PEER REVIEW EVALUATION PROCESS OF MARIE CURIE ACTIONS UNDER EU'S FP7

David Pina

Research Executive Agency, European Commission, Brussels, Belgium

Darko Hren

Department of Psychology, School of Humanities and Social Sciences, University of Split, Split, Croatia

Ana Marušić

Department of Research in Biomedicine and Health, School of Medicine, University of Split, Split, Croatia

Marie Curie Actions

- EU Fellowship programmes for researchers' mobility since 1990
- Marie Curie since 1996
- Aim: Structuring training, mobility and career development for researchers
- Under FP7 (2007-2013): €4.75 billion

Marie Curie Actions

ITN → Action 1 **ITN** Early-stage Researchers

Innovative Training Networks

Support for doctoral and early-stage training **European Training Networks, European Industrial Doctorates, European Joint Doctorates**

IEF, IOF, IIF → Action 2 **IF** Experienced Researchers

Individual Fellowships

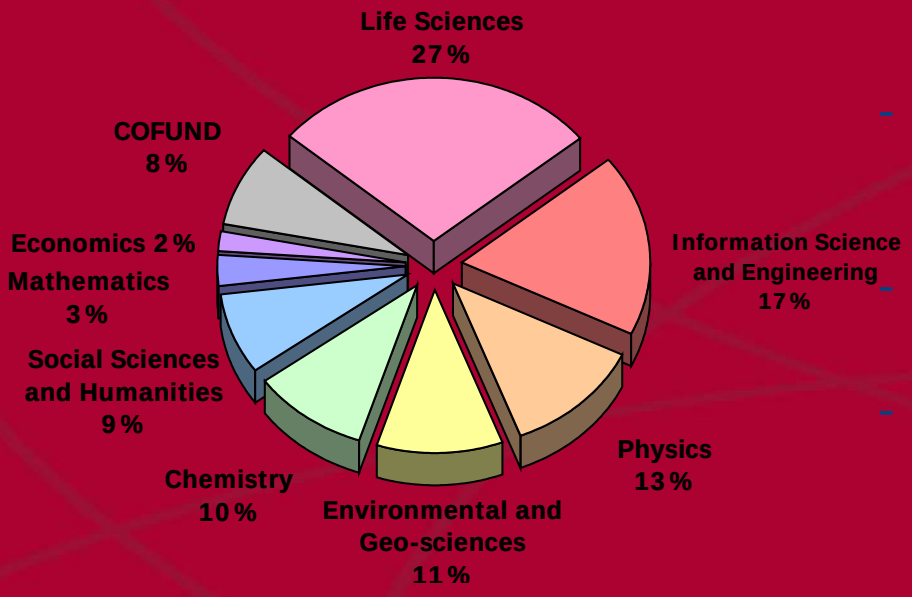
Support for experienced researchers undertaking international and inter-sector mobility: **European Fellowships** and **Global Fellowships**
Dedicated support for **career restart** and **reintegration**

