# Evidence Transmission

**Pre-Registration**

Mahvish Shaukat, Andreas Stegmann, and Mattie Toma*

## 1 Introduction

The generation of evidence about the efficacy of policy programs has grown increasingly common over the past two decades, and many think tanks and government agencies are now looking for ways to "institutionalize" evidence use across government. To aid in this endeavor, this project proposes a field experiment aimed at understanding how evidence diffuses within bureaucracies. Our project will investigate key predictors of evidence transmission. We will experimentally manipulate i) the hierarchy of the information sender; ii) the credibility of the evidence; and iii) second-order beliefs, or beliefs about colleagues' beliefs, use, and preferences related to the evidence. We will also explore the barriers to evidence transmission.

## 2 Experimental Design

### 2.1 Background and Sample

Our field experiment will be implemented across approximately 6,254 full-time employees at the World Bank headquarter office in Washington, DC.[1] In our study context, each of

*Shaukat: mshaukat@worldbank.org, The World Bank; Stegmann: andreas.stegmann@warwick.ac.uk, University of Warwick; Toma: mattie.toma@warwick.ac.uk, University of Warwick. We are grateful to John Firth for his excellent suggestion to experimentally explore second-order beliefs. We are also grateful for the thoughtful feedback from the UK's Foreign, Commonwealth, and Development Office (FCDO), and UK Behavioural Insights Team, the Global Priorities Institute at the University of Oxford, Nathan Canen, Matthew Ridley, Mateusz Stalinski, and Michael Thaler.

[1]Note that after running the experiment at the World Bank headquarter office we expect to collect an additional wave of data through some World Bank country offices. This will be a smaller sample overall,

511 divisions within our World Bank sample (e.g., the Fiscal Policy and Sustainable Growth Unit within the Macroeconomics, Trade and Investment Global Practice) represents a cluster of employees. There are on average 12 open and fixed-term employees per division in the headquarter office. On average, 6 of these employees are in senior positions in grades GG and above (Senior Specialist, Senior Officer, Manager, Senior/Lead Economist, or Director), and 6 are in junior positions in grades GF and below (Economist, Analyst, Specialist or Officer). While employees may interact with others outside their division, their division typically represents their main point of contact within the organization.

## 2.2 Design

In the first stage of the experiment, evidence will be disseminated to treated "seeds," or employees from each division within the World Bank. We will randomly select one seed per division and invite them to complete our "Seed Survey". Of note, we will give each assigned seed three days to respond; in cases where the division has more eligible seeds to randomly select, at the end of this period if the seed has not completed our survey, we will randomly select a new seed without replacement. Treated seeds will be exposed to findings from Noy and Zhang (2023), which experimentally studies the impact of ChatGPT on worker productivity. A week following evidence dissemination, we will follow up separately with the remaining employees in each seed's division to track the transmission of the evidence information across the seed's contacts within his or her division.[2] Finally, once the experiment has concluded we will follow up with all Seeds to improve our understanding of the barriers to evidence dissemination.

All information will be delivered via online surveys. To minimize experimenter demand effects, we will obfuscate the purpose of the study: Respondents will be informed that the study is about eliciting beliefs related to the use of generative AI in the workplace.[3] At no point in the survey with seed respondents do we mention that there is a follow-up survey with their team members. We will also highlight somewhat different collaboration teams when approaching the seeds versus the contacts: The seed survey will be sent on behalf of Warwick Business School and DECRG at the World Bank, while the contact surveys will be sent on behalf of the University of Warwick economics department and DECRG at the

and the inclusion of each country office will depend on the feasibility of sending Amazon gift cards in each location. Country offices in Australia and India were included in our pilot sample and will therefore be excluded from this additional wave of analysis.

[2]World Bank employees who were randomly assigned to be a Seed and who opened the survey will not be eligibile for the Contact Survey. Employees who were assigned to be a Seed but did not engage with our invitation will still be eligible for the Contact Survey.

[3]We also entered a vague description of the goal and design of the project on the AEA RCT Registry in case study participants searched for us or the study online.

World Bank.

## 2.3 Timeline

Our timeline for distributing the Seed and Contact surveys is as follows:

Seed wave 1:

- Nov 28: Seed survey
- Dec 7: Contact survey

Seed wave 2:

- Dec 1: Seed survey
- Dec 12: Contact survey

Seed wave 3:

- Dec 6: Seed survey
- Dec 15: Contact survey

Dec 21: Close all contact surveys

Follow-up surveys will be distributed in the new year.

