EXPERT PREDICTIONS AND ERRORS IN RESEARCH FUNDING DECISIONS

Chiara Franzoni, Massimiliano Guerini, Andola Stanj

Science of Science Funding Meeting, NBER Summer Institute

July 18th, 2024

Reviewers' evaluations

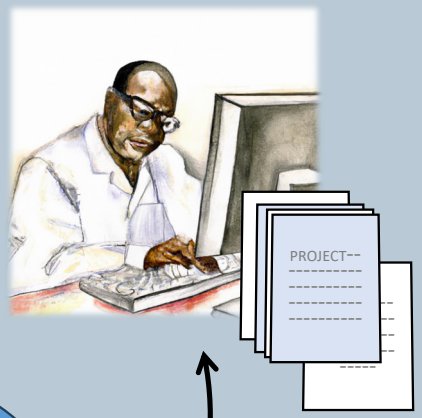
Funding agencies seeking to maximize welfare (social utility function) of limited budget

- Proposals intellectually sophisticated and requiring specialized knowledge
- Appoint a committee of peer reviewers, assuming their domain expertise provides informal advantages
- Heterogeneity regarding domain expertise

RESEARCH QUESTIONS

- Are evaluations predictive of outcomes (generally)? No
- Are experts' evaluations predictive of actual outcomes? Only if reviewer has high domain expertise
- Are experts' evaluations accurate? No, experts are biased in favor of their domain
- What errors do they make? Experts more accurate when they say "no" than when they say "yes"

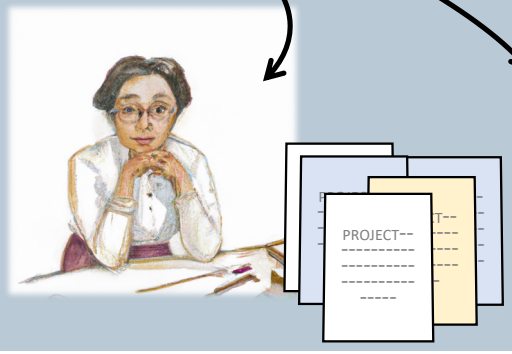
EXPERTS
INDEPENDENT
EVALUATIONS













FUNDING DECISION





Are peer review opinions predictive?

