Predicting Social Science Results

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

Leibniz Open Science Day 2025

Conducting science at scale

Decentralized model of scientific production

- Many researchers **working separately** on common questions
- **Limited input** from others on design and interpretation of results

Conducting science at scale

Decentralized model of scientific production

- Many researchers **working separately** on common questions
- **Limited input** from others on design and interpretation of results

Current challenges

- Uncertainty about **optimal allocation** of funding and resources
- High **disagreement** and low **comparability**, even within field
- **Publication bias** and **low replication rates**

Conducting science at scale

Can we do better? Potential approach:

- Incorporate input from others into research production & evaluation rather than (just) managing disagreement ex-post
- Implementable at scale by collecting **forecasts of research results**

Motivation

Recent developments

- **Increasingly common** to elicit forecasts of research results
  - e.g., Social Science Prediction Platform (DellaVigna et al., 2019)

Motivation

Recent developments

- **Increasingly common** to elicit forecasts of research results
  - e.g., Social Science Prediction Platform (DellaVigna et al., 2019)

But...

- Overall **accuracy** and **informativeness** remain unknown
- **Best practices** and **returns to elicitation** are unclear

Motivation

Recent developments

- **Increasingly common** to elicit forecasts of research results
  - e.g., Social Science Prediction Platform (DellaVigna et al., 2019)

But...

- Overall **accuracy** and **informativeness** remain unknown
- **Best practices** and **returns to elicitation** are unclear

$\rightarrow$ First **systematic evidence** on practice of predicting research results

An example

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

- "Lost" wallets given to strangers (**"target study"**)
  - Amount of money in the wallet (if any) was randomized
  - Percent of citizens who returned the wallet (**"target outcomes"**)

An example

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

- "Lost" wallets given to strangers (**"target study"**)
  - Amount of money in the wallet (if any) was randomized
  - Percent of citizens who returned the wallet (**"target outcomes"**)

? Forecasting task:

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

An example

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

- "Lost" wallets given to strangers (**"target study"**)
  - Amount of money in the wallet (if any) was randomized
  - Percent of citizens who returned the wallet (**"target outcomes"**)

? Forecasting task:

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

This study

What does our paper do?

- Investigate the history of social science predictions (***Narrative review***)
- Document current practices and accuracy (***Meta-analysis***)
- Conceptualize and measure returns to collecting forecasts

This study

What does our paper do?

- Investigate the history of social science predictions (***Narrative review***)
- Document current practices and accuracy (***Meta-analysis***)
- Conceptualize and measure returns to collecting forecasts

Today, will focus on applications of forecasting, namely:

1. **Treatment selection** and **policy choice**
2. Predicting **replicability**
3. **Null hypothesis testing** with forecast means

What counts? Inclusion criteria

1. Primarily a **social science** paper.
2. Most recent version published or publicly shared in **2015 or later**.
3. Features human **predictions** of **target outcome(s)** in a **target study**.

What counts? Inclusion criteria

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

Quantitative meta-analysis

- We identified **104** relevant papers:
  - 57 published papers, 12 in “Top-5” economics journals
- Hand-coded each paper:
  - $> 3,000$ **target outcomes**
  - $> 41,000$ **individual forecasters**
- **Variables of interest:**
  - who (researchers, forecasters); why (reasons for collecting predictions); how (elicitation details); performance (outcomes and forecast means)

Quantitative meta-analysis

- Additional **raw forecast-level** dataset - partially collected
  - # studies: 39
  - # target outcomes: 957
  - # forecasters: 18,008
  - # forecasts: 242,556

1. Treatment Selection and Policy Choice

Choosing between policies

Challenges:

- A researcher or policymaker wants to know whether **Policy *j*** or **Policy *k*** is likely to be superior without testing both
- Optimal allocation of resources between j and k is unclear

Choosing between policies

Challenges:

- A researcher or policymaker wants to know whether **Policy *j*** or **Policy *k*** is likely to be superior without testing both
- Optimal allocation of resources between *j* and *k* is unclear

Application of forecasting:

Tool for **treatment and policy selection**

- How well does it work?

Choosing between policies

Implementation overview

1. Identify **pairs of policies** j and k tested on same dependent variable
2. Observe **realized** treatment effects θ_j and θ_k ("truth")

Choosing between policies

Implementation overview

1. Identify **pairs of policies** j and k tested on same dependent variable
2. Observe **realized** treatment effects θ_j and θ_k (“truth”)
3. Calculate various **forecast** aggregations A_j and A_k:
   - Mean forecasted treatment effects of policy j and policy k
   - Median
   - Optimistic (max)
   - Pessimistic (min)
   - Plurality (share of forecasters preferring Policy j)
   - Consensus (ratio of mean to forecast SD)

