Deceptive Studies or Deceptive Answers? Competing Global Field and Survey Experiments on Anonymous Incorporation*

Michael G. Findley
University of Texas, Austin

Brock Laney
University of Chicago

Daniel L. Nielson
Brigham Young University

J.C. Sharman
Griffith University

DRAFT:
PLEASE DO NOT CIRCULATE/CITE WITHOUT PERMISSION

---

* The research design for this experiment was registered on March 2, 2011 with the Institute for Social and Policy Studies at Yale University and then later grandfathered on the Experiments in Governance and Politics registry. Of those interventions registered, we report on the Placebo, FATF, Premium, Corruption, Terrorism, and IRS conditions in this paper. All other interventions outlined in the registered document are reported in other work (see Findley, Nielson and Sharman 2013, 2014). Contact: mikefindley@austin.utexas.edu

Abstract

We compare parallel global field and survey experiments testing the availability of anonymous shell corporations, which are commonly used to hide money laundering, tax evasion, and corruption. We first performed a large field experiment using aliases and deception in asking for confidential incorporation from nearly 4,000 corporate service providers (CSPs) in more than 180 countries. We followed up with a survey experiment based on informed consent from the same CSPs, using substantively similar treatment conditions as the subjects received in the field study. More than one third of CSPs responded in the field experiment, but less than ten percent answered the survey, indicating significant selection bias. Indeed, the survey respondents systematically differed from the CSPs that answered the field-experiment inquiry. Of those that responded to the survey, 14.7 percent declared a willingness to provide an anonymous shell company. However, in the field experiment 23.7 percent of responding CSPs offered incorporation without photo identification documents, an increase of nearly two thirds suggesting a much more significant global problem. Moreover, of the overlap group that responded to both experiments, three-fourths of CSPs that had indicated they would provide anonymous shells in the field experiment dissembled in the survey experiment and claimed instead that they would demand photo ID or refuse service. Indeed, results across an array of experimental conditions provide evidence for key causal effects in the field experiment where the survey fails to show treatment effects for analogous interventions. Significant different-in-difference results buttress this finding. The experimental method chosen thus can strongly affect the substantive results obtained, for two reasons. First, field and survey experiments may select systematically different samples from the same population. Second, individuals' actual behavior in a given situation may be radically different from their survey responses about hypothetical behavior in that same situation. In this way, the paper contributes to the on-going debate on the relative validity of field and survey experiments, as well as the methodological and ethical significance of research designs premised on deception.

Introduction

In the spring of 2011 an incorporation firm in Delaware received an email from one Alberto Chabile, who expressed interest in setting up a company. Mr. Chabile explained that he was a consultant associated with the government of Guinea Bissau and was "eager to limit information disclosure as much as possible." In essence, Chabile was asking for an untraceable shell corporation – one of the commonly used devices in hiding the dirty money associated with cross-border money laundering, tax evasion, grand corruption, and other notable crimes.

International rules on corporate transparency mandate that the incorporation firm should have insisted that Chabile first supply official photo documents to prove his identity. This identification requirement was especially important here because key elements of this customer's profile indicated a high risk of criminality. Guinea Bissau is one of the world's most corrupt countries, ranking 150 out of 176 in Transparency International's 2012 Corruption Perceptions Index. The United Nations Office on Drugs and Crime has described Guinea-Bissau as the world's only true narco-state, while the US government has indicted several of the ruling military junta (who hacked the last civilian president to pieces in 2009) as drug dealers. Working with the Guinea-Bissau government, the Mexican Sinaloa drug cartel uses planes to transship cocaine, with at least one Gulfstream private jet registered to exactly the sort of Delaware shell company Mr Chilabe sought to purchase (Lb Aviation Inc) (der Spiegel, "Africa's Cocaine Hub: Guinea-Bissau 'A Drug Trafficker's Dream,'" 8 March 2013 http://www.spiegel.de/international/world/violence-plagues-african-hub-of-cocaine-trafficking-a-887306.html) Despite the international rules on customer identification

documents, and regardless of the especially elevated risk posed by Mr. Chabile's country of origin, the Delaware firm replied that "No identifying paperwork is required" to form a Delaware company.

Just over one year later, the same Delaware firm was asked in a survey whether it would require identity documentation from a consultant in Burundi (165th on TI's list) before setting up a company. The firm emphatically claimed that it would. It listed the following required documents for incorporation: "Certified copy of passport from original issuing agency, certified copy of signature, source of funds certification, bank reference letter." The survey response, indicating full compliance with corporate transparency rules, is thus directly contradicted by the firm's actual observed behavior in responding to the solicitation from Mr. Chabile, which showed blatant non-compliance.

In fact, however, there was no such person as Alberto Chabile. The original email solicitation and the later survey were both part of the same academic study comparing findings from a field experiment with a substantively similar survey experiment. The original email had been sent via a proxy server from Provo, Utah by an undergraduate research assistant as part of what to our knowledge is the first truly global field experiment, involving 21 fictitious customers making more than 7,000 solicitations to firms in 181 countries. Similar to the Chabile-Delaware example above, we saw the same pattern repeating many times: firms that expressed willingness to provide anonymous shells in the field experiment claimed they would demand photo identification in the survey experiment, raising the possibility that under certain circumstances the attitudes reported in survey experiments may provide misleading answers about subjects' actual behavior.

In this paper we capitalize on the uniquely strong internal validity made possible by random assignment to control and treatment groups, and the high external validity provided by a very realistic setting in which participants neither self-select nor know they are part of an experiment, to compare the accuracy of survey experiment techniques against the benchmark of field experiment results. We report the results from two field experiments – one on nearly 2,000 firms in 181 countries and a second on nearly 1,700 firms in the United States – and a follow-up survey experiment on all of the nearly 3,700 incorporation service providers. While the initial wave of methodological debate about experiments in the social sciences compared experimental and observational studies, this paper explores and contrasts the relative strengths and weaknesses of different kinds of experimental research designs in relation to a real-world policy problem of considerable import.

Survey experiments have sometimes been presented as providing the best of all worlds: the internal validity of the classic laboratory experiment, combined with the external validity of a representative survey, in a form that, thanks to the rise of computer-assisted techniques, is increasingly practical and cheap. Doubters have, however, continued to question the external validity of this approach. But it has been difficult to judge the merits of this criticism using observational studies, given the methodological problems of unobserved confounds and omitted variable bias – the very problems experiments have been used to overcome. If the debate simply reflects a preference for internal validity (survey experiments) versus a preference for external validity (observational studies), it is difficult to see a way forward. We seek to break this impasse

by contributing to the nascent literature on comparing different forms of social science experiments through examining parallel global survey and field experiments.

Below doing so we first provide a brief primer on the substantive topic of the study: anonymous shell companies and the businesses that form and sell them, known as Corporate Service Providers (CSPs). We then go on to review the literature on the relative external validity of laboratory, survey and field experiments, as well as efforts to use different kinds of experiments to cross-check experimental findings. The next task is to explain the research design of our parallel global field and survey experiments.

