THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
CREATE CHANGE

Discussion Paper Series
UQ School of Economics

Paper Name: Scientific Inference from Field and Laboratory Economic Experiments: Empirical Evidence
Date: June 2023
Jonathan H.W. Tan, Zhao Zichen and Daniel John Zizzo

Discussion Paper No. 663

Scientific Inference from Field and Laboratory Economic Experiments: Empirical Evidence*

Jonathan H.W. Tan,[a] Zhao Zichen,[a] and Daniel John Zizzo[b]

Draft: June 23, 2023

Abstract

Field experiments can help improve scientific inference by providing access to diverse samples that are representative in terms of demographic backgrounds, and by availing the use of assets that relate directly to the economic problem of interest. We present a study comparing claims based on laboratory and field experiments in 520 publications in 2018 and 2019 at leading general and field journals in economics. Each paper is surveyed for their key claims and matches along the dimensions of profession, age, and gender of experimental subjects; country of experiment; and experimental asset in relation to which a claim is made. We find that, particularly in the realm of policy testing, field experiments are more likely to match the key claims than laboratory experiments. However, depending on the dimension, less than 20% or only up to around 65% of field experiments including natural field experiments achieve a match. Around four out of five field experiments fail to match in at least three out of the five dimensions. We conclude that the methodological challenge of generalizing results beyond what is within the domain of the experiments themselves also applies to many papers based on field experiments, given the claims being made. In addition, we find that publications by top 20 institutions authors or with experiments conducted in Caucasian-majority countries have a substantially higher likelihood of wide generalizations.

JEL classification codes: C18, C90.

Keywords: experimental economics, lab experiments, field experiments, validity.

* Acknowledgements: We thank participants at presentations in Birmingham, East Anglia, Lancaster, and Nottingham for valuable feedback. Tan is grateful for the financial support from the Nanyang Technological University Start-up Grant and Ministry of Education Singapore AcRF Tier 1 Grant (RG126/20). The usual disclaimer applies.
[a] Department of Economics, School of Social Sciences, Nanyang Technological University.
[b] Corresponding author: d.zizzo@uq.edu.au; School of Economics, The University of Queensland.

1. Introduction

We conduct an empirical study verifying the correspondence between stated claims and actual datasets in field and laboratory experimental papers, by considering 520 publications at 21 leading general and field journals in Economics in 2018 and 2019. Inference is a key and essential step in the scientific cycle of economic inquiry, the feasibility of which is constrained by the availability of data. Published papers based on experimental economics research make *claims* that generally imply an inference from a specific experimental dataset or datasets. As an example, Khan et al. (2019) analyzed a natural field experiment dataset involving subjects from the property tax department of Punjab (Pakistan) between July 2013 and June 2015; based on this dataset, they drew a general claim for governments about the value of "incentivizing bureaucrats through performance-based postings" (p. 237). As a second example, Chen et al. (2019) ran an artefactual field experiment with employer-sponsored patients from a health screening clinic in Shanghai in April 2014. From this experiment, claims were made regarding the effect of hunger on risk aversion, price-sensitivity, decision quality and other human decision making dimensions of interest, broken down by gender. As a third example, Crockett et al. (2019) employed data from a laboratory experiment with University of Pittsburgh undergraduate students from which they draw inferences and make claims about the validity of the Lucas asset pricing model. Inference is essential in these examples, and is needed regardless of the type of experiment for any claim arguing beyond what specifically took place in experiments conducted at a given point in time in Pakistan, Texas, or Pennsylvania, respectively.

What may vary, however, is the degree to which inference is required, and this is where our study provides answers. Consider three *dimensions* which have been argued in the past as potentially making a difference for the generalizability of experiments: location, profession, and asset being employed in the experiment. We label a claim as *within-domain* if – with respect to a given dimension – it is only about or is representative of the same location, profession, or asset that characterizes the experimental dataset it is based on. If this is the case, we shall state that there is a *match* between a claim and the corresponding dataset.