IAPP → Action 3 **RISE** Exchange of Staff

Research and Innovation Staff Exchange

International and inter-sector cooperation through the exchange of staff

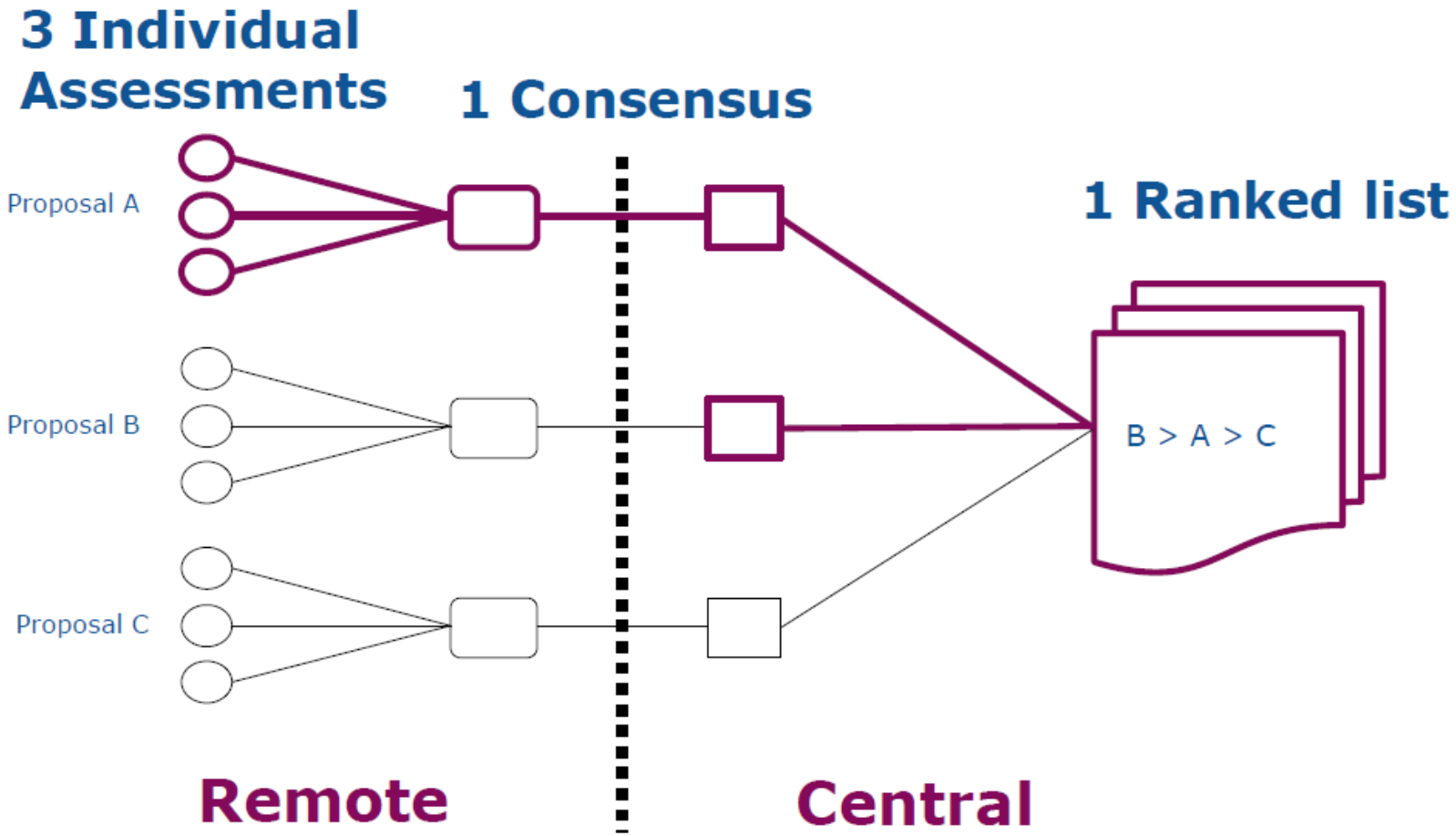
Marie Curie Actions



Budget distribution by scientific panel in FP7

- 60 000 researchers financed since the creation of the Marie Curie Actions
- More than 10 000 PhD supported in FP7
- Marie Curie researchers coming from all over the world (around 130 nationalities)
- Marie Curie host organisations in more than 80 countries
- 46% of researchers coming to EU from industrialised countries stay in Europe after the end of their IIF fellowship
- 38% women participation in FP7 MCA, close to the 40% target

Marie Curie Actions



Marie Curie Actions

Excellent. Successfully addresses all relevant aspects of the criterion in question. Any shortcomings are minor.	5		Excellent
Very Good. Addresses the criterion very well, although certain improvements are still possible.	4	4.9 – 4.0	Very Good
Good. Addresses the criterion well, although improvements would be necessary.	3	3.9 – 3.0	Good
Fair. Broadly addresses the criterion, there are significant weaknesses.	2	2.9 – 2.0	Fair
Poor. Addressed in an inadequate manner, or there are serious inherent weaknesses.	1	1.9 – 1.0	Poor
Fails to address the criterion or cannot be judged due to missing or incomplete information.	0		

Marie Curie Actions

CRITERIA

- S&T Quality
- Training (ITN, IEF) or Transfer of Knowledge (IAPP)
- Researcher (IEF)
- Implementation
- Impact