All invitations will be delivered at 9:30am. Seeds will receive daily reminders to complete the survey. The first reminder will come at 3:30pm and the second reminder will come at 9:30am. Contacts will receive three reminders in total. The first reminder will be delivered at 3:30pm two days after the invitation was sent. The second reminder will be delivered at 9:30am on December 18. The final reminder will be delivered at 9:30am on December 21.

## 2.4 Treatments

As shown in Figure 1, there will be four layers of randomization at the division level:

**Stage 1: Treatment versus Control** Three quarters of all divisions will be assigned to the Treatment condition, in which a randomly-selected employee within the division will be seeded with evidence. The remaining quarter will be assigned to the Control condition.

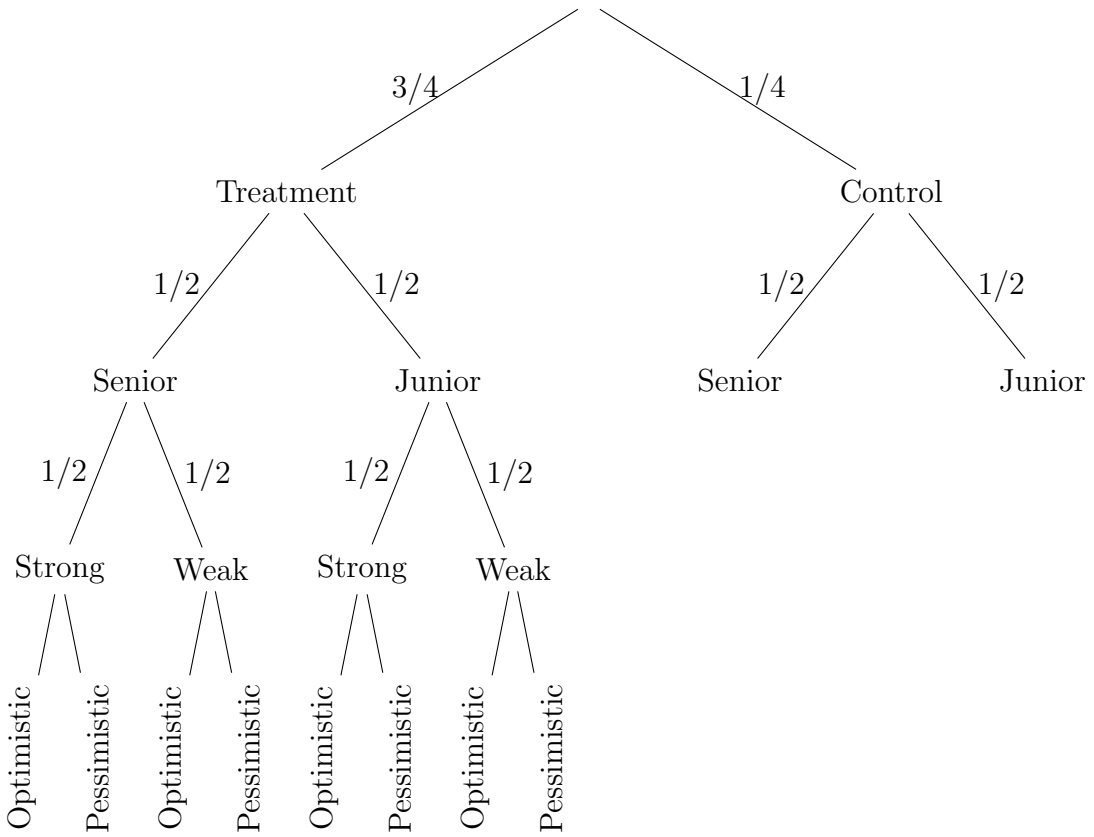


Figure 1: Levels of Randomization

Employees within the Control condition will still receive the survey materials relevant to all stages of our experiment, but no one in these divisions will actually be provided with evidence. The Control condition will serve to benchmark the transmission.

We consider a respondent "seeded" and include them in our analysis once participants reach the point in the experiment where they learn (or do not learn, for Control) about the evidence. If respondents attrit before this point, they will not be included in the analysis and we will not follow-up with them in the role of seeds.

**Stage 2: Senior vs. Junior Seed** We segment World Bank employees in our sample into two similarly-sized groups based on their pay grade: Approximately 50% of employees in our sample are "Junior," meaning they are in Grades GF and below. Approximately 50% of employees in our sample are "Senior," meaning they are in Grades GG and above.

Half of all divisions within each of the Treatment and Control conditions will be randomly assigned to "Senior Seed", while the remaining half will be assigned to "Junior Seed". Seeds in Senior Seed divisions will be randomly selected among those in Senior positions

in the division. Seeds in Junior Seed divisions will be randomly selected among those in Junior positions. The random assignment of divisions to seed rank will allow us to more robustly study the role of hierarchy in evidence transmission; that is, how evidence-sharing with Senior versus Junior individuals in a bureaucracy differentially affects transmission.

