Predicting Social Science Results

Daniel Evans — Bonn
Séverine Toussaert — Oxford
Taisuke Imai — Osaka

Introduction

- Hi! Nice to see you all!
- Today I will present a project on forecasting research results.
- Part of a larger enterprise to bring together two fields I love.

Introduction

- Hi! Nice to see you all!
- Today I will present a project on forecasting research results.
- Part of a larger enterprise to bring together two fields I love.

Behavioral science

Introduction

- Hi! Nice to see you all!
- Today I will present a project on forecasting research results.
- Part of a larger enterprise to bring together two fields I love.

Behavioral science Metascience

Introduction

- Hi! Nice to see you all!
- Today I will present a project on forecasting research results.
- Part of a larger enterprise to bring together two fields I love.

$$\text{Behavioral science} \quad \rightleftarrows \quad \text{Metascience}$$

- Dream: make them communicate to push the research frontier.

Introduction

- Hi! Nice to see you all!
- Today I will present a project on forecasting research results.
- Part of a larger enterprise to bring together two fields I love.

$$\text{Behavioral science} \quad \rightleftarrows \quad \text{Metascience}$$

- Dream: make them communicate to push the research frontier.
- Now developing an incubator for scientific research called Lab^2.

Missions of Lab^2

1. Enable experimentation at scale with many researchers and labs:
   - Replications
   - Multi-analyst studies
   - RCTs on research practices

Missions of Lab^2

1. Enable experimentation at scale with many researchers and labs:
    - Replications
    - Multi-analyst studies
    - RCTs on research practices
2. Document the life of scientific projects from A and Z
    - Combine metadata with longitudinal surveys
    - Better understand the production process of research

$\Rightarrow$ Bring crowdscience to econ and make (crowd)science less "black box".

Missions of Lab^2

1. Enable experimentation at scale with many researchers and labs:
   - Replications
   - Multi-analyst studies
   - RCTs on research practices
2. Document the life of scientific projects from A and Z
   - Combine metadata with longitudinal surveys
   - Better understand the production process of research

$\Rightarrow$ Bring crowdscience to econ and make (crowd)science less “black box”.

Fun team

Aurélien Baillon

Anna Dreber

Taisuke Imai

Magnus Johannesson

Levent Neyse

Sev Toussaert

Fun team


Aurélien Baillon


Anna Dreber


Taisuke Imai


Magnus Johannesson


Levent Neyse


Sev Toussaert

- Talav Bhimnathvala
- Raffaele Blasone
- Giulia Caprini
- **Daniel Evans**
- Avenia Ghazarian
- Adam Gill
- Vatsal Khandelwal
- Anna Popova
- Hubert Wu
- Podcast team...

Story behind the forecasting project

Sep 2020 (?) Anna Dreber hired Daniel as an RA to help on a project on peer review. Daniel eventually became a co-author.

Story behind the forecasting project

Sep 2020 (?) Anna Dreber hired Daniel as an RA to help on a project on peer review. Daniel eventually became a co-author.

Fall 2022 Taisuke pitched the project to Sev. Daniel offered to join.

Story behind the forecasting project

Sep 2020 (?) Anna Dreber hired Daniel as an RA to help on a project on peer review. Daniel eventually became a co-author.

Fall 2022 Taisuke pitched the project to Sev. Daniel offered to join.

Oct 2022 Daniel started to read. A LOT.

Story behind the forecasting project

Sep 2020 (?)	Anna Dreber hired Daniel as an RA to help on a project on peer review. Daniel eventually became a co-author.
Fall 2022	Taisuke pitched the project to Sev. Daniel offered to join.
Oct 2022	Daniel started to read. A LOT.
May 2023	First presentation by Taisuke in Berlin. Roadmap discussion.

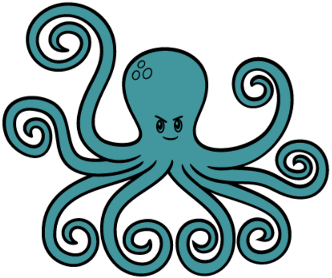
Story behind the forecasting project

Sep 2020 (?)	Anna Dreber hired Daniel as an RA to help on a project on peer review. Daniel eventually became a co-author.
Fall 2022	Taisuke pitched the project to Sev. Daniel offered to join.
Oct 2022	Daniel started to read. A LOT.
May 2023	First presentation by Taisuke in Berlin. Roadmap discussion.
Jun - Dec 2023	More presentations by Daniel and Sev. Overleaf doc growing.

Story behind the forecasting project

Sep 2020 (?)	Anna Dreber hired Daniel as an RA to help on a project on peer review. Daniel eventually became a co-author.
Fall 2022	Taisuke pitched the project to Sev. Daniel offered to join.
Oct 2022	Daniel started to read. A LOT.
May 2023	First presentation by Taisuke in Berlin. Roadmap discussion.
Jun - Dec 2023	More presentations by Daniel and Sev. Overleaf doc growing.
Today	Impromptu presentation by Sev. VERY PRELIMINARY.

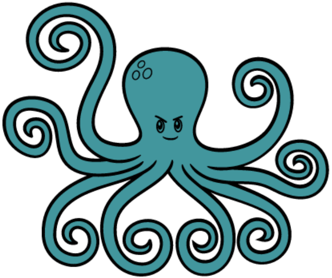
Where the story is heading next (?)



Octopus growing many arms:

- Unclear how many arms we will keep
- Will present our plan and attempts

Where the story is heading next (?)



Octopus growing many arms:

- Unclear how many arms we will keep
- Will present our plan and attempts

What I'd love to hear from you:

- Which arms you would kill
- Which arms you would grow

Motivation

- Importance of beliefs about research results:

Motivation

- Importance of beliefs about research results:
    - Billions of dollars and hours spent on research yearly.
    - Researchers use beliefs to choose projects, give advice, evaluate manuscripts.

Motivation

- Importance of beliefs about research results:
  - Billions of dollars and hours spent on research yearly.
  - Researchers use beliefs to choose projects, give advice, evaluate manuscripts.
- Usually remain implicit, but increasingly common to elicit directly.

Motivation

- Importance of beliefs about research results:
  - Billions of dollars and hours spent on research yearly.
  - Researchers use beliefs to choose projects, give advice, evaluate manuscripts.
- Usually remain implicit, but increasingly common to elicit directly.
- But ...
  - Despite stakes, overall accuracy and informativeness remain unknown.
  - Returns to direct elicitation are unclear.

Motivation

- Importance of beliefs about research results:
  - Billions of dollars and hours spent on research yearly.
  - Researchers use beliefs to choose projects, give advice, evaluate manuscripts.
- Usually remain implicit, but increasingly common to elicit directly.
- But ...
  - Despite stakes, overall accuracy and informativeness remain unknown.
  - Returns to direct elicitation are unclear.
  - Good time for a comprehensive overview

What we do

- Investigate the origins and history of forecasting
  - ⇝ narrative review

What we do

- Investigate the origins and history of forecasting
  - $\rightsquigarrow$ narrative review
- Document current practices and forecast performance
  - $\rightsquigarrow$ systematic review / meta-analysis

What we do

- Investigate the origins and history of forecasting
  - narrative review
- Document current practices and forecast performance
  - systematic review / meta-analysis
- Discuss possible paths forward

What we do

- Investigate the origins and history of forecasting
  - $\leadsto$ narrative review
- Document current practices and forecast performance
  - $\leadsto$ systematic review / meta-analysis
- Discuss possible paths forward

Work in progress $\leadsto$ comments welcome ☺

Context for the project

Civic honesty around the globe Cohn et al. (2019) *Science*

- "Lost" wallets were given to strangers around the world

Context for the project

Civic honesty around the globe Cohn et al. (2019) *Science*

- "Lost" wallets were given to strangers around the world

? What percent of strangers would attempt to return a wallet

Condition	No Money	Money ($13)	Big Money ($94)
Economists' prediction			
Actual return rate			

Context for the project

Civic honesty around the globe Cohn et al. (2019) *Science*

- "Lost" wallets were given to strangers around the world
- ? What percent of strangers would attempt to return a wallet

Condition	No Money	Money ($13)	Big Money ($94)
Economists' prediction	69%	69%	69%
Actual return rate			

Context for the project

Civic honesty around the globe Cohn et al. (2019) *Science*

- "Lost" wallets were given to strangers around the world
- ? What percent of strangers would attempt to return a wallet

Condition	No Money	Money ($13)	Big Money ($94)
Economists' prediction	69%	69%	69%
Actual return rate	39%	57%	66%

Research questions

1. Who participates in the “market” for predictions of research results?

Research questions

1. Who participates in the “market” for predictions of research results?
2. Why do researchers collect predictions of research results?