Marie Curie Actions

CRITERIA – weighting (ITN example)

- S&T Quality – 30%
- Training – 20%
- Implementation – 30%
- Impact – 30%

- Example:

$4.2\times0.3+4.7\times0.2+3.8\times0.3+4.4\times0.2=4.22$

Final score $4.22\times20=84.40$ (out of max. 100)

Aim of the study

- To examine the peer-review evaluation process in three MC Actions (ITN, IEF, IAPP)
- To assess the agreement among raters in the different phases of the evaluation workflow

Data sources

- IAPP – from 2007 to 2009 and for 2011 (4 calls)
- ITN – 2008 and from 2010 to 2012 (4 calls)
- IEF – from 2007 to 2013 (7 calls).

- Total:

n=24 897 proposals

n=74 691 individual evaluation reports – reviews

Data sources

- IAPP – from 2007 to 2009 and for 2011 (4 calls)
- ITN – 2008 and from 2010 to 2012 (4 calls)
- IEF – from 2007 to 2013 (7 calls).

- Total:

n=24 897 proposals

n=74 691 individual evaluation reports – reviews

Agreement among reviewers

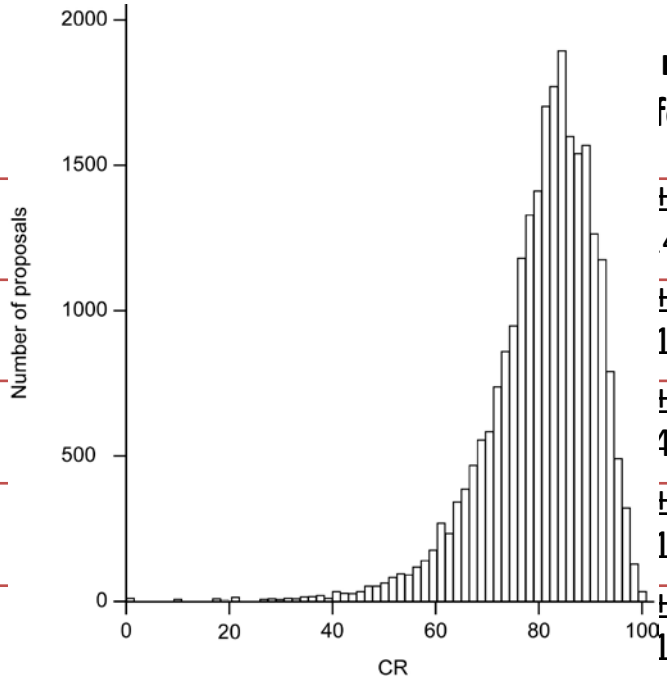
Average Deviation (AD) index

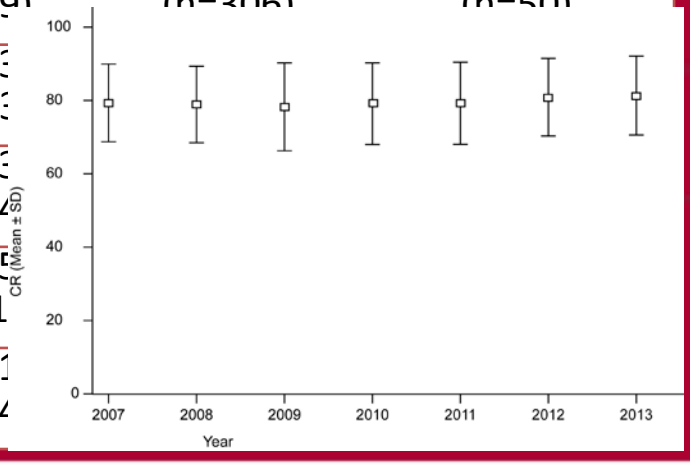
Burke MJ, Finkelstein LM, Dusig MS. On average deviation indices for estimating interrater agreement. Organizational Research Methods. 1999;2: 49-68

- Measure of disagreement that involves determining the average difference between scores of individual raters and the average scores of all raters
- Does not require the specification of null distribution
- Estimates inter-rater disagreement in the units of the original scale