**Step 3: Strong versus Weak Credibility** We will randomly assign half of divisions in each Treatment group to the "Strong Credibility" treatment, in which they will receive information that highlights the credibility of the evidence source. In particular, the information they receive will point to the fact that the evidence they are learning about was produced by researchers at MIT and was published in the academic journal *Science*. They will also learn about Dell'Acqua et al. (2023), which uses a similar experimental design and finds consistent results thereby reinforcing the credibility of Noy and Zhang (2023). The remaining half of divisions will be assigned to the "Low Credibility" treatment and will not receive any information that highlights the credibility of the evidence source. To estimate a first stage effect of this treatment, we include a question on how credible participants find the results of the paper.

This variation will allow us to causally explore the role of credibility in the dissemination of evidence.

**Stage 4: Optimistic versus Pessimistic Second-Order Perceptions** We have already completed a "Preliminary World Bank ChatGPT Survey" which asked about current levels of engagement with ChatGPT and asked about preferences regarding the use of ChatGPT. World Bank employees invited to participate in this survey will not be included in the main experiment, and include: Approximately 300 employees in Human Resources in the Washington, DC office (a division which we expected to be unusually self-contained); approximately 100 employees in the India country office; and approximately 90 employees in the Australia country office.

Key questions in the "Preliminary World Bank ChatGPT Survey" include:

1. *Status Quo Use of ChatGPT:* "Have you used ChatGPT in the past for your work" [Y/N]

2. *ChatGPT Preferences:* "ChatGPT is something that we should be using often in our work" [Strongly disagree,...,Strongly agree]

83 individuals completed our survey and responded to both survey questions. We selected two sub-samples of 40 responses to these surveys, to compile "Optimistic" and "Pessimistic" sub-samples. In our Optimistic sample, 75% report using ChatGPT for their work and 80%

either agree or strongly agree that ChatGPT is something that we should be using often in our work. In our Pessimistic sample, 20% report using ChatGPT for their work and 18% either agree or strongly agree that ChatGPT is something that we should be using often in our work.

We will randomly assign Seeds to learn about either the Optimistic or Pessimistic responses in our Preliminary World Bank ChatGPT Survey.[4] The random assignment of Seeds to differential information about their colleagues' use and approval of ChatGPT allows us to exogenously manipulate second-order beliefs about peer perceptions of ChatGPT. As such, this treatment will allow us to causally determine whether second-order beliefs that are more or less positive about the evidence affects decisions to transmit the evidence.

We ask incentivized predictions of second-order beliefs on both ChatGPT use and values, as well as predictions about the evidence, which allows us to estimate a first stage effect of this treatment.

**Saturation and stratification:** Importantly, we will employ a saturation design. There are 42 total Program Management Units (PMUs), which house the divisions over which we randomly assign treatments. We will randomly assign each PMU to a level of treatment saturation (50%, 75%, or 100%). This will determine the proportion of divisions within the PMU that will be randomly assigned to Treatment (seeded with evidence).

Finally, we will stratify treatment assignments by division size (quantiles).

# 3 Recruitment

Individuals will be invited to participate in our survey via email.

First, selected participants will receive an email from the World Bank's Development Research Group. This email will let participants know that they will be receiving an email from Qualtrics later that day inviting them to participate in a short survey. The email serves to add legitimacy to the survey.

Then, the same day, selected participants will receive an email from Qualtrics including a link to the survey.

All participants will be told that they will be paid $10 (or equivalent depending on the local currency) for participating and will be given a deadline by which to complete the survey.

[4] We will make explicit that we are only sharing a subset of responses and we will follow up with everyone once the study is complete to inform them of the responses among the full sample.

# 4 Pre-Analysis Registration Procedure

We plan to finalize all specifications (including outcome and control variables) for the survey data using a "fake" randomization procedure (Olken, 2015). After all survey data has been collected, we will conduct a "fake" randomization using the randomization code, but with a different seed than the one used for the study. Using these "fake" treatment variables, we will generate all specifications and post these results on the AEA registry. We may use a Double-LASSO procedure a la Chernozhukov et al. (2018) to refine the controls. We will generate results for the specifications using the actual randomization only once this part of the pre-analysis plan has been filed.

In the following sections, we outline initial thoughts on outcome variables and specifications. This will be refined and pre-registered using the "fake" randomization procedure.