Peer review generally has weak predictive power

Table 1. Summary of articles on the predictive validity of peer-review

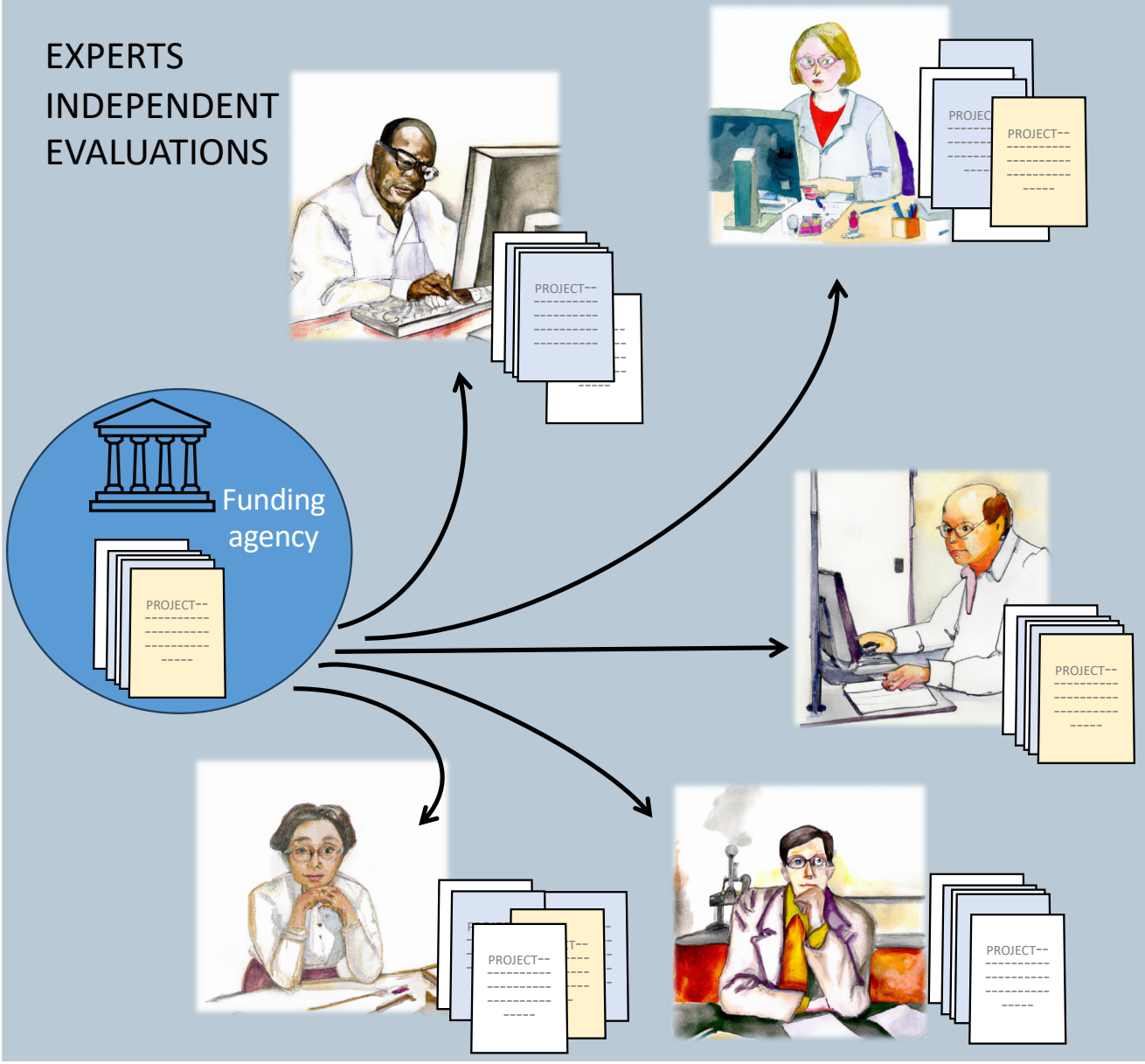
	Reinhart 2009	Danthi et al. 2014	Li and Agha 2015	Fang et al. 2016	Doyle et al. 2015
Sample of applications	Funded and unfunded	Only funded	Only funded	Only funded	Only funded
Data source	ESRC	NHLBI	NIH	NIH	NIMH
Period	1998	2001-2008	1980-2008	1980-2008	2000-2009
Country	Switzerland	US	US	US	US
Field	Biology and Medicine	Medicine	Medicine	Medicine	Medicine
Observations	4,000	1,492	137,215	102,740	1,755
Dependent variable (evaluation)	Mean Score (panel level)	Percentile ranking (panel level)	Percentile Score (panel level)	Percentile score (panel level)	Percentile ranking (panel level)
Outcome variables	Publications Citations	Publications Citations H-index	Publications Citations Patents	Publications Citations	Publications Citations
Correlation	Yes	No	Yes	Yes, but weak	No

Lack empirical evidence on expertise

Reasons:

- Final aggregated scores
- Anonymized
- Non independent (post-discussion)

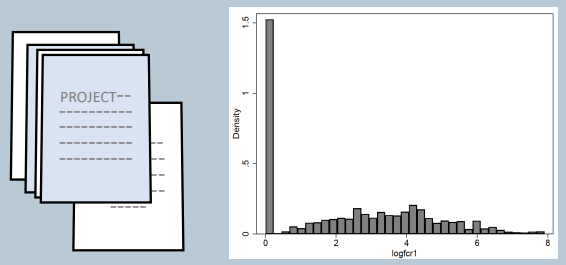
EXPERTS
INDEPENDENT
EVALUATIONS

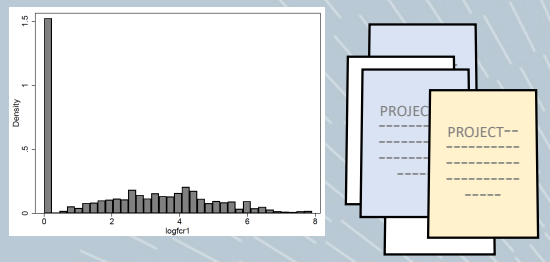


Expertise
Technical and scientific competencies on the specific domain of the proposal under evaluation.

