A Global Field Experiment Motivating INGOs to Evaluate Impact

Eliza Riley*, James Hodgson†, Michael G. Findley‡, and Daniel L. Nielson§

Prepared for APSA 2019, Washington D.C.
Panel: Transnational Politics of INGO Behavior
Friday, August 30, 2-3:30 pm

Abstract

In a global email field experiment using 51,000 development organizations as subjects, we tested the effects of different embedded encouragements on INGOs' response rates to a sincere invitation to consider a partnership in a randomized evaluation. Credibility conditions randomly assigned the authors of the email invitation as university professors touting past experience working with high-profile development organizations compared to students working with obscure groups. Information conditions included primes that the results might be negative, that the evaluation would be attractive to donors, that other INGOs have adopted the practice (social proof), that it would enhance the target INGO's international reputation, and that the INGO would be able to participate fully in the evaluation's design – all compared to control. The negative information treatment significantly decreased responses. Credibility and social proof both significantly increased responses. Mention of donor interests, international reputation, and priming the participatory aspect of the evaluation all produced null results.

Keywords: Non-Governmental Organizations, Development, Confirmation Bias, Evaluations, Randomized Control Trial, Social Proof, Credibility, Reputation

*Research Support Associate, MIT Political Methodology Lab, e_riley@mit.edu
†Senior Product Specialist, Qualtrics, jwh.oasis@gmail.com
‡Professor, UT-Austin Department of Government, mikefindley@utexas.edu
§Professor, BYU Department of Political Science, dan.nielson.byu@gmail.com

Introduction

Randomized control trials (RCTs), in which interventions are randomly assigned and desired outcomes are compared to a control group, have grown rapidly as a method of evaluating the impact of development programs worldwide. Encouraged by organizations such as MIT's Jameel Poverty Action Lab and the non-profit Innovations for Poverty Action, governments, inter-governmental agencies, and non-governmental organizations are increasingly subjecting their anti-poverty interventions to rigorous experimental testing. While critics of the method have raised high-profile and well-reasoned challenges (Deaton 2010, Rodrik 2009, Deaton and Cartwright 2018), the evaluation method appears to have increased in adoption exponentially over the prior two decades.

Arianna Legovini, manager of the World Bank Development Impact Evaluation (DIME), estimated at least 475 randomized controlled studies occurring across the bank as a whole in 2016. At BREAD, the flagship conference in development economics, the proportion of papers featuring RCTs surged from 8% in 2005 to 63% in 2010 (Banerjee, Duflo, and Kremer 2016). Despite their critics, randomized control trials are increasingly considered the "gold standard" for impact evaluations because they ostensibly minimize selection bias and otherwise guard against additional sources of error, for example shielding program managers from favoritism or corruption (Gertler et al. 2011, revised 2016). In addition to providing a valuable operational tool with inherent merits as a rationing mechanism, Athey and Imbens (2016) agree that in the realm of econometric analysis "randomized experiments [play] a special role in causal inference."

Despite the part played by RCTs in upending academic methods in social science, a relatively small proportion of international non-governmental organizations (INGOs) have adopted the evaluation practice. This has persisted in the face of a strong push from aid organizations to conduct randomized controlled trials (RCTs) and other forms of rigorous impact evaluation to assess program success (Evans and Wydick 2016). INGOs are critical actors in the funding and implementation of foreign or domestic aid projects (Dietrich 2013, 2016), and it is important to understand if, how, and why INGOs assess the effects of their programs. With limited funds to address beneficiaries' needs, INGOs must often maximize the efficiency and effectiveness of their efforts. Because INGOs are ostensibly accountable to their donors and beneficiaries, it is important to understand why so many have not implemented randomized impact evaluations, which promise evidence on the causal effects of the organizations' programs. Importantly, what encouragements might change this pattern?

Furthermore, despite the rapid increase in number and prominence of INGOs (Gourevitch et al. 2012, Aldashev and Vernier 2010), there are unsettled questions in the literature regarding NGO behavior and motivations. Conducting scientific research on INGO motivations and incentives is vital to understanding the practices and strategies of these influential development actors (Aldashev and Verdier 2010; Doh and Teegen 2002). We therefore conducted an international field experiment to examine which of the theorized motivators of INGO behavior – credibility, social proof, mention of donor interests, international reputation, confirmation bias, or participation – actually motivates an INGO's willingness to engage in an impact evaluation. To capture the effects of different encouragements on response rates, we conducted a two-stage global field experiment using more than 51,000 aid organizations from the World Association of Non-governmental Organizations (WANGO) as subjects.

The experiment occurred in the context of a sincere invitation from researchers to INGOs to consider a partnership to evaluate program impact. Specifically, we sought to answer the following questions: What information and incentives motivate INGOs to express interest in undertaking randomized impact evaluation? Using difference-in-means tests and regression analysis, we found statistically significant evidence showing that partner credibility and social proof can increase INGO willingness, though INGOs are susceptible to confirmation bias and are less likely to participate when primed that results might be negative. Information priming donor interests, international reputation, or opportunities to participate in evaluation design all produced null effects.

Our findings reveal on the one hand that development organizations seem to care more about the realized benefits of randomized impact evaluation (conducted by a credible partner) and conforming to INGO norms of practice rather than using the evaluation as a means for self-promotion (becoming more attractive to donors, increasing NGO standing in the international community, or being deeply involved in the evaluation). However, being reminded that unbiased results of evaluations can prove a program ineffective made NGOs less inclined to participate, suggesting that NGOs care more about the realized benefits of randomized evaluations when they are expected to confirm the NGOs' goals.

Literature and Theory

A growing literature examines what motivates NGO behavior (Bush and Hadden forthcoming; Peterson et al. 2018; Stroup and Wong 2017; Gourevitch et al. 2012; Ohanyan 2009; Ebrahim 2001). Some studies point to reputation as the key driver (Hyde 2012). Others ask whether credibility is most important (Gourevitch

et al. 2012). An NGO's willingness to be transparent has also been the subject of recent studies on NGO behavior. For example, instituting auditing as a standard practice for evaluation has apparently made NGOs more open and accountable (Buthe 2012). A broad literature also discusses the importance (Banjeree, Duflo and Kremer 2016), albeit imperfections (Ravallion 2012), of evaluating the impact of NGOs. While fewer studies have explored why and when NGOs choose to evaluate impact, much has been written to confirm both the importance of assessing NGO impact and on the difficulty of so doing (Hailey and James 2003). Understanding the conditions under which development organizations accept feedback and are willing to update is key to understanding the role of these influential global actors in development (Aldashev and Verdier 2010; Doh and Teegen 2002).

We will consider reasons for the relative paucity of INGO impact evaluations below, but first it is important to explore NGOs' motivations to be seen as accountable. Recent literature has examined the incentives for and circumstances under which NGOs establish and maintain credibility. In their comprehensive treatment of the topic, Gourevitch et al. (2012, Ch. 1) discuss the drivers of NGO credibility. Virtue and intrinsic motivations may matter, but external conditions often underpin the quest for transnational credibility. These may include common interests between the NGO and its audience, when claims are backed by observable costly effort, if they suffer penalties for misrepresentation (Quelch and Laidler-Kylander 2006), and external verification such as scrutiny by the media. Gourevitch et al. (2012) go on to note that NGOs "are not passive actors constrained by the conditions to be credible or not. Indeed, they are active shapers of their own images, reputations, structures, and, thus, credibility."

Increasing transparency, for example by participating in impact studies and scientifically evaluating programs through randomized field experiments, might be an important way to signal that the NGO itself believes its programs will work. As NGOs seek to increase their credibility, they may make greater efforts to invite scrutiny from skeptical audiences. Of course, such invited scrutiny can easily backfire. NGOs are all too aware that skeptics are often eager to identify hypocrisy when actions contradict virtuous claims and goals. However, as NGOs develop, they acquire interests as organizations that may deflect them from their ultimate, virtuous cause (Bob 2005, Cooley and Ron 2002, Kennedy 2004). For example, reports of high-paid executives and occasional financial scandals threaten NGOs because they darken the public's perception of NGOs' virtue and thus harm their subsequent credibility (Gibelman and Gelman 2004; O'Neill 2009).

NGOs can implement an array of accountability mechanisms to bolster their credibility (Ebrahim 2003). Nonetheless, NGO transparency and accountability have been called into question as reports of NGOs with

high overhead costs and limited impact become more widespread (Ebrahim 2005; Fry 2006; Kilby 2006). Additionally, many NGOs often compete in resource-scarce environments, which causes competition for donor funding and can even shift NGOs' focus from aiding their beneficiaries to pleasing their funders (Christensen and Ebrahim 2006; Johnson and Prakash 2007). NGOs may even act opportunistically when approached by a donor with large perceived amounts of wealth (Reed et al. 2015). Donor funding may thus have a large effect on NGO behavior. Indeed, there is a significant body of work that studies the effect of donor attractiveness on an NGO's willingness to be transparent. A dominant theory focuses on how the growing number of NGOs on the transnational stage has made the industry competitive. NGOs compete with many others for funding, causing them to be less honest in their evaluations for fear of losing their financial backing (Cooley and Ron 2010).

