External Validity in Parallel Global Field and Survey Experiments on Anonymous Incorporation*

Michael G. Findley
Associate Professor, Government, University of Texas at Austin, mikefindley@utexas.edu

Brock Laney
JD Candidate, University of Chicago Law School, blaney@uchicago.edu

Daniel L. Nielson
Professor, Political Science, Brigham Young University, dan_nielson@byu.edu

J.C. Sharman
Professor, Centre for Governance and Public Policy, Griffith University, j.sharman@griffith.edu.au

11 March 2016

---

* The research design for this study's field experiment was registered on March 2, 2011 with the Institute for Social and Policy Studies at Yale University and later grandfathered into the Experiments on Governance and Politics Registry on its inception (available at e-gap.org). The survey portion of the study was not preregistered. University and Institutional Review Board clearances were received on July 7, 2010. Replication data available on Dataverse and at the authors' webpage.

Abstract

By comparing parallel field and survey experiments testing compliance with international standards on corporate transparency we highlight potential problems in the external validity of survey experimental designs. We performed a *field* experiment using deception in which we requested anonymous business incorporation from nearly 4,000 corporate service providers in more than 180 countries. Subsequently, we conducted a *survey* experiment with the same providers using similar treatment conditions, but with informed consent to participate in a research project. Comparing responses and response rates corroborates – from a new angle and with additional implications – survey researchers' caveats about selection bias and social desirability. Our conclusions on the relative external validity and different substantive results produced by different experimental designs constitutes an important cautionary note given the increased popularity of survey experiments within international relations and political science more generally.

Introduction

Experimental designs, whether embedded in a survey, conducted in a lab, or executed in the field, are all argued to provide strong internal validity enabling estimates of causal effects. Nevertheless, questions about external validity have constrained the use of experiments in political science. But different experimental designs may provide varying purchase on the problem of external validity, as experiments differ in terms of their representativeness and degree of naturalism.

We present findings from parallel field and survey experiments to bring a new perspective to the potential pitfalls of selection bias and social desirability, and to show that naturalistic field experiments on a large and diverse sample may help overcome these problems. More generally, we argue that field experiments in which participants (1) do not self-select, (2) do not know they are being studied, and (3) comprise a large and diverse sample, can, due to greater naturalism or ecological validity coupled with representativeness, produce results less subject to bias than an equivalent survey experiment performed with the same pool. Our linked field and survey experiments thus help shed light on what difference naturalism, enabled in this case by deception, makes to the findings. These are important and current topics in light of recent controversies about the role of deception in experiments connected with social media and elections (Kramer et al. 2014; Willis 2014).

We capitalize on the strong internal validity provided by random assignment to control and treatment groups, and the high ecological validity provided by a realistic setting in which participants neither self-select nor know they are part of an experiment, to compare the results of field and survey experiments on the same pool of respondents in a more direct fashion than has been done previously. The substantive focus involves testing adherence to international standards prohibiting the formation of anonymous or untraceable shell

companies by mandating that incorporation firms obtain notarized photo identity documents from those looking to form companies. We report the results from two field experiments – one on roughly 2,000 incorporation firms in 181 countries, and a second on nearly 1,700 firms in the United States – and a follow-up survey experiment on the same pool of roughly 3,700 incorporation service providers.

Subjects in survey experiments must self-consciously opt in to the study, and hence they know that their attitudes are being probed. When deliberately reflecting on their anticipated behavior, subjects may honestly report that they would behave in a way that departs from their actual behavior if acting automatically and being observed passively. Alternatively, subjects may dissemble, especially if they know that their actual behavior would be perceived as inappropriate. The lack of realism in the survey experiment brings into the foreground the problem of ecological validity, which together with representativeness, combines to determine the external validity of the study.

These problems of surveillance, self-selection, and social desirability bias have long been known to survey researchers (e.g., Belli et al. 1999; Berinsky 2004; Tourangeau and Yan 2007), who have developed tools such as the randomized response technique or list experiment to try to counteract the resulting challenges. The present study provides an especially stark endorsement of these earlier caveats about possible biases.

This finding is especially timely and important with the recent strong growth in the popularity of survey experiments in international relations (IR). While some celebrate and others bemoan this trend (Hyde 2015: 409; Jensen et al. 2014: 291 versus Pepinsky 2014: 432, 439; Mearsheimer and Walt 2013: 448), there is general agreement that survey experiments are increasingly in vogue in IR. Thus, although they come to opposite conclusions about the desirability of this development, both Hyde and Pepinsky see survey

experiments in particular as becoming more and more prominent in graduate syllabi, workshops, and conferences; in structuring the way questions are asked and answers evaluated in the field; and in occupying a more prominent place in leading journals. Limits to the external validity of survey experiments brings their widespread adoption in IR into question.

In contrast, our field experiment provides an atypically high level of external validity due to the naturalistic setting, the authenticity of the treatment and outcome, the global coverage of thousands of actors in more than 180 countries, and the fact that these actors neither self-selected into the experiment nor knew they were under scrutiny. Because so little was known about the characteristics of businesses that form shell companies (Corporate Service Providers or CSPs) *ex ante*, a population-based survey experiment in the ideal sense was not possible (Mutz 2011), but we were able to conduct a survey experiment on hundreds of CSPs globally with a response rate comparable to many conventional studies.

We find clear evidence that different experimental techniques produce different sets of respondents. The survey experiment gave a much lower response rate (less than 10 percent) relative to the field experiment (more than one third), so there is good reason to think that the type of providers responding may vary. Moreover, in the international subject pool, more than 80 percent of respondents answering the survey that had offered anonymous shells in the field experiment reversed their stance in the survey and claimed instead that they would demand photo ID or refuse service altogether. In the U.S., the same category of dissemblers totaled 60 percent. These results suggest caution toward survey experiments that ask about socially disapproved behavior.

Before outlining the experimental designs and reporting results, we first provide a brief primer on the substantive topic of the study: anonymous shell companies and the

businesses that form and sell them. The next task is to explain the research design of our parallel global field and survey experiments, after which we introduce and discuss the results. The final section before the conclusion considers the ethics of deception in experiments – a fraught topic given recent controversies – in terms of protecting the welfare of subjects and compared to the costs of completely foreswearing deception in experiments.

Anonymous Shell Companies and Corporate Service Providers

Shell companies, those that do not engage in substantive business activity, can be formed online in hours for a few hundred dollars. Companies can own assets, hold bank accounts, and make financial transactions, but are incorporeal, expendable, and thus potentially unaccountable. While shell corporations have many legitimate business purposes, there are few justifications for *untraceable* shell companies, which hide the identity of the real owner, and thus are very useful for criminals seeking to conceal illicit transactions and assets. Unless authorities can find the real owner, the culprit is essentially invulnerable.

International rules thus stipulate that countries must be able to find the actual person in control (the beneficial owner) of all companies (FATF 2012: 22). In practice, this responsibility has been delegated to the private firms that set up and sell shell companies: corporate service providers (CSPs). These providers complete and lodge the necessary paperwork and fees required to form a company, charging their own mark-up to the client. According to the rules, these providers must collect and hold identifying documents on company owners so authorities can reference this information should the need later arise. Yet whether this global standard actually works in practice remains largely unknown. Disquieting signs suggest that CSPs routinely flout the rules, and the sort of untraceable shell companies so useful in hiding the true identity of financial criminals are therefore readily available in practice (World Bank/UN 2011; Author 2014).

Survey and Field Experiments

The relative neglect of experiments in political science until the last decade or two stems in large part from concerns about external validity – the ability to generalize from the particular experimental setting to the wider political world. Progress in extending the use of experiments in political science depends on addressing questions of practicality and external validity (Green and Gerber 2002: 818, 824-25; McDermott 2002: 39; Gaines et al. 2006: 2; Barabas and Jerit 2010: 226; Druckman 2004: 683; Druckman et al. 2006: 627; Chong and Druckman 2007: 637; Gerber and Green 2012: 10; Hovland 1959: 14; Benz and Meier 2008: 268; Levitt and List 2007; Mutz 2011).