For each reviewer-application pair







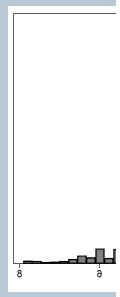
2 MODELS:

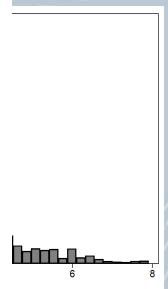
1. **PREDICTIVENESS**. How score predict outcomes

2. **ACCURACY**. Study prediction errors (distance prediction – actual)

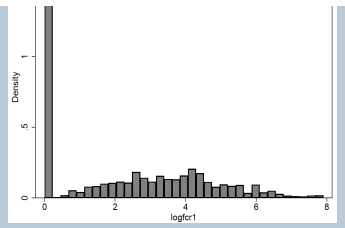
   Moderation effects of REVIEWERS' DOMAIN EXPERTISE











The value

Information processing

Expertise based on long-term memory that can be retrieved at need (Cyert and March 1963; Chase & Simon, 1973; Simon 1978, 1979).

- Recall knowledge in a reliable way (Lord and Maher 1990; Simon 1978)
- Identify/ focus on relevant cues, even if non-salient (Fiske, Kinder, and Larter 1983; McKeithen et al. 1981).
- Chunk problems in subunits that can be solved (Chi, Feltovich, and Glaser 1981a; Ericsson and Kintsch 1995)
- Metacognitive skills. Scrutinize validity of knowledge presented as factual (Alter and Oppenheimer 2007; Pintrich 2002).

..and fallacy of expertise

Social psychology

- Experts affected by cognitive biases (Camerer and Johnson 1991; Cooke 1991; Kahneman and Klein 2009) and generally overconfident
- Influenced by extraneous factors (Danziger, Levav, and Avnaim-Pesso 2011; Dushnitsky and Sarkar 2022; Hirshleifer and Shumway 2003; Kahneman and Klein 2009)

Forecasting

- Experts generally overconfident (Cooke, 1991; Ben-David et al., 2013; Bradley, 1981)
- Superforecasters not identifiable ex-ante (Mellers et al., 2015; Tetlock, 2017; Tetlock and Gardner, 2015)

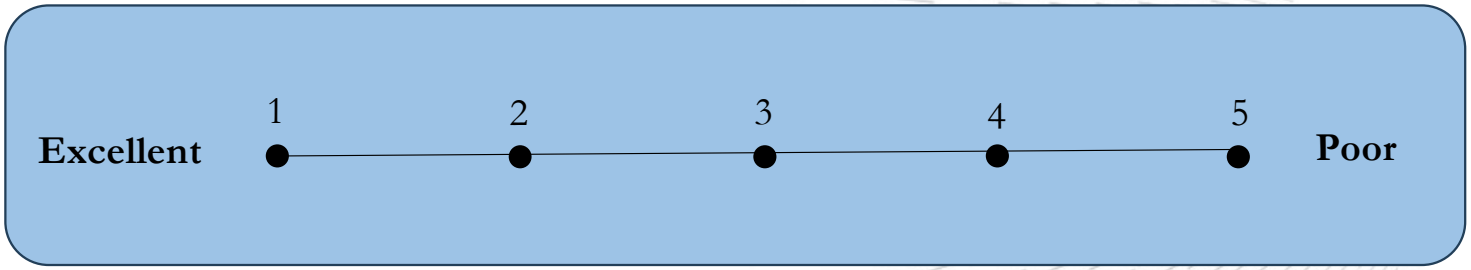
Sample: NNF funding

2012-2018 applications (accepted & rejected) with application content

16,636 evaluations (application-score pairs) of which

5,769 evaluations of application funded

Individual evaluation scores



Publication records of 75 panelists until the year of review

Domain expertise (reviewer-application)