The presence of other INGOs within transnational advocacy networks can also influence INGO accountability and, by extension, INGO credibility. The increase in the number of NGOs globally has created distortions in their incentives to be transparent (Bilodeau and Slivinski 1997; Castaneda et al 2008; Gourevitch et al. 2012). More NGOs and, as a result, more competition, may impact how organizations behave towards each other. Some scholars are optimistic that more NGOs mean more peer pressure for coordination (Mathews 1997; Simmons 1998). Stroup and Murdie (2012) have found that, in the United States, NGOs that compete for resources ultimately end up resembling each other. As a result, NGOs may imitate each other's behavior to get more funding. Alternatively, NGOs may also be less willing to share information in a competitive environment (Cooley and Ron 2002; Aldashev and Verdier 2010). The presence of other organizations may impact how an NGO presents its accountability mechanisms.

NGO behavior is affected not only by the presence of other NGOs but also by their position in the NGO community. Although NGOs rely on donors for funding, donors rely on NGOs to enhance their reputations in the development community (Ebrahim 2002). While incentives for NGOs to increase their reputations have generated limited systematic testing, literature exists on the psychology of reputation-seeking behavior. In one study, subjects were asked to donate to a development NGO. Those who were told that their names and donation amounts would be publicized donated more (Reinstein and Riener 2012). While this study would, of course, need to be replicated with NGOs as the subjects, the authors point out that, in closely knit networks, reputations most likely have a strong effect on choices to act transparently.

On the whole, NGOs may be seen as "principled actors" that have incentives to reform in order to accomplish their goals and keep their integrity and credibility intact (Parks 2013). This contention suggests that

reputation and credibility in the international community is a stronger incentive for NGO reform (Gourevitch et al. 2012). Furthermore, NGOs have been shown to respond to international pressures to increase transparency and to try to keep up with their peers who are meeting the new standards (Parks 2013). These results indicate that NGOs may be influenced by social proof and may modify their behaviors to conform to norms and conventions displayed in the behavior of their peers.

Regardless of their motivations for transparency or reforms, development organizations may update their practices as they receive more information about their effectiveness. Existing empirical evidence, however, suggests that such updating is much more likely if INGOs anticipate positive findings. In an experiment on microfinance institutions (MFIs), most of which fit comfortably under the INGO moniker, researchers found that MFIs who received negative information on the effectiveness of microfinance were less likely to elect to learn more about participating in an impact evaluation (Brigham et al. 2013). This paper seeks to replicate this study in the context of a broader sample of INGOs in order to generate more evidence on NGO behavior and the motivation to learn about practice effectiveness.

While randomized evaluations of INGO programs is not yet widespread, the trend is pointing upward, and increasing pressure is mounting for INGOs to produce credible evidence of their effectiveness. "If the [NGO] sector wants to properly serve local populations, it needs to improve how it collects evidence," Evans and Wydick (2016) asserted in a World Bank blog. External researchers and journalists are not alone in wanting to understand whether INGOs and development organizations are effective in achieving their aims.

Research Expectations

From the existing literature on INGO behavior and credibility, it is clear that motivations for INGO behavior may stem from many sources, and causal identification of the effectiveness of different incentives is underdeveloped. As the results of impact evaluations continue to challenge conventional wisdom on traditional methods of international development, it is important to know when and if INGOs consider participation in randomized evaluations in the first place.

We developed six hypotheses on the relationship between information regarding impact evaluation and NGOs' responses to our invitation to consider a possible partnership. These hypotheses form expectations around the credibility of the invitation and subsequent evaluation, the possibility of negative findings, and

information containing encouragements or motivations for pursuing a randomized evaluation: donor attractiveness, social proof of other INGO behavior, international reputation, and INGO participation in evaluation design. We develop each hypothesis below.

Professor vs. Student Credibility

For many INGOs, a lack of expertise and financial resources act as impediments to effective impact assessment. Until knowledge of how to conduct a randomized evaluation becomes more widespread, researchers based at universities are the primary practitioners of field experiments. An INGO may gain expertise and acquire rigorous assessments of program effectiveness by collaborating with such university researchers. However, and this goes to the heart of arguments made in Gourevitch et al. (2012), not all university researchers are equally credible. While INGOs may be primarily concerned about their own credibility, this concern should logically extend to the credibility of their potential partners. And researchers with greater seniority, expertise, and prior high-profile partners should naturally be more credible than researchers with less of those quantities. This rationale drove our first hypothesis.

> ***Hypothesis 1: Partner Credibility.*** *INGOs will more likely respond to partnership invitations from university professors touting many years of experience working with high-profile organizations compared to invitations from less-experienced student researchers who have worked with low-profile organizations.*

Negative Findings/Confirmation Bias

An important research stream in social and cognitive psychology holds that people are significantly more likely to update when they encounter information that conforms with their prior beliefs and to ignore or resist information that contradicts their priors (Lord, Ross and Lepper 1979; Kunda 1990; Nickerson 1998). Indeed, people's very cognitive processes are biased from the outset, attending to evidence chiefly when it reinforces their priors in a phenomenon known as motivated reasoning (Kunda 1990). Some research has even identified a backlash effect: when people encounter information that contradicts their priors, they actually dig in and come to hold their priors with greater certainty (Lord, Ross and Lepper 1979).

It is likely that this generalized phenomenon will manifest in INGO representatives. The implications here are straightforward: INGOs should be much more inclined to express interest in randomized evaluation if they believe that the results of the assessment will be positive and will therefore reinforce the INGO's goals. If, however, INGO personnel are told that the results might indeed be negative, interest in randomized

evaluation should diminish. This tendency will likely be reinforced by the fact that INGOs often rely on external funding and may be very sensitive to the reported success of their programs. Thus, some INGOs may feel pressured to avoid revealing the inefficiencies of their initiatives in an attempt to meet objectives and secure funding for the future. Negative results directly threaten their mission, giving rise to the next hypothesis.

> ***Hypothesis 2: Negative Results.*** *If an INGO is primed to expect that the results of an impact assessment will be negative, the INGO will be less inclined to express interest in randomized evaluation.*

Financial Pressure from Benefactors

As noted above, INGOs are likely very sensitive to perceptions of and signals from potential donors. If an INGO receives information that performing impact evaluations may make them more attractive to donors, it should therefore be more likely to respond and request further information. INGOs are often financially constrained. When they are forced to compete in large applicant pools for funding, they might consider undergoing impact evaluations if they perceive that doing so increases their prospects for receiving funding. If the perceived benefit of attracting donors, including large donors like the World Bank or the U.S. Agency for International Development, exceeds the anticipated costs of undergoing evaluations, INGOs should respond by requesting further information, giving rise to the next hypothesis.

> ***Hypothesis 3: Donor Attractiveness.*** *Information suggesting that randomized evaluation will improve INGOs' attractiveness to donors should increase subjects' response rates and requests for additional information.*

Emulation of Other INGOs

Also as noted above, INGOs exist in a complex international network of private donors, government aid agencies, international organizations, partner institutions, and recipient governments and citizens. To survive and thrive in this milieu, analysts have asserted that attention to transnational reputation proves key (Gourevitch et al. 2012). One shorthand way to guard and build credibility entails identifying and emulating the best practices of other prominent INGOs. INGOs may therefore be responsive to what psychologists call "social proof" (Cialdini 1993), in which descriptive reports of peers' common behavior causes conformance to those norms, motivating the next hypothesis.

> ***Hypothesis 4: Social Proof.*** *Information reporting a descriptive norm that other promi-*

nent INGOs routinely undertake randomized evaluation will increase INGOs' responsiveness and requests for additional information.

Global Reputation Building

While descriptive norms may signal how peers behave normally, such information may mask what others see as actually appropriate. While many in the same circle may take specific actions routinely, those actions may nevertheless be broadly viewed as inappropriate or in violation of higher values prescribing moral behavior. Thus, "going with the flow" may not automatically enhance reputation, and actors may search for other ways to indicate that they are paragons within the broader network. Global standards may provide such a signal (Büthe and Mattli 2013).

While the logic of international standards as a means of reputation building shares similarities with social proof, it also evinces the key distinction that the legitimacy of action emanates from compliance with specific rules rather than conformity to common behavior. Following such rules may not be normal, but may instead signal greater commitment to right action and thus may more sharply enhance reputation. This logic underlies the following hypothesis.

Hypothesis 5: International Reputation. *Information priming global standards of appropriate action centered on randomized evaluation should increase INGO responsiveness and interest in further information on RCT partnership.*

Participation in Evaluation Design

One of the great challenges in persuading organizations to undertake rigorous program evaluation centers on the possibility that the results of the assessment may be negative. However, these negative results may themselves be in error, which is especially concerning to organizations contemplating rigorous evaluation. The reasons for the negative findings may involve errors or misunderstandings of the independent research team about the aims, measures, recipient contexts, and other contingencies revolving around the intervention.

While not necessarily articulated in this way, the sentiment seems to be the following: "Why should we trust you researchers with our reputation if you do not understand what we are actually doing on the ground?" One potential means of mitigating this concern might involve the full participation of INGO personnel in the design of the evaluation. While independent researchers can still conduct the assessment using the best scientific standards, the measurement of outcomes, consideration of context, and interfacing

with program recipients can be effectively guided by the INGO that undertakes the intervention. A sincere promise for such collaborative partnership motivated the final hypothesis.

> ***Hypothesis 6: Participatory Aspect.*** *Reassurance that personnel from the target INGO will be closely involved in the design of the randomized evaluation should increase response rates to the invitation and heighten requests for additional information.*

These hypotheses guided the design of the field experiment. We sought to probe the effects of information about randomized evaluation on INGOs' expressions of interest in an RCT to assess program effectiveness.

