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

PEERE “New Frontiers of Peer Review”
www.peere.org
peereinfo@peere.org
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

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

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

**Research and Innovation Staff Exchange**
International and inter-sector cooperation through the exchange of staff

---

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

IEF → **Action 2 IF**
Experienced Researchers

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

IIF → **IAPP**

---

PEERE “New Frontiers of Peer Review”
www.peere.org
peereinfo@peere.org
Marie Curie Actions

- 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

Budget distribution by scientific panel in FP7

- Life Sciences 27%
- Information Science and Engineering 17%
- Physics 13%
- Environmental and Geo-sciences 11%
- Chemistry 10%
- Social Sciences and Humanities 9%
- Mathematics 3%
- Economics 2%
- COFUND 8%

PEERE “New Frontiers of Peer Review”
www.peere.org
peereinfo@peere.org
3 Individual Assessments 1 Consensus

Proposal A
Proposal B
Proposal C

Remote Central

1 Ranked list

B > A > C

Marie Curie Actions

PEERE "New Frontiers of Peer Review"
www.peere.org
peereinfo@peere.org
<table>
<thead>
<tr>
<th>Rating</th>
<th>Description</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td>Successfully addresses all relevant aspects of the criterion in question. Any shortcomings are minor.</td>
<td>5</td>
</tr>
<tr>
<td>Very Good</td>
<td>Addresses the criterion very well, although certain improvements are still possible.</td>
<td>4</td>
</tr>
<tr>
<td>Good</td>
<td>Addresses the criterion well, although improvements would be necessary.</td>
<td>3</td>
</tr>
<tr>
<td>Fair</td>
<td>Broadly addresses the criterion, there are significant weaknesses.</td>
<td>2</td>
</tr>
<tr>
<td>Poor</td>
<td>Addressed in an inadequate manner, or there are serious inherent weaknesses.</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Fails to address the criterion or cannot be judged due to missing or incomplete information.</td>
<td>0</td>
</tr>
</tbody>
</table>
Marie Curie Actions

CRITERIA

• S&T Quality
• Training (ITN, IEF) or Transfer of Knowledge (IAPP)
• Researcher (IEF)
• Implementation
• Impact
CRITERIA – weighting (ITN example)

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

Example:

\[ 4.2 \times 0.3 + 4.7 \times 0.2 + 3.8 \times 0.3 + 4.4 \times 0.2 = 4.22 \]

Final score \( 4.22 \times 20 = 84.40 \) (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


• 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

### Mean score (±SD) in proposals where:

| Panel                        | Total* | All raters differ | AVIER vs CR
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>rater agree</td>
<td></td>
<td></td>
<td>ICC (one-way random)</td>
</tr>
<tr>
<td>One rater differs</td>
<td></td>
<td></td>
<td>range: 0.46 – 0.64</td>
</tr>
<tr>
<td>All raters differ</td>
<td></td>
<td></td>
<td>Overall: ICC=0.67,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>95% CI=0.66-0.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(n=600)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(n=142)</td>
</tr>
</tbody>
</table>

#### Total

- **Chemistry**
  - Mean score (±SD): 79.8 ±11.0 (n=24897)
  - Mean score (±SD): 81.0 ±10.1 (n=21398)
  - Mean score (±SD): 74.0 ±13.1 (n=1424)
  - Mean score (±SD): 70.9 ±12.8 (n=2075)
  - Mean score (±SD): 69.3 ±19.8 (n=368)

- **Economic and Social Sciences/Humanities**
  - Mean score (±SD): 78.1 ±12.9 (n=4677)
  - Mean score (±SD): 79.8 ±12.4 (n=3646)
  - Mean score (±SD): 74.6 ±13.1 (n=431)
  - Mean score (±SD): 70.7 ±12.9 (n=600)
  - Mean score (±SD): 73.1 ±19.5 (n=142)

- **Information Science/Engineering**
  - Mean score (±SD): 76.9 ±11.9 (n=2983)
  - Mean score (±SD): 78.3 ±11.1 (n=2478)
  - Mean score (±SD): 70.9 ±13.7 (n=199)
  - Mean score (±SD): 69.2 ±12.7 (n=306)
  - Mean score (±SD): 62.7 ±18.0 (n=50)

- **Environment**
  - Mean score (±SD): 80.4 ±10.4 (n=3243)
  - Mean score (±SD): 81.5 ±9.4 (n=2860)
  - Mean score (±SD): 74.5 ±13.3 (n=153)
  - Mean score (±SD): 70.1 ±13.8 (n=230)
  - Mean score (±SD): 66.1 ±20.9 (n=42)