Reviewers' publications ←→ Title+summary of applications
Word embedding (SCIBERT)
Cosine similarity [reviewers's expertise] ←→ [proposal]

3 methodological problems

1. **Measuring post-funding outcomes of research**
   - Only items that reference grant (www.dimensions.ai)
   - 4 different outcomes (min 5y after funding): Publications, citations, FCR and altmetric (influence)

2. **Identification of expertise**
   - Expertise potentially correlated to prediction difficulty.
     E.g., applications in mainstream areas easier to predict and more likely to be scored by expert.
     Or experts may be in mainstream because they are better forecasters.
   - Reviewers fixed effects

3. **Sample-selection bias** (outcomes observable only for funded proposals)
   - 2-stage Heckman-correction. 1st: selection into sample. 2°: outcome

Sample selection / exclusion restriction

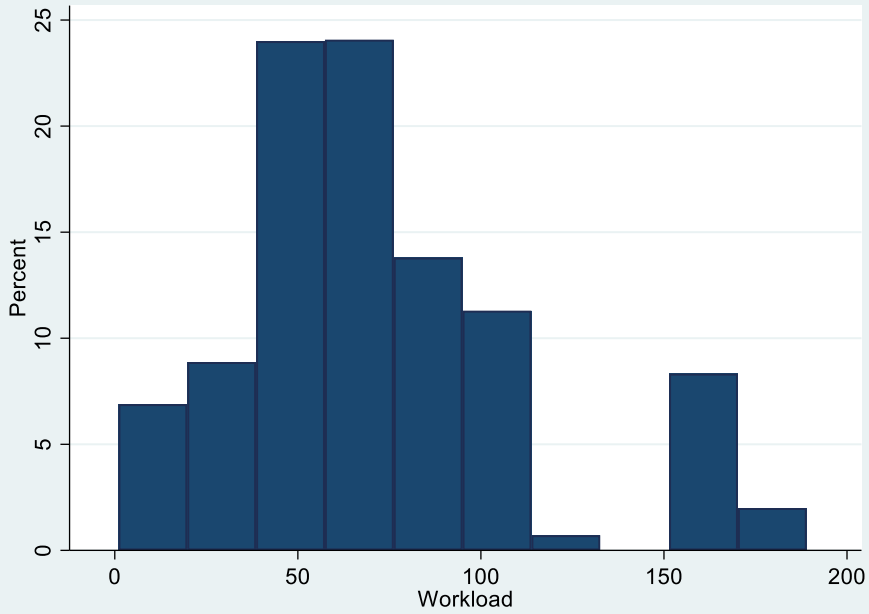
Evaluators = paid panel committee members

Receive applications to score from multiple calls

Limited ability to affect **workload.** Exogenous variation.

Workload = exclusion restriction

Heavier workload makes scores/opinions more noisy, without affecting project outcomes.



Workload:	mean= 72.7
	SD= 41.8
	min= 1
	max= 189

Predictiveness

[SELECTION]

Scores predict selection for funding

Workload predicts selection

Experts' carry more weight

Proposals evaluated by experts more likely to be funded

	(1) Outcome Publications (log)	(2) Outcome Citations (log)	(3) Outcome FCR (log)	(4) Outcome Altmetric (log)	(5) Selection Funded
Quintile score	-0.019	-0.083	-0.039	-0.061	-0.574***
	(0.030)	(0.067)	(0.053)	(0.064)	(0.019)
Expert evaluator	0.191*	0.456*	0.317	0.333	0.238***
	(0.114)	(0.256)	(0.202)	(0.244)	(0.086)
Score x Expert	-0.183***	-0.542***	-0.396***	-0.374***	-0.128***
	(0.061)	(0.136)	(0.108)	(0.130)	(0.037)
Workload					-0.009***
					(0.001)
Constant	-10.545***	-22.516***	-17.919***	-18.724***	-0.104
	(2.119)	(4.729)	(3.747)	(4.520)	(114.407)
PI characteristics	Y	Y	Y	Y	Y
Proposal size	Y	Y	Y	Y	Y
Call competition	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y
Grant type FE	Y	Y	Y	Y	Y
Reviewer FE	Y	Y	Y	Y	Y
N	2,495	2,495	2,495	2,495	16,636
Lambda	0.197**	0.437**	0.256*	0.331*	
	(0.083)	(0.184)	(0.146)	(0.176)	

Note. Heckman two-stage estimation models on post-funding outcome (columns 1-4). Selection equation in column 5. Standard errors are in parentheses. * p<0.10; ** p<0.05; *** p<0.01.