Choosing between policies

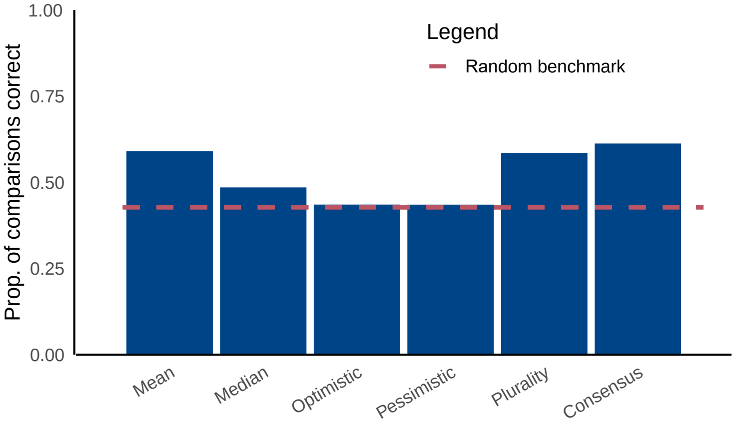
Implementation overview

1. Identify **pairs of policies** j and k tested on same dependent variable
2. Observe **realized** treatment effects θ_j and θ_k ("truth")
3. Calculate various **forecast** aggregations A_j and A_k:
   - Mean forecasted treatment effects of policy j and policy k
   - Median
   - Optimistic (max)
   - Pessimistic (min)
   - Plurality (share of forecasters preferring Policy j)
   - Consensus (ratio of mean to forecast SD)
4. Check how often forecasters **get it right**: $\theta_j > \theta_k$ AND $A_j > A_k$

Policy forecasting performance

Result 1

Some aggregation methods outperform random guessing.



2. Forecasting Replicability

Predicting whether a study will replicate

Challenges:

- Unclear which **studies to trust** until replications conducted
- Replications are **resource-intensive**
  - Which studies should be the focus of our efforts?

Predicting whether a study will replicate

Challenges:

- Unclear which **studies to trust** until replications conducted
- Replications are **resource-intensive**
  - Which studies should be the focus of our efforts?

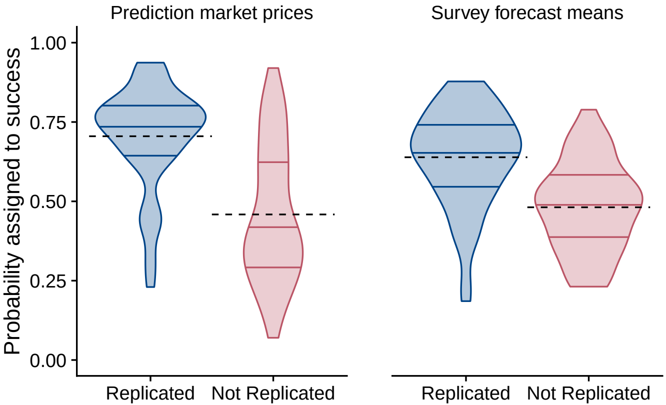
Application of forecasting:

- Assist with the **evaluation of scientific claims**
  - Assessing the **replicability** or **plausibility** of results
  - "to quickly identify findings that are unlikely to replicate" Dreber et al. (2015)

Directionality

Result 2

Forecast means are moderately **correlated with replication outcomes**.



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

3. Evaluating New Results

Evaluating the informativeness of new results

Challenges:

- Lacking in accepted measures of the **informativeness, surprisingness, novelty**, etc. of new results
- Researchers (humans) suffer from hindsight bias

Evaluating the informativeness of new results

Challenges:

- Lacking in accepted measures of the **informativeness, surprisingness, novelty**, etc. of new results
- Researchers (humans) suffer from hindsight bias

Application of forecasting:

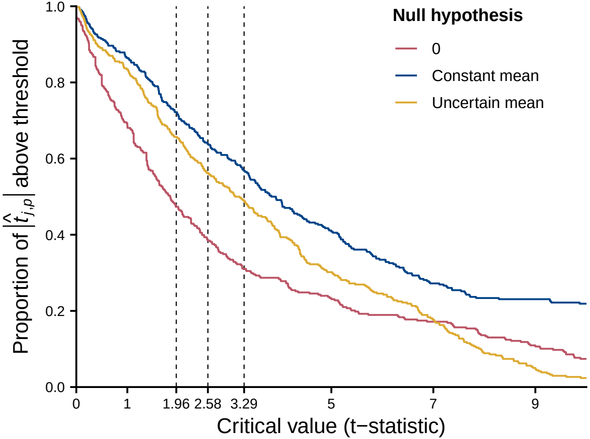
Assist with the **evaluation of scientific claims**

- Measure "information gained" from new result relative to existing scientific knowledge
- e.g., using the forecast mean as the new null hypothesis

Null hypothesis rejection rates

Result 3

Rejection rates are higher when using forecast mean as null hypothesis.



Conclusion

Potential applications of forecasting:

- **Policy choice:** improving researcher/policymaker decisions
- **Predicting replicability:** evaluating credibility and directing resources
- **Evaluating findings:** quantifying informational gains
- Grant allocation, predictions of generalizability/scalability, etc.

Conclusion

Potential applications of forecasting:

- **Policy choice:** improving researcher/policymaker decisions
- **Predicting replicability:** evaluating credibility and directing resources
- **Evaluating findings:** quantifying informational gains
- Grant allocation, predictions of generalizability/scalability, etc.

Caveats:

- Not costless (LLMs?)
- Weak-to-moderate performance in certain tasks
- Strategic incentives of forecast commissioners and forecasters
- Limited evidence on causal effects of collecting forecasts

Stefano DellaVigna, Devin Pope, and Eva Vivalt. Predict science to improve science. *Science*, 366(6464):428–429, 2019.