Research questions

1. Who participates in the “market” for predictions of research results?
2. Why do researchers collect predictions of research results?
3. How are forecasts elicited?

Research questions

1. Who participates in the “market” for predictions of research results?
2. Why do researchers collect predictions of research results?
3. How are forecasts elicited?
4. (When) Are predictions accurate and informative?

(Very short) literature review

- Earliest example $\rightsquigarrow$ "Milgram experiments" Milgram (1963)

  "[predictions] provide us a benchmark from which to see how much or little we learn through the experiment" *Milgram (1974)*

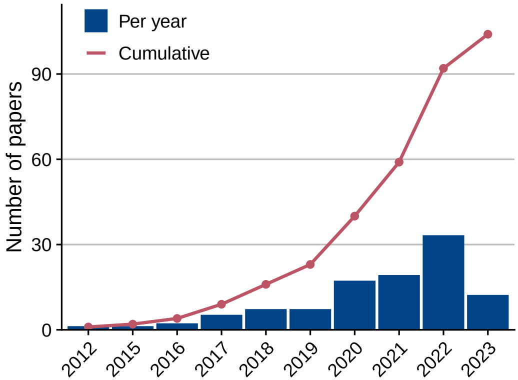
(Very short) literature review

- Earliest example $\leadsto$ "Milgram experiments" Milgram (1963)

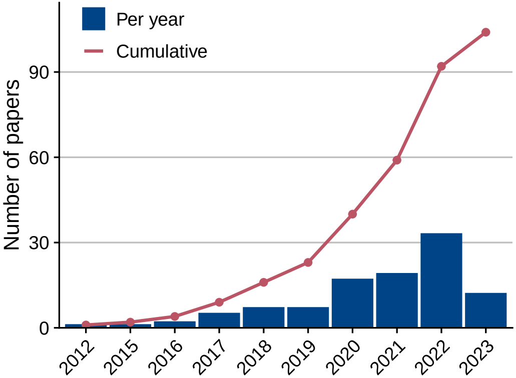
  "[predictions] provide us a benchmark from which to see how much or little we learn through the experiment" Milgram (1974)

- Difficult to obtain raw data and contact authors from old papers
- Focus efforts on more recent literature

(Very short) literature review

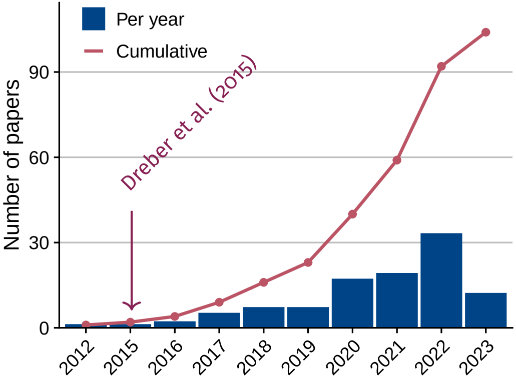


(Very short) literature review



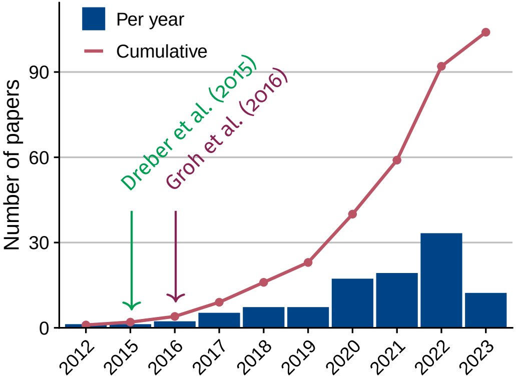
1. Growth trend
   - cutoff $\rightsquigarrow$ Aug 2023

(Very short) literature review



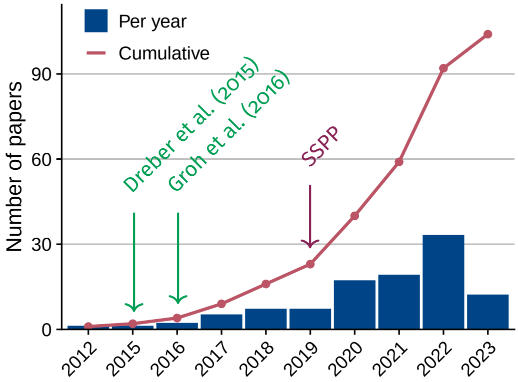
1. Growth trend
   - cutoff $\rightsquigarrow$ Aug 2023
2. Early forecasts of
   - replication outcomes Dreber et al. (2015)

(Very short) literature review



1. Growth trend
   - cutoff $\rightsquigarrow$ Aug 2023
2. Early forecasts of
   - replication outcomes Dreber et al. (2015)
   - development RCTs Groh et al. (2016)

(Very short) literature review



1. Growth trend
   - cutoff $\rightsquigarrow$ Aug 2023
2. Early forecasts of
   - replication outcomes Dreber et al. (2015)
   - development RCTs Groh et al. (2016)
3. Centralization
   - Social Science Prediction Platform DellaVigna et al. (2019)

Prediction markets on replications

One central hypothesis for each study

Will the replication result be an effect in the same direction as the original study with $p < 0.05$? Yes/No

- Participants trade contracts paying \$1 if study is replicated (\$0 o.w.).
- Prices start at \$0.50. Each participant receives \$50-100 endowment.
- Both long- and short-selling allowed
- Logarithmic scoring rule implemented by market maker.
- Price $\approx$ predicted prob. of outcome occurring (need risk neutrality)

Replication market for Camerer et al. (2016)

Market	Price	Shares Held	Investment Value	Trade
de Clippel et al. (AER 2014)	**0.76**	0.00	0.00	Trade
Duffy and Puzzello (AER 2014)	**0.81**	0.00	0.00	Trade
Dulleck et al. (AER 2011)	**0.74**	0.00	0.00	Trade
Fehr et al. (AER 2013)	**0.63**	0.00	0.00	Trade
Friedman and Oprea (AER 2012)	**0.83**	0.00	0.00	Trade
Fudenberg et al. (AER 2012)	**0.93**	0.00	0.00	Trade
Huck et al. (AER 2011)	**0.92**	0.00	0.00	Trade
Ifcher and Zarghamee (AER 2011)	**0.59**	0.00	0.00	Trade
Kessler and Roth (AER 2012)	**0.94**	0.00	0.00	Trade
Kirchler et al (AER 2012)	**0.71**	0.00	0.00	Trade
Kogan et al. (AER 2011)	**0.80**	0.00	0.00	Trade
Kuziemko et al. (QJE 2014)	**0.63**	0.00	0.00	Trade
Marzilli Ericson and Fuster (QJE 2011)	**0.62**	0.00	0.00	Trade

Replication market for Camerer et al. (2016)

Abeler et al. (AER 2011) ×

Hypothesis to bet on: Subjects exert more effort (leading to higher earnings) in a real effort task if the expectations-based reference point is increased (a comparison of the average accumulated earnings in the real effort task between the LO treatment and the HI treatment).

Price | Positions

*All times are in CET.

Tokens	100.00

Price	Shares Held	Investment Value in Tokens	Trade
0.64	0.00	0.00	Increase position by 0 tokens OK
			Decrease position by 0 tokens OK

Social Science Prediction Platform (SSPP)

SSPP ©DellaVigna and Vivalt 2019

https://socialscienceprediction.org/

Social Science Prediction Platform (SSPP)

SSPP ©DellaVigna and Vivalt 2019

https://socialscienceprediction.org/

Public Prediction Bulletin

Open Surveys

Long-run general equilbrium effects of cash transfers in Kenya ($)	Authors David Bernard, Dennis Egger, Edward Miguel, Johannes Haushofer, Michael Walker	Field Development Economics	Close Date May 1, 2023	View Details
Long-run impacts of boarding school in France ($)	Authors David Bernard, Luc Behaghel, Clément de Chaisemartin, Marc Gurgand	Field Economics Of Education	Close Date May 1, 2023	View Details
Long-run impacts of mother tongue instruction in Uganda ($)	Authors David Bernard, Julie Buhl-Wiggers, Jason Kerwin, Ricardo Montero de la Piedra, Jeffrey Smith, Rebecca L. Thornton	Field Development Economics, Economics Of Education	Close Date May 1, 2023	View Details
Long-run impacts of a Graduation program in Afghanistan ($)	Authors David Bernard, Yulia Belyajova, Aidan Coville, Guadalupe Bedoya, Thomas Escande	Field Development Economics	Close Date May 1, 2023	View Details
Long-run impacts of social signalling for vaccinations in Sierra Leone ($)	Authors David Bernard, Anne Karing	Field Development Economics,	Close Date May 1, 2023	View Details

Example: Campos-Mercade et al. (2021) on SSPP

Social Science
PREDICTION PLATFORM

ABOUT MAKE PREDICTIONS MY SURVEYS MY ACCOUNT

Behavioral interventions and vaccination uptake

Study ID sspp-2021-0021-v1

General Details

Project Behavioral interventions and vaccination uptake
Study ID sspp-2021-0021-v1
Study Title Behavioral interventions and vaccination uptake
Authors Pol Campos-Mercade, Armando Meier, Stephan Meier, Devin Pope, Florian Schneider, Erik Wengström
Completion Time 5 Minutes
Close Date Aug. 15, 2021
Discipline Economics
Field Health Economics, Behavioral Economics
Country Sweden
Abstract
Our goal is to collect predictions of experts about the effects of interventions to increase COVID-19 vaccine uptake. We have not yet analyzed the data on vaccination uptake. Your predictions will help us to contextualize the findings of our experiment.