Predictiveness

[SELECTION]

Scores NOT/WEAKLY predictive of outcomes

Lambda confirms need of Heckman correction

Continuous variable and binary variable: (75^ percentile)

Expertise moderates the predictive power of the scores for all outcomes.

	(1) Outcome Publications (log)	(2) Outcome Citations (log)	(3) Outcome FCR (log)	(4) Outcome Altmetric (log)	(5) Selection Funded
Quintile score	-0.019	-0.083	-0.039	-0.061	-0.574***
	(0.030)	(0.067)	(0.053)	(0.064)	(0.019)
Expert evaluator	0.191*	0.456*	0.317	0.333	0.238***
	(0.114)	(0.256)	(0.202)	(0.244)	(0.086)
Score x Expert	-0.183***	-0.542***	-0.396***	-0.374***	-0.128***
	(0.061)	(0.136)	(0.108)	(0.130)	(0.037)
Workload					-0.009***
					(0.001)
Constant	-10.545***	-22.516***	-17.919***	-18.724***	-0.104
	(2.119)	(4.729)	(3.747)	(4.520)	(114.407)
PI characteristics	Y	Y	Y	Y	Y
Proposal size	Y	Y	Y	Y	Y
Call competition	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y
Grant type FE	Y	Y	Y	Y	Y
Reviewer FE	Y	Y	Y	Y	Y
N	2,495	2,495	2,495	2,495	16,636
Lambda	0.197**	0.437**	0.256*	0.331*	
	(0.083)	(0.184)	(0.146)	(0.176)	

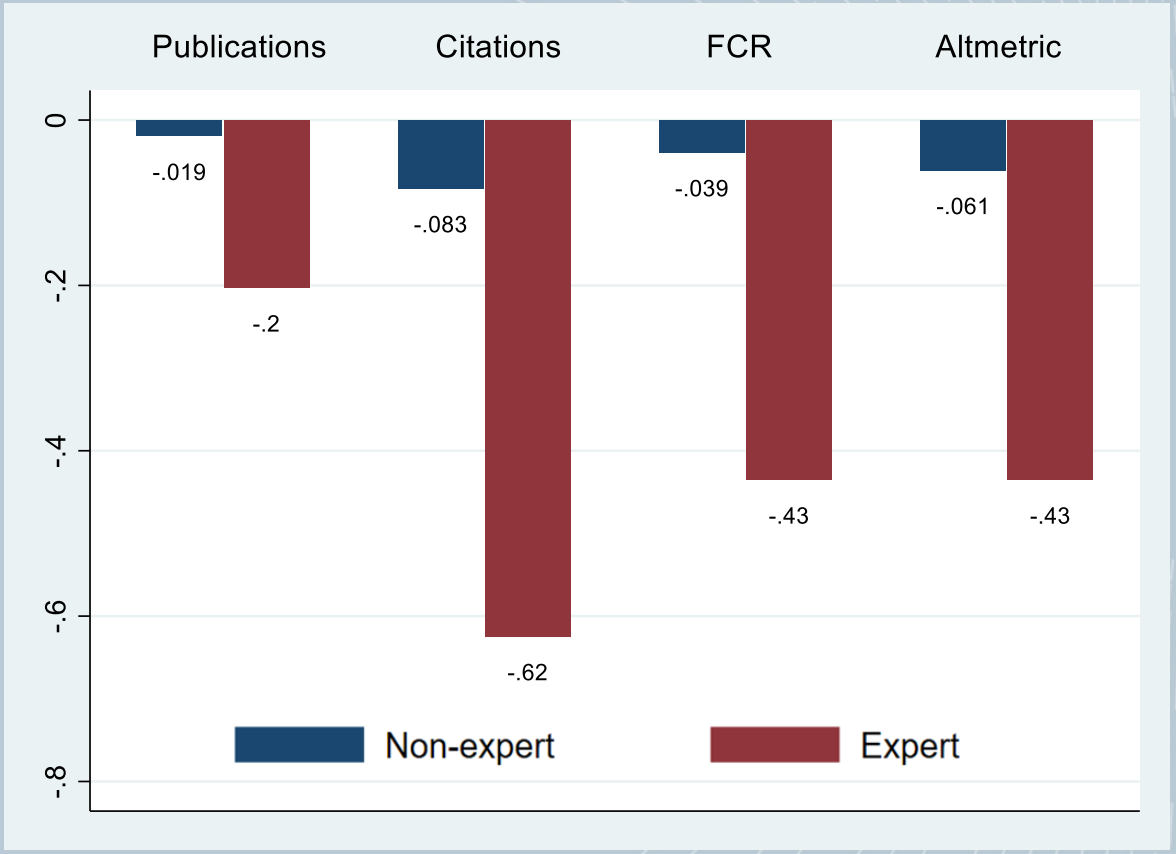
Note. Heckman two-stage estimation models on post-funding outcome (columns 1-4). Selection equation in column 5. Standard errors are in parentheses. * p<0.10; ** p<0.05; *** p<0.01.