Participants who self-select into an experiment may not know the purpose or exact nature of the experiment, but they do know they are being scrutinized, and this may systematically affect their responses (Levitt and List 2007). The experimental setting strips out the context and situational factors of real life, and once again this may systematically bias the results (Gerber and Green 2012). These problems speak to the issue of naturalism, also known as "ecological validity," i.e., the degree to which the experimental environment mirrors real-world conditions (Brewer 2000).

External validity, the ability to generalize the findings to other populations of interest, requires consideration of multiple factors, including naturalistic settings, representative participants, valid and reliable measures, and realistic treatments (Mutz 2011, 140). Generalizability, according to Mutz (2011), therefore demands both "experimental realism" – that the experiment feels "real" to the subject – and "mundane realism" – that the experiment resembles conditions in the real world (140-141). That is, external validity requires ecological validity in addition to the representativeness of subjects.

One of the purported strengths of survey experiments is that, because subjects do not see all experimental conditions in the between-subjects design, they are less primed to provide the socially desirable response. This improves survey experiments' naturalism. Moreover, participants in survey experiments can be sampled in statistically representative ways that can overcome another major source of bias (Mutz 2011). Surveys' strengths in representative sampling are widely known and relatively undisputed. But it also seems likely that survey experiments provide greater realism advantages as well by simulating scarce-information environments that routinely occur in the real world. Subjects asked to evaluate, say, a proposed anti-immigration law on its own terms may respond very differently than subjects asked to judge anti-immigration policies compared to neutral or pro-immigration measures.

However, it is possible that, for especially sensitive matters, subjects will intuitively grasp the socially desirable response even without seeing any additional information. For example, subjects in statistical minorities in relation to their attitudes toward homosexuality, racism, or religiosity may be fully aware of how their responses will be perceived by others and what the most socially desirable answer to the question should be. Indeed, this is a key reason for the invention of the list experiment (see Kuklinski et al. 1997). Hence, such subjects may dissemble toward survey-experiment items in similar ways to conventional survey questions that include more complete information. It is this possibility that the present study is especially designed to investigate.

Attempts to generalize experimental findings from lab settings to the outside world may likewise be threatened by the facts that context is often key; that people are frequently bombarded with a multiplicity of conflicting stimuli, most of which they ignore; and that most political attitudes are long-standing rather than transient (for discussion of this issue,

see McDermott 2002; Barabas and Jerit 2010; Chong and Druckman 2007; Gaines, Kuklinksi and Quirk 2007). A naturalistic setting removes the problem of the atypical, aseptic lab environment, and possibly the issues of self-selection and knowledge of scrutiny as well, depending on attrition and compliance (Green and Gerber 2002; 2012; List 2008b). The ideal setting for a field experiment is one with high "authenticity of treatments, participants, contexts, and outcome measures" (Gerber and Green 2012: 11).

A further question has been raised by Dani Rodrik (2008) in connection with a study conducted in Western Kenya which concluded that distributing mosquito nets free of charge is more effective in reducing malaria than selling them (Cohen and Dupas 2010). Presented as clinching proof of the free distribution model in general, Rodrik (2008) points out that the results may not generalize beyond Western Kenya, once more a problem of external validity. He maintains that "the only truly hard evidence that randomized evaluations typically generate relates to questions that are so narrowly limited in scope and application that they are in themselves uninteresting." (2008: 5). Mutz likewise asks "why should we be so quick to assume that results from one particular field setting will easily generalize to another, completely different, real world setting?" (Mutz 2011, 134).

The checklist has thus become dauntingly long when it comes to responding to the external validity problems that have restricted the use of experiments in political science. The experiment should be in a highly naturalistic setting, with the treatment and outcome staying close to subjects' actual routine behavior. Subjects should not self-select into the experiment, or even know that they are being observed. Experiments should closely parallel respondents' everyday choices, and they should ideally be able to be matched with these same individuals' actual choices in similar situations. Furthermore, experiments should include a sufficiently large sample of subjects from the total population of interest and, ideally, a representative

sample of that population. Below we indicate the details of our field experiment in anonymous incorporation, explaining how it better satisfies these requirements for external validity than the parallel survey experiment presented subsequently.

Research Design

Posing as consultants, researchers approached approximately 3,800 CSPs in 181 countries via email. In the field experiment, each firm was contacted at least twice and a small subset three times, separated by washout periods of six months to one year. In these emails, researchers requested information on the types of identifying documentation each firm would require (if any) before forming a corporation. Legal and logistical requirements necessitated the creation of alias email accounts from which email messages were sent to subjects. Although each of the 21 aliases hailed from a different country, all approaches identified the alias as a consultant looking to expand his business and limit liability through incorporation. In addition to explicitly identifying a country of origin for each alias, each email was signed with the most common male first and last names characteristic of the stated country of origin.

After emphasizing that the alias would prefer to maintain anonymity, each email requested information on the types of identifying documents and fees necessary to retain the firms' services. Our own prior studies had determined that email is a very common form of contact between potential clients and CSPs, so this design feature afforded strong naturalism. To avoid potential biases caused by the wording of our approach emails, we varied the grammar, diction, and syntax of our approaches (see appendix).

Treatment language was piped into predetermined, standardized sections of the approach emails. The experimental conditions either varied the information provided – mentioning international or domestic corporate transparency law or essentially offering a

bribe – or altered the country of origin and business sector of the alias to suggest a customer profile consistent with the intent to launder money from government corruption or to finance terrorist operations. Treatments were compared with a placebo condition originating from one of eight randomly assigned minor-power, low-corruption OECD countries and offering no additional information. (For treatment language see appendix.)

Each treatment was associated with different sets of four to eight aliases, one of which was randomly assigned to each subject. The corruption treatment emails, for example, were sent by aliases purporting to hail from one of eight countries with, according to Transparency International, high perceived levels of corruption: Equatorial Guinea, Guinea-Bissau, Guinea, Papua New Guinea, Kyrgyzstan, Tajikistan, Turkmenistan, or Uzbekistan (Transparency International 2014). For many Westerners – though not for millions of West Africans, Central Asians, and Pacific Islanders – the four countries in each set are relatively indistinguishable, therefore helping to control for country-specific effects (later robustness analysis detected none that altered the results reported). We dubbed this basket of countries "Guineastan." The international body governing financial transparency, the Financial Action Task Force (FATF), explicitly enjoins firms to screen potential customers from countries "identified by credible sources as having significant levels of corruption, or other criminal activity" (2006, 21).

The Guineastan corruption condition contrasts with the eight "Norstralia" countries randomly assigned in the placebo: Australia, Austria, Denmark, Finland, the Netherlands, New Zealand, Norway, and Sweden. The Norstralia and Guineastan countries contrast with the four countries in the terrorist financing condition, where aliases claimed to hail from Lebanon, Pakistan, Palestine, or Yemen (again, randomly assigned), and to consult in Saudi Arabia for Islamic charities. Again, the FATF mandates that CSPs apply special scrutiny to

customers from "[c]ountries identified by credible sources as providing funding or support for terrorist activities that have designated terrorist organisations operating within them" (2006, 21). Moreover, the FATF warns against "[c]harities and other 'not for profit' organisations which are not subject to monitoring or supervision (especially those operating on a 'cross-border' basis)" (2006, 22).

All the additional treatments altering the information provided originated from one of the Norstralia aliases. One invoked the FATF explicitly and referenced its international standard of identity disclosure upon incorporation. A second information treatment, randomly assigned only among the roughly 1,700 CSPs in the United States, attributed the ID standards to the Internal Revenue Service. And a final information treatment offered to "pay a premium" to maintain confidentiality.