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 actors in more than 180 countries, and the fact that these actors neither self-selected into the experiment nor knew they were under scrutiny. In comparing the results, we find clear evidence that different experimental techniques produce different samples. Aside from the numerical differences, the survey experiment gave a much lower response rate (less than 10 percent) relative to the field experiment (more than one third), there is good reason to think that the type of providers responding may vary. Furthermore, the survey experiment indicated a low level of providers' non-compliance with international rules mandating that they collect proof of customers' identity before establishing a company, while the level of non-compliance observed in the earlier field experiment was two-thirds greater.

Despite receiving a substantively similar treatment condition in both experiments, the CSPs who responded often claimed they behaved very differently in the survey experiment than we actually observed in those included in the field experiment .

Crucially, in the overlap group of providers included in both experiments, the survey responses proved to be a highly misleading indication as to how these CSPs had actually behaved in the field experiment. Thus, in the international subject pool more than 80 percent of respondents answering the survey that had offered anonymous shells in the field experiment claimed they would demand photo ID in the survey. In the U.S. the same category of dissemblers totaled 60 percent. One of the main conclusions of the paper is even closely-matched field and survey experiment may produce substantially different answers, first because they produce different samples, and second because survey responses may misrepresent real-world behavior.

Anonymous Shell Companies and Corporate Service Providers

In opening an Open Government Partnership conference in London 31 October 2013, British Prime Minister David Cameron explained why the problem of anonymous shell companies was his central focus:

> We need to know who really owns and controls our companies... For too long a small minority have hidden their business dealings behind a complicated web of shell companies and this cloak of secrecy has fueled all manners of questionable practice – and downright illegality. (https://www.gov.uk/government/speeches/pm-speech-at-open-government-partnership-2013).

While the limited liability company possessing its own separate legal personality is a fundamental institution of the capitalist system, the Prime Minister's speech accurately highlighted some real dangers. Companies can own assets, hold bank accounts, and make financial transactions, but are incorporeal, expendable and thus potentially unaccountable. Shell companies, those without a substantive business purpose, can be formed online in hours for a few hundred dollars. While shell

corporations have many legitimate business purposes, there are few if any justifications for *untraceable* shell companies, which can be used to screen and hide the identity of the real person behind illicit financial activity. Hence the Sinaloa drug cartel referenced earlier often holding their aircraft and bank accounts through shell companies to obscure the connection between the asset and its criminal origins. Unless the authorities can "look through" the company to 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, referred to as the beneficial owner, of all companies. In practice, this responsibility has been delegated to the private firms that set up and sell shell companies, Corporate Service Providers. These providers complete and lodge the necessary paperwork and fees necessary to set up a company, charging their own mark-up to the client purchasing the company. According to the rules, these providers must collect and hold identity documents on company owners, so that authorities can reference this information in should the need later arise. Yet the effectiveness of this global standard – whether it actually works in practice – is essentially unknown. Disquieting signs suggest that CSPs routinely flout the rules, and thus that the sort of untraceable, anonymous shell companies so useful in hiding the true identity of financial criminals are in practice readily available. This fundamental uncertainty over whether and why CSPs comply with global corporate transparency rules motivated the field and survey experiments presented below.

Before explaining these particular experiments we review the general pros and cons of different kinds of experiments, especially with reference to the question of external validity. The section below thus examines challenges to the external validity of

survey experiments in particular, summarizes recent work looking to cross-check survey experiments against other parallel experiments, and derives a list of factors necessary for a high-validity experiment.

Survey vs. Field Experiments

The relative neglect of experiments in political science until the last decade or two likely stems from concerns about practicality but – perhaps even more – so external validity, the ability to generalize from the particular experimental setting to the wider political world. Political scientists have largely acknowledged the superiority of experiments for discerning causal relationships in principle. As is widely agreed in medicine and the hard sciences, experiments are the gold standard because in expectation random assignment to control and treatment conditions balances all observed and unobserved potential confounds, meaning that researchers can more confidently attribute any subsequent difference between the experimental groups to the causal effect of the treatment.

But at the same time the conventional wisdom has long been that this method cannot tell us anything about actual political processes of interest. Experiments are best for those who can use them, but not relevant for political science (Waltz 1979: 16). A classic statement of this view is Lijphart's judgment in 1971 that "The experimental method is the most nearly ideal method for scientific explanation, but unfortunately it can only rarely be used in political science because of practical and ethical impediments" (1971: 683-684). Progress in extending the use of experiments in political science thus depends above all on addressing questions of practicality, ethics, and external validity (Green and Gerber 2002: 818, 824-25; McDermott 2002: 39; Gaines et al. 2006: 2;

Barabas and Jerritt 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 2008). How does this overriding concern affect laboratory, survey and field experiments, respectively?

Experiments in general are most closely associated with those performed in the laboratory. Here the reservations about external validity come from a number of angles, some of which are common to survey and field experiments, while others are particular to the lab setting. First is the general skepticism that the big political processes that have intrigued the field are amenable to manipulation in terms of random assignment to control and treatment groups (e.g. Lijphart 1971). Second is the related objection that the subjects of lab experiments, typically undergraduate students, self-select, are not representative of the broader population, and thus that experimental findings from the former cannot be extrapolated to the latter (as Hovland puts it, "college sophomores may not be people" 1959: 10). Third, participants in such 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 2008). Finally, the lab setting strips out the context and situational factors of real life, and once again this may systematically bias the results (Gerber and Green 2012). None of these problems is necessarily fatal, and various spirited defenses of laboratory experiments have been mounted (e.g. McDermott 2002), but it is fair to say external validity remains a key concern for lab experiments.

To what degree are survey experiments susceptible to the same critiques? As Gaines et al. recount, survey experiments sprang from what was initially regarded as an

inconvenient and frustrating quirk, according to which altering the order in which survey questions were asked or minor wording changes sometimes significantly changed respondents' answers (2006: 3), a vice that later became a virtue. Capitalizing on this phenomenon and led by scholars such as Paul Sniderman, survey experiments have been said to offer the external validity provided by representative samples of the general population coupled with the internal validity advantages of experiments. And this comes at a time when "the initial glamour of large-scale survey research had long since faded, its promise of a genuinely scientific social science long since forgotten" (Sniderman and Grob 1996: 378). Furthermore, with the rise of computer-assisted interview technologies, survey experiments became relatively cheap and easy to employ. An early example of this technique is randomly assigning information on the race, gender and other features of a hypothetical unemployment insurance claimant to test to what degree support for such insurance is sensitive to demographic factors (Sniderman and Piazza 1993). Have the promises of survey experiments been fulfilled, especially as they relate to external validity?