- **Life Sciences**
  - Mean score (±SD): 80.9±10.3 (n=7658)
  - Mean score (±SD): 82.0±9.4 (n=6785)
  - Mean score (±SD): 74.5±13.3 (n=354)

- **Mathematics**
  - Mean score (±SD): 78.2±10.2 (n=731)
  - Mean score (±SD): 79.6±8.6 (n=623)
  - Mean score (±SD): 71.1±15.2 (n=41)

- **Physics**
  - Mean score (±SD): 80.8±9.2 (n=2940)
  - Mean score (±SD): 81.6±8.5 (n=2644)
  - Mean score (±SD): 75.3±11.4 (n=114)

---

**ICC (one-way random)**

- **range**: 0.46 – 0.64
- **Overall**: ICC=0.67,

**95% CI=0.66-0.68**
## Results

<table>
<thead>
<tr>
<th>Panel</th>
<th>One rater differs</th>
<th>All raters differ</th>
<th>AVIER vs CR difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>IAPP (n=759)</td>
<td>71 (9.4%)</td>
<td>124 (16.3%)</td>
<td>23 (3.0%)</td>
</tr>
<tr>
<td>ITN (n=3545)</td>
<td>280 (7.9%)</td>
<td>415 (11.7%)</td>
<td>104 (2.9%)</td>
</tr>
<tr>
<td>IEF (n=20593)</td>
<td>1073 (5.2%)</td>
<td>1536 (7.5%)</td>
<td>241 (1.2%)</td>
</tr>
</tbody>
</table>
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

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

<table>
<thead>
<tr>
<th>Panel (No. proposals)</th>
<th>No. proposals (row %) with disagreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemistry (n=2665)</td>
<td>132 (5.0)</td>
</tr>
<tr>
<td>Economic and Social Sciences/Humanities (n=4677)</td>
<td>431 (9.2)</td>
</tr>
<tr>
<td>Information Science/Engineering (n=2983)</td>
<td>199 (6.7)</td>
</tr>
<tr>
<td>Environment/Geosciences (n=3243)</td>
<td>153 (4.7)</td>
</tr>
<tr>
<td>Life Sciences (n=7658)</td>
<td>354 (4.6)</td>
</tr>
<tr>
<td>Mathematics (n=731)</td>
<td>41 (5.6)</td>
</tr>
<tr>
<td>Physics (n=2940)</td>
<td>114 (3.9)</td>
</tr>
<tr>
<td>Total (n=24897)</td>
<td><strong>1424 (5.7)</strong></td>
</tr>
</tbody>
</table>
Results

Scenario 3: Disagreement of all three raters

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

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

### 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; \(p<0.001\))

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

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)

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

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

<table>
<thead>
<tr>
<th></th>
<th>Rater 1</th>
<th></th>
<th>Rater 2</th>
<th></th>
<th>Rater 3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S&amp;T quality</td>
<td>Training/ToK</td>
<td>Researcher</td>
<td>Implementation</td>
<td>Impact</td>
<td>S&amp;T quality</td>
</tr>
<tr>
<td>S&amp;T quality</td>
<td>1</td>
<td>0.698</td>
<td>0.600</td>
<td>0.668</td>
<td>0.693</td>
<td>0.291</td>
</tr>
<tr>
<td>Training/ToK</td>
<td>1</td>
<td>0.582</td>
<td>0.718</td>
<td>0.740</td>
<td>0.282</td>
<td>0.361</td>
</tr>
<tr>
<td>Researcher</td>
<td>1</td>
<td>0.582</td>
<td>0.646</td>
<td>0.217</td>
<td>0.231</td>
<td>0.293</td>
</tr>
<tr>
<td>Implementation</td>
<td>1</td>
<td>0.740</td>
<td>0.281</td>
<td>0.330</td>
<td>0.247</td>
<td>0.360</td>
</tr>
<tr>
<td>Impact</td>
<td>1</td>
<td>0.278</td>
<td>0.325</td>
<td>0.251</td>
<td>0.318</td>
<td>0.341</td>
</tr>
</tbody>
</table>

### Observations

- **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.

---

**PEERE “New Frontiers of Peer Review”**

[www.peere.org](http://www.peere.org)
[peereinfo@peere.org](mailto:peereinfo@peere.org)
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