If, for example, Khan et al. (2019) only wished to make claims about what happens in property tax departments in Punjab or perhaps even Pakistan (generously assuming Punjab being representative of Pakistan), then we would be able to label the claim *with respect to the location dimension* as *within-domain* as the claim refers to the same location as the experimental data it is based on. As a second example, whereas Chen et al.'s (2019) data was collected from a specific sample of employer-

sponsored patients from a particular company, their claim referred to the general populations, so their claims are not *within-domain with respect to the profession*. As a third example, the asset pricing model is about money and because Crockett al.'s (2019) experiment is about money as the asset being used, we can note that the claims made in this paper are *within-domain with respect to the asset dimension*. Crucially, we are *not* arguing here whether out-of-domain inference is justified or not justified. Such an out-of-domain inference will be justified in many cases in the lack of a plausible confound based on the existing body of scientific evidence (Zizzo 2013). For example, in the case of Chen et al. (2019), one can argue that the physiological impact of hunger is unlikely to vary depending on the specific sample being employed.

The question of whether inference is within-domain or out-of-domain is nonetheless an important one. For one, it speaks to the way that claims are formulated in experimental papers and the need to be self-reflective of how ambitious we expect such claims to be. This is in a context where the publishing incentives in top journals are about persuading reviewers and editors that the paper has significant value added. In particular, nowadays desk rejection is a common practice in Economics journals that at least partially relates to how significant the expected contribution of a paper is.

For another, whether experimental claims are within-domain or out-of-domain speaks to the vexed issue of generalizability from experimental data, and specifically to testing the view that field experiments – particularly natural field experiments – are strong in their ability to make claims about the natural world in a way that laboratory experiments are not. In a series of papers with collaborators, List has argued for field experiments acting as a bridge between laboratory experiments and the "natural world" (e.g., al-Ubaydli and List 2015a, 2015b; List 2007, 2011; Levitt and List 2007). For example, al-Ubaydli and List (2015b) argue that, in general, natural field experiments will be superior to laboratory experiments, although the latter have an unavoidable residual role where natural field experiments cannot be properly constructed.[1] There is also a common view of a trade-off between between internal and external validity in experiments, and in this context field experiments are seen as stronger in external validity (e.g., among others, Campbell and Stanley 1963; Brinberg and McGrath 1985; Schram 2005; Cartwright 2007; Roe and Just 2009; Samek 2019; Lin et al. 2021).[2]

[1] They also recognize that laboratory experiments may have other virtues such as normally enabling the identification of qualitative effects, a point strongly stressed by Kessler and Vesterlund (2015). Herbst and Mas (2015) is a counterexample of quantitatively comparable evidence of productivity spillovers from experimental labor markets set in the lab and field.
[2] For various criticisms of this dichotomy or claims of trade-offs between internal and external validity, see for example Jimenez-Buedo and Miller (2010), Manski (2013), Jimenez-Buedo and Russo (2021) and Trafimow (2022). For robust defenses of the value of laboratory experiments, see e.g. Harrison (2013) and Camerer (2015).

Almost by definition, we expect laboratory experiments to be low on within-domain inference. For example, insofar as they use university students, any conclusion that is not specific to university students will involve out-of-domain inference. However, to what extent is it the case that actually published field experimental papers engage in within-domain inference and are therefore *prima facie* more plausible in terms of generalizability? And to what extent does the alternative view that "the issue of realism… is not a distinctive feature of lab versus field data" (Falk and Heckman 2009, p. 536) hold in practice?

We answer these questions by systematically analyzing the relationship between claims and experimental data in all the 520 experimental papers published in 21 top journals over the last two years prior to the COVID-19 pandemic, 2018 and 2019. Our study includes the world's top 5 general economics journals, other leading general economics journals (such as *Review of Economics and Statistics*), and top field journals both in experimental and behavioral economics (such as *Experimental Economics* and *Journal of Economic Behavior and Organization*) and in other fields (such as *Journal of Development Economics* and *Journal of Public Economics*). The full list is provided in section 2.2. Following Harrison and List (2004), experiments are classified as conventional lab experiments or artefactual, framed, and natural field experiments. Artefactual field experiments differ from laboratory experiments by not using university students. Framed field experiments add a natural frame, asset, or task to artefactual field experiments. In natural field experiments (such as randomized control trials), the environment is one where subjects naturally operate and are even unware that they are part of an experiment. We analyze the extent to which there are matches between claims and experimental datasets in each study across five dimensions: profession, age, gender of experimental subjects, country of experiment, and experimental asset in relation to which a claim is made.

