Developing and validating a method of coherence-based judgment aggregation

Emily H Ho[1] and David Budescu[2]

[1] Northwestern University, Department of Medical Social Sciences [2] Fordham University, Department of Psychology

INTRODUCTION

- Forecasting, or the prediction of future events, is often evaluated by **correspondence**, the extent to which judgments are accurate, and **coherence,** the extent to which judgments follow logical and probabilistic axioms (Hammond, 1996)
- Example of coherence: unitarity (probabilities of mutually exclusive and exhaustive events add up to 1)
- Example of correspondence: correctly predicting the outcome of the 2020 Georgia's run-off Senate race
- Recent research has suggested there is a link between these two concepts
  - A global forecasting tournament finds that those who are highly accurate also tend to score higher on coherence measures (Mellers et al., 2018)
  - The 'wisdom of the select crowd' suggests forecasting accuracy can be improved by aggregating only a select few (Mannes, Soll, and Larrick, 2014)
  - Statistically coherentizing judgments makes them more accurate (Karvetski et al., 2013)
  - Despite coherence being central to accuracy, there is no unified measure of construct

RESEARCH QUESTION AND METHODS

Aim 1: psychometrically validate a measure of coherence (CFS; Coherence Forecasting Scale)

1) Develop a scale that measures five features of coherence: Binary probabilities, trinary probabilities, time horizon, spatial distance, and probability intervals
2) Used a new method of Automatic Item Generation (AIG) to design multiple forms measuring same construct

Aim 2: Use individual coherence weights as a new, empirically derived weight for judgment aggregation from a forecasting platform, Good Judgment Open (gjopen.com)