Results

Panel	Mean score (±SD) in proposals where:			Mean score (±SD) in proposals with AVIER vs CR difference
		rater fers	All raters differ	
Total		±13.1 424)	ICC (one-way random) range: 0.46 – 0.64 Overall:ICC=0.67, 95%CI=0.66-0.68	
Chemisty		±13.2 132)		
Economic and Social Sciences/Humanities		±13.1 431)	(n=600)	(n=142)
Information Science/Engineering		±13.7 199)	69.2±12.7 (n=306)	62.7±18.0 (n=50)
Environment		±13 153		
Life Sciences	80.9±10.3 (n=7658)	82.0±9.4 (n=6785)	74.5±13 (n=354	
Mathematics	78.2±10.2 (n=731)	79.6±8.6 (n=623)	71.1±15 (n=41	
Physics	80.8±9.2 (n=2940)	81.6±8.5 (n=2644)	75.3±11 (n=114	

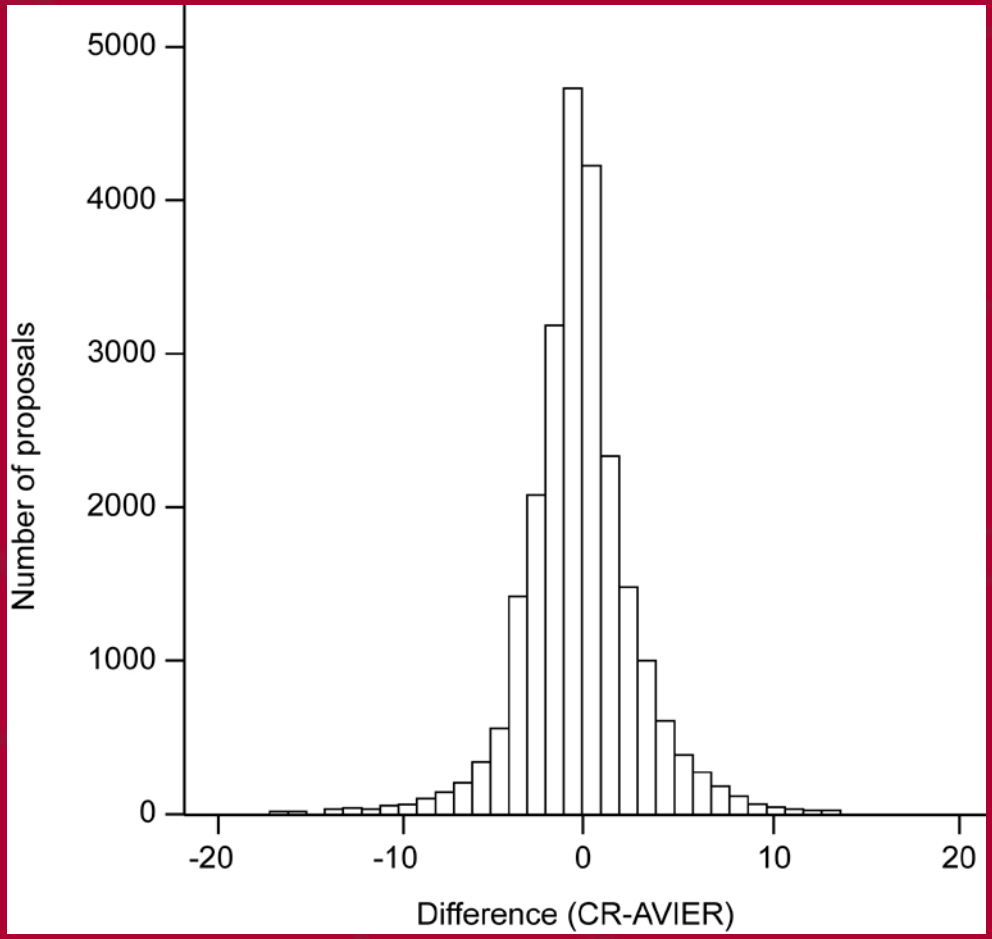




Results

Panel	Disagreement (No. Proposals, row %)		
	One rater differs	All raters differ	AVIER vs CR difference
IAPP (n=759)	71 (9.4%)	124 (16.3%)	23 (3.0%)
ITN (n=3545)	280 (7.9%)	415 (11.7%)	104 (2.9%)
IEF (n=20593)	1073 (5.2%)	1536 (7.5%)	241 (1.2%)