For some psychologists, it may be sufficient for survey experiments to illuminate states of mind (though many others prefer measures of actual behavior), but for political scientists the goal is to extrapolate to real political processes, and thus once again we face the question of whether experimental causal inferences drawn from surveys translate to equivalent processes in the outside world (Gaines et al. 2006: 2). Despite their relative novelty, however, difficult questions about the external validity of survey experiments were raised as early as the 1950s. Thus Hovland (who designed experiments for the U.S. Army during World War II concerning the impact of motivational films on the troops)

questioned the premise of extrapolating from experimental settings, where respondents are secluded from their usual context, given a single, strong stimulus, and then almost immediately tested to see what effect this stimulus might have. Attempts to extrapolate these findings to the outside world are threatened by the facts that context or situation is often key; that people are often bombarded with a multiplicity of conflicting stimuli, most of which they ignore; and that most political attitudes are long-standing rather than transient.

More than 50 years after Hovland, in directly asking whether survey experiments are in fact externally valid, Barabas and Jerit raise largely the same questions (2010; see also Chong and Druckman 2007; Gaines, Kuklinksi and Quirk 2007; Kinder 2007). Their findings are not especially reassuring. Barabas and Jerit aim to test relative external validity by running survey experiments in parallel with what they term a natural experiment. The substantive focus is on Medicare funding and a new U.S. citizenship test. The substantive significance for each experiment are similar, but here we refer only to the first case. In Spring 2007 the U.S. government announced that trust funds for Medicare had fallen below a key threshold, attracting some media attention and calls for a response. The treatment embedded in Barabas and Jerit's Internet survey experiment summarized this announcement of the funding shortfall and then asked about respondents' knowledge of and attitude towards Medicare – specifically how soon it would run out of funds – and whether or not the program was in crisis. The control group was simply asked the same questions without the summary of the announcement. The survey experiment showed a significant change in the knowledge and attitudes of the treatment group.

What the authors style as a natural experiment was a survey conducted the month before (March) and the month after (May) the government announcement and attendant media coverage. Directly contradicting the survey experiment, this exercise showed no change in either knowledge or attitude. A sub-set of this group, those with high media exposure (or with college education, used as a proxy for high media exposure), did show a change in knowledge from March to May 2007, but not a change in attitude. At least in this case, many of the reservations about the external validity of survey experiments seem to have been borne out: the findings of the survey experiment were contradicted by the results derived from respondents in the naturalistic setting.

Yet critics of this negative verdict might reasonably question whether in fact this is a case of comparing one kind of experimental evidence against another. Does a survey before and after a government press release really count as a natural experiment? Referencing Shadish, Cook and Campbell's example of property values before and after an earthquake (2002: 17) the authors maintain that it is (2010: 227 fn 2). We can conclusively rule out that the possibility that both an earthquake and a fall in property values reflect a common underlying cause acting independently on each factor. In contrast, it seems much harder to conclusively reject the proposition that both the government announcement about Medicare *and* popular knowledge and attitudes towards this program might be shaped by a common, underlying causal factor or factors (ageing population, sharply rising medical costs, budget deficit, etc.). More broadly, Gerber and Green, as well as King, Keohane and Verba, are adamant that without randomization, there simply is no experiment (2012: 17; 1994: 7) (a forthcoming study by Jerit, Barabas and Clifford in *Public Opinion Quarterly* aims to address this point) There certainly may

be value in comparing experimental and non-experimental studies (see Cook, Shaddish and Wong 2008) in checking external (and internal) validity, but given the problems inherent in studies based on observational data, some proponents of survey experiments will not be convinced.

Turning from political science to economics, it is possible to find scholarship that tests survey experiments against actually observed behavior, as well as against field experiments. The latter are defined by the same sort of random allocation to control and treatment groups as in any other experiment, but are distinguished by being set in a natural environment (List 2008: 205; Gerber and Green 2012: xv). Proponents of field experiments anchor their arguments for the utility of this method in claims about the superior external validity of this approach relative to survey and laboratory experiments (Green and Gerber 2002; 2012; List 2008b). Intuitively, if we are trying to explain what goes on in the real world rather than the laboratory, then experiments conducted in the real world might be more helpful than those conducted in the lab. 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. The ideal setting for a field experiment is one with high "authenticity of treatments, participants, contexts and outcome measures" (Gerber and Green 2012: 11). In economics this is where the environment and design of the experiment "cannot be reasonably distinguished from the tasks the agent has entered the marketplace to complete," and ideally where the subjects do not know they are in an experiment (List 2008: 205-6).

One such example concerns parallel field and survey experiments on determinants of charitable giving in a Costa Rican national park (Alpizar, Carlsson and Johansoon-

Stenman 2008). Survey experiments indicate that pro-social behavior, like hypothetical donations, is affected by the degree of anonymity: the greater the anonymity, the less the pro-social the behavior, akin to Gerber's and Green's findings on voter turnout (2004). Similarly, the level of hypothetical donations is sensitive to information on the average donation, e.g. many respondents are reluctant to give much more or less than the average. To what extent do these effects show up when it is actual donations at stake in natural settings?

The authors trained a team to become accredited as guides in a Costa Rican national park. These guides then approached foreign tourists at the park, either surveying them on how much they would be prepared to donate for the upkeep and maintenance of the park, or actually requesting such a donation. The first set of treatments for both the hypothetical and actual donation request was varying information about average donation figures. The second was to either have respondents place a sealed envelope in a private ballot box with their donation (if any), or the write figure they would be prepared to donate in the envelope, in contrast to donating or indicating their preferred figure in such a way that the guide could obviously see.

The setting is clearly realistic, the subjects do not self-select, they do not know they are part of an experiment, and hence we might assume the external validity to be quite high. Do the experimental data on actual behavior confirm the experimental survey data on hypothetical behavior? The authors describe "the most striking finding" as follows: "In the actual contribution treatment, 48 percent of subjects chose to contribute and the average contribution was $2.43, while in the hypothetical contribution treatment, 87 percent of the subjects stated that they would contribute an average of $7.58."

Perhaps not so surprisingly, it turns out that people were much more willing to give away hypothetical money in the survey than real money in the field experiment. However, the authors emphasize the extent to which the same treatments create similar effects in the survey and field experiments: more anonymity means fewer and lower donations in both experiments, and so too higher average figures lead to higher donations of both real and hypothetical money, even if the magnitudes of these effects are different.

Another test of the external validity of experiments again looks at charitable giving, but this time compares results from the lab with the previously recorded actual behavior of the same individuals. The central question is "whether the same individuals act in experiments as they would in the field" (Benz and Meier 2008: 268). The respondents are students at the University of Zurich who are asked each semester on their enrollment forms (i.e. privately) whether or not they wish to donate CHF 5-7 to one of two charitable causes. Unbeknownst to the students, the records of whether or not they had chosen to donate for the two years before and after the experiment was conducted were made available to these scholars. In the laboratory, students were given CHF 12 and the option to contribute some, all or none of this to the same two charities listed on their enrollment forms. The students who had never donated in the previous two years donated 65 percent of their funds in the lab exercise, giving some support to Levitt and List's criticisms of the lab environment boosting pro-social behavior. More positively, those students who had always donated gave the highest amount on average, whereas those who had sometimes donated gave less, but more than those who had never donated.