In the survey experiment, we evaluated nine conditions from the field experiment. The results for other field experimental conditions are reported elsewhere (see Author 2013; 2014; 2015). To summarize, the nine experimental conditions we consider here are:

1. **Placebo** – originating from the Norstralia countries and offering no additional information.
2. **FATF** – invoking the Financial Action Task Force and its rules for identification of the beneficial owner.
3. **Premium –** offering to pay more money for confidential incorporation.
4. **Corruption** – originating from one of the Guineastan countries identified by Transparency International as high in perceived corruption.
5. **Terrorism** – originating from Lebanon, Pakistan, Palestine, or Yemen associated by Pape (2005) and others with terrorism.
6. **US Origin** – originating from the United States.
7. **Penalties** – citing possible legal penalties for non-compliance.
8. **Norms** – noting that most countries have signed onto FATF and that "reputable businessmen should do the right thing."

9. **IRS** – noting the rule for identity disclosure and attributing it to the Internal Revenue Service.

For those that did not respond to our first email, we randomly assigned six different follow-up email letters that we sent to firms that remained non-responsive after seven days from our initial contact. Follow-up emails provided little additional text apart from an expression of continued interest in hearing from the subject and a reference to the original email, which was copied immediately below the follow up.

Sampling and Randomization

As no sampling frame existed when we began the study, we created a sample of corporate service providers listed on the Internet through systematic, country-by-country inquiries using a common search engine. These service providers could exist as law firms with a web presence, specialized corporate service providers with a physical office but also a website, or Internet-only entities specializing in incorporation services. The common requirement was that they offered incorporation services for some fee, typically ranging between $500 and $3,000. Given that CSPs can exist without a web presence, our sample was not random, but likely represented firms that were, on average, more open to public scrutiny and therefore more likely to be compliant relative to firms attempting to stay "off the radar." Whatever compliance we identify in the results would thus likely be an overestimate of the actual level of compliance were we able to treat all providers, or a fully random sample.

We employed a block randomization strategy for assigning treatment conditions to subjects (see appendix). Within each block, we randomly assigned subjects to treatment conditions in equal proportions. To dampen potential multiple comparisons problems, we assigned more subjects to the control condition compared to any single treatment condition.

In the US sample, 16% of subjects were assigned to each treatment condition and 36% to the control, and in the international sample 11% of subjects were assigned to each treatment condition and 23% to the control. During the random assignment of conditions for services that we treated two or three times, we performed the same randomization strategy but set conditions disallowing the assignment of the same treatment more than once to any subject. This strategy became necessary to avoid detection; although we waited at least six months before contacting a service for a second time, we feared that subjects might have detected an exact duplicate of treatment conditions received previously. No subject firm implied in correspondence that it suspected it was involved in a social science experiment.

Research assistants sent emails (from proxy servers to avoid detection) through alias accounts in nine waves beginning in March 2011 and ending in May 2012. The size of each wave varied, but ranged from 600 to 1200 subjects. The low response rate in the US sample prompted us to send two rounds of follow-up emails to non-responsive firms.

Corresponding with Subjects

Because subjects sometimes responded without providing information on identifying documents, we established a standardized system for responding to subjects' emails and questions. With a few exceptions, subject responses fell into one or more of 26 scenario categories for which we drafted standardized basic responses. If we did not receive an outcome of interest from the firm's initial response, researchers followed up until the CSP either offered anonymous incorporation, specified the required documents, refused service, ceased communication, or it became clear an outcome measure could not be obtained (i.e., the firm requested payment up front or information sufficiently specific that we could not provide it within the parameters of our general approaches).

Coding

As mentioned previously, research assistants coded responses based on the types of identifying documents subjects required before proceeding with incorporation. Using the FATF recommendation of identifying the beneficial owner as the standard for compliance, we coded subjects as noncompliant, partially compliant, or fully compliant. The type of photo identification was our primary metric for determining compliance level. Subjects that required no photo identification were coded as noncompliant. Partially compliant subjects included those that required a photocopy of a government-issued identification. To be classified as fully compliant, subjects must have required a certified, notarized, or apostilled copy of government identification bearing a photograph, or an in-person meeting. Two research assistants separately coded each response and a third arbitrated any coding disagreements. The research assistants assigned the same codes 80 percent of the time, meaning that a senior coder made a final determination in 20 percent of the observations. Of the 80 percent in which there was agreement, a senior coder also randomly checked to ensure accuracy and in nearly all cases found the codes to be correctly assigned.

We have good reason to believe that even though no money changed hands and no shell companies were actually set up, providers accurately communicated the documents they would need to incorporate a company. An.earlier, related audit study went through every stage of the incorporation process with 45 providers, barring actually transferring any money. In every instance, CSPs were consistent from beginning to end of the process with regards to the identity documents required, including cases in which no such documentation was requested. The audit study also involved paying for three shell companies to be incorporated in the U.S., England, and the Seychelles; once again, proof of identity requirements did not vary from the initial contact (Author 2011). Furthermore, interviewing

CSPs and observing them at trade shows strongly suggest that giving would-be clients contradictory information would be commercially counter-productive for CSPs.

The design of this field experiment gives us a fairly high level of confidence in the external validity for four reasons. First, the experiment takes place in a naturalistic setting given that the incorporation business is a highly internationalized, Internet-dependent industry. Client profiles and the main elements of the approaches were culled from many interviews with CSPs and participant-observation work at their trade shows in London, Miami, Singapore, Hong Kong, Geneva, and the Caribbean. The treatments, different solicitations for shell companies, the outcome, and customer due diligence procedures in responding to client requests to form a company are all part of the workday routine for CSPs. Second, subjects did not self-select into the experiment, nor did they know they were being scrutinized. Third, although there is no definitive global count of CSPs, we captured thousands of such firms from almost every country in the world, suggesting that extrapolation based on our sample is justified. Fourth, because our sample consisted of firms with some Internet presence, any noncompliance we find may understate actual rates of noncompliance in the world, as particularly unscrupulous firms may attempt to stay off the grid. In sum, these positive elements suggest that this field experiment matches a high level of internal validity and a high level of external validity.

Survey Experimental Design

In the survey experiment, we approached subjects as researchers investigating incorporation practices and mailed our correspondence through a survey-distributing platform (Qualtrics). In our recruitment email, we provided a brief introduction to ourselves, background information on the scope and size of our study, a standard statement requesting informed consent, and a request that subjects complete a brief survey. To incentivize

completion of the survey, we offered to make the results from our study available to any CSP that completed it, while also assuring them that we would anonymize their responses.

Little information beyond the type of firm and its country location was available for a large set of CSPs. This fact prevented us from employing a population-based survey experiment in which we could be confident that the subject pool was representative of the general population of CSPs (Mutz 2011). Nevertheless, the method of contact and self-selected response is characteristic of many survey experiments that are not population-based and thus serves as a relevant, though perhaps weaker, comparison to the field experiment.

The survey opened with questions designed to obtain information on the firms themselves (e.g., in which business areas they specialized). We then presented a hypothetical situation patterned after the actual situation we presented to each subject under the alias guise earlier in the field experiment. With some modifications to the language used in the treatments meant to reduce the likelihood of detection, we randomly assigned a substantively similar survey experimental condition to the one used in the field experiment. Recalling that we performed two to three rounds in the field experiment, if subject A, for example, received treatments 1, 2, and 3 in the field experiment, we randomly selected one of those three treatments for the hypothetical situation in the survey experiment. Thus, subjects read a hypothetical wherein the potential clients are "planning to incorporate their business in your country and would like to procure the help of your firm. They indicate that they want to get things started as quickly and anonymously as possible." After this prompt, we included the treatment language and an indication of the client's country of origin.