Results

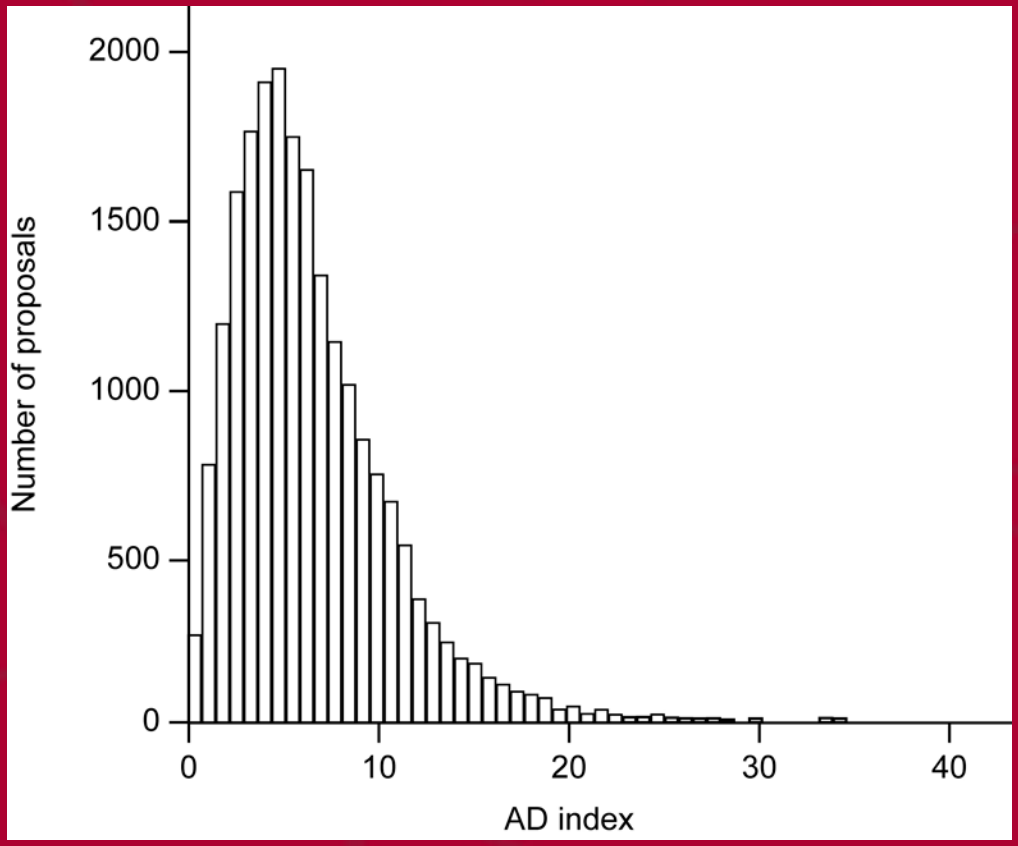


Distribution of differences between Consensus Reports (CR) and average Individual Evaluation Reports (AVIER) scores
Mean = -0.3
SD = 3.19

61.4% of all proposals had less than 2 points difference between AVIER and CR scores