Students' responses in the lab thus did indicate *something* of their propensity to actually donate to the same causes, both before and after the experiment. The authors

note, however, that the correlations between behavior in the two settings is between 0.25 and 0.4, or relatively low. They explain this by the importance of situational factors: rather than a propensity for pro-social behavior being inherent in individuals and expressed in a consistent manner across different settings, situational factors more often predominate (Benz and Meier 2008: 279-280). This same point on the importance of context is also stressed by Alpizar et al. (2008: 311-312).

These two studies clearly represent some strong advances in experimental design. In both cases, actual behavior can be observed in an environment where subjects do not self-select into the experiment and do not know they are being scrutinized. The environment is very natural, even the lab experiment presents Zurich students with the exact same choice they make every semester, which can then be matched with each individual's actual past and future behavior. The verdict on the validity of survey and laboratory experiments is mixed, with some obvious deviations and some important similarities in the results. The common conclusion of the importance of context and situational factors, however, does raise some doubts about external validity, given the small stakes in play and even more so the localized settings. Presumably the point of studying undergraduates in Zurich and foreign visitors to a Costa Rican national park is not to know more about these two very specific groups, but rather to make some contingent generalizations about larger populations. Yet if specific context is so important, for all their realism, what can these experiments actually tell us about the wider world?

This point about extrapolating from very specific contexts has been raised by Dani Rodrik (2008) in connection with the field experiment conducted in Western Kenya

concluding 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 points out that the results may not generalize beyond Western Kenya, once more a problem of external validity. Speaking of policy, Rodrik's objection might apply to theory questions even more strongly: "Randomized evaluation did *not* yield hard evidence when it comes to the actual policy questions of interest. This should not have been a surprise: 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. The 'hard evidence' from the randomized evaluation has to be supplemented with lots of soft evidence before it becomes usable" (2008: 5). Are we back to Lijphart's position that experiments are excellent in principle (high internal validity), but impractical for any political question we actually care about (low generalizability)?

Taking into account both the positive examples summarized above and the critiques, the checklist of the ideal experimental study has become dauntingly long, especially when it comes to surmounting the external validity problems that have so 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 should ideally be able to be matched with these same individuals' actual choices in similar situations. Furthermore, experiments should include a large number of respondents in a large number of countries comprising a significant section of the total

population of interest, and they should relate to actual questions of policy and theoretical interest. Below we indicate the details of our field experiment in anonymous incorporation, explaining how it largely satisfies these requirements for external validity, before then detailing the design of our parallel survey experiment.

Research Design

Setup

Posing as international businessmen, we approached approximately 3,800 law firms and incorporation services in 181 countries via email. We approached all firms at least twice and a small subset three times, separated temporally by six months to one year. In these emails, we requested information on the types of identifying documentation each firm would require from us (if any) before forming a corporation on our behalf. 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 an international businessperson looking to expand his consulting business and limit liability through incorporation. Additionally, 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.

To avoid potential biases caused by the wording of our approach emails, we varied the grammar, diction, and syntax of our approaches to subjects. We wrote 33 different versions of the approach letter, with each one containing the basic information outlined in the previous paragraph. Of the 33 emails, the eight originating from English-speaking countries were written without grammatical or spelling errors, while the other

25 had one or two small errors per letter to enhance authenticity. In addition to modifying the wording of our letters, we also randomly assigned one of 10 slightly different subject lines to accompany each email correspondence. (Please refer to the appendix for more details on emails and subject lines.)

Treatment language was piped into predetermined, standardized sections of the approach emails. The experimental conditions either varied the information provided – priming either international or domestic corporate transparency law or offering essentially 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. All treatments were compared to a placebo condition originating from one of eight randomly assigned minor-power, low-corruption OECD countries and offering no additional information. Specific treatment language can be found in the appendix.

Each treatment was associated with different sets of 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 relatively low-profile 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 2011). For most Westerners – though of course not for millions of West Africans, Central Asians, and Pacific Islanders – the four countries in each set of two are relatively indistinguishable. For ease of reference, we dubbed this basket of countries "Guineastan." Despite the unpleasantness associated with outgroup profiling of any kind, 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. And both the Norstralia and Guineastan countries contrast with the four countries in the terrorist financing condition, where aliases claimed to hail from (again, randomly assigned) Lebanon, Pakistan, Palestine, or Yemen 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 of the information treatments originated from one of the randomly assigned Norstralia countries and aliases. One invoked the FATF explicitly and specifically referenced its international standard of identity disclosure upon incorporation. A second, randomly assigned only to the 1,701 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. We report the results from the six experimental conditions we replicated from the field experiment to the survey. The results for other field experimental conditions are reported elsewhere (see Findley,

Nielson, and Sharman 2013, 2014). To summarize, the six experimental conditions reported below are:

1. **Placebo** – originating from the Norstralia countries and offering no additional information.
2. **FATF** – evoking 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 the Guineastan countries identified by Transparency International as high in perceived corruption.
5. **Terrorism** – originating from Lebanon, Pakistan, Palestine, and Yemen associated by Pape (2005) and others with suicide terror.
6. **IRS** – noting the rule for identity disclosure and attributing it to the Internal Revenue Service.

In addition to explicitly identifying a country of origin for each alias, each email was signed with the most common first and last names characteristic of the stated country of origin. The most commonly used male name in Uzbekistan, for example, is Abdullo and the most common last name is Ogorodov, so Abdullo Ogorodov served as the Uzbekistani alias.

Anticipating that many of the subjects might not respond to our first email, we drafted and 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.

Randomization

We employed a block randomization strategy for assigning treatment conditions to subjects. We created 10 blocking categories in our international sample and 14 in our U.S. sample. In the international sample, the categories were based on company type (incorporation service or law firm) and country stratification based on country categorization as OECD members, tax havens, or developing countries further subgrouped into three categories by their rankings for ease of doing business: low, medium, and high business friendliness (World Bank 2011). In the U.S. sample, we again created blocks based on company type and an ease of incorporation ranking – once more subgrouping by low, medium, and high ease of incorporation (Beacon Hill 2010). Additionally, we created separate strata for California and Delaware (the two states with the most incorporations) and Wyoming and Nevada (the two most notorious tax-haven states aside from Delaware).

Within each block, we randomly assigned subjects to each treatment condition in equal proportions. To dampen potential multiple comparisons problems, we assigned roughly twice as many subjects to the control condition as to any single treatment condition. 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 suspected that subjects might have detected an

exact duplicate of treatment conditions received previously. No subject firm implied that it suspected it was involved in a social science experiment, though many, as intended, expressed concern about our approaches.

Research assistants sent emails through alias accounts in nine waves beginning in March 2011 and ending in May 2012. The size of each wave varied, but consisted of anywhere from 600 to 1200 subjects. The low response rate in the US sample prompted us to send an additional round 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 a 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 from the firm.

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 ultimate beneficial owner as a 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 (such as a passport). To be classified as fully compliant, subjects must have required a certified, notarized, or apostilled copy of identification bearing a photograph or an in-person meeting. All responses were coded separately by two research assistants. A third research assistant arbitrated any coding disagreements. As an extra step meant to increase the accuracy and consistency of our coding, research assistants performed a second round of blind coding and arbitration after all correspondence with the subjects had ended.