Example: Campos-Mercade et al. (2021) on SSPP

Please give an estimate of the difference in share of people getting vaccinated between each treatment and the Control condition (in percentage points).

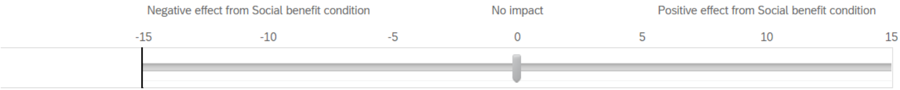
Remember that in the Control condition, we only encourage participants to take the COVID-19 vaccine as soon as possible and provide a link to a website where they find information of how to book a vaccination appointment. The encouragement statement and the link are also included in all other except the Minimal condition.

Note: Based on actual current vaccination rates and earlier representative surveys, our best guess will be that **around 70% of people in the Control condition will vaccinate** within the first month of availability.

Social benefits condition

Remember that in the Social benefits condition, we tell participants that the COVID-19 vaccine not only protects them, but also protects people around them. We then ask them to make a list of 4 people who would benefit from the vaccine.

Difference in vaccination uptake between Social benefits condition and Control condition (percentage points):



Data

Inclusion criteria

1. Primarily a social science paper.
2. Most recent version published or publicly shared in 2015 or later.

Inclusion criteria

1. Primarily a social science paper.
2. Most recent version published or publicly shared in 2015 or later.
3. Paper presents predictions of $\geqslant 1$ target outcome in a target study.

Inclusion criteria

1. Primarily a social science paper.
2. Most recent version published or publicly shared in 2015 or later.
3. Paper presents predictions of $\geqslant 1$ target outcome in a target study.
4. Forecast elicitation cannot affect the target outcome(s) predicted.

Inclusion criteria

1. Primarily a social science paper.
2. Most recent version published or publicly shared in 2015 or later.
3. Paper presents predictions of $\geqslant 1$ target outcome in a target study.
4. Forecast elicitation cannot affect the target outcome(s) predicted.
5. Forecasts elicited by or in cooperation with the author(s) of the target study.

Search

- We identified 104 relevant papers:

Search

- We identified 104 relevant papers:
  - 57 published papers, 12 in “Top-5” journals
  - 47 working papers

Search

- We identified 104 relevant papers:
  - 57 published papers, 12 in "Top-5" journals
  - 47 working papers
- Hand-coded each paper:

Search

- We identified 104 relevant papers:
  - 57 published papers, 12 in "Top-5" journals
  - 47 working papers
- Hand-coded each paper:
  - $> 3,000$ target outcomes
  - $> 41,000$ individual forecasters

Coding

- What $\rightsquigarrow$ Type of the “target” study and outcome

Coding

- What $\rightsquigarrow$ Type of the “target” study and outcome
- When / How $\rightsquigarrow$ Prediction elicitation method

Coding

- What $\rightsquigarrow$ Type of the “target” study and outcome
- When / How $\rightsquigarrow$ Prediction elicitation method
- Who $\rightsquigarrow$ Participant characteristics

Coding

- What $\rightsquigarrow$ Type of the “target” study and outcome
- When / How $\rightsquigarrow$ Prediction elicitation method
- Who $\rightsquigarrow$ Participant characteristics
- Why $\rightsquigarrow$ Reasons for collecting predictions

coding_2023-04_26

Home Insert Page Layout Formulas Data Review View Share

A1 key

	A	AZO	AZP	AZQ	AZR	AZS	AZT	AZU	AZV	AZW	AZX	AZY	AZZ	BAA	BAB	BAC	BAD	BAE	BAF	BAG	
1	key	THELV_2020	THELV_2020	THELV_2020	WOEWW_2	ACO_2021_1	ACO_2021_1	ACO_2021_1	AACR_2021_	AACR_2021_	AACR_2021_	AACR_2021_	AACR_2021_	AACR_2021_	AACR_2021_	AVSB_2021_	AMRRY_202	AMRRY_202	BKR_2021_(	BRSST_2021	BRS
2	citation	Tierney et al	Tierney et al	Tierney et al	Wein et al. (	Abebe et al.	Abebe et al.	Abebe et al.	Advani et al.	Advani et al.	Advani et al.	Advani et al.	Advani et al.	Advani et al.	Advani et al.	Alister et al.	Allen et al. (	Allen et al. (	Balietti et al	Bessone et a	Bes
3	id																				
4	id_target_exp	1	1	1		1	1	1	1	1	1	1	1	1	1		1	1		1	
5	id_target_outcome	24	24	24		1	2	3	1	2	3	4	5	6	7		1	2		1	
6	id_forecast	3	4	5		1	1	1	1	1	1	1	1				1	1		1	
7	forecast_presented_in_paper	0	0	0		1	1	1	1	1	1	1	1	0	0		1	1		1	
8	title	Creative Des	Creative Des	Creative Des	Primed, But	The Selectio	The Selectio	The Selectio	Race-Relate	Race-Relate	Race-Relate	Race-Relate	Race-Relate	Race-Relate	Race-Relate	How Do We	Correcting P	Correcting P	Optimal Des	The Econom	The
9	author_lastname	Tierney; Har	Tierney; Har	Tierney; Har	Wein; Owsle	Abebe; Caria	Abebe; Caria	Abebe; Caria	Advani; Ash;	Advani; Ash;	Advani; Ash;	Advani; Ash;	Advani; Ash;	Advani; Ash;	Advani; Ash;	Alister; Vick	Allen; Mahu	Allen; Mahu	Balietti; Klei	Bessone; Ra	Bes
10	author_firstname	Warren; Jay	Warren; Jay	Warren; Jay	Tom; Nichol	Girum; A Ste	Girum; A Ste	Girum; A Ste	Arun; Elliott;	Arun; Elliott;	Arun; Elliott;	Arun; Elliott;	Arun; Elliott;	Arun; Elliott;	Arun; Elliott;	Manikya; Ra	James; Arlet	James; Arlet	Stefano; Bre	Pedro; Gauta	Ped
11	year	2020	2020	2020	2020	2021	2021	2021	2021	2021	2021	2021	2021	2021	2021	2021	2021	2021	2021	2021	
12	journal	Organizatior	Organizatior	Organizational Behavior a		American Ec	American Ec	American Economic Review								Advances in Methods and Practices in			Experimenta	Quarterly Jo	Qua
13	published	1	1	1	0	1	1	1	0	0	0	0	0	0	0	1	0	0	1	1	
14	prior_wp	1	1	1	0	1	1	1	1	1	1	1	1	1	1		1	1		1	
15	prior_wp_forecasts	1	1	1		1	1	1	1	1	1	1	1	1	1		1	1		0	
16	link	https://doi.c	https://doi.c	https://doi.c	https://osf.i	https://doi.c	https://doi.c	https://doi.c	https://pape	https://pape	https://pape	https://pape	https://pape	https://pape	https://pape	https://doi.c	https://www	https://www	https://doi.c	https://doi.c	http
17	exclude	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	
18	exclude_reason				Forecast data not mappable to target effects											forecasts not elicited, they just ask for			Forecasting is the part of the		
19	source	gs_cite_dp2	gs_cite_dp2	gs_cite_dp2	gs_cite_dp2	gs_cite_dp2	gs_cite_dp2	gs_cite_dp2	gs_key_sspp	gs_key_sspp	gs_key_sspp	gs_key_sspp	gs_key_sspp	gs_key_sspp	gs_key_sspp	gs_key_rep	sspp archive	sspp archive	gs_cite_dp2	gs_cite_dpv	gs_
20	data_public	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	0	0		1	
21	data_public_link	https://osf.i	https://osf.i	https://osf.i	https://osf.i	https://www	https://www	https://www.openicpsr.org/openicpsr/project/127401/version/V1/view								https://osf.io/dj2fx/				https://data	http
22	data_request																				
23	target_data_type	Replication (	Replication (	Replication (	Lab-in-field (	Field experir	Field experir	Field experir	Observation	Observation	Observation	Observation	Observation	Observation	Observational data on sh		Field experir	Field experiment data stu		Field experir	Fiel
24	target_data_own	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	1		1	
25	target_data_type_lab	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0		0	
26	target_data_type_lab_in_field	0	0	0	1	0	0	0	0	0	0	0	0	0	0		0	0		0	
27	target_data_type_field	0	0	0	0	1	1	1	0	0	0	0	0	0	0		1	1		1	
28	target_data_type_online	1	1	1	0	0	0	0	0	0	0	0	0	0	0		0	0		0	
29	target_data_type_mixed	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0		0	
30	target_data_type_rct	1	1	1	1	1	1	1	0	0	0	0	0	0	0		1	1		1	
31	target_data_type_replication	1	1	1	0	0	0	0	0	0	0	0	0	0	0		0	0		0	
32	replication_definition	Original resu	Original resu	Original result replicated in the original direction																	
33	target_data_type_obs	0	0	0	0	0	0	0	1	1	1	1	1	1	1		0	0		0	
34	target_effect_description	Having parti	Effect size fo	Effect size for the interac		Application (	Average Rav	Average of t	Share of rac	Share of rac	Share of rac	Share of rac	Share of rac	Share of rac	Share of race-related res		Treatment e	Treatment effect of lead		Effect of sle	Effe
35	target_effect	0	-0.0554	-0.0554		0.53	40	54									0.00425	-0.00541		-0.040	
36	target_effect_se																0.014	0.0137		0.02	
37	target_effect_pvalue																p>0.1	p>0.1		0.05	
38	target_effect_null																1	1		1	
39	target_effect_sem																				
40	target_effect_sd																				
41	sspp	0	0	0		0	0	0	1	1	1	1	1	1	1	0	1	1	0	0	
42	sspp_link								https://socia	https://socia	https://socia	https://socia	https://socia	https://socia	https://socialsciencepred		https://socia	https://socialscienceprediction.org/s/dx4z			
43	forecast_collection	Online surve	Online surve	Online survey		Interviews w	Interviews w	Interviews w	Social Scien	Social Scien	Social Scien	Social Scien	Social Scien	Social Scien	Social Science Prediction		SSPP	SSPP			
44	forecaster_rec_channel	Posted link t	Posted link t	Posted link to academic w		Firm manag	Firm manag	Firm manag	Personal em	Personal em	Personal em	Personal em	Personal em	Personal em	Personal email; Twitter;		SSPP	SSPP		68 personali	68 p
45	forecaster_n	194	194	194		195	195	195	240	240	240	240	240	240	240		71	71		122	