# 5 Outcomes

## 5.1 Primary Outcome: Standardized index of awareness of evidence

In the follow-up contact survey administered to all eligible World Bank employees in each division except the seed, we will elicit a number of outcomes related to the contacts' awareness of the evidence. This will provide a measure of the degree to which evidence disseminated among the contacts.

We will ask the following questions in the survey:

1. On a scale from -10 to 10, where -10 stands for extremely negative and 10 stands for extremely positive, what do you think is the impact of ChatGPT on skilled workers' productivity?

2. On a scale from 0 to 100, where 0 stands for not at all confident and 100 stands for extremely confident, how confident are you in your previous answer?

3. Have you recently discussed (either online or in person) the impact of generative AI tools such as ChatGPT on workers' productivity with your colleagues? [Binary yes/no response]

4. Are you familiar with any research studies discussing the impact of ChatGPT on productivity? [Binary yes/no response]

Our primary outcome will then be a composite measure of the responses to these four questions. To incorporate our confidence question, we will flip the value of all responses when the respondent's answer to (1) was negative; for instance, if the respondent answered -5 for (1) and 70 for (2), then we would input "-70" as the measure of confidence. We plan to standardize responses to each question and take the average of the four. Finally, we plan to standardize this resulting average.

## 5.2 Secondary Outcome: Standardized index of behavior change

In the follow-up contact survey, we will also elicit several measures of behavior change in response to evidence transmission.

First, we will ask the two questions about use of ChatGPT:

1. Do you have a ChatGPT account? [Y/N]

2. Have you used ChatGPT in the past for your work? [Y/N]

Second, we will ask individuals to express interest in exploring opportunities or collaborating with working groups at the World Bank on artificial intelligence.

Third, we will introduce an incentivized Becker-Degroot-Marshak elicitation to elicit a willingness to pay for GPT-4, up to $10.

We will include a secondary outcome as the composite measure of the responses to these questions. In particular, we plan to standardize responses to each question and take the average of the four responses. Finally, we plan to standardize this resulting average.

## 5.3 Secondary Outcome: Infographic engagement

We will additionally attach unique Google Analytics UTM tags for an infographic and journal article PDF which we share with each Treated Seed. The infographic summarizes the evidence (see Appendix Section A for the infographic we will share). The journal article is the PDF copy of Noy and Zhang (2023). By examining the number of clicks per unique link, we will be able to identify differential engagement with the evidence source. This secondary outcome will be reflected by the number of clicks on the links provided by each seed.

Importantly, we will not be able to disentangle clicks generated by the seed versus by their contacts. As such, this outcome more broadly reflects engagement with the infographic and journal article, rather than transmission per se.

### 5.4 Secondary Outcome: Awareness of colleagues' beliefs about ChatGPT

Finally, we will ask the same incentivized questions that we asked Seeds to elicit beliefs about colleagues' views on the use of ChatGPT. We are interested in whether this "evidence" about colleagues' beliefs transmitted to Contacts.

We ask contacts to predict their colleagues' responses to the following two questions:

1. *Status Quo Use of ChatGPT:* "Have you used ChatGPT in the past for your work" [Y/N]
2. *ChatGPT Preferences:* "ChatGPT is something that we should be using often in our work" [Strongly disagree,...,Strongly]

We will create a standardized index based on responses to these questions, as described above.

## 6 Analysis

### 6.1 Primary Specification

We plan to estimate the following linear probability model across all contacts surveyed to identify whether our treatment conditions influence evidence transmission among contact $i$ in division $d$:

$$Aware_{id} = \alpha TreatmentA_d + \beta TreatmentB_d + \gamma X_{id} + \delta T' + FE_{size} + \epsilon_{id} \quad (1)$$

where $Aware_{id}$ is the standardized index of evidence awareness as described in Section 5.1. $TreatmentA_d$ and $TreatmentB_d$ are indicators equal to one for the two treatment conditions in each branch in Figure 1: Senior, Junior; Strong, Weak; and Optimistic, Pessimistic. $X_{id}$ is a vector of control variables, including contact rank; seed and contact priors, status quo, and preferences regarding ChatGPT use; seed and contact gender; Seed time spent in the office; and seed and contact education. $T'$ is a vector of all other treatment indicators not included in $TreatmentA_d$ and $TreatmentB_d$. $FE_{size}$ captures fixed effects for division size. We will cluster robust standard errors at the division level.

Our primary hypothesis of interest is whether the coefficients on $TreatmentA_d$ and $TreatmentB_d$ are equal.