Why might the design of this field experiment give us a fairly high level of confidence in the external validity? Recalling the checklist described earlier: 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. 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 workaday routine for CSPs. Subjects did not self-select into the experiment, nor did they know they were being scrutinized. Though there is no definitive global count of CSPs, we captured thousands of such firms from almost every country in the world; there should therefore be limited worry of trying to extrapolate from very local results to global conclusions. These features all suggest that this field experiment matches a high level of internal validity with a high level of externality validity.

Survey Experimental Design

In the survey portion of our 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, background information on the scope and size of our study, and a request that subjects complete a brief survey. As a form of compensation, we offered to make the results from our study available to any CSP that completed it, while also assuring them that we would anonymize all responses they might provide.

The survey opened with a few questions designed to obtain information on the firms themselves. We asked, for example, in which business areas they specialized and how many people their firm employed. We also asked about the types of documents they felt should be required from clients looking to incorporate and whether or not they would require a personal meeting with a client before incorporating.

After acquiring this initial information, we presented a hypothetical situation patterned after the actual situation we presented to each subject under the alias guise earlier in the experiment. With some modifications to the treatments meant to reduce the likelihood of detection, we randomly assigned a substantively similar survey experimental condition to 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 experiment, we randomly selected one of those three treatments for the hypothetical situation in the survey. 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, as in the experiment, paired each treatment with an indication of the hypothetical client's country of origin.

In addition to modifying the treatment language to avoid detection, we implemented three other precautions to reduce the probability that subjects would associate our survey request with the experimental approach made earlier. First, we waited at least six months after finishing our correspondences with subjects before distributing the survey. Most of our correspondences were so brief that many subjects likely did not remember them for very long after ending our interaction, especially given that most CSPs receive many inquiries from potential customers each month. Second, we did not include subjects in the survey with whom we carried out long or notable correspondence, since those subjects were arguably more likely to remember our prior contact. Finally, we changed 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 instead of Lebanon, Palestine, Pakistan, or Yemen as in the field experiment. Countries of origin were randomly assigned from within the country lists in a manner identical to the treatment randomization.

We distributed the surveys through Qualtrics and a non-response follow up email went out from the same distributing platform to any firm that did not finish the survey within seven days. Research assistants coded responses using the same procedures established for coding field experimental responses. Code rounds in disagreement were arbitrated by a project lead. See the appendix for the survey language. The parallel designs of the field and survey experiments enable a relatively close comparison of

observed behavior in a natural environment with expressed attitudes in a setting where subjects knew they were being studied. The results among the overlap group that responded both to the field and survey experiments suggest that, at least in this setting where subjects often behave inappropriately according to international rules, words do not match actions very closely.

Results

Response Rates

The divergence between the field and survey experiments first manifests with the most basic descriptive statistics. Low response rates appear to plague survey experiments, and our study lends additional evidence for concern in this area: only 255 of 1,987 CSPs, or 12.8 percent, in the international subject pool completed the survey. The response rate for CSPs in the U.S. subject pool was considerably worse: 70 of 1,699 or 4.1 percent. This contrasts to the field experiment, where CSPs believed they were engaged in business development and thus proved much more likely to reply: we received 2,091 responses to our 4,365 inquiries for a 47.9 percent response-rate in the international subject pool, and we obtained replies to 592 of 2,986 inquiries (19.8 percent) for the U.S.-based CSPs. In the U.S., reply rates for incorporation services were similar to the international subject pool (246 responses to 466 inquiries for 52.8 percent) but law firms proved much less likely to answer (346 replies to 2,520 requests for 13.7 percent). The combined response rate was 8.8 percent for the survey but 36.5 percent for the field experiment, representing a more than three-fold increase.

Later follow-up emails to CSPs in both subject pools suggest that the non-responders were not in large measure screening based on risk. After all field-experiment

rounds were completed, we contacted all CSPs that failed to respond to any inquiry and sent an email from a different Norstralia alias that 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). Essentially, this follow-up asked if the firm was still in business and assisting international customers. This non-response check received replies from merely an additional 5.8 percent of CSPs in the international pool and 3.9 percent of U.S. CSPs. 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 the respondents to the survey. Logistic regression analysis (see Table 1 for the international sample and Table 2 for the U.S. sample) suggests that the subjects answering the survey were not a representative sample of the CSPs responding to the inquiries from aliases in the field experiment. Subjects were significantly less likely to complete the survey if they had refused service in the field experiment compared to the other outcome conditions as baselines. Law firms (coded 0 for Company Type) were significantly more likely to complete the survey than incorporation services (coded 1). And CSPs in tax havens and OECD countries were significantly less likely to complete the survey compared to CSPs in developing countries as the base condition. These results are precisely the opposite of the field experiment, where 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 the aliases.

Table 1: Logistic Regression Results for Selection into Survey Response (Int'l.)

	Survey Reply	Survey Reply	Survey Reply	Survey Reply	Exp Reply
Experiment Reply	0.885***	1.270***	1.184***	0.0619	
	(0.192)	(0.218)	(0.193)	(0.324)	
Exp. Noncompliant	0.385		0.0858	1.208***	
	(0.243)		(0.242)	(0.356)	
Exp Part-Compliant		-0.385	-0.299	0.823**	
		(0.243)	(0.215)	(0.340)	
Exp. Compliant	0.299	-0.0858		1.122***	
	(0.215)	(0.242)		(0.339)	
Exp. Refusal	-0.823**	-1.208***	-1.122***		
	(0.340)	(0.356)	(0.339)		
Company Type	-0.458***	-0.458***	-0.458***	-0.458***	0.551***
	(0.150)	(0.150)	(0.150)	(0.150)	(0.0667)
Tax Haven	-0.466**	-0.466**	-0.466**	-0.466**	0.695***
	(0.190)	(0.190)	(0.190)	(0.190)	(0.0809)
OECD	-0.769***	-0.769***	-0.769***	-0.769***	0.206***
	(0.187)	(0.187)	(0.187)	(0.187)	(0.0759)
Constant	-1.960***	-1.960***	-1.960***	-1.960***	-0.609***
	(0.130)	(0.130)	(0.130)	(0.130)	(0.0531)
Observations	1,988	1,988	1,988	1,988	4,365

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 2: Logistic Regression Results for Selection into Survey Response (U.S.)