Sheet1

Cell A45 commented by tc={9ED56AE4-801B-45FE-9AC2-98AF0CA89010}

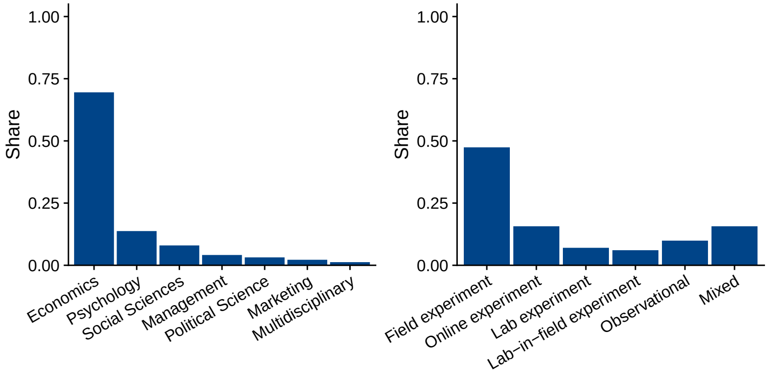
100%

Who participates in the market for forecasting?

Demand-side characteristics

Result 1

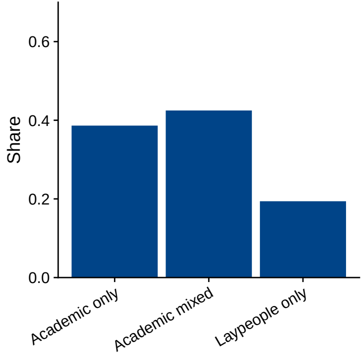
The practice of collecting forecasts is far more widespread among economists and for field experiments.



Supply-side characteristics

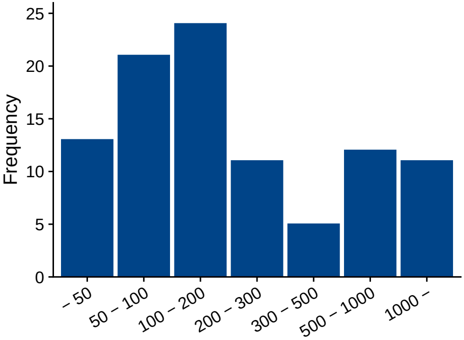
Result 2

Forecasters are recruited from a variety of pools with different levels and types of expertise. However, the focus remains on academic expertise.



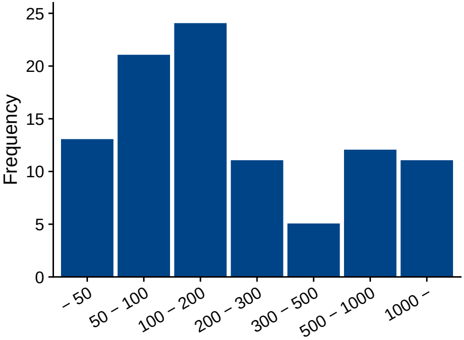
- 70 with outreach to academic experts
- 24 studies recruited via SSPP
- 19 MTurk/Prolific

Supply-side characteristics



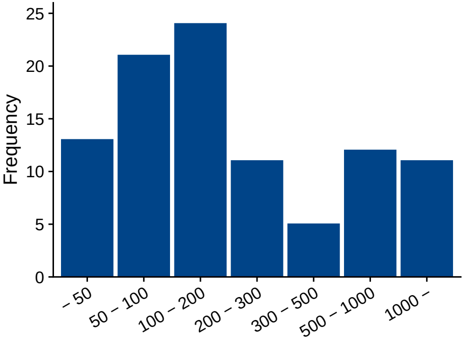
- Large heterogeneity in sample size

Supply-side characteristics



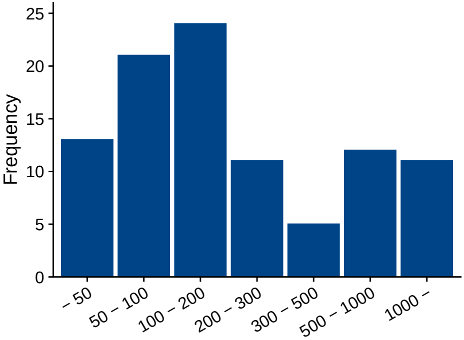
- Large heterogeneity in sample size
- Partly reflects different goals

Supply-side characteristics



- Large heterogeneity in sample size
- Partly reflects different goals
- ⚠ Objectives are not always made clear

Supply-side characteristics



- Large heterogeneity in sample size
- Partly reflects different goals
- ⚠ Objectives are not always made clear

⇒ Next stop: understand the goals.

Why do researchers collect forecasts?

Why using forecasts?

- Assist with the evaluation of scientific claims

Why using forecasts?

Assist with the evaluation of scientific claims

- Contextualizing research findings within existing scientific knowledge

Why using forecasts?

Assist with the evaluation of scientific claims

- Contextualizing research findings within existing scientific knowledge
- Combating hindsight bias
- Inoculating against publication bias
  - “Surprising” null results might be more publishable
  - Null effects insignificant against $H_0: \theta = 0$, but possibly significant against $H_0: \theta = \mu$ for some $|\mu| \gg 0$.

Why using forecasts?

Assist with the evaluation of scientific claims

- Contextualizing research findings within existing scientific knowledge
- Combating hindsight bias
- Inoculating against publication bias
    - "Surprising" null results might be more publishable
    - Null effects insignificant against $H_0: \ \theta = 0$, but possibly significant against $H_0: \ \theta = \mu$ for some $|\mu| \gg 0$.
- Assessing the replicability or plausibility of results

Why using forecasts?

- Understanding-the-world motives
  - Beliefs influence choices
    - e.g. policymaker beliefs might affect program adoption

Why using forecasts?

- Understanding-the-world motives
  - Beliefs influence choices
    - e.g. policymaker beliefs might affect program adoption
- Tool for study and treatment selection
  - "to quickly identify findings that are unlikely to replicate" Dreber et al. (2015)
  - identify which treatment arm will be most impactful

Why using forecasts?

- Different statistics taken from the distribution of forecasts may matter depending on the goal(s) of the forecasting exercise

Why using forecasts?