Research Design

This study was conducted as a global email field experiment. Subjects included more than 51,000 aid organizations from the World Association of Non-governmental Organizations (WANGO), a large, credible, publicly available database. In the database, the NGOs have been separated into the region of the world where they operate. These regions include (Sub-Saharan) Africa, Asia, Central America, Europe, Middle East/North Africa (MENA), North America, the Pacific, and South America. We focused specifically on organizations categorized as international organizations, government institutions, civil society organizations, training and research centers, information providers, and grant providers. These categories represent the types of organizations closest to the traditional definition of INGOs, usually not-for-profit organizations dedicated to serving a target community. We excluded other organizations that did not fit this description, such as development consulting firms and private-sector support organizations. While we recognize that WANGO members likely do not constitute a fully representative sample of all INGOs, the breadth and sheer size of WANGO's membership suggests that it broadly reflects INGOs globally.

The experiment is a 2 x 2 x 5 full factorial design with a credibility treatment, a confirmation bias treatment, and four information treatments built into the language of the email. The 2 x 2 x 5 factorial design produces 20 possible combinations of treatments and control groups. The credibility condition varies whether the email is sent by (0) an inexperienced undergraduate student compared to an (1) experienced university professor. The confirmation bias condition suggests that the results from the impact evaluation might be (1) negative compared to (0) control with no additional information included about results.

The information conditions are randomly assigned statements designed to encourage response, including: (0) a control/placebo with general encouragement information that essentially repeats the information

already provided in different words, or statements that (1) "donors value rigorous impact evaluations," (2) "other INGOs are engaging in impact evaluation," (3) "INGOs that engage in rigorous impact evaluation are more likely to be certified as responsible by INGO standards bodies," or (4) "any [partner] organization... will be deeply involved in the design of the evaluation." After being separated into their respective randomization blocks based on world region, the INGOs were randomly assigned to one of the 20 possible combinations of experimental conditions. Each of the three factorial conditions were force balanced within blocks so that a nearly equal number of each condition was evenly distributed across regions. It is worth emphasizing here that one third of all subject INGOs (17,860 of 51,931) were assigned to the encouragement information placebo condition in order to mitigate concerns about multiple comparisons bias.

After completing the randomization, the INGOs received an email with one of the randomly assigned conditions. The email invited INGOs to click on an embedded link to express interest in the evaluation partnership and to complete a short survey providing basic information about the organization. In the body of the email invitation, we explained that we were gauging the INGO's interest in possibly working to undergo an impact evaluation. The invitation offered no promises about future partnerships but rather made clear that this was a preliminary contact to explore mutual interest. In the invitation email, we defined randomized evaluations and explained their importance. INGOs were instructed to click on the link and fill out the survey should they be interested in receiving additional information about an evaluation. INGOs that chose not to continue further correspondence were deleted from our email list, and INGOs expressing continued interest in potential impact evaluations were noted. Follow-up contacts are on-going.

In constructing the treatment emails, we feared that our writing style might be inaccessible for non-native English speakers. However, many of these INGOs are international organizations and may operate at least partially in English. As with all randomized control trials, our treatments may provide evidence on causal effects of encouragements for impact evaluations on INGO behavior. However, since we did not have much background information on the INGOs, it proved difficult to effectively identify how specific characteristics of an INGO might correlate with its response. An INGO may or may not be interested because of hidden factors, such as a lack of funds, an opaque internal evaluation system, prior exposure to impact evaluations, or a mistrust of our organization. We rely on the logic of randomization, which anticipates that such confounds will be balanced across conditions in expectation. The large number of subjects involved reassures us on this score, but we acknowledge that imbalances may have occurred on these unobserved possible confounds.

We hoped to clarify some of these causal mechanisms during qualitative follow-up exchanges. Thus, in

addition to email communication, we sought to gather qualitative information from interested INGOs to help shed light on causal mechanisms with a Qualtrics survey. We asked all interested respondents to fill out an online Qualtrics survey, which asked questions about organizational features – the number of full-time employees, the volume of revenues, the identification of principal donors, the number of individuals they serve, and experience with past evaluation. Including control variable for these features did not qualitatively alter results.

Our experiment took place in two rounds over a 14-month period between spring of 2017 and 2018. A washout period of one year was inserted between rounds. Round 1 was conducted in the spring of 2017 at the University of Texas at Austin and Round 2 occurred in the spring of 2018 at Brigham University University in Utah. The names of the professors and students sending the email invitations were changed between rounds accordingly. The email approaches were made to the same set of WANGO members across rounds, but all subjects that responded in the Texas round were removed from the subject pool before contacts began in the BYU round. The same set of treatment assignments were used across experiment rounds. Because the research design across the rounds was substantively identical, we pooled the results shown in the main text. In the interest of transparency, we report the individual phases I and II in tables and displayed on a coefficient plot in the Appendix (Appendix Table A1, Figure A1). Results between rounds were substantively similar (though not identical).[1]

Base Email and Treatment Language

Experimental treatments were embedded in an invitation email whose text we reproduce here.

> Dear [NGO NAME]:
>
> I am writing to gauge your interest in a potential partnership for impact evaluation. I am a [student/professor] at the University of Texas at Austin in the United States. With prior partner organizations including [USAID and the World Bank/Deniva and PEDL] I have performed research to improve program assessment and performance.
>
> One of my specializations is randomized impact evaluation, in which one group of people that participates in a program is compared to a control group that does not participate. This method is similar to medical research with treatments and placebos.
>
> [treatment paragraph(s)]

[1]The key exception was the estimated effects for the Participatory Aspect treatment, as reported in the appendix. For that condition, the sign of the coefficient flipped from positive in the Texas round to negative in the BYU round; both were significant statistically, though in the BYU round only at the 0.1 level. We are uncertain about the cause of the discrepancy in effects and continue analysis to enable more informed speculation for the unanticipated difference across rounds.

> Organizations often express interest in better evaluations of impact but lack experienced partners who can help them in design and execution. If you are interested in learning more about randomized impact evaluation and potentially pursuing a partnership, I invite you to click on the link below to complete a short survey (5-10 minutes) to tell me more about your organization and level of interest.
>
> I emphasize that this initial communication is merely exploratory and not yet an invitation for partnership. We will need to learn more about each other to see if there is a good fit. To that end, I invite you to visit my website at [michael-findley.com/william-mathias.com] (through the embedded link or through an internet search for ["Michael Findley Texas Austin" / "William Mathias Texas Austin"]).
>
> I look forward to your response.
>
> Sincerely,
>
> [Michael Findley / William Mathias] [Professor, UT-Austin / Student, UT-Austin]
>
> Please click on this link to complete the survey: [Unique URL to survey embedded here]

The precise language of our intervention text can be found in Table 1 below:

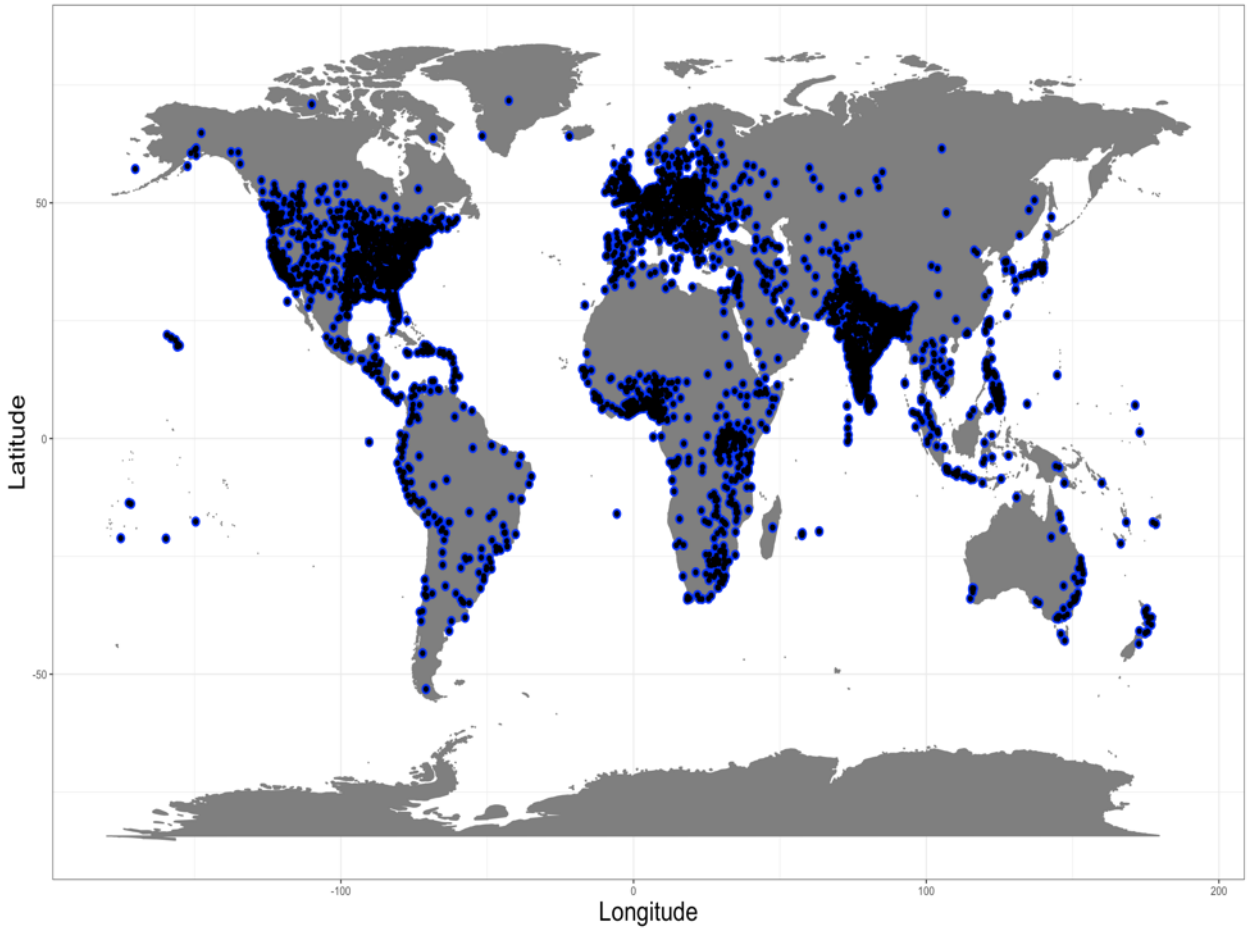
Table 1: Text of Experimental Conditions