Our dimensions largely map onto five of six factors that determine a field context as proposed by Harrison and List (2004), which Harrison (2013, p. 104) argues "could serve as contaminants of the inferences drawn from laboratory environments." Namely, we consider 1) the nature of the subject pool and 2) the nature of the information that the subjects bring to the task, which are captured by matching the profession, age, and gender of the subjects; as well as the 3) nature of the commodity; and 4) the nature of the task or trading rules applied, which are captured by matching the experimental asset. The experimental classification discussed above captures 6) the nature of the environment that

the subject operates in, which we additionally capture by matching to country and which is also correlated with the experimental asset.[3]

The choice of dimensions was guided by what was practically possible to identify consistently across papers, and, even so, mean age and gender ratio could only be identified respectively for 33% and 46% of papers in our sample. It was also determined by research showing that these dimensions may matter for decision making, at least in some settings. For example, Croson and Gneezy (2009) review gender differences in preferences emerging from experimental evidence. Belot el. (2015) reviews experimental research on differences between student and non-student samples; profession and age may explain any such differences. The behavioral peculiarities of subjects of Western, educated, industrialized, rich and democratic (WEIRD) societies and non-generalizability to other samples (Henrich et al. 2010a, 2010b) implies the potential importance of country as a dimension. Also, traditional research on the crowding out of motivation (e.g. Frey and Oberhozer-Gee 1997; Festré and Garrouste 2014) implies that conclusions found with money as an asset may or may not generalize to conclusions found when another asset is at stake. Of course, this is not to say that these dimensions may necessarily be relevant for the claim being made in an experiment. To mention two examples, Hackethal et al. (2022) argue for the qualitative comparability of risk behaviors across experiments with or without monetary incentives and conducted on private investors, professional investors, and students, suggesting that matching profession and stakes might not be so important; and Cappelen et al. (2016) argue for the robustness across lab and field samples of experimental tests of fairness responses. Taking a different tack, Camerer (2015) argues that generalizability is irrelevant for theory-testing experiments. Whatever the stance one takes on Camerer's (2015) position, our empirical analysis will test whether it is the case that we are more likely to find out-of-domain inference in theory-testing experiments, and whether conversely we find less of it in policy testbed experiments.

Our summary results are as follows. Conventional lab experiments are with standard university student samples, and while they generally do not achieve matches for profession and country, their asset match rate of around 20% compares to that of artefactual and framed field experiments. Similarly, for those papers that provide gender breakdown, conventional lab experiments do roughly as well as any field experiment category (at 44-67% each). Natural field experiments have the highest asset match

[3] We did not consider 5) the nature of the stakes as a dimension, as it was not possible to identify this in a consistent way across 520 studies, given differences in the type of asset, countries, subject pools and information provided about the experiment. However, we did identify as a control variable for our analysis whether the experiment was financially incentivized or not. Also, we initially thought to include the experimental setting as a direct proxy for nature of the environment, but (other than in relation to the type of experiments) we were not able to identify a clear, unambiguous and workable taxonomy for how a setting should be identified.

rate, but this is still only around 60% and appears driven by policy testbed experiments. Profession match is highest with artefactual and framed field experiments, but still again only around 60%. Country match is always lower than 20%, and age match is always lower than 50%. Only one out of five field experiments achieves a match in at least 3 out of 5 dimensions. As conjectured, theory testing and policy testbed experimental papers have a lower and a higher match rate, respectively, Our analysis also controls for other covariates: papers with experiments run in countries with a majority-White population (a partial proxy for WEIRD countries), and papers with authors at the top 20 institutions in the world (in the RePEc ranking), tend to generalize their claims more (i.e. have a lower match rate) relative to the experimental datasets they actually have.

Section 2 presents the methodology. Section 3 presents the results. Section 4 discusses and concludes.