Predictiveness
Magnitude

[SELECTION]

1 st.dev. increase in quintile score is associated with 20% fewer publications, 62% fewer citations, 48% lower FCR and 43% lower public influence.

Predictiveness contingent on expertise!



Predictiveness
Top scholars/ scientific prestige

[SELECTION]

Reviewers with prestigius publications are also more influential in funding decisions.

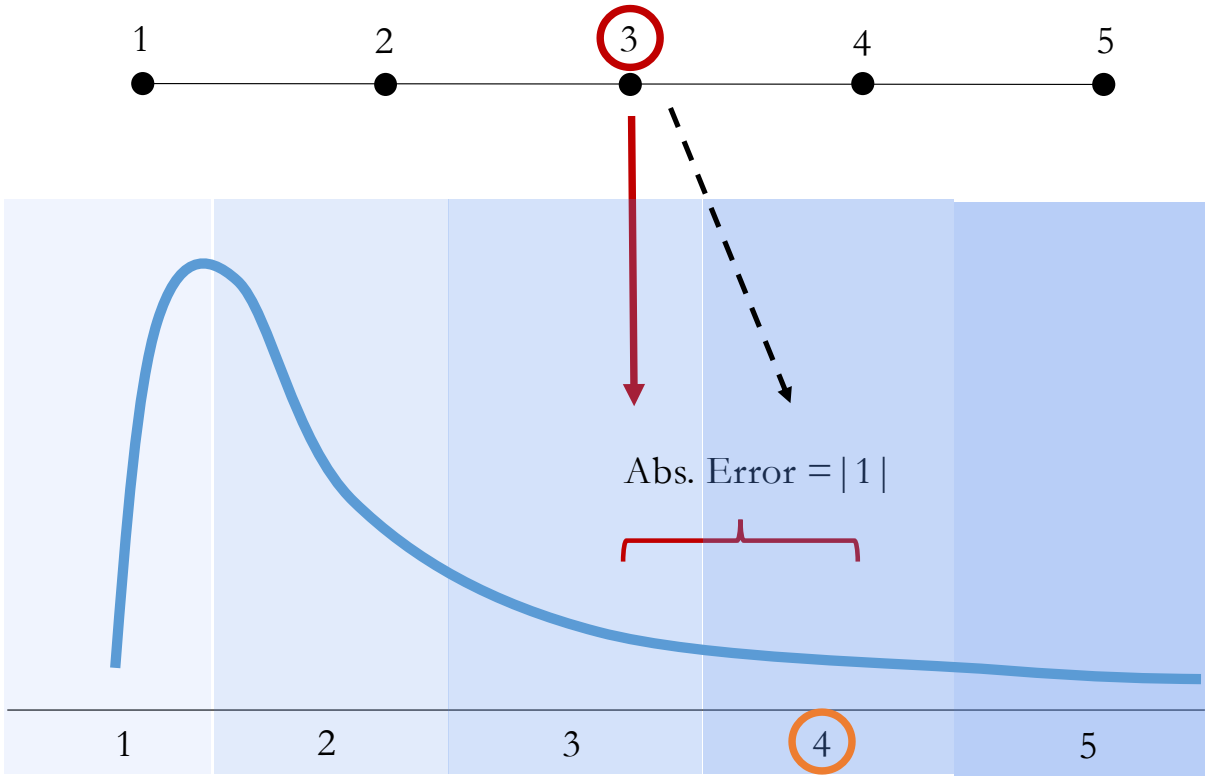
But no predictive power!

