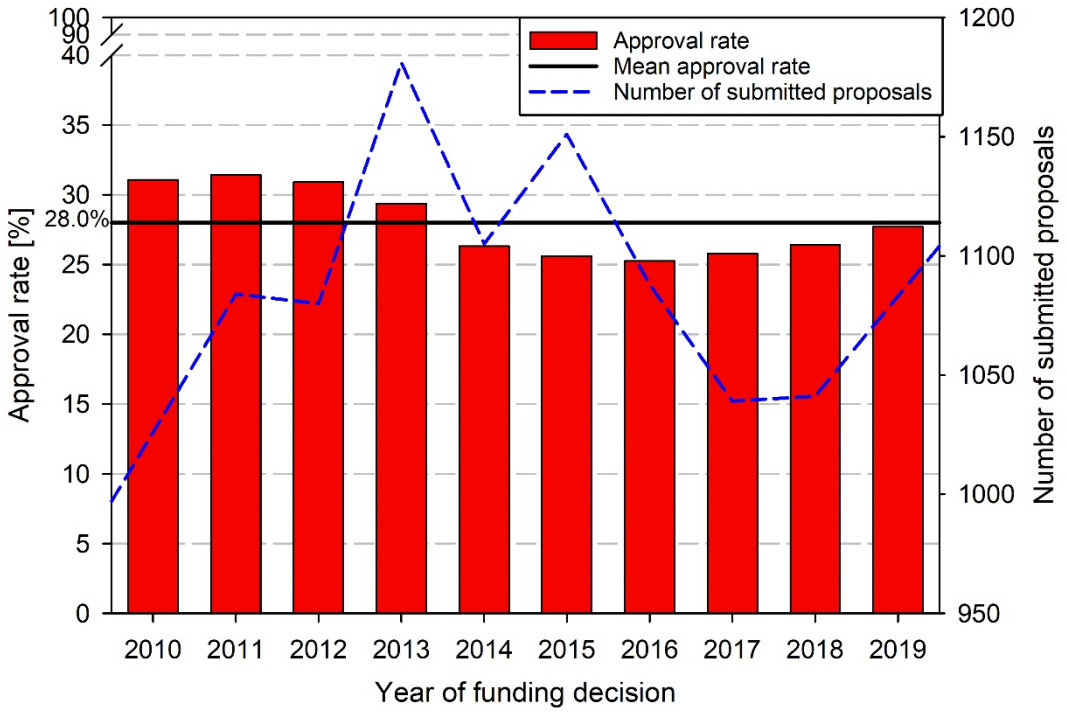

FIGURE 24: Approval rates separated for different years of funding decisions

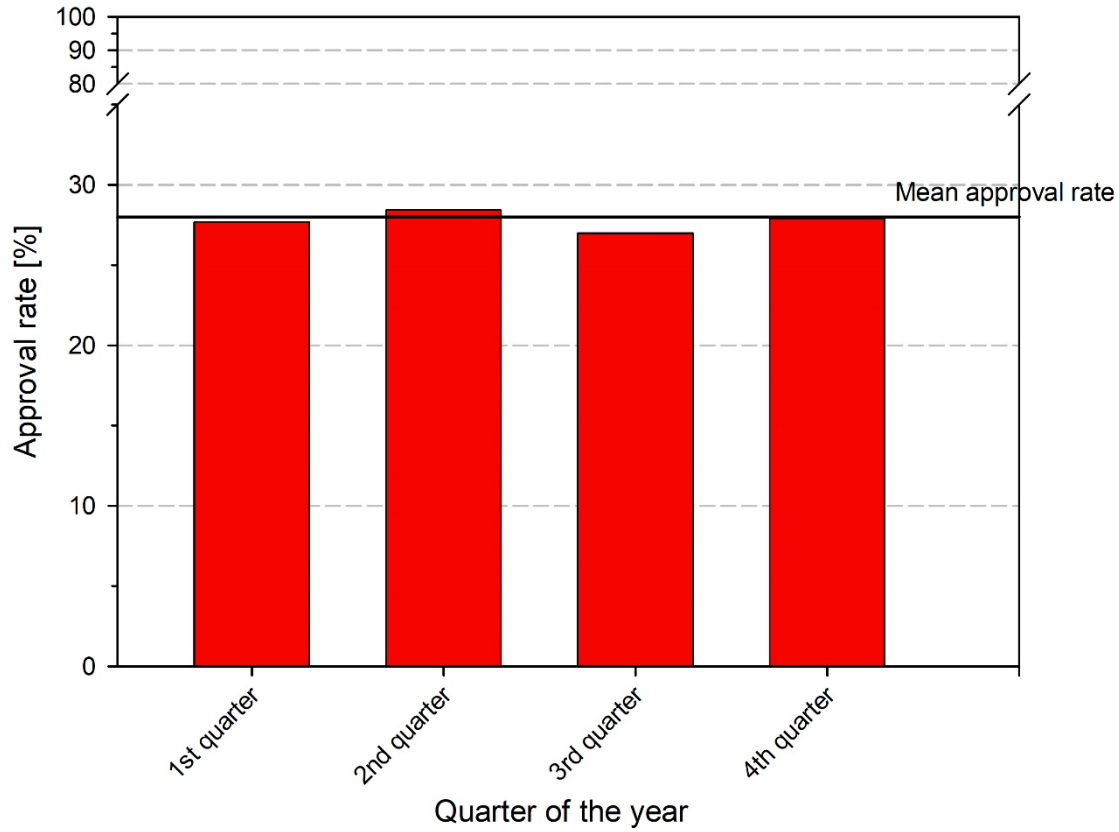

FIGURE 25: Approval rates separated for the quarters of the years of funding decisions

4.4.2 Mediation analysis

The summarised results of mediation analysis are presented in **Table 16** and **Table 17** in the form of odds ratios with confidence intervals in brackets. The results of the underlying individual regression analyses are not reported. ORs with no statistically significant divergence from 1.0 (no bias) are marked in green; ORs with higher odds for the group assigned 1, e.g. women, are marked in red, and ORs with lower odds for the group assigned 1 are marked in blue.

Table 16: Natural direct, indirect and total effect of a set of Austrian universities as bias variables in terms of odds ratios of approval rates with 95% confidence intervals in brackets (**green** = "OR=1", **blue** = "OR<1", **red** = "OR>1", statistically significant)

Bias variable (0/1)	Adjustment	Natural direct effect	Natural indirect effect	Total effect
Random variable (1=assigned)	No	0.97 [0.90, 1.04]	0.98 [0.92, 1.04]	0.95 [0.87, 1.03]
	Yes	0.97 [0.90, 1.03]	0.98 [0.92, 1.04]	0.94 [0.86, 1.03]
Gender (1=female)	No	1.08 [1.01, 1.16]	0.85 [0.80, 0.91]	0.92 [0.83, 1.02]
	Yes	1.04 [0.95, 1.14]	0.90 [0.81, 0.98]	0.93 [0.82, 1.05]
Age at submission (1=older than 42)	No	0.91 [0.86, 0.96]	1.12 [1.06, 1.18]	1.02 [0.94, 1.11]