	Survey Reply	Survey Reply	Survey Reply	Survey Reply	Exp Reply
Experiment Reply	2.263***	1.383***	1.457	0.553	
	(0.505)	(0.334)	(1.139)	(0.432)	
Exp. Noncompliant	-0.880*		-0.0731	0.830*	
	(0.510)		(1.133)	(0.466)	
Exp Part-Compliant		0.880*	0.807	1.710***	
		(0.510)	(1.201)	(0.603)	
Exp. Compliant	-0.807	0.0731		0.903	
	(1.201)	(1.133)		(1.184)	
Exp. Refusal	-1.710***	-0.830*	-0.903		
	(0.603)	(0.466)	(1.184)		
Company Type	1.086***	1.086***	1.086***	1.086***	1.950***
	(0.297)	(0.297)	(0.297)	(0.297)	(0.109)

Constant	-3.874***	-3.874***	-3.874***	-3.874***	-1.838***
	(0.189)	(0.189)	(0.189)	(0.189)	(0.0579)
Observations	1,699	1,699	1,699	1,699	2,986

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

It should be noted that in the field experiment the response rate was a critical outcome measure subject to treatment – 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. Thus below we do not emphasize differences in response rates across experimental conditions even if the overall response-rate differences between the two study types are large and likely meaningful.

In relation to how the samples of the population captured by the two techniques might differ, it could be expected that the motivation of those replying to the field experiment solicitation would contrast with those replying to the survey. Clearly, those replying to the solicitation are hoping for extra business, whereas completing the survey brings a much more diffuse benefit. For large service providers especially, the emails might be answered by different parts of the business.

Outcome Tabulations

Additional descriptive statistics deepen the concern that the answers to the survey experiment are radically different than the field experiment, thanks to differing samples and the tendency of the overlap group to falsely claim more compliance in the survey than they exhibited in the field experiment. Panel 3A in Table 3 shows the frequency and

proportion of subjects that responded in the different outcome categories for the field experiment compared to how the same CSPs responded in the field experiment. If the survey experiment were to reflect the field experiment with some accuracy, then the number of CSPs should be concentrated along the diagonals – indicating that they responded similarly to the substantively similar treatment conditions across the two study platforms. But this is emphatically not what occurred. As might be expected given the low overall response rates, the vast majority of subjects across all categories of outcomes in the field experiment simply did not respond to the survey. But many others claimed in the survey experiment that they would behave differently than they actually did when faced with a substantively similar treatment condition in the field experiment.

For example, as shown in the top row in Panel 3A of Table 3, of the 170 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), 129 failed to answer the survey. While 8 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 months earlier we observed the same firms offer anonymous shells under effectively indistinguishable treatment conditions in the field experiment. In this sub-section of the population that responded both to the solicitation and the survey, the survey experiment simply is not recovering similar reactions to the treatment conditions across study platforms.

This is shown even more starkly in Table 4. Fully 33 of the 41 CSPs – 80.5 percent – that answered the field-experiment inquiries in a non-compliant way and thus offered anonymous shells 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 are large and likely quite meaningful, with eye-opening implications for survey experiments seeking to reveal information on sensitive topics.

Table 3: Cross-Tabulation of Subjects

Panel 3A: International

			Survey Outcome			
Experiment Outcome	Non-compliant	Part-compliant	Compliant	Refusal	Non-response	Total
Non-compliant	8 (4.7%)	22 (12.9%)	8 (4.7%)	3 (1.8%)	129 (75.9%)	170
Part-compliant	4 (1.3%)	35 (11.4%)	9 (2.9%)	4 (1.3%)	254 (83%)	306
Compliant	3 (0.9%)	26 (7.8%)	27 (8.1%)	7 (2.1%)	271 (81.1%)	334
Refusal	0 (0.0%)	6 (4.1%)	5 (3.4%)	1 (0.7%)	135 (91.8%)	147
Non-response	15 (1.5%)	43 (4.2%)	19 (1.8%)	10 (1.0%)	943 (91.6%)	1030
Total	30	132	68	25	1732	1987

Panel 3B: United States

			Survey Outcome			
Experiment Outcome	Non-compliant	Part-compliant	Compliant	Refusal	Non-response	Total
Non-compliant	9 (5.7%)	10 (6.4%)	1 (0.6%)	3 (1.9%)	134 (85.3%)	157
Part-compliant	0 (0.0%)	5 (19.2%)	1 (3.9%)	1 (3.9%)	19 (73.1%)	26
Compliant	0 (0.0%)	0 (0.0%)	1 (16.7%)	0 (0.0%)	5 (83.3%)	6
Refusal	2 (1.3%)	5 (3.3%)	0 (0.0%)	0 (0.0%)	144 (95.4%)	151
Non-response	6 (0.4%)	15 (1.1%)	4 (0.3%)	7 (0.5%)	1327 (97.6%)	1359
Total	17	35	7	11	1629	1699

Note: Table 3 is a cross tabulation showing how subjects behaved in the experiment vs. the survey. The rows represent the outcome in the experiment whereas the columns represent the outcome in the survey. This shows that, for example, of the 170 noncompliant subjects from the experiment on international CSPs, only 8 were non-compliant in the survey, 22 part-compliant, and so forth. Panel A contains the results for the international sample and panel B shows the US results. Also note that this comparison considers subjects that received the same treatment in both experiment and survey.

Table 4: Cross Tabulations by Proportion of Respondents across Outcomes in the Field and Survey Experiments

Panel 4A: International

			Survey Outcome		
Experiment Outcome	Non-compliant	Part-compliant	Compliant	Refusal	Total
Non-compliant	8 (19.5%)	22 (53.6%)	8 (19.5%)	3 (7.3%)	41
Part-compliant	4 (7.7%)	35 (67.3%)	9 (17.3%)	4 (7.7%)	52
Compliant	3 (4.8%)	26 (41.3%)	27 (42.9%)	7 (11%)	63
Refusal	0 (0.0%)	6 (50.0%)	5 (41.7%)	1 (8.3%)	12
Non-response	15 (17.2)%	43 (49.4%)	19 (21.8%)	10 (11.5%)	87
Total	30	132	68	25	255

Panel 4B: United States

			Survey Outcome		
Experiment Outcome	Non-compliant	Part-compliant	Compliant	Refusal	Total
Non-compliant	9 (39.1%)	10 (43.5%)	1 (4.4%)	3 (13.0%)	23
Part-compliant	0 (0.0%)	5 (71.4%)	1 (14.3%)	1 (14.3%)	7
Compliant	0 (0.0%)	0 (0.0%)	1 (100%)	0 (0.0%)	1
Refusal	2 (28.6%)	5 (71.4%)	0 (0.0%)	0 (0.0%)	7
Non-response	6 (18.8%)	15 (46.9%)	4 (12.5%)	7 (21.9%)	32
Total	17	35	7	11	70

Note: Table 4 refines the cross-tabulation to instead show the percentage of outcomes among those that respond. It shows, for example, that of those responding to the survey, nearly 55% of the previously non-compliant responses become partly compliant. Panels 4A and 4B show these results for the International and United States samples respectively.

Treatment Effects and Difference-in-Differences

We already observed that even when ignoring specific information about treatment conditions, non-compliance rates drop dramatically in moving from the experiment to the survey. 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 in the experiment compare to the differences in the survey? We take up these two questions by identifying the basic differences in proportions followed by a difference-in-differences approach. In short, we find that differences between the experiment and follow-up survey again manifest themselves when examining the treatment effects for randomly assigned interventions.