RESULTS

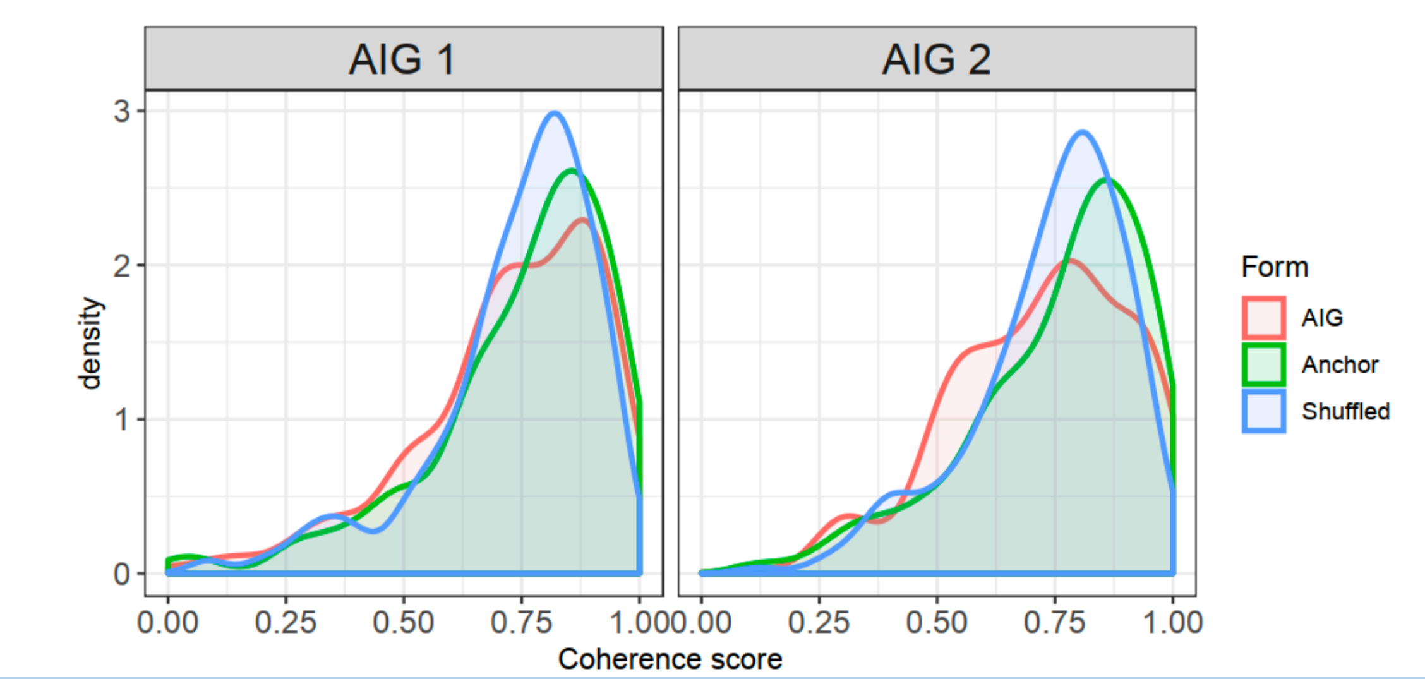
Aim 1: Creating a coherence measure

Psychometric statistics

- Created two sets of coherence items (Anchor, AIG forms)
- The coherence measure resulted in, for each form, five scores measuring knowledge of (1) binary probabilities, (2) trinary probabilities, (3-5) probability with respect to time horizon, spatial distance, and probability intervals.
- Cronbach's alpha = ; Test-retest reliability across the three coherence scale forms, after a week lag, ranged from 0.66-0.76
- CFS was related to active open-minded thinking and cognitive reflection

Feasibility of Automatic Item Generation

- To determine the interchangeability of two forms, I created, for each participant, a set of hybrid forms
- Compared *M* and *SD* of estimated hybrid score compared with individual's actual anchor and AIG score
- Each participant completed $k = 2$ forms of $p = 5$ items, Each individual had $2^5 = 32$ form profiles. Density of scores looks similar across three forms