RESEARCH PRESTIGE
not a substitute of
EXPERTISE

Table 9. Evaluator's scientific prestige
Panel A – Predictiveness

	(1) Outcome Publications (log)	(2) Outcome Citations (log)	(3) Outcome FCR (log)	(4) Outcome Altmetric (log)	(5) Selection Funded
Quintile score	-0.044	-0.137*	-0.086	-0.103	-0.550***
	(0.032)	(0.072)	(0.057)	(0.069)	(0.021)
Expert evaluator	0.186	0.444*	0.307	0.321	0.237***
	(0.114)	(0.255)	(0.202)	(0.244)	(0.086)
Score x Expert	-0.182***	-0.536***	-0.394***	-0.369***	-0.128***
	(0.061)	(0.136)	(0.108)	(0.130)	(0.037)
Prestige	0.030	-0.072	0.054	-0.023	0.255*
	(0.185)	(0.412)	(0.327)	(0.394)	(0.149)
Score x Prestige	0.081*	0.182*	0.155*	0.144	-0.098**
	(0.045)	(0.100)	(0.079)	(0.095)	(0.038)
Workload					-0.009***
					(0.001)
Constant	-10.710***	-22.745***	-18.221***	-18.933***	-0.491
	(2.125)	(4.743)	(3.757)	(4.535)	(111.851)
Controls	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y
Grant type FE	Y	Y	Y	Y	Y
Reviewer FE	Y	Y	Y	Y	Y
N	2,495	2,495	2,495	2,495	16,636
Lambda	0.202**	0.448**	0.266*	0.334*	
	(0.082)	(0.184)	(0.146)	(0.176)	

Accuracy (prediction errors)