❗ Different statistics taken from the distribution of forecasts may matter depending on the goal(s) of the forecasting exercise

- Select most successful intervention

  $\leadsto$ aggregate forecasts into a single prediction

  - “crowd average” often outperforms individual forecasts

Why using forecasts?

❗ Different statistics taken from the distribution of forecasts may matter depending on the goal(s) of the forecasting exercise

- Select most successful intervention

  $\rightsquigarrow$ aggregate forecasts into a single prediction

  - "crowd average" often outperforms individual forecasts

- Assess riskiness of intervention

  $\rightsquigarrow$ measure expert disagreement

  - robustness concerns $\rightsquigarrow$ go with lowest disagreement
  - novelty considerations $\rightsquigarrow$ go with most disagreement

Motives for data collection

Result 3

Researchers cite the desire to contextualize their results with respect to the prior academic consensus.

Motives for data collection

Result 3

Researchers cite the desire to contextualize their results with respect to the prior academic consensus.

- Predominance of the word "result" in stated rationales

Motives for data collection

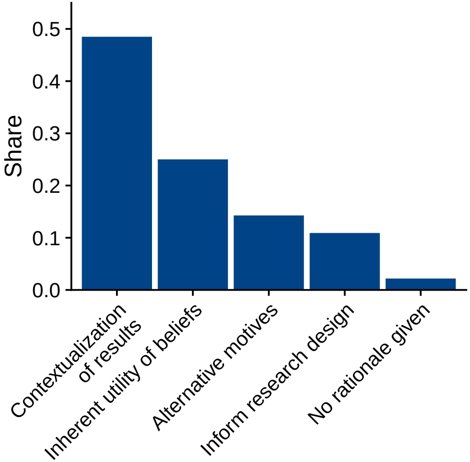
Result 3

Researchers cite the desire to contextualize their results with respect to the prior academic consensus.

- Predominance of the word “result” in stated rationales
- Other keywords
  - “hindsight (bias)”
  - “replication”
  - “publication (bias)”
  - “surprise”

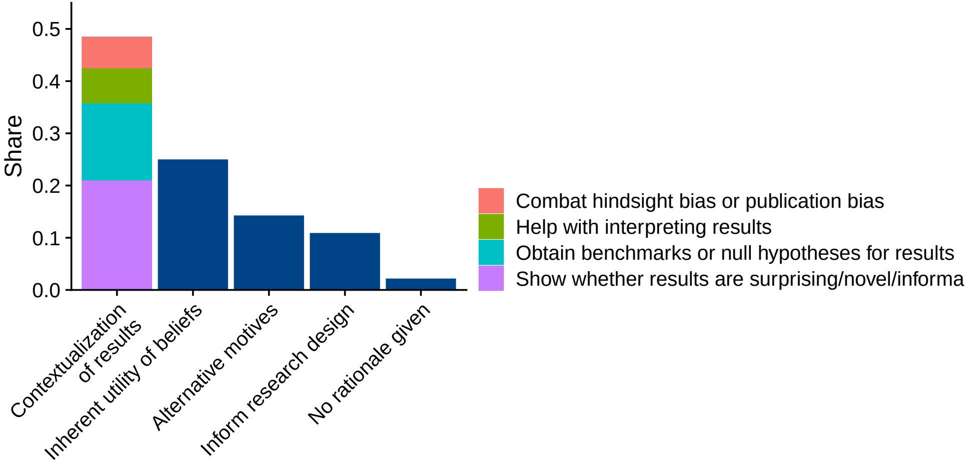
Motives for data collection

- Hand coding identified 149 rationales across the 104 papers



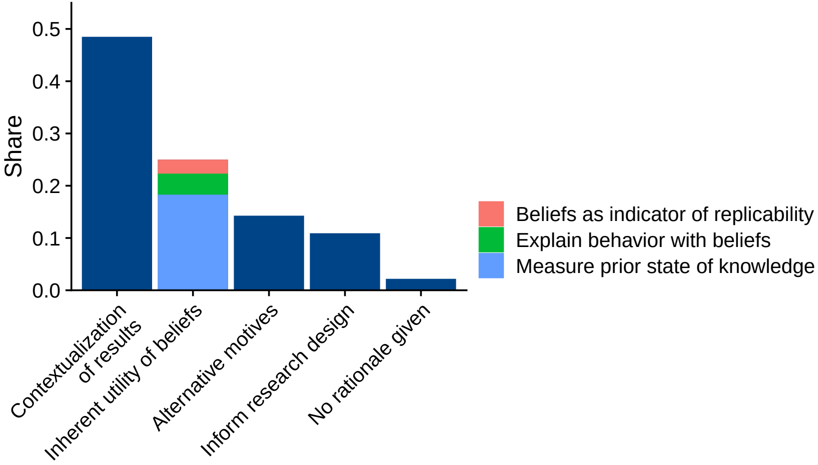
Motives for data collection

- Hand coding identified 149 rationales across the 104 papers



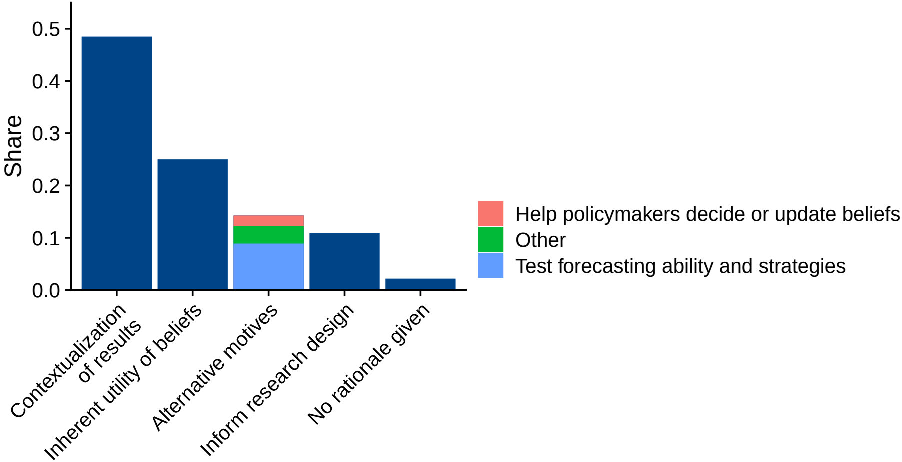
Motives for data collection

- Hand coding identified 149 rationales across the 104 papers



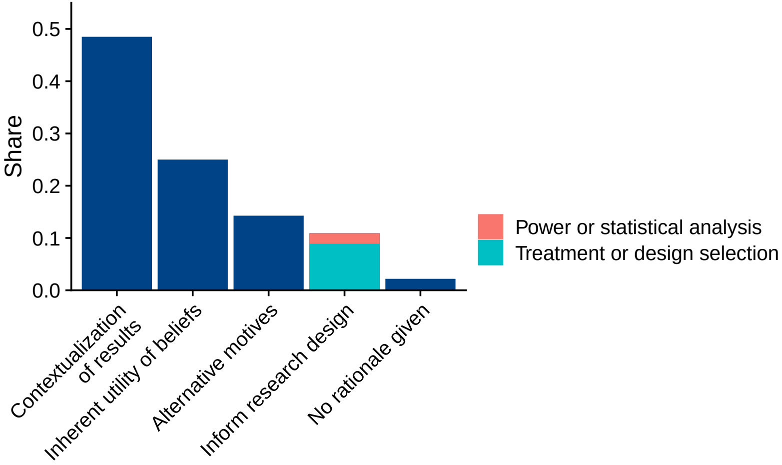
Motives for data collection

- Hand coding identified 149 rationales across the 104 papers



Motives for data collection

- Hand coding identified 149 rationales across the 104 papers



Motives for data collection

- Very few attempts to quantify the value of information contained in experiment.

Motives for data collection

- Very few attempts to quantify the value of information contained in experiment.
- Need to compare prior and posterior beliefs after seeing the data.
- Virtually no paper presents information of this kind.

Motives for data collection

- Very few attempts to quantify the value of information contained in experiment.
- Need to compare prior and posterior beliefs after seeing the data.
- Virtually no paper presents information of this kind.
- Two approaches:
  1. Modeling in a Bayesian framework (cf work of Rachael Meager) $\Rightarrow$ normative benchmark
  2. Direct elicitation of prior and posteriors $\Rightarrow$ positive statement

Motives for data collection

- Very few attempts to quantify the value of information contained in experiment.
- Need to compare prior and posterior beliefs after seeing the data.
- Virtually no paper presents information of this kind.
- Two approaches:
  1. Modeling in a Bayesian framework (cf work of Rachael Meager) $\Rightarrow$ normative benchmark
  2. Direct elicitation of prior and posteriors $\Rightarrow$ positive statement
- Divergence of posterior from prior captures both surprisingness and learning. Distinction between surprisingness and novelty?