2. Methodology

2.1 Matches over five dimensions

The key building block of our analysis is the notion of a *match* between an experimental claim and the experimental data for a given *dimension* (such as experimental asset). We define a match as *positive or negative* for a given dimension in which there was a correspondence or otherwise, respectively, between the specific experiment being conducted and the claims made for that dimension.[4] Specifically:

- ***Profession***: a match was positive (a *profession match*) if the experimental data was drawn from the specific profession(s) or group(s) (e.g., students; doctors; patients) for which the claim was made. Otherwise, a match was negative if the claims were generalizable to other professions, groups, or the general population.
- ***Country***: a match was positive (a *country match*) if the claim referred to the country or countries where the experiment was conducted without extrapolating these results to the rest of world.
- ***Asset*****:** a match was positive (an *asset match*) if the asset(s) employed in experiment matched that (those) of the claims made. Examples of *asset* include money, effort, human capital, and health.

[4] Whenever we are not explicit whether a match is positive or negative (or indeterminate), we refer to a positive match.

- ***Age*:** a match (an *age match*) was positive in two cases. First, if the study was about comparing different age cohorts (there were 4 in the dataset), there is a match if the relevant age cohorts are represented in the sample. Second, a positive match was deemed if the mean age of the experimental data was close enough to the representative age of the relevant population, i.e., $\pm$ 5 years. If the claim was made for a specific cohort, it would be recorded as a missing observation unless the representative age for that cohort is known. Otherwise, for claims about general population, we used the country-specific mean age as the representative age (Central Intelligence Agency 2021).
- ***Gender*:** a match (a *gender match*) was positive in two cases. First, if the study was about gender effects (there were 18 in the dataset), there is a match if the relevant genders are represented in the sample. Second, a positive match was deemed if the mean gender ratio of the experimental data was close enough to the representative gender ratio of the population, i.e., $\pm$ 5%: if the claim was made for a specific cohort, it would be recorded as a missing observation since the gender ratio is unknown. Otherwise, for claims about the general population, we considered a 45-55% recorded male-to-female ratio in the experimental data as representative.

Seven out of 520 papers include two types of experiments, e.g. Haggag et al. (2019) includes both a conventional and a framed field experiment. Whenever multiple experiments are included in a paper, whether from the same type or from different types of experiments, we deem a positive match to exist for the study if a positive match holds for at least one of the experiments.

We provide a more detailed discussion of our classification process in the next section. One could argue that our match test related to profession, country, and asset is quite a conservative test in that any claim that is not strictly related to the given profession(s), country(ies), and asset(s) leads to a negative match in our classification exercise. In the robustness analysis described in section 3.3, we relax this test by applying what we label as the "rule of 2" and the "rule of 3". According to the rule of 2 (rule of 3), if an experiment presents data from two (three) professions, countries or assets, we deem a positive match to be met, even if the claim is more general. Obviously, these robustness tests err on the opposite side of being over-generous (e.g. one could hardly label an experiment run in Denmark, France, and Spain representative of non-WEIRD countries), but the point of considering them is to show that our basic findings are robust to allowing for them.

To illustrate our methodology, we list 5 examples of how we determine matches over the five dimensions in Table 1.

Example 1. Buyukboyaci et al. (2019) ran both a theory testing and policy testbed conventional laboratory experiment in Turkey. They draw hypotheses based on theory and claim that "individual investment choices decrease in response to an increase in the probability of bankruptcy or an increase in risk aversion; total investment difference between proportionality and either of the other two principles is independent of the probability of bankruptcy as long as both induce an interior equilibrium" (p.607). These claims are not restricted to by country, so there is a negative match on this regard as well as obviously on profession, given that university students are employed. Students are younger on average than the general population, so there is a negative match in this dimension as well. However, the experiment employs money and bankruptcy environments are about money, so there is an asset match. There is also a gender match.

Example 2. Candelo et al. (2018) have an artefactual field experiment on transmission of information within transnational social networks. The key claim of the paper relates to immigrants and their home-based social network members, and is about "how information transmitted via social networks across national boundaries influences financial decisions of individuals in the home country " (p. 906): there is a profession match. The paper is clear that it is about financial risky decisions and the experiment is about money, so there is an asset match. We do not know what the mean age and gender is in the relevant population, so we cannot determine age or gender match. There is no qualification of the results by country, so there is a negative match on this dimension.