Recall that we matched the treatments between experiment and survey so that the very same subjects received the same experimental interventions with only the research context in flux. As the subjects and conditions were identical, this setup allows us to consider the effects of changing the methodological approach from a field experiment to a survey experiment, first in terms of the different samples captured, then in the contrasts in the overlap group of their self-reported hypothetical response as compared with their actual behavior. As detailed above, we expected the interventions to have quite different effects depending on whether subjects knew they were being studied. After all, we were enticing subjects to run afoul of international standards. Willingness to abet corruption and terrorism or otherwise flout international law ought to be far more prevalent in the experiment than in the survey.

Table 5 and 6 displays the basic differences both between experiment and survey as well as between treatments and control (for the international and U.S. sample

respectively). Consider, for example, noncompliance in the international sample Terrorism condition. Of the 422 subjects assigned to the Terrorism condition in the experiment, 24 (5.7%) were noncompliant. Two points of comparison are especially instructive. First, when comparing against the Placebo for the experiment, we learn that non-compliance is lower in the Terrorism condition (in the Placebo 97 of 1,110 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 experiment are statistically different from the Placebo. The survey on the other hand shows few differences between the treatments and the Placebo. The few exceptions occur in the U.S. survey sample: for the Terrorism condition, part compliance decreases, compliance increases, and for the IRS the non-response proportion and noncompliance levels change substantially. A survey that hoped to understand how Corruption and Terrorism affect compliance with international financial transparency standards might thus reach a conclusion that none of these factors matter. And yet the experiment offers strong evidence that such a conclusion would be erroneous.

Table 5: Comparative Treatment Effects for Field and Survey Experiments – International CSPs

	N	No response	Noncomp	Part comp	Comp	Refusal
Placebo Exp	1110	532	97	183	209	93
Proportion		47.6%	8.7%	16.5%	18.8%	8.4%
Placebo Surv	630	563	9	37	16	5
Proportion		89.4%	1.4%	5.9%	2.5%	0.8%
Terror Exp	422	258***	24**	47***	64**	29
Proportion		61.10%	5.70%	11.10%	15.20%	6.90%
Terror Surv	204	173	2	12	9	7
Proportion		85.30%	1.00%	5.90%	4.40%	3.40%
Corrupt Exp	429	236***	38	61	64**	30
Proportion		55.00%	8.90%	14.20%	14.90%	7.00%
Corrupt Surv	211	180	4	20	6	1
Proportion		85.30%	1.90%	9.50%	2.80%	0.50%
Prem Exp	385	210***	24*	66	56**	29
Proportion		54.50%	6.20%	17.10%	14.50%	7.50%
Prem Surv	188	162	4	14	6	2
Proportion		86.20%	2.10%	7.40%	3.20%	1.10%
FATF Exp	391	199	35	63	67	27
Proportion		50.90%	9.00%	16.10%	17.10%	6.90%
FATF Surv	209	182	5	11	8	3
Proportion		87.10%	2.40%	5.30%	3.80%	1.40%
PCT Exp	1236	705***	87*	174*	184***	88
Proportion		57.00%	7.00%	14.10%	14.90%	7.10%
PCT Surv	603	516	10	46	21	10
Proportion		85.60%	1.70%	7.60%	3.50%	1.70%

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

Note: Table 5 compares the experimental and survey results for the Placebo, four conditions of the experiment, and a combined category (premium, corruption, terrorism; labeled PCT), all for the international sample. The proportions for the experiment and survey can be compared against each other for given conditions. And the results for the conditions can be compared against the Placebo. Statistical significance denotes difference between treatment and placebo. Difference-in-difference tests to compare the experiment and survey are shown in Tables 7 and 8.

Table 6: Comparative Treatment Effects for Field and Survey Experiments – U.S. CSPs

	N	No response	Noncomp	Part comp	Comp	Refusal
Placebo Exp	815	618	91	13	3	90
Proportion		75.8%	11.2%	1.6%	0.4%	11.0%
Placebo Surv	466	450	1	12	1	2
Proportion		96.6%	0.2%	2.6%	0.2%	0.4%
Terror Exp	548	467***	33***	8	3	37***
Proportion		85.20%	6.00%	1.50%	0.50%	6.80%
Terror Surv	325	315	3	3*	4**	0
Proportion		96.90%	0.90%	0.90%	1.20%	0.00%
Corrupt Exp	532	430***	54	8	1	39***
Proportion		80.80%	10.20%	1.50%	0.20%	7.30%
Corrupt Surv	326	315	3	5	0	3
Proportion		96.60%	0.90%	1.50%	0.00%	0.90%
IRS Exp	552	453***	42***	12	2	43**
Proportion		82.10%	7.60%	2.20%	0.40%	7.80%
IRS Surv	311	284**	10*	12	1	4
Proportion		91.30%	3.20%	3.90%	0.30%	1.30%
FATF Exp	544	429	54	11	2	48
Proportion		78.90%	9.90%	2.00%	0.40%	8.80%
FATF Surv	315	300	4	5	1	5
Proportion		95.20%	1.30%	1.60%	0.30%	1.60%
ICT Exp	1632	1350***	129***	28	7	119***
Proportion		82.70%	7.90%	1.70%	0.40%	7.30%
ICT Surv	962	914	16	20	5	7
Proportion		95.00%	1.70%	2.10%	0.50%	0.70%

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

Note: Table 6 compares the experimental and survey results for the Placebo, four conditions of the experiment, and a combined category (IRS, corruption, terrorism; labeled ICT), all for the US sample. The proportions for the experiment and survey can be compared against each other for given conditions. And the results for the conditions can be compared against the Placebo. Statistical significance denotes difference between treatment and placebo. Difference-in-difference tests to compare the experiment and survey are shown in Tables 7 and 8.

Second, of interest in this study is the difference between the experiment and the survey for a given condition. To continue the non-compliant Terrorism example, in the survey only 2 of 204 subjects (1%) were non-compliant, which stands in stark contrast to the 5.7% figure (24 of 422) for the experiment. Examining the tables for the international and U.S. samples, it is evident that the survey garners a much lower percentage than the experiment in nearly every outcome for every condition, except of course non-response which is much higher than the experiment. As mentioned, the low percentages in the survey are further compounded by the problem that the treatments in the survey are rarely statistically different from the placebo for the survey. This suggests an important comparison to make is whether the differences between treatment and control in the experiment are similar to the differences between treatment and control in the survey.

Tables 7 and 8 show the difference in difference estimates for the international and U.S. samples respectively. For each set of comparisons, we first display the difference between treatment and control for the treatment in the experiment and survey. The third row in each set contains the difference between the two differences. In the international sample, the difference-in-differences estimates are modest, but nonetheless show a number of significant differences especially for the terrorism, corruption, and joint conditions. The estimates are stronger for the U.S. sample. In particular, the difference-in-differences estimates for non-compliance in the Terrorism, IRS, and combined conditions are all large. Moreover, refusal rates are also distinct in the same conditions.