	Yes	0.95 [0.89, 1.01]	1.07 [1.01, 1.13]	1.01 [0.92, 1.09]
First (=1) vs. last quarter of a year	No	0.99 [0.92, 1.07]	0.97 [0.89, 1.06]	0.96 [0.86, 1.09]
	Yes	0.95 [0.88, 1.02]	1.04 [0.95, 1.13]	0.98 [0.87, 1.11]

Table 17: Natural direct, indirect and total effect of a set of Austrian universities as bias variables in terms of odds ratios of approval rates with 95% confidence intervals in brackets (**green** = "OR=1", **blue** = "OR<1", **red** = "OR>1", statistically significant)

Bias variable (0/1)	Adjustment	Natural direct effect	Natural indirect effect	Total effect
University of Vienna	No	0.95 [0.90, 1.02]	1.35 [1.25, 1.47]	1.29 [1.17, 1.42]
	Yes	0.99 [0.94, 1.06]	1.32 [1.21, 1.43]	1.31 [1.19, 1.45]
MedUni Vienna	No	1.25 [1.12, 1.38]	0.66 [0.60, 0.73]	0.82 [0.70, 0.96]
	Yes	1.16 [1.04, 1.29]	0.77 [0.69, 0.85]	0.89 [0.76, 1.03]
TU Vienna	No	0.92 [0.83, 1.02]	1.22 [1.10, 1.35]	1.12 [0.97, 1.29]
	Yes	0.95 [0.85, 1.06]	1.08 [0.96, 1.22]	1.02 [0.88, 1.20]
University of Innsbruck	No	0.92 [0.83, 1.03]	1.06 [0.95, 1.19]	0.98 [0.84, 1.16]
	Yes	0.92 [0.83, 1.02]	1.05 [0.93, 1.19]	0.96 [0.82, 1.14]
University of Graz	No	0.96 [0.86, 1.08]	0.96 [0.87, 1.05]	0.92 [0.79, 1.07]
	Yes	1.01 [0.90, 1.15]	0.94 [0.85, 1.03]	0.95 [0.82, 1.12]
University of Natural Resources	No	0.95 [0.82, 1.07]	0.72 [0.62, 0.83]	0.68 [0.55, 0.82]
	Yes	0.90 [0.77, 1.03]	0.75 [0.64, 0.88]	0.69 [0.55, 0.83]