Example 3. Cilliers et al. (2018) employ two natural field experiments in Ugandan primary schools to test the effect of paying for locally monitored performance. Their key claim is that "we demonstrate that pay for locally monitored performance passes both welfare and fiscal sustainability test" (p. 69), and this is more general than just with reference to Ugandan primary school teachers or indeed education, implying a negative match in terms of profession, country, and asset. Because of the nature of the relevant sample being public service workers, we do not know what the representative mean age and gender ratio is, and so we cannot assess age and gender matching.

Example 4. Haggag et al. (2019) run an artefactual field experiment (with US MTurk workers) and a conventional laboratory experiment (with US university students). Their key claim is a general one that "people misattribute the influence of a temporary state to a stable quality of the consumption good" (p. 2136), and, because of its generality, there is a negative match for country and profession (unless we were to use the rule of 2, since the experiment tests both MTurk workers and university students). The gender ratio is outside the 45-55% range for both experiments, entailing a negative match for gender. The MTurk experiment however is within ± 5 years of the US population mean age, implying

a positive age match. The experiment uses a self-made drink but draws an inference for general consumption goods, and so there is a negative match for asset.

Example 5. Muralidharan et al. (2019) present a policy testbed natural field experiment with "a technology-led supplementary instruction program in post-primary grades in urban India" (p. 1456) and are clear that their key claim is about the positive effect of this specific program. Indeed, the experiment is held in India, with the data being collected from five public middle schools. And the experimental asset they examine is human capital as proxied by test scores, which fall into the same domain as their claim. There is therefore a positive match for profession, country and asset. We do not know what the age range and gender ratio of India middle school students is, and so we cannot assess the age and gender match (even though the information about the sample is provided).

Table 1. *Examples of five dimensions and matches*

		Profession	**Country**	**Asset**	**Age**	**Gender**
Buyukboyaci et al. (2019)	claim	general	general	money	32.2	45-55
	data	students (-)	Turkey (-)	money (+)	21.69 (-)	47 (+)
Candelo et al. (2018)	claim	immigrants & social network members	general	money	missing	missing
	data	immigrants & social network members (+)	USA & Mexico (-)	money (+)	37.5 (?)	47 (?)
Cilliers et al. (2018)	claim	public service workers	general	general public services	missing	missing
	data	primary school teachers (-)	Uganda (-)	teaching (-)	34.92 (?)	41 (?)
Haggag et al. (2019)	claim	general	general	general consumption goods	38.5	45-55
	data	Mturk workers, students (-)	USA (-)	self-made drink (-)	34.09 (+) 30.54 (-)	59 (-) 43 (-)
Muralidharan et al. (2019)	claim	middle-school students	India	human capital	missing	missing
	data	middle-school students (+)	India (+)	human capital (+)	12 (?)	76 (?)