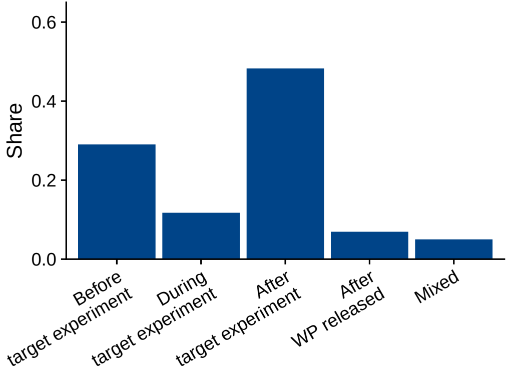
Motives for data collection

- Very few attempts to quantify the value of information contained in experiment.
- Need to compare prior and posterior beliefs after seeing the data.
- Virtually no paper presents information of this kind.
- Two approaches:
    1. Modeling in a Bayesian framework (cf work of Rachael Meager) $\Rightarrow$ normative benchmark
    2. Direct elicitation of prior and posteriors $\Rightarrow$ positive statement
- Divergence of posterior from prior captures both surprisingness and learning. Distinction between surprisingness and novelty?

Motives for data collection

Result 4

A large fraction of researchers collect forecasts after observing the findings of their study, reflecting a desire to make sense of their results.



Implications of forecast timing (1)

- **Q:** Does timing predict distributions of effect sizes/null effects?
  e.g., authors see null results and collect forecasts ex-post.

Implications of forecast timing (1)

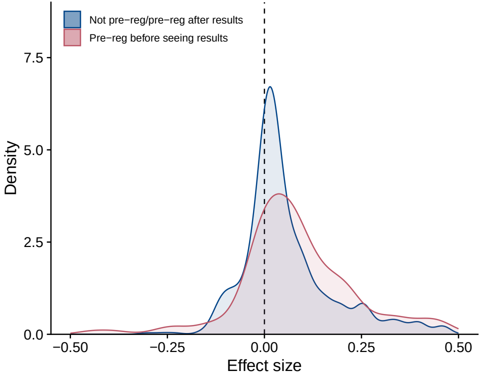
- **Q:** Does timing predict distributions of effect sizes/null effects? e.g., authors see null results and collect forecasts ex-post.
- Difficulties of measurement:
  - Lack of information about forecast timing.
  - Lag between decision to collect forecasts and collection date.
  - “Pre”-registration before forecasts, but after seeing target results.

Implications of forecast timing (1)

- **Q:** Does timing predict distributions of effect sizes/null effects? e.g., authors see null results and collect forecasts ex-post.
- Difficulties of measurement:
  - Lack of information about forecast timing.
  - Lag between decision to collect forecasts and collection date.
  - “Pre”-registration before forecasts, but after seeing target results.
- **Approach:** identify papers pre-registered before forecasts *and* results

Implications of forecast timing (2)

- $N = 667$ (blue) vs. $N = 167$ (red) outcomes.
- Failure to pre-register predicts concentration of effects ~ 0 ($p < 0.001$)



Motives for data collection

On the to-do list:

- Compare the distribution of null results for papers with and without forecasts.
- Are papers with null results more likely to contain forecasts relative to close neighbors?
- Understand how selection affects inference.

How are forecasts elicited?

Elicitation of forecasts

Result 5

Authors primarily elicit forecasts of treatment effects and use surveys rather than markets. However, considerable heterogeneity in survey elicitation methods exists.

Elicitation of forecasts

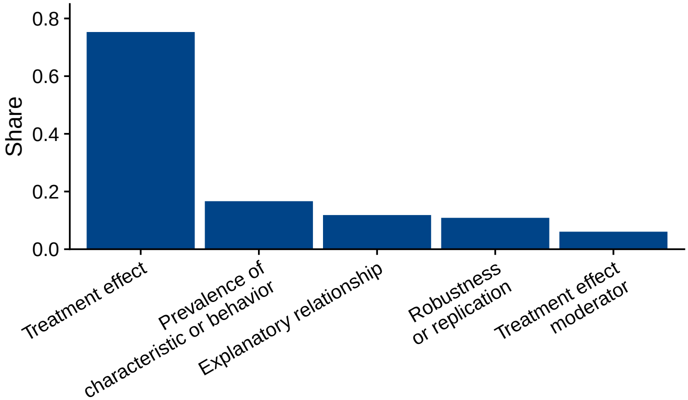
Result 5

Authors primarily elicit forecasts of treatment effects and use surveys rather than markets. However, considerable heterogeneity in survey elicitation methods exists.

- Heterogeneity in
  - type $\rightsquigarrow$ probability, proportion, raw mean, standardized effect, ...
  - procedure $\rightsquigarrow$ individual vs. market, incentives for accuracy, framing, ...

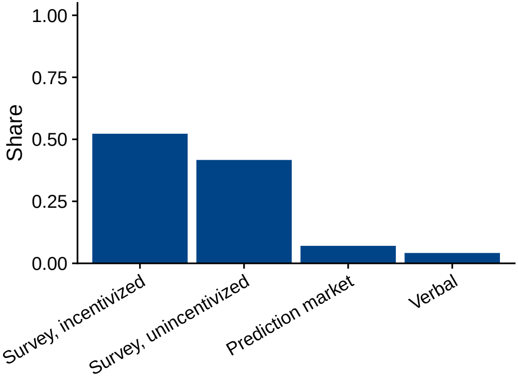
Elicitation of forecasts

- Primary focus on the forecasting of treatment effects
- Huge variation in terms of standardization, benchmark info, ...



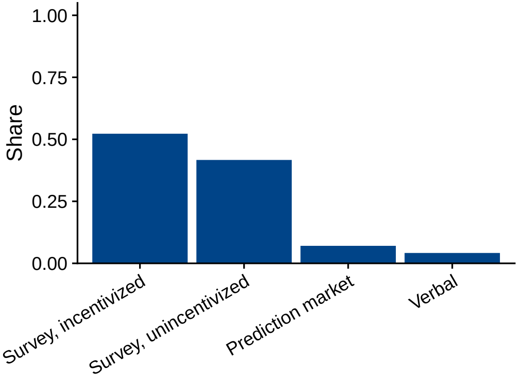
Elicitation of forecasts

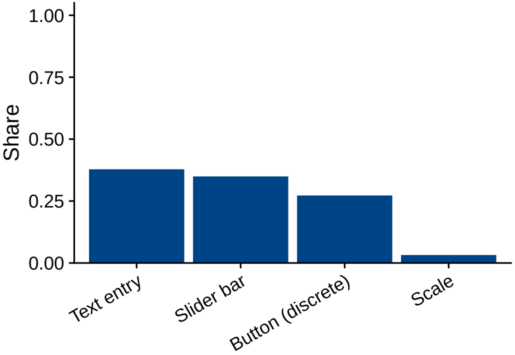
- All 104 used individual elicitation
- 7 also used prediction markets



Elicitation of forecasts

- All 104 used individual elicitation
- 7 also used prediction markets
- Surveys use a mix of text, sliders and buttons





(When) Are predictions accurate and informative?


Preliminary!

Individual-level forecaster dataset

- Based on a subset of papers for which we have the individual-level raw forecast data
    - # studies: 34
    - # forecasters: 15,336
    - # forecasts: 228,246

Individual-level forecaster dataset

- Based on a subset of papers for which we have the individual-level raw forecast data
  - # studies: 34
  - # forecasters: 15,336
  - # forecasts: 228,246
- For a subset of analyses below, we separate
  - treatment effect SDs
  - binary outcomes

Forecast evaluation

1. Accuracy
    - Multiple dimensions (directional or size of deviations)
    - Necessity of benchmarking, but sensitivity to the choice of benchmark

Forecast evaluation

1. Accuracy
    - Multiple dimensions (directional or size of deviations)
    - Necessity of benchmarking, but sensitivity to the choice of benchmark
2. Bias
    - Forecasters can be very close to the truth but also biased.
    - On average, do they over- or underpredict effects?

Struggles with standardization and aggregation

=

Struggles with standardization and aggregation

=

Meaningless means...

DATA COLADA

Thinking about evidence, and vice versa

HOME TABLE OF CONTENTS FEEDBACK POLICY SEMINAR ABOUT

[104] Meaningless Means: Some Fundamental Problems With Meta-Analytic Averages

Posted on November 1, 2022 by Uri, Joe, & Leif

This post is an introduction to a series of posts about meta-analysis [1]. We think that many, perhaps most, meta-analyses in the behavioral sciences are invalid. In this introductory post, we make that case with arguments. In subsequent posts, we will make that case by presenting examples taken from published meta-analyses.