Exp. Condition	Intervention Text
Partner Credibility	*Professor Condition: I am writing to gauge your interest in a potential partnership for impact evaluation. I am a professor at [the University of Texas at Austin (Phase 1)/ Brigham Young University (Phase 2)]. With prior partner organizations including USAID and the World Bank, I have performed research to improve program assessment and performance.* *Student Condition: I am writing to gauge your interest in a potential partnership for impact evaluation. I am a student at [the University of Texas at Austin (Phase 1)/ Brigham Young University (Phase 2)]. With prior partner organizations including [Deniva and IPD - UTexas/Liahona], I have performed research to improve program assessment and performance.*
Negative Results	*In the interest of full disclosure, however, it is worth noting that the unbiased results of randomized evaluations are sometimes negative. Such findings would mean that the program may be ineffective or perhaps even counterproductive.*
Placebo/Control	*An impact evaluation may be helpful to your organization as you consider changes to existing services. Past partners have used the data to guide decision making. The detailed data from the evaluation can be employed by your organization's leaders to make assessments, to develop methods of service delivery, and to facilitate outreach to the organization's constituents.*
Donor Attractiveness	*Rigorous impact evaluation may make your organization more attractive to donors. The online charity evaluator GiveWell.org reserves its highest ratings for non-profits that show credible evidence of impact through randomized evaluation. Large-scale donors such as the World Bank, USAID, and the UK's DFID increasingly emphasize rigorous impact evaluation in their programs and sub-contracts.*
Social Proof	*NGOs are increasingly trending to randomized impact evaluations. For example, the Abdul Latif Jameel Poverty Action Lab at the Massachusetts Institute of Technology has performed nearly 800 impact evaluations in 65 different countries. They have worked with more than 500 partners, including organizations such as the International Rescue Committee, Oxfam International, and FINCA International.*
International Reputation	*Accountability and transparency help NGOs abide by international standards for best practices. Participating in an impact evaluation enhances credibility in the NGO community and indicates concordance with standards institutions like the INGO Accountability Charter. Members of the INGO Accountability Charter include prominent NGOs such as Transparency International, Oxfam, Care, and Amnesty International.*
Participatory Aspect	*I am particularly interested in using a participatory approach for evaluation. This means that any organization we work with will be deeply involved in the design of the evaluation, including the questions we investigate, the locations studied, and the outcomes measured. This should increase the skills and capacity of the partner organization for creating new knowledge in future evaluations that its personnel might undertake.*

Data & Distribution

Figure 1 depicts the international distribution of our sample. The global field experiment spanned six continents and 233 unique countries (including territories separately). A sample listing of NGO names, country, and regions can be found in the appendix (Table A1), and a complete listing would be found among our replication files as well as on the WANGO website. We took advantage of advances in technology and globalization, which enable this Internet-enabled approach to field experiments and thus takes full advantage of the web's vast economy of scale. Indeed, an experiment on such a macro scale is one of the first of its kind, taking an important step in experimental research methods in international relations.

Figure 1. Distribution and Concentration of Global INGO Sample



We created this map using Google Maps' API to geocode all unique INGO city names to their exact longitudes and latitudes.

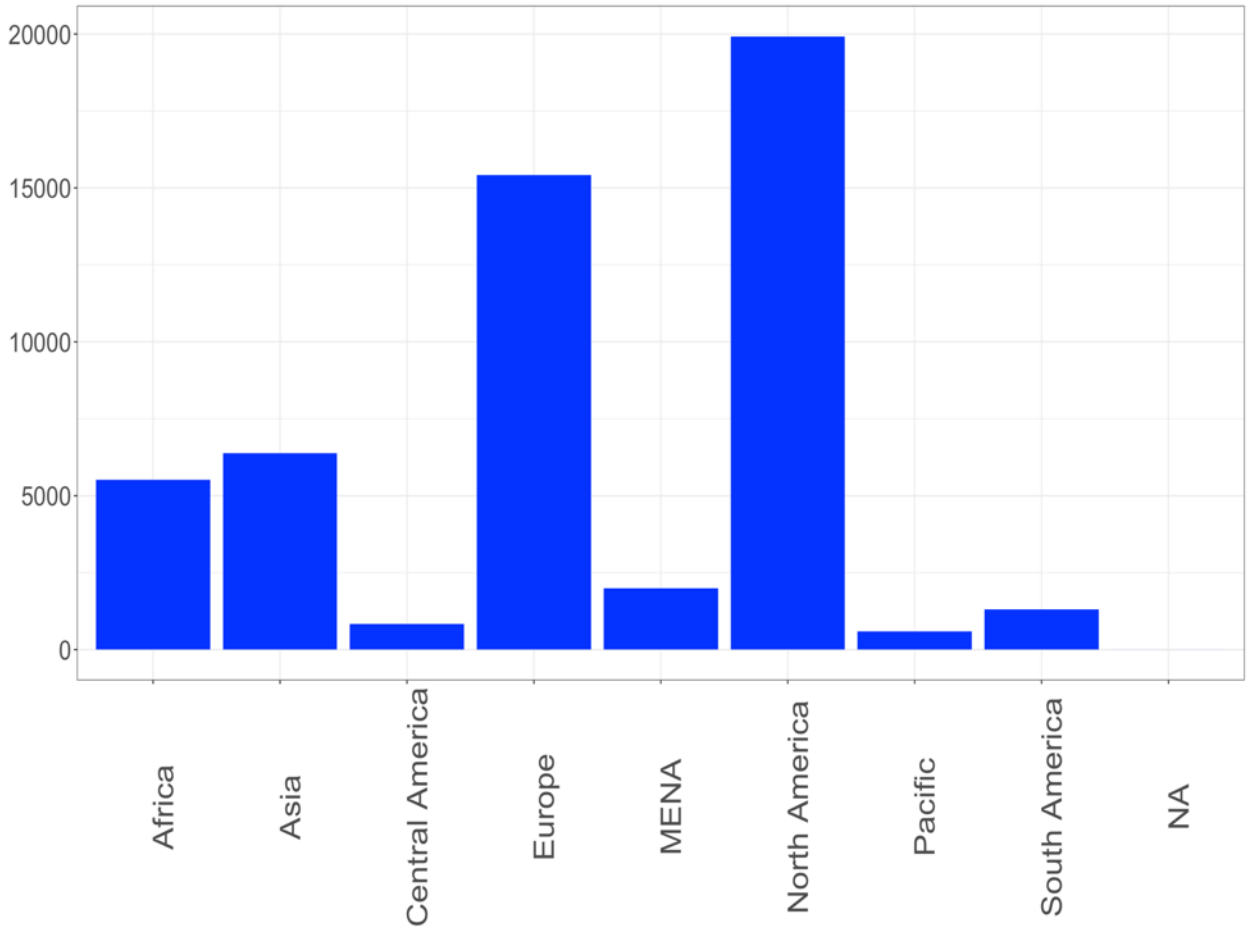
We see a heavy concentration of INGOs in North America, particularly in the Eastern United States and on the West Coast. We also see a heavy concentration in both Central and Western Europe and across all of India. Eastern and Western Africa and Southeast Asia also have notable concentrations of INGOs in our sample. Central and Southern South America are more sparsely populated with INGOs in our sample, as is Central and Western Australia, Northern and Southwestern Africa, and Eastern Europe and Central Asia. These patterns mostly track the combination of population density and per-capita income, with INGOs concentrating in densely populated and wealthy areas, and especially the combination of both.

Table 2. INGO Count by Region

Region	INGO Count
Africa	5512
Asia	6373
Central America	833
Europe	15414
MENA	1987
North America	19923
Pacific	586
South America	1302
Total	51928

Table 2 and Figure 2 also show the INGO count by broad geographic region, revealing that our sample concentrates strongly in North America, followed by Europe. Not surprisingly, the fewest INGOs are based in the Pacific, though we still include 586 INGOs from that least-populated region in our sample.