We implemented three precautions in addition to modifying the treatment language to reduce the probability that subjects would associate our survey request with the field experiment. First, we waited at least six months after finishing our field-experiment

correspondence with subjects before distributing the survey. Second, we did not include subjects in the survey with whom we carried out long or notable correspondence, which amounted to roughly 50 firms (thus under 4 percent of the approximately 1,300 CSPs that responded in the field experiment). These lengthy conversations occurred fairly evenly across conditions and thus by dropping them we likely did not introduce systematic bias into the remaining sample. Finally, we changed and randomly assigned the countries of origin for each treatment, but followed the same criteria for country selection as in the experiment. Attached with our terrorism treatment, for example, hypothetical clients in the survey hailed from the West Bank, Oman, or Turkey. Countries for the corruption condition in the survey experiment included Burundi, Chad, and Angola; countries in the placebo and additional information conditions were Iceland, Belgium, and Luxembourg.

The outcome measure asked what the respondent CSP would do when faced with the assigned request for incorporation. The answer space was open-ended, allowing free response similar to the email replies received in the field experiment. We note here that a list experiment would not have preserved the parallel structure to the field experiment; by disallowing free response it would have artificially restricted response options. Moreover, the actual quantity of interest was not the more truthful responses that selection of the sensitive item on a list experiment might provide but the *treatment effects across experimental conditions* on the subjects' average propensity to select the sensitive item in treatment compared to placebo. Such a list-experiment-within-a-survey experiment was not part of the conventional social science toolkit so, unfortunately, we did not think to employ it. But such an approach might reveal treatment effects on less-biased outcomes in future survey experiments probing topics prone to social desirability bias, and we therefore encourage its use (on list experiments, see Kuklinski et al. 1997; Imai 2011; De Jonge, Kiewiet & Nickerson 2014).

We distributed the surveys through Qualtrics and sent a non-response follow up email from the same platform to any firm that did not finish the survey within seven days. Research assistants coded survey responses using the same procedures established for coding responses from the field experiment with nearly identical inter-coder reliability rates. Maintaining these parallel designs for the field and survey experiments enabled us to compare observed behavior in a natural environment with expressed attitudes in a setting where subjects knew they were being studied.

Results

To provide some basic context, Figure 1 displays the different response rates across field and survey experiments in the international (left panel) and U.S. (right panel) samples. The Venn diagrams are drawn to scale and show the patterns of overlap (or not) in which firms responded to the field and survey experiments. The figure demonstrates that the field experiment elicited much higher response rates than the survey experiment. Interestingly, however, some subjects responded only to the survey and not the field experiment, indicating that each method elicits responses from different sets of subjects. The international sample produced a higher response rate for both field and survey experiments relative to the U.S. sample, but within each case the overlapping set is fairly similar in size.

[FIGURE 1 ABOUT HERE]

Response Rates

The divergence between the field and survey experiments first manifests with basic descriptive statistics. Low response rates appear to plague survey experiments, especially in the absence of direct incentives, and our study lends additional evidence for concern: only 267 of 2,149 CSPs, or 12.4 percent, in the international subject pool completed the survey.

The response rate for CSPs in the U.S. subject pool was considerably worse: 75 of 1,762, or 4.3 percent. Compared directly with the observations matched in the field experiment, CSPs proved much more likely to reply: we received 1,037 responses to our 2,149 inquiries for a 48.3 percent response rate in the international subject pool, and we obtained replies to 376 of 1,762 inquiries (21.3 percent) for the U.S.-based CSPs. Thus, the combined response rate was 8.7 percent for the survey but 36.1 percent for the field experiment.

It is important to note here that the two categories of responses were to very different types of inquiries, so differences in response rates should be expected. Survey-experiment respondents were answering by clicking on a link to a Qualtrics instrument that they understood would be probing their opinions and attitudes. Field-experiment subjects believed they were undertaking business communication that might lead to profits. Both subject pools were thus self-selected and therefore disposed to bias; neither was sampled at random and hence fully representative. Given this self-selection, the key point is that the 9-percent sample in the survey experiment, by sheer force of numbers, was less likely to represent the normal behavior and attitudes of the population than the 36-percent sample in the field experiment.

An additional measure we undertook reinforces this point. After all field experiment rounds were completed, we contacted all CSPs that failed to respond to all inquiries and sent an email from a Norstralia alias with which each CSP had no prior contact in the field experiment. These test emails made no mention of the need for confidentiality, worries about taxes, or the desire to reduce legal liability (each a key element of emails across all experimental conditions) and simply asked instead for information about incorporation. These test emails received replies from only 5.8 percent of CSPs in the full international pool and 3.9 percent of CSPs in the full U.S. pool. This suggests that the field experiment

achieved responses from very near the upper bound of CSPs willing to assist foreign customers and thus should be seen as relatively representative – or at least a very large share – of the set of CSPs available through Internet contact.

The same, however, cannot be said of survey respondents. Logistic regression analysis (see Table 1) of several variables suggests that the subjects answering the survey were not altogether representative of the CSPs responding to the inquiries from aliases in the field experiment. On the one hand, subjects that responded in the field experiment were more likely to respond to the survey experiment as noted in row 1 for the international and U.S. samples with rotation of excluded categories. This result appears in three out of four rotations for both the international and U.S. samples and is substantively meaningful as indicated by large percent changes in predicted probabilities.

And yet, on the other hand, the regression analysis showed that subjects who refused service in the field experiment were significantly less likely to complete the survey. Also, incorporation service providers (coded 1) were significantly less likely to complete the survey than law firms (coded 0). And providers in tax havens and OECD countries were significantly less likely to complete the survey compared to CSPs in developing countries. These latter results for the international sample are precisely the opposite of the field experiment, in which incorporation services were significantly more likely to respond compared to law firms, and CSPs in tax havens and OECD members were likewise significantly more likely to reply to the inquiries from our aliases. As shown by percent changes in predicted probabilities, the results are substantively meaningful in that the effects for the international sample (for refusal, service providers, and country categories) range from 33%-60% changes of the probability of a survey reply. In the U.S. sample, the changes range from 53% to 186% for field refusal and company type.

We note here that the relatively low proportion of survey responses and its self-selective nature may have created two types of known survey error. The lower statistical power of the survey experiment means that random noise in the sample might be making the estimates imprecise and thus less able to identify a significant treatment effect. However, it is also the case that, because the survey sample was highly self-selected, this creates not only statistical noise and imprecision but actual bias. The few respondents to the survey – perhaps motivated by reasons other than profit-seeking – are unlikely to be representative of the broader population of corporate service providers.

[TABLE 1 ABOUT HERE]

In the field experiment, the response rate was a critical outcome measure and we saw significant effects for several of the treatments, especially the corruption, terrorism, and premium conditions. However, in the survey experiments, subjects responded to the experimental conditions after having completed several prior questions, and only one of the 325 respondents dropped out after seeing the critical question with the embedded experiment, so response rates were likely not sensitive to treatment in the survey, an advantage of survey experiments (see Mutz 2011). Thus, we do not emphasize differences in response rates across experimental conditions below, even if the overall selection-effect differences between the two study types are large and likely meaningful.

Outcome Tabulations

Additional descriptive statistics strengthen the impression that the answers to the survey experiment are substantially different than the field experiment, thanks to differing numbers and characteristics of respondents and the tendency of the overlap group to falsely claim more compliance in the survey than in the field experiment. Panel 2A in Table 2 shows the frequency and proportion of subjects that responded in the different outcome categories

for the field experiment compared to the same CSPs in the survey experiment. If the survey experiment were to mirror the field experiment, then the number of CSPs should be concentrated along the principal diagonals – indicating that they responded similarly to the substantively similar treatment conditions. But this is emphatically not what occurred. As might be expected given low overall response rates, the vast majority of subjects simply did not respond to the survey altogether. But many still claimed in the survey that they would behave differently than they actually did when faced with a substantively similar treatment condition in the field experiment, raising serious questions about the external validity of our survey experiment.