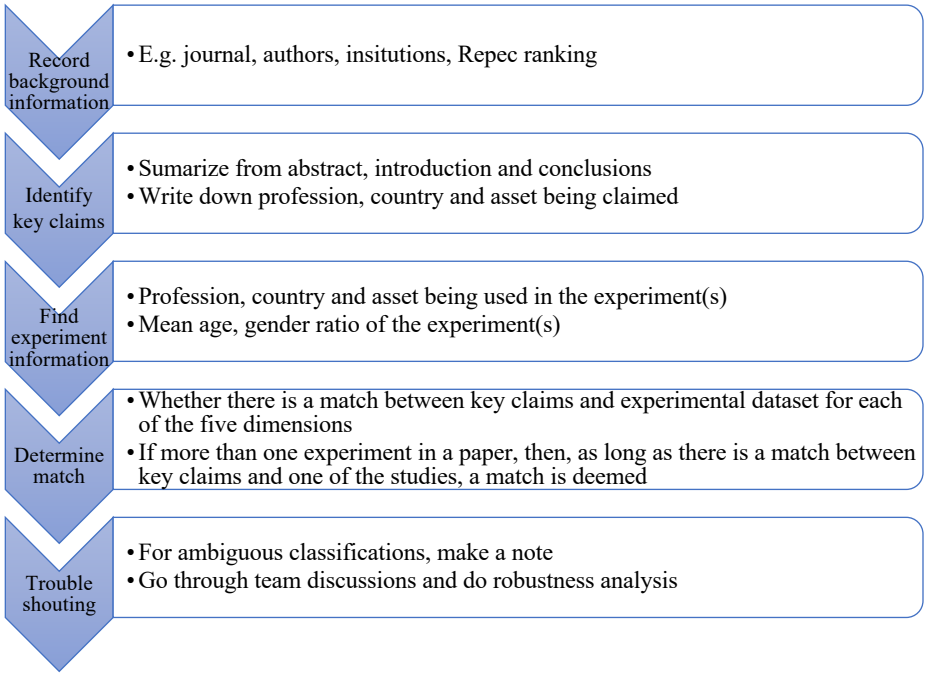
Note: We use the term "general" to denote that a claim applies generally, at least implicitly, across professions, countries, assets, genders or ages. (+), (-) and (?) denote a positive, negative and undetermined match, respectively. Papers 1 and 4 were published in top 5 journals (*AER* and *Rev Ec Stud*), whereas papers 2, 3 and 5 were published in top field journals (*JEBO*, *J Pub Econ* and *Exp Econ*).

2.2 Procedure

We tested our methodology in a pilot stage where one of the authors collected data on 100 papers and the other two authors reviewed 15% of the work. After several iterations of meetings to discuss how to address the problems encountered and how to refine the protocol, we formulated a systematic procedure as shown in Figure 1 and, in relation to matching, finalized as described in the previous sub-

section. Following this procedure, we reviewed all experimental papers that we could find published in 2018 and 2019 in 21 leading journals (see Table 2 for the list). As in Reuben et al. (2022), we excluded simulation studies, pure theoretical studies, meta-analyses, and natural experiments where treatment assignment is not due to an active choice by the experimenter.

Figure 1. *Flowchart of the procedure*



Reviewing each paper was a time-consuming exercise that took on average 20 minutes per paper. This was because of the need to identify the key experimental claims and ensure that any qualifications that could matter for the matching protocol were identified; the need to identify what the claims entail in terms of the five dimensions; as well as the need to identify what is entailed from the experimental dataset in relation to the dimensions.[5] We thus collected a dataset of 520 papers from a range of leading journals; while this number reflectes our resources constraints, further research could of course look at more papers including further top journals and years of publication.

Perhaps because of the journals being all top journals and so placing a premium on clarity of the value added, once the protocol was finalized, it was normally straightforward to identify the key claims and any associated qualification. As discussed further in section 3.3, there were nonetheless a small

[5] It takes about 10 minutes for papers with clear claims and information of the experiment being explicitly stated. Nevertheless, it takes up to 45 minutes if the papers have ambiguous claims (or qualifications), or experimental data needs to be extracted from online appendix or supplementary materials.

minority of cases (6.7%) where judging the matches was not straightforward; the Appendix B1 re-does the analysis excluding these cases to check robustness.

We constructed a dataset with a total of 520 papers as described in Table 2. In our dataset, 60.4% of the papers used conventional laboratory experiments, 8.3% used artefactual field experiments, 12.1% used framed field experiments, 17.9% used natural field experiments, and the remaining (just seven) used mixed experimental methods, i.e. more than one type of experiment (specifically, two). Our dataset is largely derived from top field journals that are popular outlets for experimental economics, such as the *Journal of Economic Behavior and Organization* (35.8%), *Games and Economic Behavior* (12.3%), and *Experimental Economics* (10.8%), while 9.2% are from world top 5 economics journals. About one-tenth of the experiments are conducted online and nine-tenths are incentivized. 46% of the papers test a theoretical model, while 25.2% were testbeds for a policy (the two objectives can co-exist: see e.g. Example 1 above). For each paper, we recorded the highest rank of the authors' institutional affiliations based on the RePEc ranking for the last 10 years, as of December 2022. Slightly less than one-tenth of the papers are done by authors from top 20 institutions. There is no established global list of WEIRD countries, and so, rather than creating our own, we proxied this by focusing on the W: 74.6% of experiments were conducted in majority Caucasian (White) population countries (World Population Review 2023).