Figure 2. INGO Count by Region



Results & Discussion

We began analysis of the data using simple difference-in-means tests, shown in Table 3. For each intervention in the factorial design, we compared mean values in the treatment condition to the relevant comparison baseline. The results suggest that the Negative Results condition caused a decrease of 0.3 percentage points from the mean 3.9-percent response rate in the control condition in which no additional information about anticipated findings was presented. This difference is significant at the 0.1 threshold. While the decrease is relatively small in substantive terms and its statistical significance clears only the least rigorous conventional threshold, it nevertheless represents an 8 percent decrease from the base rate and suggests, albeit mildly, that INGOs may be sensitive to a prime warning about possible negative results. This aligns with psychologists' concerns about confirmation bias.

The Partner Credibility condition in which a professor sent the email invitation caused a 0.7 percentage-point increase in responses, to 4.1 percent, from the base rate of 3.4 percent in the student comparison condition. This represents a nearly 22 percent increase in response rate and suggests a substantively meaningful effect that is significant statistically at the 0.001 level. This result appears to indicate that INGOs respond positively to signals of potential partners' credibility as operationalized by professional status and prior experience. Additionally, the Social Proof condition raised the response rate to 4.1 percent from the base rate in the placebo information condition of 3.7 percent, for a 0.4 percentage-point increase that represents an 11 percent bump over the base rate. This difference is significant at the 0.1 threshold and suggests tenuously that priming subject INGOs about other INGOs' participation in randomized evaluation increases target INGO interest in the possible partnership.

Table 3. Condition Frequencies, Means, Differences and T-Test *P*-Values

	Compar. Mean	Compar. *N*	Treat Mean	Treat *N*	Difference	(p-value)
Negative Results	0.039	30,637	0.036	21,294	0.003	0.056
Partner Credibility	0.034	25,968	0.041	25,963	-0.007	0.000
Donor Attractiveness	0.037	17,860	0.036	8,515	0.001	0.660
International Reputation	0.037	17,860	0.036	8,522	0.001	0.831
Social Proof	0.037	17,860	0.041	8,519	-0.004	0.097
Participatory Aspect	0.037	17,860	0.039	8,515	-0.002	0.459

Next, we analyzed the data using multiple regression to compare the mean differences in response rates between intervention groups while controlling for the other experimental conditions. Figure 3 displays coefficient plots with point estimates depicted as dots and 95-percent confidence intervals shown as error bars.

Confidence intervals that overlap the zero line (at the mean of the control/comparison group) are not significant statistically at conventional levels.

Our binary dependent variable, whether the development organization filled out the survey as an expression of interest, meant that the data was suited to maximum likelihood estimation, and thus we elected to employ logistic regression in analysis. Coefficients, standard errors in parentheses, and significance thresholds for the OLS and logit models are reported in Table 4. The results largely corroborate the difference-in-means estimates in Table 3, save that the Negative Results and Social Proof coefficients are now significant at the more-exacting 0.05 threshold.

The relatively low proportion of responders to the email invitations might suggest that the estimates may be subject to bias since they are rare events (King et al. 2001). Our very large sample size helps to mitigate concern here, given that even the smallest treatment groups with lowest response rate (Donor Attractiveness) received more than 300 total responses. Nevertheless, to address the possible bias, we employed the method suggested by Firth (1993), which uses penalized likelihood as a more general approach to address and reduce problems of small-sample bias in maximum likelihood estimation. The Firth method produces results that are qualitatively similar to the reported findings in Table 4.

Care should be taken in interpreting the coefficient estimates reported in Table 4, however. We report in the main text what Muralidharan et al. (2019) call the "short" model in which interactions across factorial conditions are omitted. However, this approach may be problematic and lead to incorrect inference if the interaction effects between interventions are not zero. Thus, it is important to stipulate that the estimated coefficients in the models for each of the treatments include *both* the main effects of the intervention *along with* a composite, weighted-average effect of the given treatment interacted with each of the other experimental conditions.

Our experiment, as noted, was designed as a 5 x 2 x 2 factorial, and a "long" model with all possible interactions included is not recommended for such complicated designs. Nevertheless, to provide readers a sense of how this might matter, in Appendix Table A3 we show estimates for a "longer" model that includes all binary interactions across conditions (but omits all triple interactions). As can be seen in the table, the coefficient for the Partner Credibility/Professor condition is qualitatively similar as in the short model (and retains significance at the .01 level). The coefficient for Social Proof in the longer model increases in magnitude significantly and is now significant at the .01 level. But the coefficient for Negative Results, while

substantively similar to the short model, is no longer significant statistically at conventional levels. The models suggest that these changes resulted from the interaction of the Negative Condition with Social Proof, which bears a negative sign and is statistically significant with a p-value of 0.053.