For example, as shown in the top row in Panel 2A of Table 2, of the 173 CSPs in the field experiment that responded to inquiries and indicated that they would be willing to provide an anonymous shell (and therefore were coded non-compliant), 131 failed to answer the survey. While 9 CSPs remained consistent and indicated they would not require any photo ID whatsoever, another 22 claimed they would in fact require (non-notarized) photo ID, 8 maintained they would require notarized photo ID, and an additional 3 declared they would refuse service altogether – and this despite the fact that we observed the same firms offer anonymous shells under substantively similar treatment conditions in the field experiment just months earlier.

[TABLE 2 ABOUT HERE]

This is shown even more starkly in Table 3, which considers only survey subjects that responded. Fully 33 of the 42 CSPs that answered the field-experiment inquiries in a non-compliant way and thus offered anonymous shell companies dissembled in the survey and claimed that they would demand photo ID or refuse service altogether when facing a substantively similar treatment condition. These disparities have broader implications both

for the importance of deception in research designs that target learning about inappropriate behavior, and for the increasing popularity of survey experiments in IR and political science writ large.

[TABLE 3 ABOUT HERE]

Treatment Effects in the Survey versus the Field Experiment

Even ignoring variation across treatment conditions, non-compliance rates drop dramatically in moving from the field to the survey experiment, which is expected given a social desirability bias. When we unpack and analyze the specific treatment conditions developed for the study, what do we learn? And how do differences between treatment and control conditions in the field experiment compare to the differences in the survey? We take up these two questions by identifying the basic differences in proportions. We find that differences between the experiment and follow-up survey again manifest themselves in the treatment effects for the randomly assigned interventions.

Tables 4 and 5 display the basic differences both between the field experiment and survey experiment as well as between treatments and placebo (for the international and U.S. samples respectively) for six of the key conditions. Given that the low response rates in the survey experiment led to a significant loss of power, we also include a pooled condition for the Premium, Terror, and Corruption conditions in the international subject pool and the IRS, Terror, and Corruption conditions in the U.S. sample to increase cell sizes for this set of high-risk treatment conditions.

Of the 424 subjects assigned to the Terrorism condition in the international field experiment, 24 (5.7%) proved noncompliant. When comparing against the Placebo for the field experiment, we learn that non-compliance is significantly lower in the Terrorism condition (in the Placebo 97 of 1,112, or 8.7%, were non-compliers). For each of the

treatments – (1) Terrorism, (2) Corruption, (3) Premium, (4) FATF, and (5) Terrorism, Corruption, and Premium jointly – the differences between treatment and control are contained in the table. Indeed, many of the treatments in the field experiment are statistically different from the Placebo.

The survey experiment, on the other hand, shows few differences between the treatments and the Placebo. This may owe in part to the smaller subject pool and thus greater imprecision given random noise, but most of the substantive differences are also small, suggesting few meaningful effects even if power were improved. One exception occurs in the international subject pool in which the proportion of partial compliance goes down (though not significantly) in the field experiment, yet goes up significantly in the survey experiment. Likewise, the terrorism condition appears to decrease the refusal rate (again not significantly) in the field experiment, yet increase refusals significantly (at the 0.1 level) in the survey experiment. Additional exceptions occur in the U.S. survey sample: for the Terrorism condition, compliance increases significantly in the survey experiment where no such change occurs in the field experiment. Further, in the FATF condition, the non-compliance rate is unchanged statistically in the field experiment but increases significantly (at the 0.1 level) in the survey experiment.

For both the international and U.S. samples, we also considered the treatment effects when we drop non-response from the survey, as with the analysis in Table 3. (See third row entries of Tables 4 and 5 for each condition.) The results are broadly similar to those in which we keep survey non-response as an option, with on key exception. In the U.S. sample, with non-response excluded there are still fewer treatment effects than the field experiment, but only slightly, and nonetheless more than when non-response is included.

Importantly, for the IRS condition, non-compliance levels decrease significantly in the field experiment, yet increase significantly in the survey experiment. Here, the effects are statistically significant in the *opposite* direction from field to survey experiment. The same pattern is true for the pooled IRS/Corruption/Terrorism condition: a significant decrease for non-compliance in the field experiment, but an increase in non-compliance in the survey experiment (likely due to the IRS condition). A survey experiment that hoped to understand how high-risk requests affect compliance with international financial transparency standards might thus reach conclusions that the field experiment suggests are inaccurate, indeed, opposite. This supports the idea that different experimental designs matter for estimating the significance of treatment effects, even when applied to the same sample, in keeping with recent studies like Hainmuller et al. (2015), but contradicting others (Berlinsky et al. 2012; Weinberg et al. 2014).

[TABLES 4 & 5 ABOUT HERE]

Comparing the results in Tables 4 and 5 demonstrates stark differences between the field and survey experiments. (Appendix Tables A1 & A2 provide an alternative illustration of the results.) Given the types of data, formal tests of the differences are not straightforward. We nonetheless conducted one test that may provide additional evidence about the differences in the types of experiments: the Spearman's rho. (See Appendix Tables A3 and A4).

A Spearman's rho test calculates a correlation of the coefficients in the field experiment relative to the survey experiment. Strong positive, significant correlations indicate that the field and survey experiments produce nearly identical results, whereas strong negative, significant correlations indicate that the two types of experiments produce nearly opposite results. In the results reported in the appendix for the international and U.S.

samples, with the exception of part compliance, which is modestly positive and significant (at 0.1 and 0.05 levels respectively), all other results indicate no strong relationship between the field and survey experiments. These tests corroborate the differences observed above.

Ethics

The deception in the field experiment enhances confidence in the results obtained, but also raises important ethical implications. The justification for the field experiment is founded upon the *Belmont Report* (1979) principle of beneficence: the risks and costs for the participants in the study are strongly outweighed by the benefits of learning about vital patterns in corporate secrecy that harm many people throughout the world. Rather than being a purely intellectual exercise, the study helps to increase knowledge on existing vulnerabilities in systems designed to reduce financial crime and the associated human suffering. Indeed, before this study there had been no large-scale audit of CSPs' adherence to global transparency standards, so the results contributed significantly to what was known about this important policy area. Even so, we were careful to safeguard respondents' welfare. Because our approaches closely mimicked providers' everyday routines, the exercise involved little inconvenience to the participants. We estimate that providers' email replies took 3-5 minutes on average, and thus the cost in respondents' time was less than most surveys. All names of individuals and firms were permanently deleted from the dataset to ensure that none could come to harm on account of their replies, and to guard against this information being extracted from the authors under duress (e.g., through subpoena). When using deception sometimes scholars debrief their subjects. Given the low costs for subjects, but the sensitivity of the issue, we chose not to debrief subjects so as not to draw additional attention. In discussing these ethical and practical issues with others in the scholarly and university communities, feedback suggested that debriefing may do more harm than good.

The divergences between the field and survey experiments illustrate the impact of deception in the research design. Though scholars may still legitimately oppose all uses of deception, they should be aware that this stance entails a price to be paid in terms of better understanding a range of serious policy and social problems. We expect that the ethical discussion over the use of deception will continue, as it should; we hope that this study helps to inform the debate.

Conclusion

Those advocating for the greater use of experiments in political science must overcome objections centered on external validity. By comparing parallel survey and field experiments, we have provided new experimental evidence to confirm survey researchers' warnings about social desirability and other biases. The survey response rate was a small fraction of that in the field experiment, suggesting pronounced selection bias. Among the sub-sample of CSPs that responded to both experiments, there was a marked difference in levels of hypothetical non-compliance in the survey compared with the level of actual non-compliance in the field experiment. Although there are obvious limits to the conclusions that can be drawn from any one study, our findings constitute an important cautionary note about the external validity of survey experiments, and thus pose a challenge to the increasing popularity of this research design.

References

Author. 2011.

Author. 2013.

Author. 2014.

Author. 2015.

Barabas, Jason and Jennifer Jerit. 2010. "Are Survey Experiments Externally Valid?" *American Political Science Review* 104 (2): 226-242.