Table 7: Difference in Differences for International CSPs

Differences	No response	Noncomp	Part comp	Comp	Refusal
Terr vs. Placebo Diff Exp	13.6%***	-3.1%**	-5.3%***	-3.7%**	-1.50%
Terr vs. Placebo Diff Surv	-4.10%	-0.40%	0%	1.90%	2.60%
Diff-in-diffs	-17.6%***	2.60%	5.4%*	5.5%*	4.1%*
Corr vs. Placebo Diff Exp	7.4%***	0.10%	-2.30%	-3.9%**	-1.40%
Corr vs. Placebo Diff Surv	-4.10%	0.50%	3.60%	0.30%	-0.30%
Diff-in-diffs	-11.5%***	0.40%	5.9%*	4.20%	1.10%
Prem vs. Placebo Diff Exp	7%***	-2.5%*	0.70%	-4.3%**	-0.80%
Prem vs. Placebo Diff Surv	-3.20%	0.70%	1.60%	0.70%	0.30%
Diff-in-diffs	-10.2%**	3.20%	0.90%	4.90%	1.10%
FATF vs. Placebo Diff Exp	3.30%	0.20%	-0.40%	-1.70%	-1.50%
FATF vs. Placebo Diff Surv	-2.30%	1%	-0.60%	1.30%	0.60%
Diff-in-diffs	-5.60%	0.80%	-0.20%	3%	2.10%
PCT vs. Placebo Diff Exp	9.4%***	-1.8%*	-2.4%*	-3.9%***	-1.30%
PCT vs. Placebo Diff Surv	-3.80%	0.20%	1.80%	0.90%	0.90%
Diff-in-diffs	-13.2%***	2%	4.2%*	4.9%**	2.10%

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

Note: Table 7 shows the difference tests between each treatment condition and placebo, followed by a test of the difference in those differences, all for the international sample. If the difference in those differences is large, it suggests the experiment and survey are producing quite distinct results.

Table 8: Difference in Differences for U.S. CSPs

Differences	No response	Noncomp	Part comp	Comp	Refusal
Terr vs. Placebo Diff Exp	9.4%***	-5.1%***	-0.10%	0.20%	-4.3%***
Terr vs. Placebo Diff Surv	0.40%	0.70%	-1.7%*	1.0%**	-0.40%
Diff-in-diffs	-9.0%***	5.9%***	-1.50%	0.80%	3.9%*
Corr vs. Placebo Diff Exp	5.0%***	-1.00%	-0.10%	-0.20%	-3.7%***
Corr vs. Placebo Diff Surv	0.10%	0.70%	-1.00%	-0.20%	0.50%
Diff-in-diffs	-4.90%	1.70%	-1.00%	0	4.20%
IRS vs. Placebo Diff Exp	6.2%***	-3.6%***	0.60%	0%	-3.3%**
IRS vs. Placebo Diff Surv	-5.2%**	3.0%*	1.30%	0.10%	0.90%
Diff-in-diffs	-11.5%***	6.6%***	0.70%	0.10%	6.4%*
FATF vs. Placebo Diff Exp	3%	-1.20%	0.40%	0%	-2.20%
FATF vs. Placebo Diff Surv	-1.30%	1.10%	-1%	0.10%	1.20%
Diff-in-diffs	-4.40%	2.30%	-1.40%	0.10%	3.40%
ICT vs. Placebo Diff Exp	6.9%***	-3.3%***	0.10%	0%	-3.8%***
ICT vs. Placebo Diff Surv	-1.60%	1.40%	-0.50%	0.30%	0.30%
Diff-in-diffs	-8.4%***	4.7%***	-0.60%	0.30%	4.0%**

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

Note: Table 8 shows the difference tests between each treatment condition and placebo, followed by a test of the difference in those differences, all for the US sample. If the difference in those differences is large, it suggests the experiment and survey are producing quite distinct results.

It is important to point out that although a number of results demonstrate significant differences, still a large number of comparisons are quite similar. Notably, the lack of differences between experiment and survey tend to occur when there is no difference (between Placebo and treatment) within either the experiment or survey. And thus at a broad level the experiment and survey are not altogether different, especially in the null comparisons.

Conclusion

Those advocating for the greater use of experiments in political science must overcome the central objections that have previously limited the use of this approach in our field: doubts about practicality, ethics, and external validity. While survey experiments are clearly practical, critics have asserted that they offer suffer from problematic external validity. We share the view of others in political science and economics that the best way to judge this claim is to test survey experiments against other experiments.

The particular features of our anonymous incorporation field experiment give good reason to believe it has high external validity. Comparing CSPs' actual behavior against their survey responses tends to confirm some of the doubts raised about survey experiments. The survey response rate was a small fraction of that of the field experiment, indicating a pronounced selection bias. Different forms of experiments generated different samples of the population. Even more importantly, there was a massive difference in levels of hypothetical non-compliance compared with the level of

non-compliance indicated by the field experiment among the sub-sample that responded to both experiments.

Finally, we maintain that the use of deception is justified in this instance, both because methodologically we would have been unable to establish an accurate benchmark against which to judge the validity of our survey absent deception, but more importantly because anonymous shell companies pose a clear and present danger thanks to their role in enabling serious transnational crime.

References

Alpizar, Francisco, Frederik Carlsson, Olof Johansoon-Stenman. 2008. 'Does Context Matter More for Hypothetical than for Actual Contributions? Evidence from a Natural Field Experiment' *Experimental Economics* 11 (3): 299-314

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

Benz, Matthias and Stephan Meier. 2008. 'Do People Behave in Experiments as in the Field? Evidence from Donations' *Experimental Economics* 11 (3): 268-281

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, no. 4: 724-750.

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.

Findley, Michael G., Daniel L. Nielson, and J.C. Sharman. 2013. "Using Field Experiments in International Relations: A Randomized Study of Anonymous Incorporation." *International Organization* 67(4): 657-693.

Findley, Michael G., Daniel L. Nielson, and J.C. Sharman. 2014. *Global Shell Games: Experiments in Transnational Relations, Crime, and Terrorism.* Cambridge, UK: Cambridge University Press.

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. 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 W.W Norton New York.

Hovland, Carl V. 1959. 'Reconciling Results Derived from Experimental and Survey Studies of Attitude Change' *American Psychologist*, 14, No. 1: 8-17.

Kinder, Donald R. 2007. Curmudgeonly Advice. *Journal of Communication* 57 (1): 155-162.

King, Gary, Robert O. Keohane and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton: Princeton University Press.

List, John A. 2008, '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

List, John A and Levitt. 2008. "What do Laboratory Tests Measuring Social Preferences Tell Us about the Real World?" *Journal of Economic Perspectives* 21 (2): 153-174.

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

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

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.

Sniderman, Paul M and Thomas Piazza. 1993. *The Scar of Race*. Cambridge, Mass.: Harvard University Press.

Sniderman, Paul and Douglas Grob. 1996. Innovations in Experimental Designin Attitude Surveys. *American Review of Sociology* 22: 377-399.

Waltz, Kenneth N. 1979. *Theory of International Politics*. Reading: Addison-Wesley.