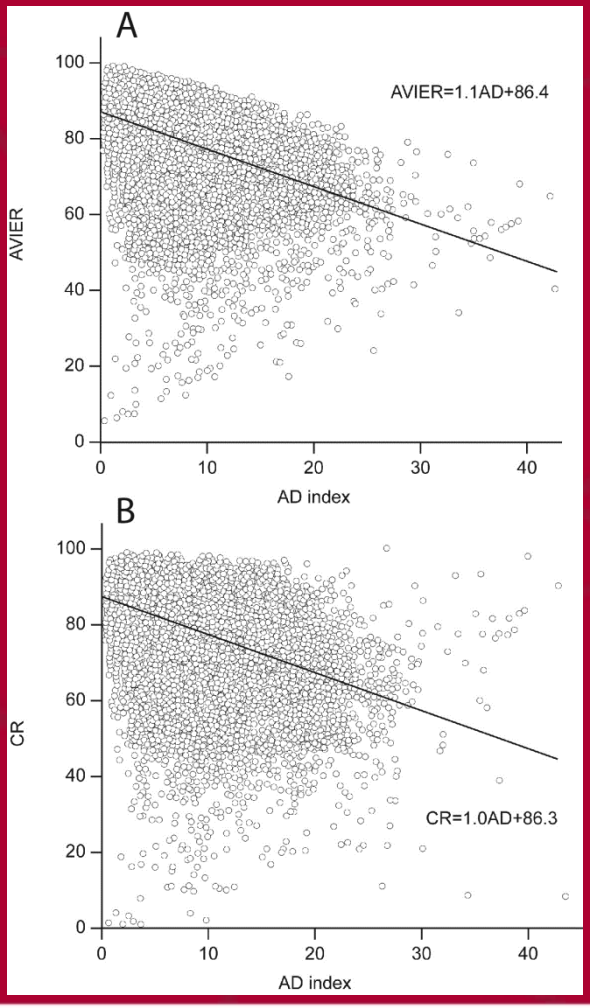
IER – individual evaluation report
AVIER – average IER from remote ev.
CR – consensus report

Results



Overall median AD index = 5.4 points (on a scale 0-100)
For three quarters of all proposals equal or below 8.3 points

Results



More disagreement for proposals with lower scores

IER – individual evaluation report
AVIER – average IER from remote ev.
CR – consensus report
AD – average difference

Results

	No. proposals(row %) with disagreement
Panel (No. proposals)	One rater differs
Chemistry (n=2665)	132 (5.0)
Economic and Social Sciences/Humanities (n=4677)	431 (9.2)
Information Science/Engineering (n=2983)	199 (6.7)
Environment/Geosciences (n=3243)	153 (4.7)
Life Sciences (n=7658)	354 (4.6)
Mathematics (n=731)	41 (5.6)
Physics (n=2940)	114 (3.9)
Total (n=24897)	**1424 (5.7)**

Scenario 1: one rater scores a proposal in a completely different way than the other two raters

a) two agree (difference between their scores less than or equal to 5 points – because 5.4 was the median AD for all proposals)

b) One disagrees for ≥10 points - because this would put the difference above 3rd quartile for all AD indices for IER scores

Results

	No. proposals(row %) with disagreement	
Panel (No. proposals)	One rater differs	All raters differ
Chemistry (n=2665)	132 (5.0)	171 (6.4)
Economic and Social Sciences/Humanities (n=4677)	431 (9.2)	600 (12.8)
Information Science/Engineering (n=2983)	199 (6.7)	306 (10.3)
Environment/Geosciences (n=3243)	153 (4.7)	230 (7.1)
Life Sciences (n=7658)	354 (4.6)	519 (6.8)
Mathematics (n=731)	41 (5.6)	67 (9.2)
Physics (n=2940)	114 (3.9)	182 (6.2)
Total (n=24897)	1424 (5.7)	**2075 (8.3)**

Scenario 3: Disagreement of all three raters

a) difference between eeach pair of IER scores ≥10 points (on a scale 0-100)

Results

	No. proposals(row %) with disagreement		
Panel (No. proposals)	One rater differs	All raters differ	Difference in AVIER vs CR
Chemistry (n=2665)	132 (5.0)	171 (6.4)	32 (1.2)
Economic and Social Sciences/Humanities (n=4677)	431 (9.2)	600 (12.8)	142 (3.0)
Information Science/Engineering (n=2983)	199 (6.7)	306 (10.3)	50 (1.7)
Environment/Geosciences (n=3243)	153 (4.7)	230 (7.1)	42 (1.3)
Life Sciences (n=7658)	354 (4.6)	519 (6.8)	71 (0.9)
Mathematics (n=731)	41 (5.6)	67 (9.2)	5 (0.7)
Physics (n=2940)	114 (3.9)	182 (6.2)	26 (0.9)
Total (n=24897)	1424 (5.7)	2075 (8.3)	368 (1.5)

Scenario 3: absolute difference between CR and AVIER scores ≥10 (scale 0-100)