We have recently written a short article for Nature Reviews Psychology in which we briefly described some fundamental problems with meta-analysis, and proposed an alternative way to generate more productive and less misleading literature reviews (.htm). Because of space constraints, in that article we couldn't fully articulate our concerns with meta-analysis, and we were unable to include many examples. But we can do that here, over the course of a few posts.

GET COLADA EMAIL ALERTS.

you@email.com

Subscribe

Join 8,779 other subscribers

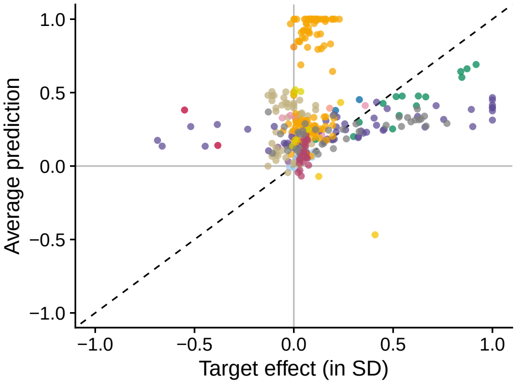
SOCIAL MEDIA

Bluesky We announce posts on Bluesky

And link to them on our Facebook page

Directionality: Continuous outcomes

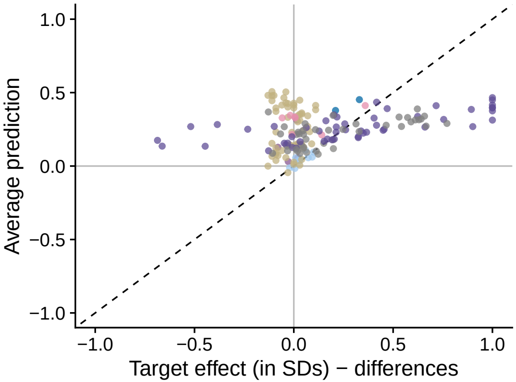
- Do forecasters get the direction of effects right?
- Standardized effect sizes

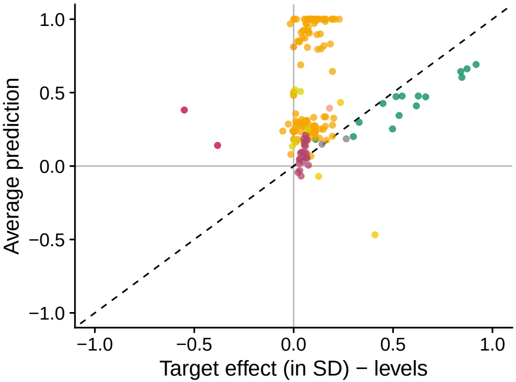


- Weak correlation ($\rho = 0.28$)
- Study-specific features may influence performance.

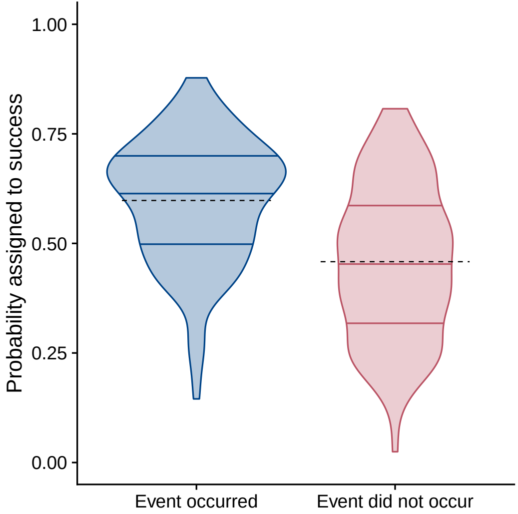
Directionality: Continuous outcomes

- Do forecasters get the direction of effects right?
- Standardized effect sizes by type





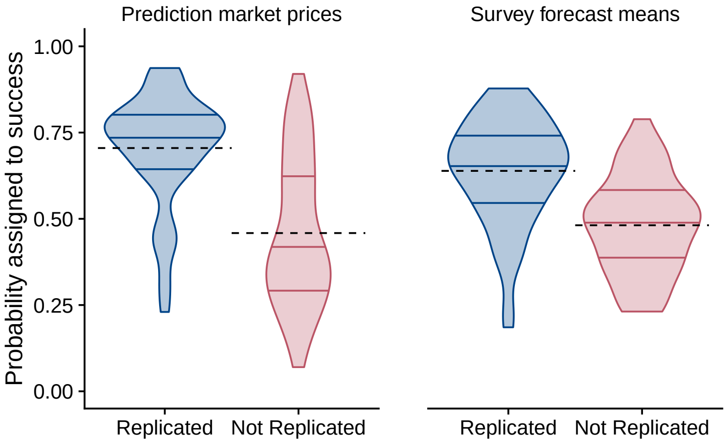
Directionality: Discrete outcomes



	$E = 1$	$E = 0$
$p \geqslant 0.50$	0.75	0.43
$p < 0.50$	0.25	0.57

- Good discriminatory power
- Type I errors more frequent

Directionality: Binary replication outcomes



- $\rho^{PM} = 0.54$
- $\rho^{SF} = 0.47$

Point accuracy

- Forecasters can get point estimates very off even if they are right about the direction.
- Various ways of measuring prediction error

  $\Rightarrow$ Today: mean-squared error of average forecast
- Performance relative to two benchmarks
  1. random ("monkey") benchmark (all outcomes equally likely)
  2. uninformed ("null") model (e.g., no effect of intervention; 50% replication)

Point accuracy

- Forecasters can get point estimates very off even if they are right about the direction.

- Various ways of measuring prediction error

  $\Rightarrow$ Today: mean-squared error of average forecast

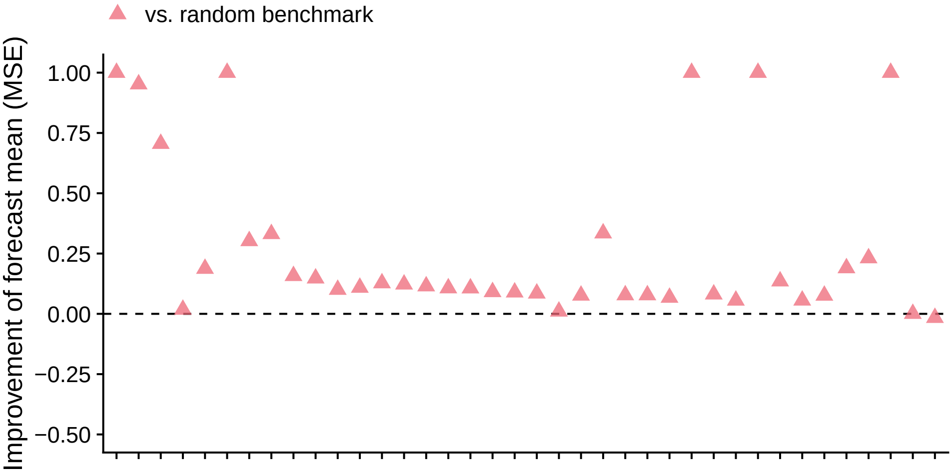
- Performance relative to two benchmarks
  1. random ("monkey") benchmark (all outcomes equally likely)
  2. uninformed ("null") model (e.g., no effect of intervention; 50% replication)

Point accuracy

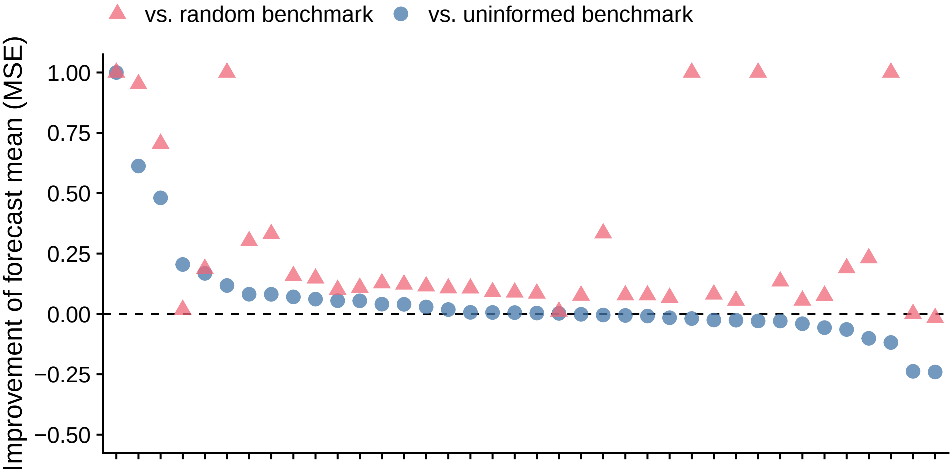
- Forecasters can get point estimates very off even if they are right about the direction.
- Various ways of measuring prediction error

  $\Rightarrow$ Today: mean-squared error of average forecast
- Performance relative to two benchmarks
  1. random (“monkey”) benchmark (all outcomes equally likely)
  2. uninformed (“null”) model (e.g., no effect of intervention; 50% replication)

Point accuracy



Point accuracy



Point accuracy - other benchmarks

Exploring two other benchmarks:

- LLM benchmark: takes into account the published literature up to the forecast data collection date.
- Omniscient benchmark: knows sample estimate but accounts for sampling error.