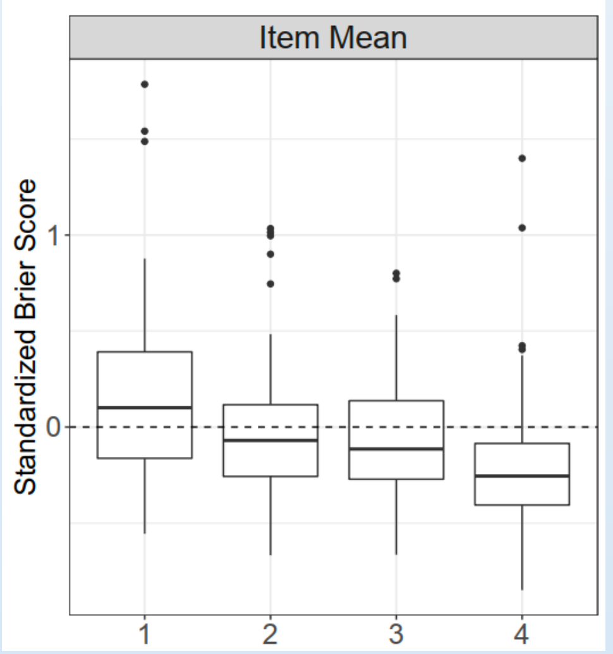


Aim 2: Validating coherence measure aggregation weights

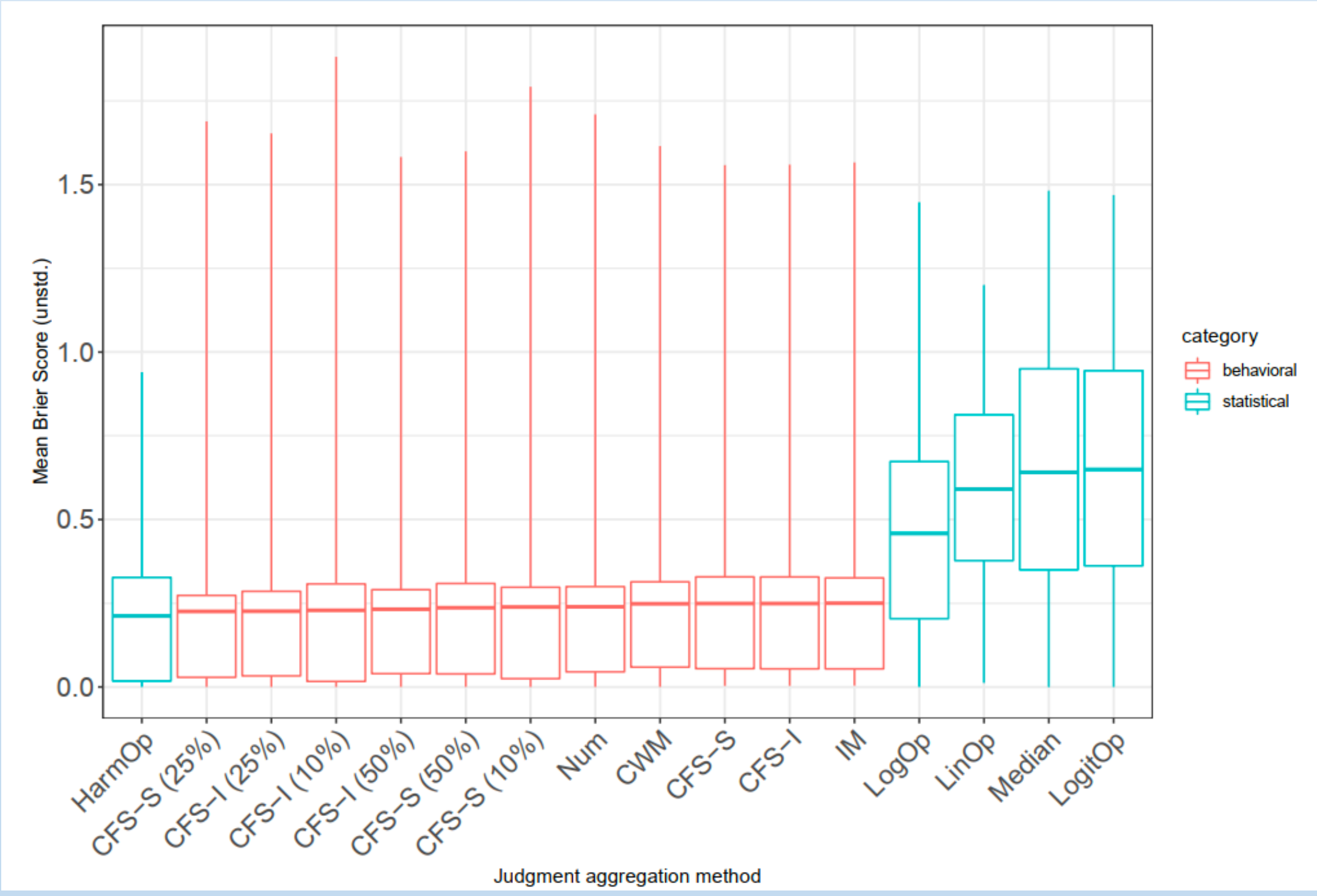
- Compared accuracy against a variety of statistical-based aggregation methods (linear, logarithmic, harmonic, logit mean, and median) and behavioral methods (coherence measure, incoherence metric and contributed-weighted scoring, and numeracy scores)
- Accuracy was calculated using the Brier score and the multinomial form of the Brier score for more than two categories

RESULTS (cont'd)

- Study 2: Survey links were completed by 243 Good Judgment Open forecasters (Age M = 50.8, SD = 15, 83% Male, Mean CFS score (M = 0.88, SD = 0.12)
- Correlation between coherence scores and accuracy was r = -0.41, higher than numeracy and cognitive reflection



Quartile of ooherence scores



CFS scoring taking a subset of the highly coherent was the highest performing behavioral method, and second highest overall

CONCLUSIONS

- A coherence measure using a new psychometric framework of Automatic Item Generation yields similar scores
- Coherence varies systematically across individuals and can be used as an empirical weight to procure more accurate judgment aggregates
- Correlation between CFS and accuracy is high relative to existing estimates in the literature

ACKNOWLEDGEMENTS

This work was supported by a National Science Foundation Dissertation Improvement Grant, #1919055. I gratefully acknowledge the assistance of Eva Chen, Luis Enrique Urturbey, and Philip Rescobar at Good Judgment Open for allowing me access to the forecasters on which Study 2 is based. Finally, I thank my external reader for this dissertation, David R Mandel, for his many helpful comments.

Contact: Emily-ho@northwestern.edu

zoom

https://fordham.zoom.us/j/87015639640