METASCIENCE
2021 CONFERENCE

Does a zero inter-rater reliability mean grant peer review is arbitrary?

Patrícia Martinková
Institute of Computer Science, of the Czech Academy of Sciences
www.cs.cas.cz/martinkova

COS CENTER FOR OPEN SCIENCE

RoRI RESEARCH ON RESEARCH INSTITUTE

aimos

Motivation

- Grant peer review allocates billions of dollars of research funding
- Most selection mechanisms rely on peer review for assessing the quality and potential impact of proposed research

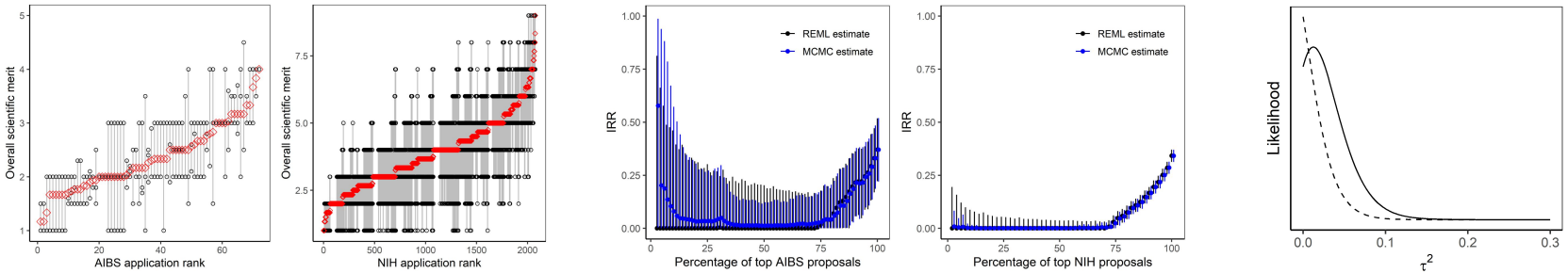
Inter-rater reliability (IRR):

- Measure of consistency among raters (>0,6 good, 0.35-0.6 fair, <0.3 low)
- Pier et al. (2018) found **zero** IRR in mock peer review of funded NIH proposals
- Does a zero IRR mean **grant peer review is arbitrary?**

Erosheva, E., Martinková, P., & Lee, C. J. (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, 184*(3), 904-919, doi:10.1111/rssa.12681

Does a zero IRR mean grant peer review is arbitrary?

- Real data from a complete range of submissions to National Institutes of Health (NIH) and American Institute of Biological Sciences (AIBS)
  - Multi-rater IRR of about 0.3 for complete range of samples (0.6 for average rating of 3 raters)
  - Estimating local IRR from subsets of restricted-quality proposals will likely result in zero estimates under many scenarios
  - Zero estimates of IRR are possible even when the true value is not zero



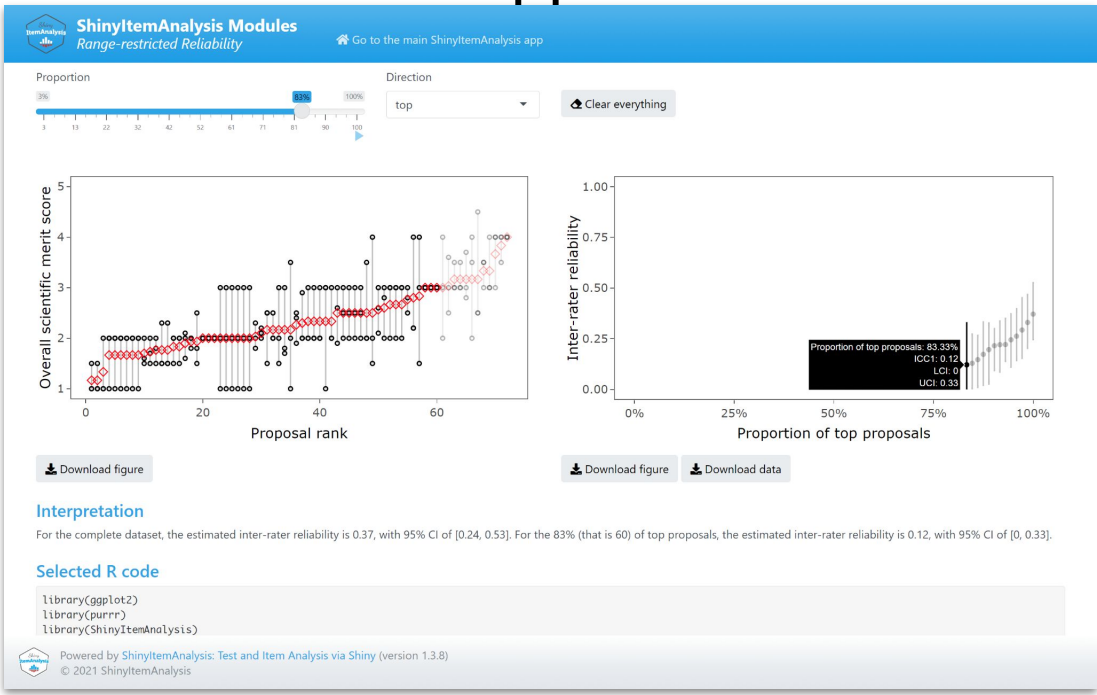
Erosheva, E., Martinková, P., & Lee, C. J. (2021). When zero may not be zero: A cautionary note on the use of inter-rater reliability in evaluating grant peer review. *Journal of the Royal Statistical Society — Series A, 184*(3), 904-919, doi:10.1111/rssa.12681

Software implementation and interactive app

- ShinyItemAnalysis R package
- **`ICCrestricted()`** function

Module on Range-restricted IRR:

- AIBS data
- Top & bottom restriction
- Restricted ratings (left figure)
- IRR estimates (right figure)
- Interpretation of results
- Sample R code



Martinková, P., & Drabinová, A. (2018) ShinyItemAnalysis for teaching psychometrics and to enforce routine analysis of educational tests. *The R Journal, 10*(2), 503—515, doi:10.32614/RJ-2018-074. See www.ShinyItemAnalysis.org

Discussion and Conclusion

- Zero IRR estimates are quite plausible in restricted-range scenarios
- **Question:** Is it valid to interpret range-restricted IRR estimates as indicators of peer review quality when reviewers are asked to score grant proposals across the whole range of submissions? **Answer:** Not from the measurement standpoint.

- Interactive SIA modules to support dissemination, replicability, open science

Thank you for your attention!

www.cs.cas.cz/martinkova

@PMartinkova

References

- Erosheva, E., Martinková, P., & Lee, C. J. (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, 184*(3), 904-919, doi:10.1111/rssa.12681
- Martinková, P., & Drabinová, A. (2018) ShinyItemAnalysis for teaching psychometrics and to enforce routine analysis of educational tests. *The R Journal, 10*(2), 503—515, doi:10.32614/RJ-2018-074. See www.ShinyItemAnalysis.org
- Pier, E. L., et al. (2018). Low agreement among reviewers evaluating the same NIH grant applications. *Proceedings of the National Academy of Sciences, 115*(12), 2952-2957.
- ShinyItemAnalysis module on Restricted-range IRR, https://shiny.cs.cas.cz/ShinyItemAnalysis-module-IRRrestricted/

Acknowledgements

- Elena Erosheva & Carole Lee (University of Washington)
- František Bartoš, Jan Netík (COMputational PSychometrics group, ICS CAS)
- Czech Science Foundation, Grant Number 21-03658S
- National Science Foundation, Grant Number 1759825
    - Disclaimer: Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the Czech Science Foundation.