Belli, Robert F., Michael W. Traugott, Margaret Young and Katherine A. McGonagle. 1999. "Reducing Voting Overreporting in Surveys: Social Desirability, Memory Failure, and Source Monitoring." *Public Opinion Quarterly* 4 (1): 90-108.

Belmont Report. 1979. U.S. Department of Health and Human Services. Accessed 17 November 2014 at http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html.

Berinsky, Adam J. 2004. "Can We Talk? Self-Presentation and the Survey Response." *Political Psychology* 25 (4): 643-659.

Berinsky, Adam J., Gregory A. Huber and Gabriel S. Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis* 20 (3): 351-368.

Brewer, Marilynn B. 2000. "Research Design and Issues of Validity." In Reis, Harry T., and Charles M. Judd, eds. *Handbook of Research Methods in Social and Personality Psychology,* pp. 3-16. New York: Cambridge University Press.

Chong, Dennis and James N. Druckman. 2007. "Framing Public Opinion in Competitive Democracies." *American Political Science Review* 2007 101 (4): 637-655.

Cohen, Jessica and Pacaline Dupas. 2010. "Free Distribution or Cost Sharing: Evidence from a Randomized Malaria Prevention Experiment." *Quarterly Journal of Economics* 125 (1): 1-45.

Cook, Thomas D., William R. Shadish, and Vivian C. Wong. 2008. "Three Conditions under which Experiments and Observational Studies produce Comparable Causal Estimates: New Findings from Within-Study Comparisons." *Journal of Policy Analysis and Management* 27 (4): 724-750.

Coppock, Alexander and Donald P. Green. 2015. "Assessing the Correspondence between Experimental Results Achieved in the Lab in the Field: A Review of Recent Social Science Research." *Political Science Research and Methods* 3 (1): 113-131.

Druckman, James N. 2004. "Political Preference Formation: Competition, Deliberation, and the Ir(relevance) of Framing Effects." *American Political Science Review* 98 (4): 671-684.

Druckman, James N., Donald P. Green, James H. Kuklinski, and Arthur Lupia. 2006. "The Growth and Development of Experimental Methods in Political Science." *American Political Science Review* 100 (4): 627-635.

FATF. 2006. "The Misuse of Corporate Vehicles, Including Trust and Corporate Service Providers." Paris.

FATF. 2012. International Standards on Combating Money Laundering and the Financial of Terrorism & Proliferation. Paris.

Gaines, Brian J., James H. Kuklinski, and Paul J. Quirk. 2006. "The Logic of the Survey Experiment Re-examined." *Political Analysis* 2006 15 (1): 1-20.

Gerber, Alan S., Donald P. Green, and Christopher W. Larimer. 2008. "Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment." *American Political Science Review* 101 (1): 33-48.

Gerber, Alan S. and Donald P. Green. 2012. *Field Experiments: Design, Analysis and Interpretation*, New York: W.W. Norton.

Green, Donald P. and Alan S. Gerber. 2002. "Reclaiming the Experimental Tradition in Political Science." 805-832 in *Political Science: State of the Discipline* edited by Ira Katznelson and Helen V. Milner New York: W.W. Norton.

Hainmuller, Jens, Dominik Hangartner, and Teppei Yamamoto. 2015. "Validating Vignette and Conjoint Survey Experiments against Real-World Behavior." *Proceedings of the National Academy of Sciences* 112 (8): 2395-2400.

Hyde, Susan D. 2015. "Experiments in International Relations; Lab, Survey and Field." *Annual Review of Political Science* 18 (1): 403-424.

Imai, Kosuke. (2011). "Multivariate Regression Analysis for the Item Count Technique." *Journal of the American Statistical Association*, 106 (494): 407-416.

Jensen, Nathan M., Bumba Mukherjee, and William T. Bernhard. 2014. "Introduction: Survey and Experimental Research in International Political Economy." *International Interactions* 40 (3): 287-304.

Jerit, Jennifer, Jason Barabas and Scott Clifford. 2013. "Comparing Contemporaneous Laboratory and Field Experiments on Media Effects." *Public Opinion Quarterly* 77 (1): 256-282.

Kramer, Adam D.I., Jamie E. Guillory, and Jeffrey T. Hancock. 2014. "Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks." *Proceedings of the National Academy of Science* 111 (29): 8788-8790.

Kuklinski, James H., Paul M. Sniderman, Kathleen Knight, Thomas Piazza, Philip E. Tetlock, Gordon R. Lawrence, and Barbara Mellers. 1997. "Racial Prejudice and Attitudes toward Affirmative Action." *American Journal of Political Science* 41, 2 (April): 402-419.

Levitt, Stephen D. and John A. List. 2007."What do Laboratory Tests Measuring Social Preferences Tell Us about the Real World?" *Journal of Economic Perspectives* 21 (2): 153-174.

List, John A. 2008a "Introduction to Field Experiments in Economics with Applications to the Economics of Charity." *Experimental Economics* 11 (3): 203-212.

List, John A. 2008b. "Field Experiments in Economics: The Past, Present and Future." National Bureau of Economic Research Working Paper 14356

McDermott, Rose. 2002. "Experimental Methods in Political Science." *Annual Review of Political Science* 5: 31-61.

Mearsheimer, John J. and Stephen M. Walt. 2013. "Leaving Theory Behind: Why Simple Hypothesis Testing is Bad for International Relations," *European Journal of International Relations* 19 (3): 427-457.

Mutz, Diana. 2011. *Population-Based Survey Experiments*. Princeton, NJ: Princeton University Press.

De Jonge, Chad P. Kiewiet, and David W. Nickerson. 2014. "Artificial Inflation or Deflation? Assessing the Item Count Technique in Comparative Surveys." *Political Behavior* 36 (3): 659-682.

Pape, Robert A. 2005. *Dying to Win*. New York: Random House.

Pepinsky, Thomas B. 2014. "Surveys, Experiments and the Landscape of International Political Economy." *International Interactions* 40 (3): 431-442.

Rodrik, Dani. 2008. "The New Development Economics: We Shall Experiment, But How Shall we Learn?" Unpublished paper, John F. Kennedy School of Government, Harvard University.

Tourangeau, Roger and Ting Yan. 2007. "Sensitive Questions in Surveys." *Psychological Bulletin* 133 (5): 859-883.