Table 2. *Descriptive statistics of the dataset*

Type of Experiment	Conventional laboratory: 314 (60.4%), artefactual FE: 43 (8.3%), Framed FE: 63 (12.1%), Natural FE: 93 (17.9%), Mixed: 7 (1.3%)
Characteristics	Policy testbeds: 131 (25.2%), Theory testing: 239 (46%), Online: 61 (11.7%), Incentivized: 468 (90%)
Institution Ranking	Top 20: 45 (8.7%), Top 21-100: 100 (19.2%)
Country	Caucasian countries: 388 (74.6%), Multiple counties: 21 (4%)
Experimental Economics Journals	Experimental Economics: 56 (10.8%), Games and Economic Behavior: 64 (12.3%), Journal of Economic Behavior and Organization: 186 (35.8%), Journal of Risk and Uncertainty: 17 (3.3%)
Top 5 Journals	American Economic Review: 18 (3.5%), Econometrica: 4 (0.8%), Journal of Political Economy: 5 (1%), Review of Economic Studies 11 (2.1%), The Quarterly Journal of Economics: 10 (1.9%)
Other Journals	International Economic Review: 7 (1.3%), Journal of Agricultural Economics: 16 (3.1%), Journal of Development Economics: 23 (4.4%), Journal of Economic Theory: 8 (1.5%), Journal of Environmental Economics and Management: 26 (5%), Journal of Health Economics: 13 (2.5%), Journal of Labor Economics: 4 (0.8%), Journal of Monetary Economics: 3(0.6%), Journal of Public Economics: 31 (6%), Journal of the European Economic Association: 11 (2.1%), The Journal of Law and Economics: 4 (0.8%), The Review of Economics and Statistics: 3 (0.6%),

Note: The number stated before each parenthesis refers to the number of papers, followed by its percentage of the 520 studies enclosed in the parenthesis. FE stands for field experiment.

3. Results

In this section, we first present some descriptive statistics and analyze the extent to which papers based on different type of experiments achieve positive matches over the five dimensions. Next, we consider the features of publications that achieve matches in at least three dimensions. Finally, we discuss some robustness checks. Further analysis is included in the online appendix.

3.1 Descriptive statistics

Table 3 shows the (positive) match rates over the five dimensions, i.e. the ratio between the number of papers for which there is a positive match rate and the number of papers for which a positive or negative match rate can be determined. For the age and gender dimensions, we get a large proportion of missing values (67.3% and 54.2%, respectively) either due to unknown representative age and gender, or the relevant information not being provided in the paper. Among the papers with no missing values, the gender match rate is 51.7%. Instead, profession and asset match rates are just around 25%, with the mean plummeting to 6% for the country match, as papers rarely make claims specific to a country.

Table 3. *Descriptive statistics of positive match variables*

Variables	Frequency	Percentage
Profession match (=1)	120	23.1%
Country match (=1)	31	6.0%
Asset match (=1)	139	26.7%
Age match (=1)	31	18.2%
Missing value	350	67.3%
Gender match (=1)	123	51.7%
Missing value	282	54.2%

Note: Where there are missing values, percentages are computed as a proportion of the papers for which values are available.