4.5 CONTEXT EFFECTS: BIG-FISH-LITTLE-POND EFFECT

4.5.1 PRELIMINARY REMARKS

- A funding decision may depend not only on the quality of a proposal, but also on the specific context in which the decision was taken, i.e. the specific conditions of a session of the Board of Trustees. In this context, the quality level of the proposals discussed in a session of the Board of Trustees is particularly important. Accordingly, proposals of average quality may yet be approved in a session with below-average quality level, while the same proposals would be rejected in a session with above-average quality level, and vice versa. Such effects are even likelier if approval rates remain relatively constant across sessions, since even if quality level is high, a session cannot approve more proposals than permitted by the approval rate. If such context effects appear, they constitute distortion/bias, as the benchmark for a funding decision is no longer the proposal's quality alone, but also the context of the decision. This is a so-called compositional effect, as the total effect of the ratings of reviewers on the approval rate is split into an individual effect of the proposal and an effect of the session. With reference to H.W. Marsh, we speak of a Big-Fish-Little-Pond effect (BFLP effect) in this context. Marsh demonstrated such context effects for schools and universities.
- A necessary condition for such an effect is that sessions differ with regard to the mean quality level of proposals.

4.5.2 METHODOLOGICAL APPROACH

- In order to demonstrate such context effects, the mean ratings of reviewers per proposal are split into two variables, a variable with the mean ratings per session on the one hand, and, on the other hand, a variable with the mean ratings per proposal with respect to the grand mean centred. Logistic regression analysis is used to examine if the variable with the mean ratings per session is a statistically significant predictor of approval rates. In that case, the regression coefficient indicates the compositional effect, i.e. the effect of the session regardless of the quality of the individual proposal.
- Only the 50 sessions of the Board of Trustees were taken into account. Other sessions, e.g. meetings of the Executive Board, were excluded. Three sessions of the Board of Trustees were eliminated because they included proposals rated on the old and new scales. Therefore, a total of 47 sessions were analysed.
- The BFLP effect was estimated using logistic regression with random effects of the binary funding decision on the two variables as predictors, taking into account in the model the grading scale as a main effect and interaction effect, as well as differences in approval rates across sessions.

4.5.3 POLICY-RELEVANT SUMMARY OF THE FINDINGS

- The sessions of the Board of Trustees differ both in terms of mean quality level of proposals and in terms of approval rates. However, the differences are comparatively small as measured by the standard deviations of the two variables.
- As expected, we found that the better a proposal's mean peer review rating as compared to the mean quality of proposals discussed in a session, the higher the probability of its approval. Separately, however, a Big-Fish-Little-Pond effect (BFLPE) was also found, at least for the old scale (till the year 2015). For proposals rated on the

old scale (till the year 2015), probability of approval was higher for proposals discussed in sessions of the Board of Trustees with above-average quality level than for those discussed in sessions with below-average level, and this regardless of the actual quality of the proposal. This effect was not observed for the new scale (from 2015 onwards).

- The effect can be quantified as an odds ratio. The odds ratio puts the ratio of approval rates $p_{+1}/(1-p_{+1})$ for a session with a quality level increased by one unit in relation to the ratio of approval rates of the mean of all sessions $p_0/(1-p_0)$ in turn: $p_{+1}/(1-p_{+1})$ / $p_0/(1-p_0)$. If the odds ratio is 1.0, there is no distortion.
- If a session's quality level was higher by 1 unit on the rating scale as compared to the mean of all sessions, this yielded an odds ratio of 1.19 for the old scale and an odds ratio of 1.05 for the new scale. In other words, the odds of a session with a quality level higher by one scale unit entailed higher odds by 1.19 (1.05), or 19% (5%), than the odds of a session with an average quality level of proposals, regardless of the actual quality of a proposal. For the new scale, there was no statistically significant deviation of the odds ratio, at 1.05, from 1.0, i.e. a quality level increase by 1 scale unit shows no effect for the new scale.