Biasedness

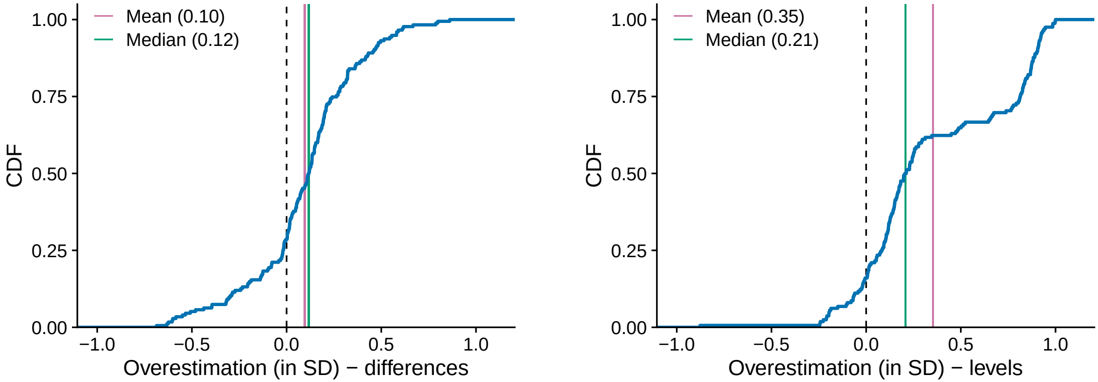
Result 6

Predictions of treatment effect sizes and replicability tend to be biased upwards.

⚠ We do not know the distribution of true effects

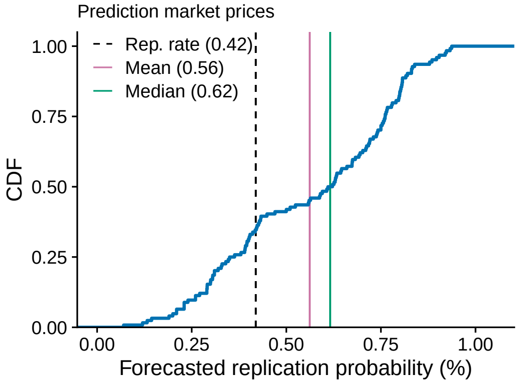
Biasedness

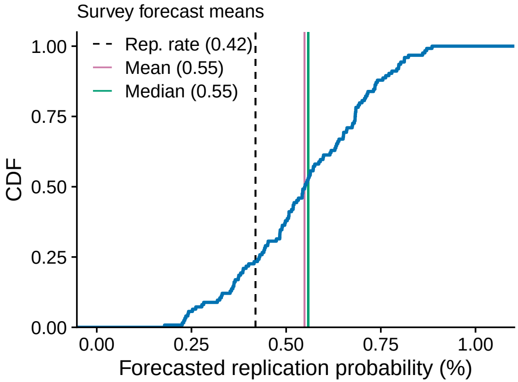
- Overestimation = forecast mean − effect size



Biasedness

- Mean predicted replication probabilities





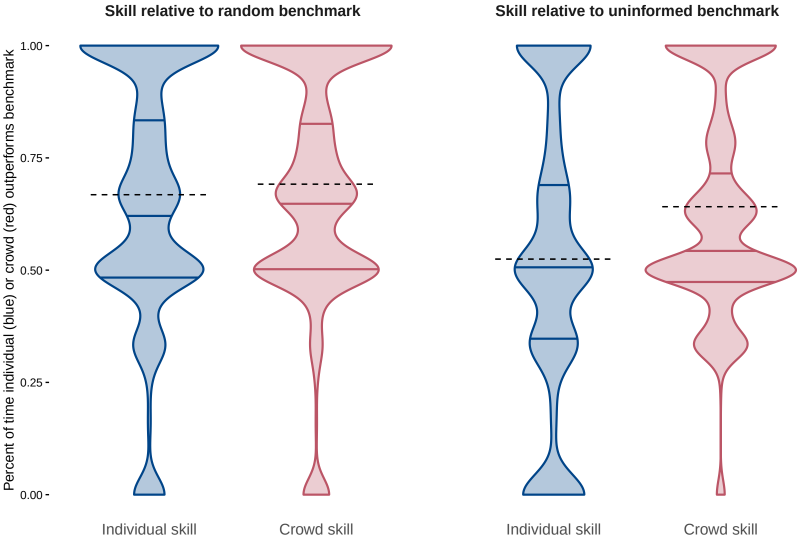
"Wisdom of Crowds" (WoC)

Result 7

(Tentative) Individual forecasts are very noisy and WoC estimates significantly improve upon individual forecasts. Most of the improvement emerges for crowds as small as $N = 5$.

- Conduct bootstrap simulations with 1,000 samples
- Calculate the WoC estimate for crowds of size N
- Today: will just contrast performance of full-size crowd to individuals.

Skill of individuals vs. crowds



Preliminary summary on performance

1. Forecasts are informative but median is an overestimate
    - In line with literature on overconfidence/overoptimism
    - Conjecture: authors seek forecasts for null results + forecasters not conditioning on this?

Preliminary summary on performance

1. Forecasts are informative but median is an overestimate
   - In line with literature on overconfidence/overoptimism
   - Conjecture: authors seek forecasts for null results + forecasters not conditioning on this?
2. WoC estimates improve quickly with crowd size N
   - If goal is to get accurate estimates, no need to collect 1,000 forecasts
   - To do: understand when WoC does worse and why

Preliminary summary on performance

1. Forecasts are informative but median is an overestimate
   - In line with literature on overconfidence/overoptimism
   - Conjecture: authors seek forecasts for null results + forecasters not conditioning on this?
2. WoC estimates improve quickly with crowd size N
   - If goal is to get accurate estimates, no need to collect 1,000 forecasts
   - To do: understand when WoC does worse and why
3. Other to do's: individual-level determinants of forecasting accuracy
   - Characteristics of superforecasters?
   - Understand trade-off between quality and quantity

Discussion

Looking forward (1)

- Too early for definitive conclusions
- Some thoughts on this use of this practice:

Looking forward (1)

- Too early for definitive conclusions
- Some thoughts on this use of this practice:

- Importance of collecting forecasts before seeing results (?)
- Less arbitrary selection rules for how to sample forecasters
- Elicit predictions and confidence jointly
- Proper statistical testing that accounts for uncertainty
- Forecasts for theory/macro papers?
- More usage of forecasts for study design/selection

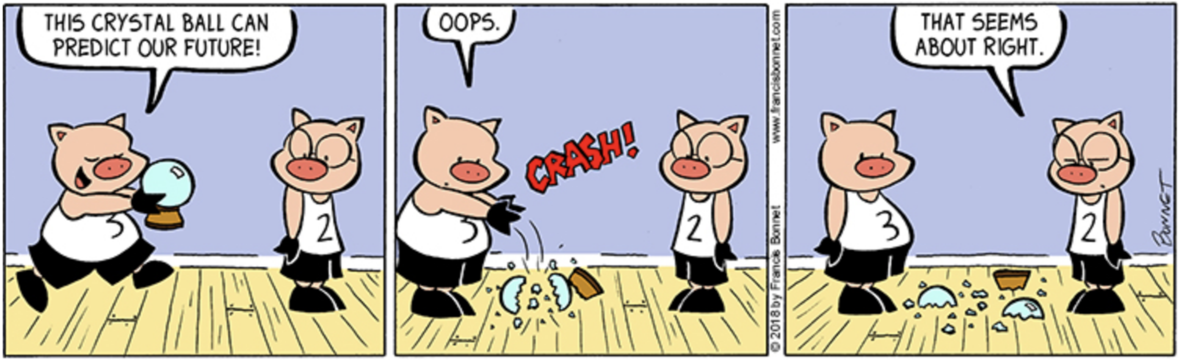
Looking forward (2)

- More thoughts on challenges and unknowns:

Looking forward (2)

More thoughts on challenges and unknowns:

- How to solve the public good problem of forecast production? ML/hybrid models?
- Scientific value of forecast production? Helpful for null results?
- How to address the incentive problem re timing? Should we worry?
- How to improve forecast accuracy? What is an “expert”?
- Broadening the use of forecasts to study QRPs, research impact, or for peer review?



severine.toussaert@economics.ox.ac.uk
https://www.severinetoussaert.com
https://labsquare.net