Figure 3. Factors Motivating INGOs to Evaluate Impact, Coefficient Plot

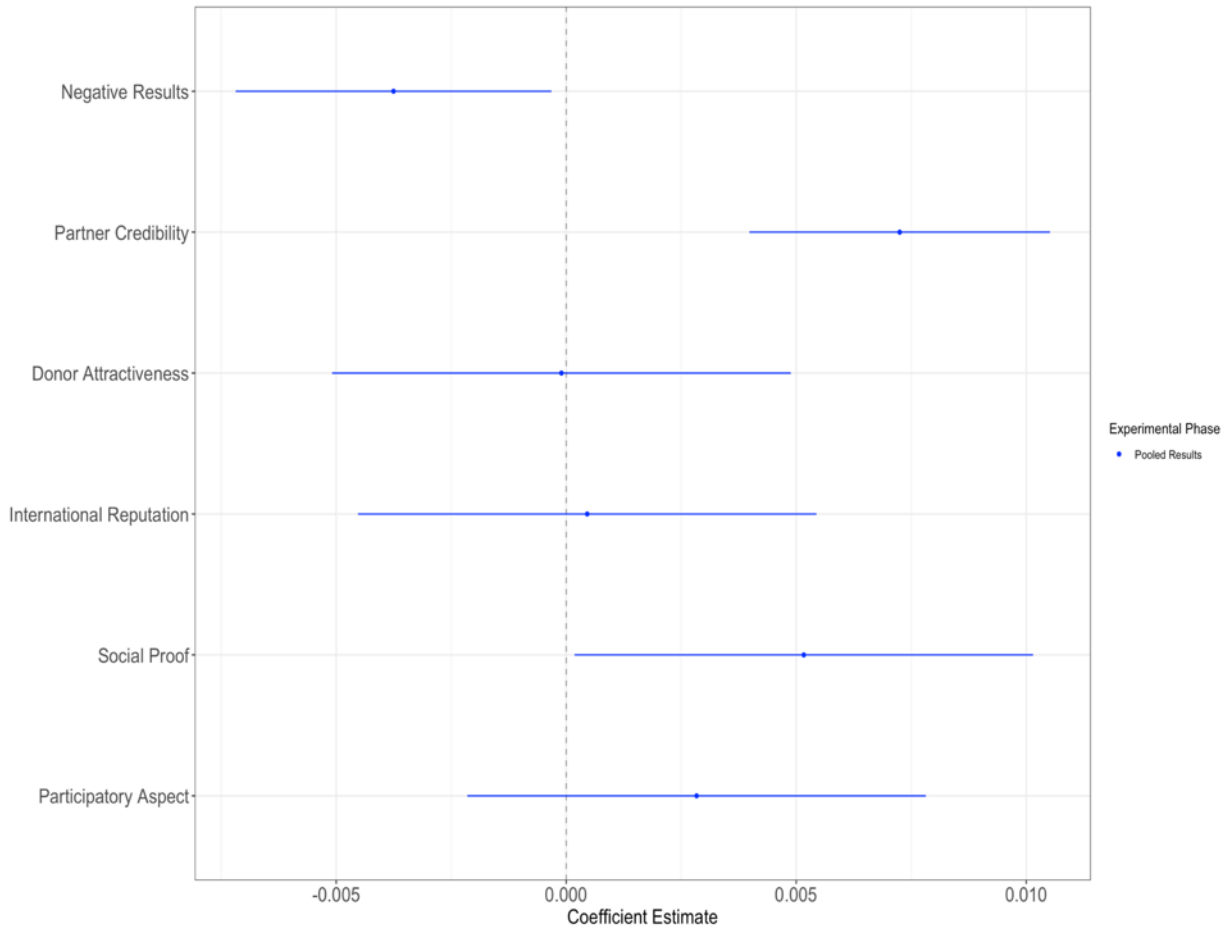


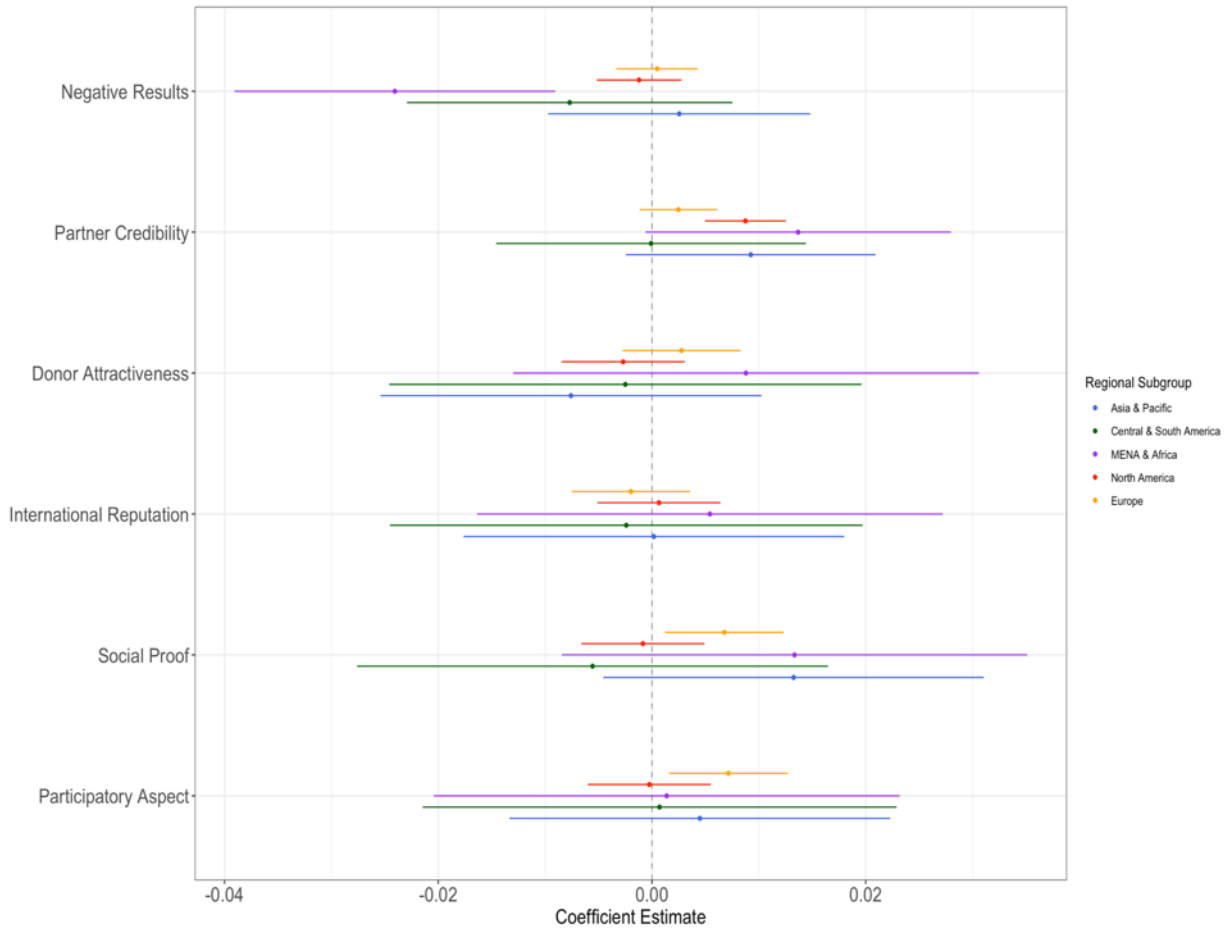
Table 4. Factors Motivating INGOs to Evaluate Impact, Regression Results

	Dependent Variable: Responded	
	OLS	*logistic*
	(1)	(2)
Negative Results	−0.004**	−0.105**
	(0.002)	(0.049)
Partner Credibility	0.007***	0.201***
	(0.002)	(0.046)
Donor Attractiveness	−0.0001	−0.004
	(0.003)	(0.072)
International Reputation	0.0005	0.012
	(0.003)	(0.071)
Social Proof	0.005**	0.139**
	(0.003)	(0.069)
Participatory Aspect	0.003	0.078
	(0.003)	(0.070)
Constant	0.034***	−3.346***
	(0.002)	(0.048)
Observations	51,931	51,931
R^2	0.001	
Adjusted R^2	0.0004	
Log Likelihood		−8,292.509
Akaike Inf. Crit.		16,599.020
Residual Std. Error	0.190 (df = 51924)	
F Statistic	4.669*** (df = 6; 51924)	
Note:	*p<0.1; **p<0.05; ***p<0.01	

Subgroup Analysis

In addition to analyzing the data on the whole, we also completed subgroup analysis, dividing our data into subsets based on the organizations' regions, so as to better understand how world region may have moderated the treatments' effects on the organizations' willingness to undergo an impact evaluation. Subsetting the analysis by geographic region revealed regional variation in INGO willingness to evaluate impact in response to the various treatments, which has interesting implications for further investigation of both cultural and geopolitical differences across and between INGOs around the world.

Figure 4. Regional Variation in Motivators to Evaluate Impact



Our subgroup analysis reveals some interesting differences in results across the regional subsets. First, there is no evidence that INGOs in Asia and Central/South America were swayed by any of the interventions, though smaller sample sizes in both regions reduce power to estimate effects and it is important to remember that absence of evidence is not the same as evidence of absence. INGOs in MENA/Africa, however, are significantly deterred by the prime that results might be Negative. North American INGOs are significantly more likely to express interest in participating in an impact evaluation when contacted by the professor in the Partner Credibility condition. This suggests that perhaps the social network of partners is more salient for that largest set of INGOs that resides in North America. The subgroup analysis also suggests that the Participatory Aspect condition seemed to particularly motivate INGOs in Europe to express interest in an evaluation partnership. European INGOs also seem to be the main subgroup driving the significant result for the Social Proof condition.

Perhaps the strongest takeaway from these subgroup results centers on what appears to be fairly broad

regional heterogeneity across INGOs in how they react to invitations for possible research collaborations. INGOs in some regions seem much more swayed by certain messages than their counterparts in other regions. This reinforces prior research on INGOs suggesting that treating them as a single type of actor behaviorally may be problematic (Reed et al. 2015). This clearly reinforces a call for more systematic research identifying the sources of behavioral heterogeneity among INGOs globally.

Table 4. Regional Subgroup Analysis, OLS Regression Results

	Dependent variable: Responded				
	Asia & Pacific	Central & South America	MENA & Africa	North America	Europe
	(1)	(2)	(3)	(4)	(5)
Negative Results	0.003	−0.008	−0.024***	−0.001	0.0005
	(0.006)	(0.008)	(0.008)	(0.002)	(0.002)
Partner Credibility	0.009	−0.0001	0.014*	0.009***	0.002
	(0.006)	(0.007)	(0.007)	(0.002)	(0.002)
Donor Attractiveness	−0.008	−0.002	0.009	−0.003	0.003
	(0.009)	(0.011)	(0.011)	(0.003)	(0.003)
International erputation	0.0002	−0.002	0.005	0.001	−0.002
	(0.009)	(0.011)	(0.011)	(0.003)	(0.003)
Social Proof	0.013	−0.006	0.013	−0.001	0.007**
	(0.009)	(0.011)	(0.011)	(0.003)	(0.003)
Participatory Aspect	0.004	0.001	0.001	−0.0003	0.007**
	(0.009)	(0.011)	(0.011)	(0.003)	(0.003)
Constant	0.059***	0.035***	0.110***	0.015***	0.010***
	(0.006)	(0.008)	(0.007)	(0.002)	(0.002)
Observations	6,959	2,135	7,499	19,923	15,414
R^2	0.001	0.001	0.002	0.001	0.001
Adjusted R^2	0.0001	−0.002	0.001	0.001	0.001
Residual Std. Error	0.248 (df = 6952)	0.171 (df = 2128)	0.315 (df = 7492)	0.136 (df = 19916)	0.115 (df = 15407)
F Statistic	1.169 (df = 6; 6952)	0.259 (df = 6; 2128)	2.393** (df = 6; 7492)	3.723*** (df = 6; 19916)	2.662** (df = 6; 15407)
Note:					*p<0.1; **p<0.05; ***p<0.01

This subgroup examination certainly paves the way for further exploration of these regional differences, particularly of the cultural or geopolitical approaches of INGO networks and regulations unique to the respective regions. For example, perhaps there are differences in the reporting of programs and funding in African INGOs that would make them particularly vulnerable to concerns about negative evaluation results. Of course, other distinguishing characteristics of INGOs are likely driving differences in willingness to respond as well, for example organizational size or funding endowment, which may also influence attitudes toward participating in an impact evaluation.

Multiple Testing Adjustment

In order to account for a possible false discovery rate, we checked the robustness of our model by adjusting for multiple comparisons. Indeed, the probability of erroneously identifying a statistically "significant" re-

sult simply due to chance increases as more hypotheses are tested. The familywise error rate calculates the probability of coming to at least one false conclusion when testing multiple hypotheses. We attempted to mitigate this problem through design by increasing the size of the placebo information condition to roughly twice the size of each of the information treatment groups. Nevertheless, to account for the familywise error rate, we chose to run the conservative Bonferonni test. The Bonferonni correction is used to reduce the chances of obtaining false-positive results when multiple pair-wise tests are performed on a single dataset.

The familywise error rate (FWER) is calculated as follows:

$$\text{FWE} \leq 1 - (1 - \mathbf{a})^{\text{c}} \tag{1}$$

Where a = the alpha significance level (i.e. .05) and c = the number of comparisons.

We used the Bonferroni correction to deal with the FWER for our multiple hypothesis tests. The Bonferroni correction reduces the possibility of getting a statistically significant result when performing multiple tests by dividing the original alpha value (in our case, .05) by the number of analyses on the dependent variable. Table 4 displays both the original p-value and the adjusted p-value for each of the intervention covariates.

Table 5. Treatment Covariates with Bonferri-corrected P-Values

Treatment Covariate	**Unadjusted P-Value**	**Bonferri-corrected P-Value**
Negative Results	.0321**	.1926
Partner Credibility	.0000128***	.0000828***
Donor Attractiveness	.9671	1.000
International Reputation	.8575	1.0000
Social Proof	.0423**	.2538
Participatory Aspect	.2652	1.000

After correcting for the familywise error rate by calculating adjusted P-values using the Bonferri correction, we see that only the Partner Credibility condition (professor versus student) retains its statistical significance.