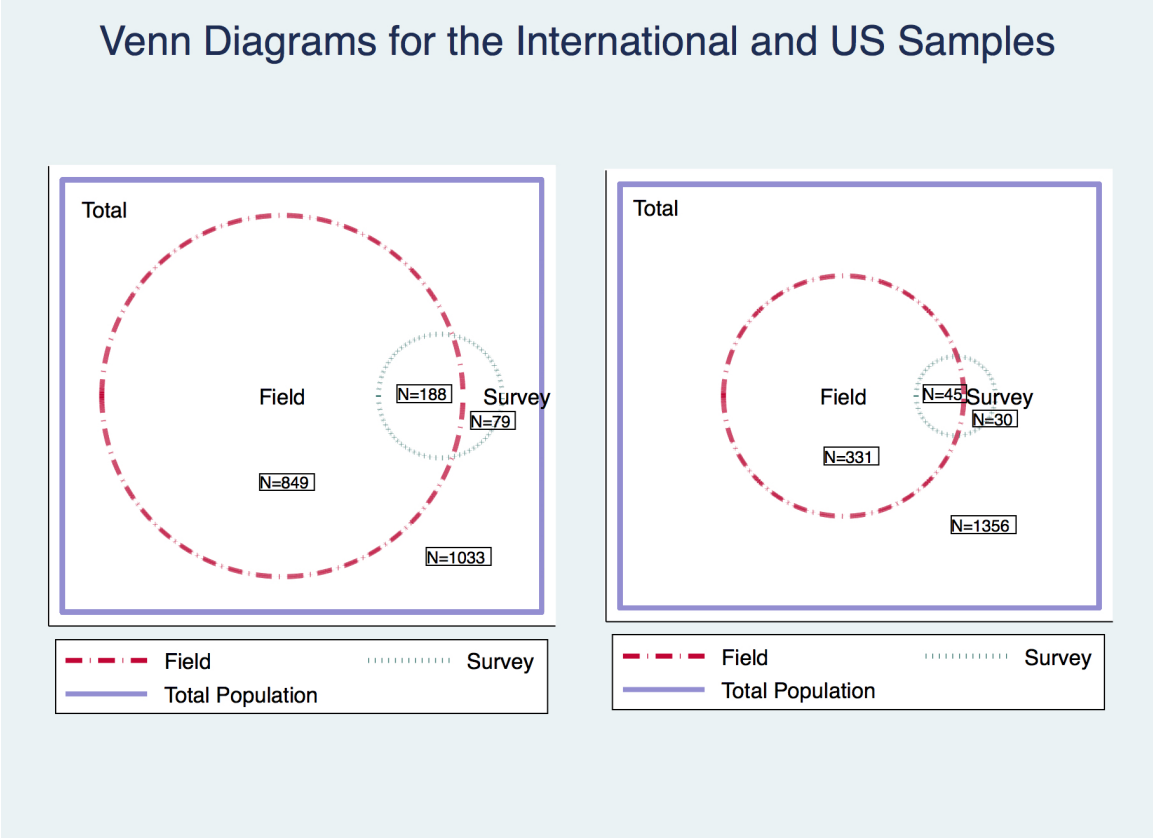
Weinberg, Jill D. Jeremy Freese, and David McElhattan. 2014. "Comparing Data Characteristics and Results of an Online Factorial Survey between a Population-Based and Crowdsource-Recruited Sample." *Sociological Science* 1 (7): 292-310.

Willis, Derek. 2014. "Professors' Research Project Stirs Outrage in Montana," *New York Times*, 28 October 2014, accessed 17 November 2014 at www.nytimes.com

World Bank. 2011. "Ease of Doing Business Survey." Washington D.C.

World Bank/United Nations Office on Drugs and Crime. 2011. *The Puppet Masters: How the Corrupt use Legal Structures to Hide Stolen Assets and what to Do About It*. Washington D.C.

Figure 1: International corporate service providers (left panel) and U.S. corporate service providers (right panel) scaled by response rate.



Note: "Total" represents all possible inquiries for the field and survey; "Field" represents responses in the field experiment; "Survey" represents responses in the survey experiment. The overlapping field-survey set represents the subjects who responded to both field and survey. The results show that very few subjects responded to both the field and survey experiments, indicating that the two methods produce very different samples for study.

Table 1: Logistic Regression Results for Selection into Survey Response

International	Survey Reply	Survey Reply	Survey Reply	Survey Reply	Field Reply
Field Exp. Reply	0.907***	1.307***	1.233***	0.227	
	(0.197)	(0.221)	(0.196)	(0.265)	
Prob %Δ in 0→1	123%	217%	198%	22%	
Field Noncomp.	0.4		0.074	1.08***	
	(0.243)		(0.242)	(0.301)	
Prob %Δ in 0→1	42%		7%	145%	
Field Part-Comp.		-0.4	-0.326	0.68**	
		(0.243)	(0.215)	(0.284)	
Prob %Δ in 0→1	---	-30%	-25%	80%	
Field Compliant	0.326	-0.074		1.006***	
	(0.215)	(0.242)		(0.282)	
Prob %Δ in 0→1	33%	-6%		134%	
Field Refusal	-0.68**	-1.08***	-1.006***		
	(0.284)	(0.301)	(0.281)		
Prob %Δ in 0→1	-46%	-63%	-60%		
Corp Service Provider	-0.455***	-0.455***	-0.455***	-0.455***	0.527***
	(0.15)	(0.15)	(0.15)	(0.15)	(0.066)
Prob %Δ in 0→1	-33%	-33%	-33%	-33%	30%
Tax Haven	-0.474**	-0.474**	-0.474**	-0.474**	0.607***
	(0.19)	(0.19)	(0.19)	(0.19)	(0.081)
Prob %Δ in 0→1	-35%	-35%	-35%	-35%	32%
OECD	-0.773***	-0.773***	-0.773***	-0.773***	0.198***
	(0.187)	(0.187)	(0.187)	(0.187)	(0.075)
Prob %Δ in 0→1	-50%	-50%	-50%	-50%	10%
Constant	-1.996***	-1.996***	-1.996***	-1.996***	-0.443***
	(0.135)	(0.135)	(0.135)	(0.135)	(0.052)
Observations	1,989	1,989	1,989	1,989	4,435
United States	**Surv. Reply**	**Survey Reply**	**Survey Reply**	**Survey Reply**	**Field Reply**
Field Exp. Reply	2.3***	1.468***	1.493	0.686*	
	(0.506)	(0.333)	(1.14)	(0.393)	
Prob %Δ in 0→1	761%	305%	314%	93%	
Field Noncomp.	-0.834		-0.025	0.782*	
	(0.509)		(1.133)	(0.428)	
Prob %Δ in 0→1	-56%		-2%	112%	
Field Part-Comp.		0.834	0.809	1.616***	
		(0.509)	(1.201)	(0.576)	
Prob %Δ in 0→1		122%	117%	352%	
Field Compliant	-0.809	0.025		0.807	
	(1.201)	(1.133)		(1.17)	
Prob %Δ in 0→1	-55%	2%		117%	
Field Refusal	-1.616***	-0.782*	-0.807		
	(0.576)	(0.428)	(1.17)		
Prob %Δ in 0→1	-80%	-53%	-54%		
Corp Service Provider	1.099***	1.099***	1.099***	1.099***	1.877***
	(0.294)	(0.294)	(0.294)	(0.294)	(0.108)
Prob %Δ in 0→1	186%	186%	186%	186%	247%
Constant	-3.92***	-3.92***	-3.92***	-3.92***	-1.666***
	(0.195)	(0.195)	(0.195)	(0.195)	(0.054)
Observations	1,701	1,701	1,701	1,701	2,996

Standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1; Note: These results show that the subjects completing the survey were not qualitatively similar to those captured in the field experiment. For examples, providers in tax havens and OECD countries were significantly less likely to complete the survey compared to providers in developing countries. And law firms were significantly more likely to complete the survey in the international sample, but less likely to complete the survey in the U.S. sample, relative to other corporate service providers. Percent changes in predicted probabilities show the results are substantively meaningful in most cases.

Table 2: Cross-Tabulation of Subjects

Panel 2A: International

			Survey Outcome			
Field Outcome	Non-compliant	Part-compliant	Compliant	Refusal	Non-response	Total
Non-compliant	9 (5.2%)	22 (12.7%)	8 (4.6%)	3 (1.7%)	131(75.7%)	173
Part-compliant	4 (1.3%)	40 (12.9%)	9 (2.9%)	4 (1.3%)	254 (81.7%)	311
Compliant	3 (0.9%)	26 (7.7%)	30 (8.9%)	8 (2.4%)	272 (80.2%)	339
Refusal	2 (0.9%)	10 (4.7%)	7 (3.3%)	3 (1.4%)	192 (89.7%)	214
Non-response	13 (1.2%)	39 (3.5%)	17 (1.5%)	10 (0.9%)	1033 (92.9%)	1112
Total	31	137	71	28	1882	2149

Panel 2B: United States

			Survey Outcome			
Field Outcome	Non-compliant	Part-compliant	Compliant	Refusal	Non-response	Total
Non-compliant	10 (6.3%)	11 (6.9%)	1 (0.6%)	3 (1.9%)	134 (84.3%)	159
Part-compliant	0 (0.0%)	6 (22.2%)	1 (3.7%)	1 (3.7%)	19 (70.4%)	27
Compliant	0 (0.0%)	0 (0.0%)	1 (16.7%)	0 (0.0%)	5 (83.3%)	6
Refusal	2 (1.1%)	5 (2.7%)	1 (0.5%)	3 (1.6%)	173 (94%)	184
Non-response	6 (0.4%)	15 (1.1%)	3 (0.2%)	6 (0.4%)	1356 (97.8%)	1386
Total	18	37	7	13	1687	1762

Note: Table 2 is a cross tabulation showing how subjects behaved in the experiment vs. the survey. Panel 2A contains the results for the international sample and panel 2B shows the US results. The rows represent the outcome in the experiment whereas the columns represent the outcome in the survey. This shows that, for example, of the 173 noncompliant subjects from the experiment on international CSPs, only 9 were non-compliant in the survey, 22 part-compliant, and so forth. If the field and survey experiments produced identical responses, then all observations would occur along the principal diagonal, which we do not see in these results. Also note that this comparison considers subjects that received the same treatment in both experiment and survey.