4.5.4 Findings

4.5.4.1 Big-fish-little-pond effect

Table 18: Descriptive statistics of the reviewers' mean score (percentage scale) and approval rate for the N = 47 sessions of the Board of Trustees

Scale	Variable	N	Mean	SD	Min	Max
Old	Reviewers' mean score	26	20.01	0.95	18.38	22.93
	Approval rate	26	0.31	0.02	0.26	0.33
New	Reviewers' mean score	21	27.79	1.45	24.85	31.23
	Approval rate	21	0.27	0.02	0.24	0.31

The sessions of the Board of Trustees differ in terms of the quality level of proposals and the level of approval rates (**Table 18**). The standard deviations are 0.95 and 1.45 for "reviewers' mean score" and 0.02 for "approval rate".

The results of logistic regression (**Table 19**) show that the better reviewers' mean score for a proposal as compared to the mean level of the session of the Board of Trustees (negative values), the higher the probability that the proposal will be approved (b_1 = -0.42*). However, the context of the funding decision plays a role, too. The higher the quality level of proposals in a session (negative values), the higher the probability that a proposal will be approved (b_2 = -0.17*). The scale has no effect on approval (b_5 = -0.05).

TABLE 19: Parameter estimates for the logistic regression model (fixed effects model) of the probability of approval of a proposal

Effect	Parameter	Estimate	SE	t-value
Intercept	b_0	-1.18	1.02	-1.16
Deviation from session mean	b_1	-0.42	0.01	-33.14*
Session mean	b_2	-0.17	0.05	3.42*
Scale (0 = old / 1 = new)	b_3	-0.05	1.50	-0.03
Deviation × scale (=1)	b_4	0.19	0.02	12.87*
Session mean × scale (=1)	b_5	0.14	0.06	2.11*
Compositional eff. old scale ($b_1 + b_4$)	b_6	-0.17	0.05	-3.42*
Compositional eff. new scale ($b_2 + b_5$)	b_7	0.04	0.04	-0.95*

*p<.05, df = 9,814

Overall, these context effects are expressed in the odds ratios, which can be represented in two ways. We can express, on the one hand, what happens if the quality level of a session or of a proposal decreases by one unit (or the mean score increases by one unit) (**TABLE 20**), and on the other hand, what happens if the quality level of a session or of a proposal increases by one unit (or the mean score decreases by one unit) (**TABLE 21**). The odds ratio expresses the ratio of the approval rate $p_{+1}/(1-p_{+1})$ for proposals/sessions with a quality level increased (decreased) by one unit in relation to the mean approval rate $p_0/(1-p_0)$ as follows: $p_{+1}/(1-p_{+1})/p_0/(1-p_0)$.

On both scales, proposals have an odds ratio well above 1 (old scale: 1.52, new scale: 1.25) if the quality of the proposals is above average as compared to the session mean (**TABLE 21**). The better,i.e., the lower the mean peer review rating for a proposal as compared to the mean quality level of proposals discussed in a session, the higher the proposal's probability of approval. However, the mean quality level of proposals of a session also has an influence, i.e. there is a Big-Fish-Little-Pond effect. If a session's quality level was higher by 1 unit on the scale as compared to the mean of all sessions, this yielded an odds ratio of 1.19 for the old scale and an odds ratio of 1.05 for the new scale. In other words, the odds of a session with a quality level higher by one scale unit entailed higher odds by 1.19 (1.05), or 19% (5%), than an average session. For the new scale, there was no statistically significant deviation of the odds ratio, at 1.05, from 1.0, i.e. a quality level increase by 1 unit shows no effect. If quality level increases by 2 units on the scale, the odds ratio is e.g. 1.08 [0.91, 1.25] for the new scale, i.e. the odds are higher by 8% than for a session with average quality level, but not statistically significant.

TABLE 20: BFLP effects in terms of odds ratios, if a *mean score increases* by one unit, i.e. the quality of a proposal decreases by one unit, separated for old and new scale (95% confidence interval in brackets)

Scale	Deviation from session mean	Session mean	Odds ratios
Old (2009-2015)	0 (average)	+1 score unit	0.841 [0.757, 0.924]
	+1 score unit	0 (average)	0.658 [0.642, 0.675]
New (2015-2019)	0 (average)	+1 score unit	0.963 [0.888, 1.038]
	+1 score unit	0 (average)	0.798 [0.785, 0.810]

Reading example: If the mean reviewers' score of a proposal equals the average of the session (deviation from session mean = 0), an increase of the session mean by +1 unit results in an odds ratio of 0.841 for the old scale.