Conclusion

The findings from the experiment offer four general lessons we wish to highlight. Our results suggest that NGOs are indeed interested in learning, but moreso under certain circumstances than others. First, while not surprising, the evidence here corroborates the contention that credibility matters substantially to IN-

GOs. This is arguably the strongest finding in the study. While INGOs are certainly expected to guard their own reputations and seek to enhance organizational credibility (Gourevitch et al. 2012), the findings here suggest that the credibility logic extends to how NGOs – particularly in North America – screen their potential partners by monitoring researchers' credibility signals, at least as measured by professional rank and experience.

Second, we see that the "social proof" or accredited participation in evaluations of other prominent INGOs significantly increases responses, particularly among European INGOs. The findings here are especially interesting because the Social Proof condition was substantively quite similar to the International Reputation condition, including in the model INGOs named as global models of behavior. One "paragon" INGO, Oxfam, was actually named explicitly in both conditions. Yet there is no evidence that INGOs on average responded to primes promising enhanced global reputation but significant evidence that they were moved by the evaluation behavior of other prominent INGOs.

Third, we find that being told of potential negative consequences has, on average, a significant deterrent effect on INGO proclivity to take part in an impact evaluation. This suggests that INGOs are significantly more interested in learning about their program effectiveness when they are not reminded that the results could be negative. At least that is one way to read the Negative Results finding. A more positive spin might observe that 3.6 percent of INGOs expressed interest in a randomized evaluation *even if* they were primed that the results might suggest that their program is ineffective. The negative prime only reduced the proportion responding by 8 percent / 3 percentage points from the control-group base rate. This means that a roughly equal number of INGOs express interest in evaluations regardless.

Finally, geographic region seems to moderate willingness to respond, which could be due to cultural or geopolitical factors unique to these regions beyond also reflecting the limited statistical power in some regions to detect treatment effects. For example, credibility is significant only in North America, and the negative information has significant negative effects only in Africa and MENA regions. Regional heterogeneity among NGOs – and organizational variation in general – clearly deserves more systematic study.

With important implications for both development policy and academic research, this global field experiment provides fresh evidence on INGO behavior and organizations' motivation to learn about practice effectiveness. Our original data quantifying the factors most likely to motivate development organizations to participate in impact evaluations contribute novel empirical evidence to answer a question widely discussed

in the INGO behavior literature: what motivates non-governmental organizations to evaluate their projects? Our analysis suggests avenues for further exploration of when and why INGOs are interested in program learning.

References

Aldashev, Gani and Thierry Verdier. 2010. "Goodwill Bazaar: NGO Competition and Giving to Development." *Journal of Development Economics* vol. 91, no. 1 (January): 48-63.

Athey, Susan and Guido Imbens. 2016. "Recursive Partitioning for Heterogeneous Causal Effects." *Proceedings of the National Academy of Sciences of the United States of America* vol. 113, no. 27 (July): 7353-60.

Banerjee, Abhijit V. and Esther Duflo. 2009. "The Experimental Approach to Development Economics." *Annual Review of Economics*, Annual Reviews vol. 1, no. 1 (May): 151-78.

Banerjee, Abhijit V., Esther Duflo, and Michael Kremer. 2016. "The Influence of Randomized Controlled Trials on Development Economics Research and on Development Policy." Paper prepared for "The State of Economics, The State of the World." Conference proceedings volume.

Bob, Clifford. 2006. *The Marketing of Rebellion: Insurgents, Media, and International Activism.* Cambridge, UK: Cambridge University Press.

Bush, Sarah and Jennifer Hadden. "Density and Decline in the Founding of International NGOs in the United States." *International Studies Quarterly*, forthcoming.

Buthe, Tim and Walter Mattli. 2011. *The New Global Rulers: The Privatization of Regulation in the World Economy.* Princeton University Press, Princeton, NJ.

Cameron, Drew B., Anjini Mishra, and Annette N. Brown. 2015. "The Growth of Impact Evaluation for International Development: How Much Have We Learned?" *Journal of Development Effectiveness* vol. 8, no. 1 (April): 1-21.

Cialdini, Robert B. 1993. *Influence: The Psychology of Persuasion.* New York: Harper Collins.

Cooley, Alexander and James Ron. 2002. "The NGO Scramble: Organizational Insecurity and the Political Economy of Transnational Action." *Harvard Kennedy School Quarterly Journal: International Security.*

Deaton, Angus, and Nancy Cartwright. 2018. "Understanding and misunderstanding randomized controlled trials." *Social Science Medicine* vol. 210: 2-21.

Deaton, Angus. 2010. "Instruments, randomization, and learning about development." *Journal of Economic Literature* vol. 48, no. 2: 424-55.

Dietrich, Simone. 2013. "Bypass or engage? Explaining donor delivery tactics in foreign aid allocation." *International Studies Quarterly* vol. 57, no. 4: 698-712.

Dietrich, Simone. 2016. "Donor political economies and the pursuit of aid effectiveness." *International Organization* vol. 70, no. 1: 65-102.

Doh, Jonathan P. and Hildy Teegan. 2003. *Globalization and NGOs: Transforming Business, Government,*

and Society. Westport, CT: Praeger Publihsers.

Ebrahim, Alnoor. 2010. "The Many Faces of Nonprofit Accountability." *Harvard Business School working paper series.*

Elsig, Manfred. 2011. "Principal–agent theory and the World Trade Organization: Complex agency and 'missing delegation."' *European Journal of International Relations* vol. 17, no. 3 (September): 495-517.

Evans, David and Bruce Wydick. 2016. "Is My NGO Having a Positive Impact?" World Bank Blogs (February).

Firth, David. 1993. "Bias reduction of maximum likelihood estimates." *Biometrika* vol. 80:27-38.

Gertler, Paul J., Sebastian Martinez, Patrick Premand, Laura B. Rawlings, and Christel M. J. Vermeersch. 2011. "Impact Evaluation in Practice." The World Bank Office of the Publisher, Washington, DC.

Gourevitch, Peter A., David A. Lake, and Janice Gross Stein. 2012. *The Credibility of Transnational NGOs: When Virtue Is Not Enough.* Cambridge, UK: Cambridge University Press.

Gibelman, Margaret and Sheldon R. Gelman. 2004. "A Loss of Credibility: Patterns of Wrongdoing Among Nongovernmental Organizations." *Voluntas: International Journal of Voluntary and Nonprofit Organizations.* vol. 15, no. 4 (December): 355-81.

Hailey, John. and James, Rick. 2003. "Learning Leaders: The Key to Learning Organizations." In Rober, Laura, Pettit, Jethro and Eade, Deborah, eds. *Development and the Learning Organization*, Oxfam, UK: 190-204.

Hawkins, Darren G., David A. Lake, Daniel L. Nielson, and Michael J. Tierney (eds.). 2006. *Delegation and Agency in International Organizations.* Cambridge, UK: Cambrige University Press.

Hyde, Susan D. 2012. "Why believe international election monitors?" In Peter A. Gourevitch, David A. Lake, and Janice Gross Stein, eds. *The Credibility of Transnational NGOs: When Virtue Is Not Enough.* Cambridge, UK: Cambridge University Press. Pp. 37-61.

Karlan, Dean S. and Jonathan Zinman. 2008. "Credit Elasticities in Less-Developed Economies: Implications for Microfinance." *American Economic Review* vol. 98, no. 3 (June): 1040-68.

Kennedy, David. 2004. *The Dark Sides of Virtue: Reassessing International Humanitarianism.* Princeton University Press, Princeton, NJ.

King, Gary, and Langche Zeng. 2001. "Logistic regression in rare events data." *Political Analysis* vol. 9, no. 2: 137-163.

Kunda, Ziva. 1990. "The case for motivated reasoning." *Psychological Bulletin* vol. 108, no. 3: 480.

Lord, Charles G., Lee Ross, and Mark R. Lepper. 1979. "Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence." *Journal of Personality and Social Psychology* vol. 37, no. 11: 2098.

Muralidharan, Karthik, Mauricio Romero, and Kaspar Wuthrich. 2019. "Factorial designs, model selection, and (incorrect) inference in randomized experiments." Unpublished manuscript.

Nickerson, Raymond S. 1998. "Confirmation bias: A ubiquitous phenomenon in many guises." *Review of General Psychology* vol. 2, no. 2: 175-220.

Nielson, Daniel L. and Michael J. Tierney. 2003. "Delegation to International Organizations: Agency Theory and World Bank Environmental Reform." *International Organization* vol. 57, no. 2 (Spring): 241-76.

Ohanyan, Anna. 2009. "Policy Wars for Peace:Network Model of NGO Behavior." *International Studies Review* vol. 11, no. 3 (September): 475-501.

O'Neill, Onora. 2009. "Ethics for Communication?" *European Journal of Philosophy* (May).

Peterson, Timothy M., Amanda Murdie, and Victor Asal. 2018. "Human Rights NGO Shaming and the Exports of Abusive States." *British Journal of Political Science* vol. 48, no. 3 (June): 767-86.

Pollack, Mark A. 1997. "Delegation, Agency, and Agenda Setting in the European Community." *International Organization* vol. 51, no. 1 (Winter): 99-134.

Quelch, John A. and Nathalie Laidler-Kylander. 2006. "The New Global Brands: Managing Non-Government Organizations in the 21st Century." South-Western College Publishing Group, Nashville, TN.