Positive and negative differences were equally distributed (180 or 48.9% positive and 188 or 51.1% negative differences)

Significantly lower CR scores than other proposals (69.3±19.8 vs 79.8±11.0; $p<0.001$)

Results

Panel (No. proposals)	No. proposals(row %) with disagreement		
	One rater differs	All raters differ	Difference in AVIER vs CR
Chemistry (n=2665)	132 (5.0)	171 (6.4)	32 (1.2)
Economic and Social Sciences/Humanities (n=4677)	**431 (9.2)**	**600 (12.8)**	**142 (3.0)**
Information Science/Engineering (n=2983)	199 (6.7)	306 (10.3)	50 (1.7)
Environment/Geosciences (n=3243)	153 (4.7)	230 (7.1)	42 (1.3)
Life Sciences (n=7658)	354 (4.6)	519 (6.8)	71 (0.9)
Mathematics (n=731)	41 (5.6)	67 (9.2)	5 (0.7)
Physics (n=2940)	114 (3.9)	182 (6.2)	26 (0.9)
Total (n=24897)	1424 (5.7)	2075 (8.3)	368 (1.5)

Scenario 3: absolute difference between CR and AVIER scores ≥10 (scale 0-100)

Positive and negative differences were equally distributed (180 or 48.9% positive and 188 or 51.1% negative differences)

Significantly lower CR scores than other proposals (69.3±19.8 vs 79.8±11.0; $p<0.001$)

Results

Pearson's inter-correlations of IER criteria of different raters

		Rater 1					Rater2					Rater 3				
		S&T quality	Training/ToK	Researcher	Implementation	Impact	S&T quality	Training/ToK	Researcher	Implementation	Impact	S&T quality	Training/ToK	Researcher	Implementation	Impact
Rater 1	S&T quality	1	0.698	0.600	0.668	0.693	0.291	0.279	0.231	0.278	0.274	0.296	0.290	0.231	0.289	0.282
	Training/ToK		1	0.582	0.718	0.740	0.282	0.361	0.248	0.319	0.324	0.270	0.357	0.236	0.324	0.320
	Researcher			1	0.582	0.646	0.217	0.231	0.293	0.230	0.241	0.234	0.246	0.306	0.249	0.251
	Implementation				1	0.740	0.281	0.330	0.247	0.360	0.328	0.282	0.335	0.254	0.367	0.330
	Impact					1	0.278	0.325	0.251	0.318	0.341	0.277	0.327	0.260	0.328	0.341
Rater 2	S&T quality												0.286	0.230	0.285	0.276
	Training/ToK												0.369	0.250	0.335	0.328
	Researcher												0.240	0.294	0.244	0.244
	Implementation												0.332	0.245	0.367	0.330
	Impact												0.322	0.256	0.329	0.342
Rater 3	S&T quality												0.695	0.606	0.665	0.690
	Training/ToK												1	0.589	0.710	0.737
	Researcher													1	0.573	0.645
	Implementation														1	0.733
	Impact															1

Low correlations between different rater's scores for the same criterion and the same proposal
High correlations of the same rater's scores of different criteria for the same proposal

➢ Raters scored proposals in a more holistic way and, generally, assessed each criterion in relation to the other criteria of the same proposal

Results

Principal components analysis with the evaluation criteria – to investigate latent structure that underlies a set of items (criteria scored by three raters)

- Three components, each representing a single rater
- Confirmed our conclusion that criteria scores reflected the rater's global score rather than specific aspects of the proposal.
- The three-component solution explained large portion of variance (73%) and component loadings were very high (all above 0.7).

Conclusions

- Good internal consistency and overall high agreement among expert reviewers
- Disagreement was greater for proposals with lower scores
- At least for some of the proposals, the remote assessments and its average score (AVIER) can provide reliable final judgment of the proposal (especially for IF)

Conclusions

- About 15% of the proposals' population that may need more discussion in order to reach consensus on the final score
- IAPP and ITN calls had a greater number of proposals with disagreements, demonstrating that the evaluation of complex proposals, involving partnerships of several research groups with multidisciplinary and inter-sectorial features, require a more elaborate review procedure