To account for multiple hypothesis testing, we will apply a bootstrap-based procedure to control the Family-Wise Error Rate, where our "family" will consist of these three tests, where $TreatmentA_d$ and $TreatmentB_d$ correspond to:

1. Senior, Junior
2. Strong, Weak
3. Optimistic, Pessimistic

Note that for benchmarking purposes we will also run a version of Equation 1 where we include only one treatment indicator equal to one if the contact is in a division with a treated seed. This will provide an estimate of the degree to which the evidence transmitted overall, provided exposure to the evidence.

## 6.2 Exploratory Analyses

The design of our field experiment allows for addressing a rich set of exploratory questions in addition to our primary tests described above. We are less powered to detect effects for these analyses and therefore don't expect to include them among our primary tests, where we will apply multiple hypothesis corrections. Because our treatment of these analyses is less stringent than that for our primary tests, we will indicate in any project reports that the analyses are exploratory.

These exploratory analyses include nine broad categories:

**1. Secondary Outcome: Behavior Change** Our primary analyses explore the impact of different treatments on a standardized index of awareness of evidence. We will also conduct exploratory analyses using our standardized index of behavior change, as described in Section 5.2. We will use the same specification and tests described in Section 6.1 but with behavior change rather than awareness as the outcome of interest.

**2. Secondary Outcome: Infographic and journal article PDF engagement** We will further conduct exploratory analyses using the number of clicks on the infographic and journal article PDF links provided to each seed, as described in Section 5.3. We will use the same specification and tests described in Section 6.1 but with clicks as the outcome of interest. Importantly, this analysis will be done at the seed rather than contact level.

**3. Secondary Outcome: Awareness of colleagues' beliefs about ChatGPT** We will conduct another exploratory analysis using awareness of colleagues' beliefs about ChatGPT as an outcome, as described in Section 5.4. We will use the same specification and tests described in Section 6.1 but with awareness of colleagues' beliefs as the outcome of interest.

**4. Follow-up Survey** We will also send a follow-up survey to Seeds to understand the barriers to evidence use. We will produce a descriptive analysis based on the findings from this survey. If there is sufficient evidence sharing in our main experiment, we will also relate the responses in this follow-up survey to decisions to share evidence.

**5. Predictors of Evidence Dissemination** Our treatment conditions in which we vary characteristics of the evidence allow us to causally estimate determinants of evidence transmission. However, we also collect rich data on other seed and contact characteristics that allow us to explore further predictors of transmission. In particular, we plan to include key employee characteristics—for instance, prior beliefs or World Bank tenure—in place of the Treatment indicator in Equation 1 to explore the association between these characteristics and evidence dissemination.

**6. Heterogeneous Treatment Effects** We are able to explore not just how these seed and contact characteristics correlate with evidence transmission, but also how they interact with our treatment conditions. Of particular interest to us is the interaction between seed and contact rank, allowing us to answer questions such as whether evidence transmits relatively more to Junior contacts when the seed is also Junior.

**7. Instrumental Variables Analysis** Two of our treatments are intended to update beliefs: Our "Strong" versus "Weak" credibility treatments are intended to update beliefs about the credibility of Noy and Zhang (2023), while our "Optimistic" versus "Pessimistic" treatments are intended to update second-order beliefs about colleagues' status quo ChatGPT use and preferences around ChatGPT. We elicit beliefs about credibility as well as second-order beliefs. As such, we can use these elicitations as a first stage in an IV regression exploring the impact of our treatments on the outcomes identified above.

**8. Saturation** Our saturation design allows us to explore not just the impact of being in a treated division, but also the impact of being in a PMU with more or fewer treated divisions. We will use this variation to create a Treatment Intensity variable. We can use this for a number of exploratory analyses. For instance, we can use this design to measure

spillover effects. We can also use it to look at transmission to employees in director-level positions with reports from across multiple divisions.

**9. Email contacts** After we implement our experiment—and assuming there is sufficient evidence transmission—we plan to reach out to the IT team at the World Bank to ask whether they could use our data to indicate how often each Contact communicates with a Treated Seed via email. This data will serve as a new Treatment Intensity variable, and we can run all of our primary and secondary specifications with this instead.

# References

**Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins**, "Double/Debiased Machine Learning for Treatment and Structural Parameters," *Econometrics Journal*, 2018, *21*, C1–C68.

**Dell'Acqua, Fabrizio, Edward McFowland, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani**, "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality," Harvard Business School Technology & Operations Mgt. Unit Working Paper 24-013 September 15 2023.

**Noy, Shakked and Whitney Zhang**, "Experimental evidence on the productivity effects of generative artificial intelligence," *Science*, 2023, *381*, 187–192.

**Olken, Benjamin**, "Promises and Perils of Pre-Analysis Plans," *Journal of Economic Perspectives*, 2015, *29* (3), 61–80.

# A Infographic