Ravillion, Martin. 2012. "Fighting Poverty One Experiment at a Time: Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty: Review Essay." *Journal of Economic Literature* vol. 50, no. 1 (March): 103-14.

Reed, Brian, Ryan Bakow, Alex Egbert, Michael Findley, Billy Matthias, and Daniel Nielson. 2015. "Reputation, Information Asymmetry and NGO Opportunism: A Randomized Global Field Experiment." Unpublished manuscript presented at the Research Seminar in the Department of Political Science at the University of California – San Diego. April 30, 2015, La Jolla, CA.

Rodrik, Dani. 2009. "The New Development Economics: We Shall Experiment, But How Shall We Learn?" In J. Cohen and W. Easterly, eds., *What Works in Development? Thinking Big and Thinking Small.* Washington, DC: Brookings Institution Press.

Stroup, Sarah S. and Wendy H. Wong. 2017. *The Authority Trap: Strategic Choices of International NGOs.* Cornell University Press, Ithaca, NY.

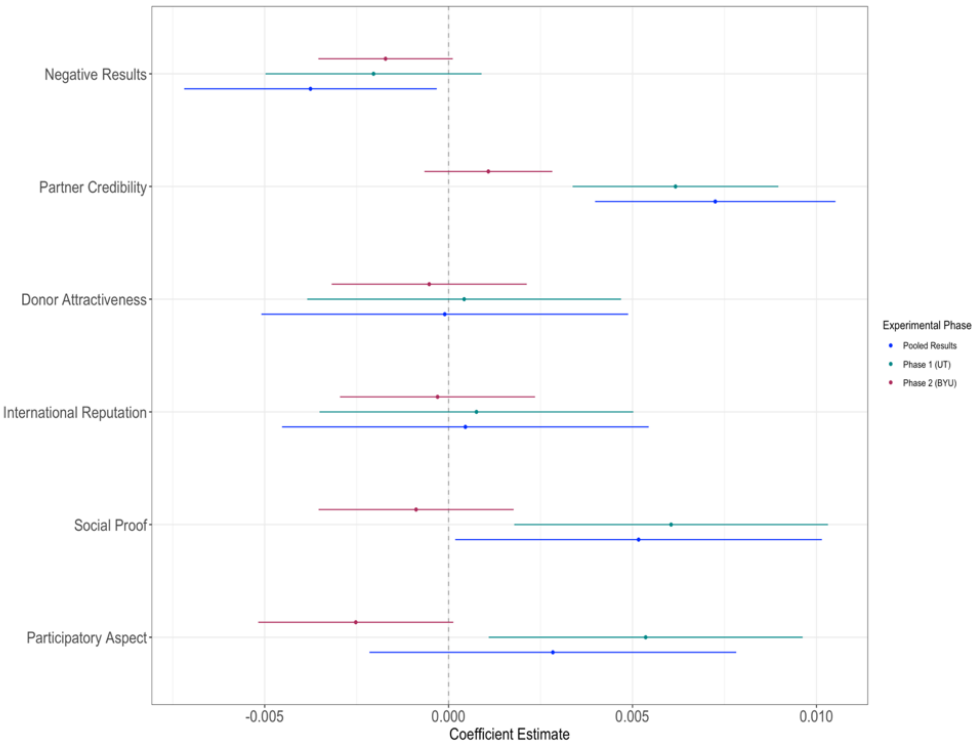
APPENDIX

Table A1. Sample of NGO Listings (Organization Name, Country, Region)

	FirstName	Country	Region
1	ADA	Afghanistan	MENA
2	Afghan human rights training organisation	Afghanistan	MENA
3	Afghanistan Solidarity for Social and Humanity Organization	Afghanistan	MENA
4	Afro-Asian Council of Opthalmology	Afghanistan	MENA
5	ANAFAE (Afghan National Association For Adult Educatio)	Afghanistan	MENA
6	CARE International	Afghanistan	MENA
7	Caretakers of the Environment Tanzania	Afghanistan	MENA
8	CARPO	Afghanistan	MENA
9	chmkvovhto	Afghanistan	MENA
10	Community Empowerment and Development Organization	Afghanistan	MENA
11	Cooperation Center for Afghanistan (CCA)	Afghanistan	MENA
12	Cooperation for Peace and Development	Afghanistan	MENA
13	Cooperation for Peace and Development (CPD)	Afghanistan	MENA
14	CrisisLink	Afghanistan	MENA
15	cwmbsbzozi	Afghanistan	MENA
16	Development and Public Awareness (DPA)	Afghanistan	MENA
17	Feed the Birds International	Afghanistan	MENA
18	Fight Against Fraud Crimes and Corruption	Afghanistan	MENA
19	Green Life Empowerment	Afghanistan	MENA
20	Gulf Training	Afghanistan	MENA
21	Hands in Hand Organization	Afghanistan	MENA
22	Human Rights and Eradication of Violence Organization (HREVO)	Afghanistan	MENA
23	Humanitarian Programme of Biodiversity (PHB)	Afghanistan	MENA
24	Johnf988	Afghanistan	MENA
25	Johnk943	Afghanistan	MENA
26	Mercy Organization for Viable Empowerment (MOVE)	Afghanistan	MENA
27	Moms Against Messy Rooms	Afghanistan	MENA
28	Organization for Development of Human Resources Capacity	Afghanistan	MENA
29	Organization for Sustainable Development Afghanistan	Afghanistan	MENA
30	oymuvgbqng	Afghanistan	MENA

Table A1 shows a sample of the list of INGOs in our sample, arranged alphabetically by organization name and country. The organization name was entered manually, so we leave room for some human entry error and inconsistencies in spelling. A complete list is available with our replication files as well as on the official WANGO website.

Figure A2. Pooled Results (Phases 1, 2, and Pooled)



Our experiment had two phases separated by a one-year washout period to avoid contamination. Because of no meaningful differences in our implementation (same Qualtrics survey, changed only the institution name of sender from University of Texas to Brigham Young University), we pooled the results in our main analysis. However, it is curious to examine the separate phases in comparison to the pooled results. We see, for example, that the participatory aspect is significantly negative in the BYU round, significantly positive in the UT round, which yields the null result in our pooled findings. As the difference in institutional name is likely not driving this discrepancy alone, this raises questions about different interpretations of and responses to the participatory aspect. Further research could try to isolate factors such as close inclusion with the partner versus deep integration in the evaluation process to better identify what, if anything, in the participatory realm is influencing NGO willingness to evaluate impact, and in which direction. It is also curious that social proof was not significant and was in fact slightly negative in the BYU phase. The precise coefficient values are displayed in Table A3.

Table A2: Pooled Results (Phases 1, 2, and Pooled)

	Dependent variable: Responded		
	Phase 1: Texas	Phase 2: BYU	Pooled Phase
	(1)	(2)	(3)
Negative Results	−0.002	−0.002*	−0.004**
	(0.001)	(0.001)	(0.002)
Partner Credibility	0.006***	0.001	0.007***
	(0.001)	(0.001)	(0.002)
Donor Attractiveness	0.0004	−0.001	−0.0001
	(0.002)	(0.001)	(0.003)
International Reputation	0.001	−0.0003	0.0005
	(0.002)	(0.001)	(0.003)
Social Proof	0.006***	−0.001	0.005**
	(0.002)	(0.001)	(0.003)
Participatory Aspect	0.005**	−0.003*	0.003
	(0.002)	(0.001)	(0.003)
Constant	0.023***	0.011***	0.034***
	(0.001)	(0.001)	(0.002)
Observations	51,931	51,931	51,931
R^2	0.001	0.0002	0.001
Adjusted R^2	0.001	0.0001	0.0004
Residual Std. Error (df = 51924)	0.163	0.101	0.190
F Statistic (df = 6; 51924)	5.345***	1.667	4.669***
Note:			*p<0.1; **p<0.05; ***p<0.01

Table A3: Factorial Interaction Effects

	Model 1: OLS b/se	Model 2: Logit b/se
Negative=1	-0.003	-0.106
	(0.004)	(0.110)
Professor=1	0.008***	0.218***
	(0.003)	(0.083)
Negative=1 × Professor=1	-0.000	0.003
	(0.004)	(0.098)
Donor Attract.=1	0.001	0.045
	(0.004)	(0.120)
Negative=1 × Donor Attract.=1	0.001	0.023
	(0.005)	(0.151)
Reputation=1	-0.003	-0.091
	(0.004)	(0.124)
Negative=1 × Reputation=1	0.003	0.098
	(0.005)	(0.150)
Social Proof=1	0.012***	0.308***
	(0.004)	(0.112)
Negative=1 × Social Proof=1	-0.010*	-0.251*
	(0.005)	(0.147)
Participatory=1	-0.000	-0.004
	(0.004)	(0.121)
Negative=1 × Participatory=1	0.005	0.148
	(0.005)	(0.148)
Professor=1 × Donor Attract.=1	-0.004	-0.112
	(0.005)	(0.144)
Professor=1 × Reputation=1	0.003	0.099
	(0.005)	(0.144)
Professor=1 × Social Proof=1	-0.003	-0.105
	(0.005)	(0.137)
Professor=1 × Participatory=1	0.001	0.015
	(0.005)	(0.141)
Constant	0.034***	-3.356***
	(0.002)	(0.064)
R-sqr	0.001	
Obs	51931	51931
	*p<0.1; **p<0.05; ***p<0.01	