Table 3: Cross Tabulations by Proportion of Respondents across Outcomes in the Field and Survey Experiments

Panel 3A: International			**Survey Outcome**		
Field Outcome	Non-compliant	Part-compliant	Compliant	Refusal	Total
Non-compliant	9 (21.4%)	22 (52.4%)	8 (19.1%)	3 (7.1%)	42
Part-compliant	4 (7 %)	40 (70.2%)	9 (15.8%)	4 (7%)	57
Compliant	3 (4.5%)	26 (38.8%)	30 (44.8%)	8 (11.9%)	67
Refusal	2 (9.1%)	10 (45.5%)	7 (31.8%)	3 (13.6%)	22
Non-response	13 (16.5)%	39 (49.4%)	17 (21.5%)	10 (12.7%)	79
Total	31	137	71	28	267

Panel 3B: United States			**Survey Outcome**		
Field Outcome	Non-compliant	Part-compliant	Compliant	Refusal	Total
Non-compliant	10 (40%)	11 (44%)	1 (4%)	3 (12%)	25
Part-compliant	0 (0.0%)	6 (75%)	1 (12.5%)	1 (12.5%)	8
Compliant	0 (0.0%)	0 (0.0%)	1 (100%)	0 (0.0%)	1
Refusal	2 (18.2%)	5 (45.5%)	1 (9.1%)	3 (27.3%)	11
Non-response	6 (20%)	15 (50%)	3 (10%)	6 (20%)	30
Total	18	37	7	13	75

Note: Table 3 refines the cross-tabulation to show the percentage of outcomes *among those that responded*. Panel 3A contains the results for the international sample and panel 3B shows the US results. As with Table 2, if the field and survey experiments produced identical responses, then all observations would occur along the principal diagonal (excluding non-response for the field experiment). The table shows, for example, that of the 42 non-compliant respondents in the international field experiment, only 9 (21.4%) of the responders continued to be non-compliant in the survey.

Table 4: Comparative Treatment Effects for Int'l Field / Survey Experiments

	N	No Response	Non-Compliant	Part Compliant	Compliant	Refusal
Placebo Field	1112	495	97	184	210	126
Proportion		44.5%	8.7%	16.5%	18.9%	11.3%
Terror Field	424	247***	24**	46***	64*	43
Proportion		58.2%	5.7%	10.8%	15.1%	10.1%
Corrupt Field	428	225***	38	61	64*	40
Proportion		52.6%	8.9%	14.3%	15%	9.3%
Prem Field	385	191*	24	66	56*	48
Proportion		49.6%	6.2%	17.1%	14.5%	12.5%
FATF Field	390	190	35	62	66	37
Proportion		48.7%	9%	15.9%	16.9%	9.5%
Prem/Corr/Terr Field	1237	663***	86	173*	184***	126
Proportion		53.4%	7%	14%	14.9%	10.6%
Placebo Survey	618	548	8	37	20	5
Proportion		88.7%	1.3%	6%	3.2%	0.8%
(% of Responders)	70	---	11.4%	52.9%	28.6%	7.1%
Terror Survey	198	170	1	11	10	6**
Proportion		85.9%	0.5%	5.6%	5.1%	3%
(% of Responders)	28	---	3.6%	39.3%	35.7%	21.4%**
Corrupt Survey	206	176	3	20*	6	1
Proportion		85.4%	1.5%	9.7%	2.9%	0.5%
(% of Responders)	30	---	10%	66.7%	20%	3.3%
Prem Survey	186	160	4	14	5	3
Proportion		86%	2.2%	7.5%	2.7%	1.6%
(% of Responders)	26	---	15.4%	53.8%	19.2%	11.5%
FATF Survey	207	181	5	10	8	3
Proportion		87.4%	2.4%	4.8%	3.9%	1.4%
(% of Responders)	26	---	19.2%	38.5%	30.8%	11.5%
Prem/Corr/Terr Sur	590	506	8	45	21	10
Proportion		85.8%	1.4%	7.6%	3.6%	1.7%
(% of Responders)	84	---	9.5%	53.6%	25%	11.9%

***p<0.01, **p<0.05, *p<0.1

Note: Table 4 compares four treatments (including a combined condition—premium, corruption, & terrorism) to the Placebo for the field & survey experiments. Statistical significance denotes a difference between treatment and placebo proportions using a two-sided test. The results demonstrate that there are a number of treatment effects in the field experiment, but far fewer in the survey experiment, including when limiting comparisons to the responders.

Table 5: Comparative Treatment Effects for U.S. Field / Survey Experiments

	N	No Response	Non-Compliant	Part Compliant	Compliant	Refusal
Placebo Field	816	602	92	13	3	106
Proportion		73.8%	11.3%	1.6%	0.4%	13.0%
Terror Field	550	458***	32***	8	2	50**
Proportion		83.3%	5.8%	1.5%	0.4%	9.1%
Corrupt Field	532	417*	54	8	1	52*
Proportion		78.4%	10.1%	1.5%	0.2%	9.8%
IRS Field	552	442***	42**	12	2	54*
Proportion		80.1%	7.6%	2.2%	0.4%	9.8%
FATF Field	546	417	54	11	2	62
Proportion		76.4%	9.9%	2%	0.4%	11.4%
IRS/Corr/Terr Field	1634	1317***	128***	28	5	156***
Proportion		80.6%	7.8%	1.7%	0.3%	9.5%
Placebo Survey	481	465	1	12	1	2
Proportion		96.7%	0.2%	2.5%	0.2%	0.4%
(% of Responders)	16	---	6.3%	75%	6.3%	8.6%
Terror Survey	314	304	3	3	4*	0
Proportion		96.8%	1%	1%	1.3%	0.0%
(% of Responders)	10	---	30%	30%**	40%**	0.0%
Corrupt Survey	299	291	3	4	0	1
Proportion		97.3%	1%	1.3%	0.0%	0.3%
(% of Responders)	8	---	37.5%*	50%	0.0%	12.5%
IRS Survey	301	278***	6***	12	1	4
Proportion		92.4%	2%	4%	0.3%	1.3%
(% of Responders)	23	---	26.1%	52.2%	4.3%	17.4%
FATF Survey	306	292	4*	5	1	4
Proportion		95.4%	1.3%	1.6%	0.3%	1.3%
(% of Responders)	14	---	28.6%	35.7%**	7.1%	28.6%
IRS/Corr/Terr Survey	914	873	12**	19	5	5
Proportion		95.5%	1.3%	2.1%	0.5%	0.5%
(% of Responders)	41	---	29.3%*	46.3%*	12.2%	12.2%

***p<0.01, **p<0.05, *p<0.1

Note: Table 5 compares four treatments (including a combined condition—IRS, corruption, and terrorism) to the Placebo for the field & survey experiments. Statistical significance denotes a difference between treatment and placebo proportions using a two-sided test. The results demonstrate that there are a number of treatment effects in the field experiment, but fewer in the survey experiment, including when limiting comparisons to the responders.