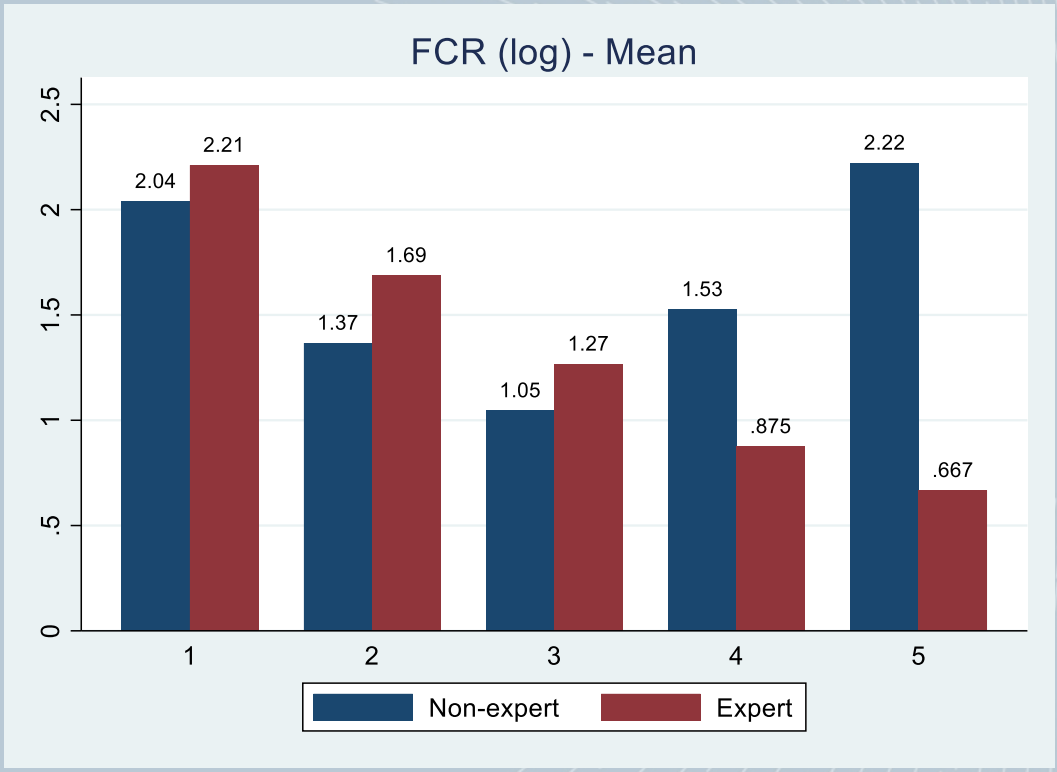


Absolute error =
= | actual_rank$_{jk}$ – pre-funding_rank$_{ji}$ |

Accuracy (prediction errors)

→ Expert's errors are larger for good scores and smaller for bad scores (biased?)

→ Non-experts' errors are volatile (random?)



Accuracy (prediction errors)

Negative moderation effect expertise X score

→ the prediction errors of experts are smaller for worst scores (high score indicate negative opinion).

Domain experts make larger absolute errors than non-experts

	(1) Error Publications (log)	(2) Error Citations (log)	(3) Error FCR (log)	(4) Error Altmetric (log)
Quintile score	0.146***	-0.029	0.022	0.029
	(0.038)	(0.037)	(0.038)	(0.038)
Expert evaluator	0.490***	0.532***	0.478***	0.401***
	(0.143)	(0.141)	(0.144)	(0.145)
Score x Expert	-0.270***	-0.264***	-0.236***	-0.242***
	(0.076)	(0.075)	(0.077)	(0.077)
Constant	11.412***	10.446***	10.865***	10.097***
	(2.654)	(2.614)	(2.663)	(2.684)
PI characteristics	Y	Y	Y	Y
Proposal size	Y	Y	Y	Y
Call competition	Y	Y	Y	Y
Year FE	Y	Y	Y	Y
Grant type FE	Y	Y	Y	Y
Reviewer FE	Y	Y	Y	Y
N	2,495	2,495	2,495	2,495
Lambda	-0.680***	-0.277***	-0.423***	-0.399***
	(0.103)	(0.102)	(0.104)	(0.105)

Note. Heckman two-stage estimation models on prediction errors (columns 1-4). Standard errors are in parentheses. * p<0.10; ** p<0.05; *** p<0.01.

Robustness

- OLS instead of the Heckman two-stage models.
- Poisson or Negative Binomial to model outcome variables
- Different thresholds of cosine similarity for *expert evaluators* (70^{th}, 80^{th} and 90^{th} perc. cosine similarity.
- Add supplementary control variables, such as the number of evaluators per application, or evaluators' research productivity, measured by their publication counts before the application date.
- Quadratic error (Davis-Stober et al. 2014), instead of the absolute error, in models of accuracy

Conclusions

- Predictive validity of evaluations contingent on reviewers' expertise.
  - Specialized domain expertise needed (>=75^ percentile)
  - Experts more influential in funding decisions: expertise carries weight
  - Scientific excellence not a substitute for expertise: prestigious scholars more influential, but not predictive

- Experts not very accurate: make large prediction errors
  - Errors not randomly-distributed across the scale (biased)
  - Experts are accurate when giving negative evaluations
  - Inaccurate when giving positive evaluations

- Mechanisms?
  - No evidence that experts prioritize more novel, more volatile projects (if anything they are conservative)
  - Overestimate gains in their domain (in good faith)?
  - Confirmation bias?
  - Strategic behaviour?



Feedback welcome!

chiara.franzoni@polimi.it