TABLE 21: BFLP effects in terms of odds ratios, if *mean score decreases* by one unit, i.e. the quality of a proposal increases by one unit, separated for old and new scale (95% confidence interval in brackets)

Scale	Deviation from session mean	Session mean	Odds ratios
Old (2009-2015)	0 (average)	-1 score unit	1.19 [1.07, 1.31]
	-1 score unit	0 (average)	1.52 [1.48, 1.56]
New (2015-2019)	0 (average)	-1 score unit	1.05 [0.98, 1.13]
	-1 score unit	0 (average)	1.25 [1.23, 1.27]

Reading example: If the mean reviewers' score of a proposal equals the average of the session (deviation from session mean = 0), a decrease of the session mean by -1 unit results in an odds ratio of 1.19 for the old scale.

4.5.4.2 APPROVED PROPOSALS WITH THE LOWEST RATINGS OF REVIEWERS

TABLE 22: The 5 approved proposals with the lowest average peer reviewers' score for old and new scale on the raw scales and the percentage scale (1=excellent, 100=poor)

Proposal	Old scale (till 2015)			New scale (from 2015 onwards)		
	RID	Raw scale	Percentage scale	RID	Raw scale	Percentage scale
1	R_134072	48.0	53.0	R_198603	2	25
2	R_143080	48.0	53.0	R_160583	2.33	33.33
3	R_142832	43.25	57.75	R_188553	2.33	33.33
4	R_146299	43	58.0	R_159205	2.5	37.5
5	R_151217	1	100	R_186252	2.67	41.67

5 INTERDISCIPLINARITY

5.1 PRELIMINARY REMARKS

- Interdisciplinarity is defined on the basis of the number of disciplines listed in a project proposal. If only one discipline is listed, the project is mono-disciplinary. Interdisciplinarity depends on the number of disciplines to be taken into account. Six disciplines are taken into account for WD1 and 21 disciplines for FWF21.
- To estimate the effect of interdisciplinarity, we used the mediation analysis from chapter 4.

5.2 POLICY-RELEVANT SUMMARY OF THE FINDINGS

Overall, the odds of approval of proposals for interdisciplinary projects as compared to mono-disciplinary projects are lower by ~0.87 for 21 disciplines (FWF21) and by ~0.75 for 6 disciplines (WD1). It is a slight effect (~5% divergence in approval rates).

5.3 FINDINGS

TABLE 23: Proportion of proposals with more than 1 discipline (WD1, FWF21), total and separated for the main disciplines (WD1)

Main disciplines	N	21 disciplines FWF21	6 disciplines WD1
Natural Sciences	5,676	0.57	0.25
Technical Sciences	527	0.72	0.72
Human Medicine, Health Sciences	1,753	0.66	0.41
Agricultural Sciences, Veterinary Medicine	142	0.80	0.80
Social Sciences	1,087	0.50	0.35
Humanities	1,686	0.70	0.35
Total	10,871		

TABLE 24: Natural direct, indirect and total effect of a set of Austrian universities as bias variables in terms of odds ratios of approval rates with 95% confidence intervals in brackets (green = "OR=1", blue = "OR<1", red = "OR>1", statistically significant)

Bias variable (0/1)	Adjustment	Natural direct effect	Natural indirect effect	Total effect
Interdisciplinarity FWF21 (1=inter.)	No	0.93 [0.88, 0.98]	0.94 [0.88, 1.00]	0.87 [0.80, 0.94]
	Yes	0.92 [0.87, 0.97]	0.94 [0.88, 1.01]	0.86 [0.79, 0.94]
Interdisciplinarity WD1 (1=inter.)	No	0.97 [0.91, 1.03]	0.77 [0.72, 0.83]	0.75 [0.68, 0.82]
	Yes	0.96 [0.90, 1.02]	0.79 [0.74, 0.85]	0.76 [0.68, 0.85]

6 REFERENCES

Ersoheva, E. A., Martinkova, P., & Lee, C.L. (2021). When zero may not be zero: A cautionary note on the use of inter-rater reliability in evaluating grant peer review. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 184*(3), 904-919. https://doi.org/10.1111/rssa.12681

Jayasinghe, U. W., Marsh, H. W., & Bond, N. (2003). A multilevel cross-classified modelling approach to peer review of grant proposal: the effects of assessor and researcher attributes on assessor ratings. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 166(3), 279-300.

Lee, C. J., & Sugimoto, C., Zhang, G., & Cronin, B. (2013). Bias in peer review. Journal of the American Society for Information Science and Technology, 64(1), 2-17.

Mutz, R., Bornmann, L., & Daniel, H.-D. (2012). Heterogeneity of Inter-Rater Reliabilities of Grant Peer Reviews and Its Determinants: A General Estimating Equations Approach. *PLoS ONE, 7*(10): e48509. doi:10.1371/journal.pone.0048509.