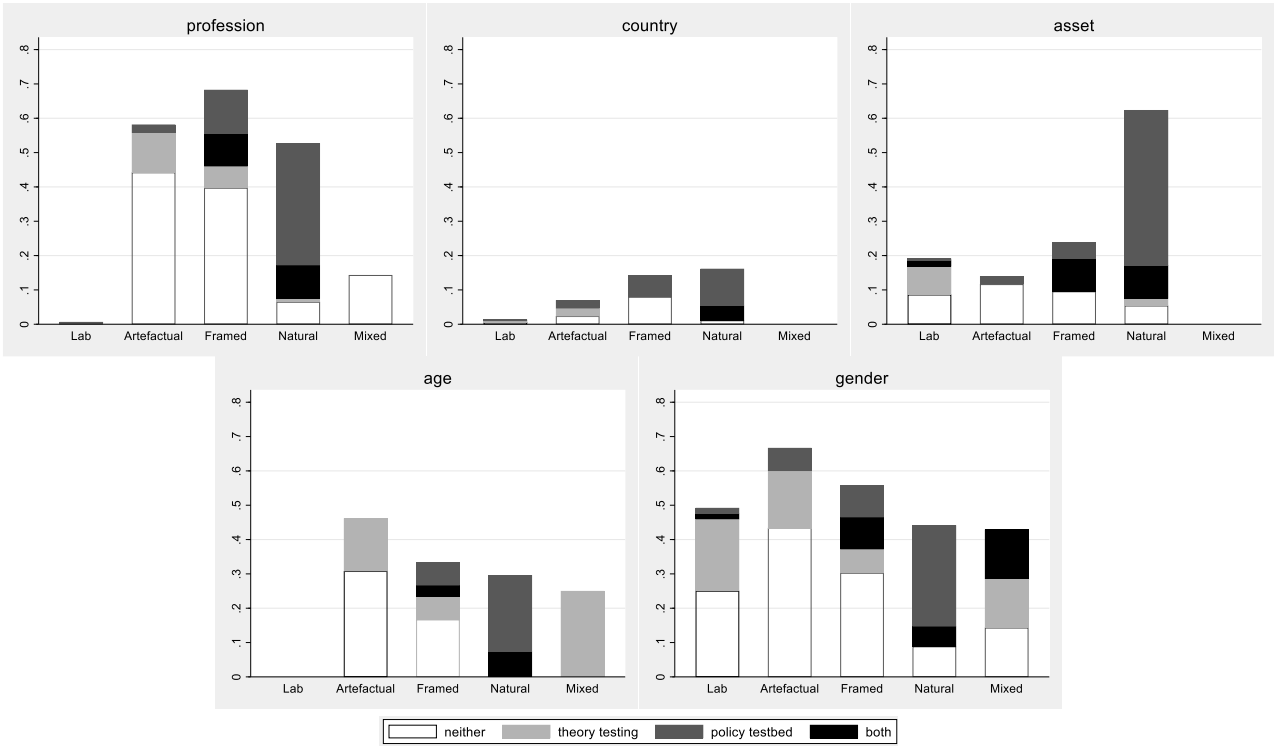
Figure 2 depicts the match rates in five dimensions of each experiment category.[6] Given the use of standard university student samples, it is unsurprising that conventional lab experiments generally do not achieve positive matches on profession and age; the country match is also very low. However, their asset match rate is at around 20%, comparable to that of artefactual and framed field experiments

[6] While included for completeness, we will not comment specifically on the Mixed category, given the tiny nature of the sample involved (n = 7).

(14% and 24% respectively).[7] The gender match rate is also comparable to that of any field experiment category (at 44-67%).

Result 1. *Conventional laboratory experiments achieve asset match rates of around 20% and gender match rates of around 50%.*

Figure 2. *Match rates of claims based on each experiment type for each dimension*



Note: Lab: conventional lab experiments; Artefactual, Framed, Natural: artefactual, framed, natural field experiments, respectively. Mixed: more than 1 experiment type.

At the other extreme, papers based (purely) on natural field experiments only have a significantly higher match rate than the other field experiment dimensions in relation to the asset match (Mann-Whitney both $p < .01$, two-tailed), although even then this is bounded at around 60%. This higher asset match rate for natural field experiments is entirely driven by policy testbed experiments; Figure 2 shows that the point asset match rate is actually lower than the other categories, including conventional laboratory experiments, if any policy testing experiments are excluded. For other categories, the match rates for natural field experiments are either not significantly higher (country match, at 16%) or lower (at 53%, 30% and 44% for profession, age and gender respectively) than other field experiment categories.

[7] This is because a number of papers in conventional lab experiments make claims about monetary decisions (e.g., behaviors in financial market/context, studying a game without external inference), so employing money in the experiment can obtain a match.

Result 2. *Natural field experiments have a higher asset match rates than other categories of field experiments only for asset match. This is entirely driven by policy testbed experiments. Match rates are the same as or lower than other categories of experiments for the other dimensions.*

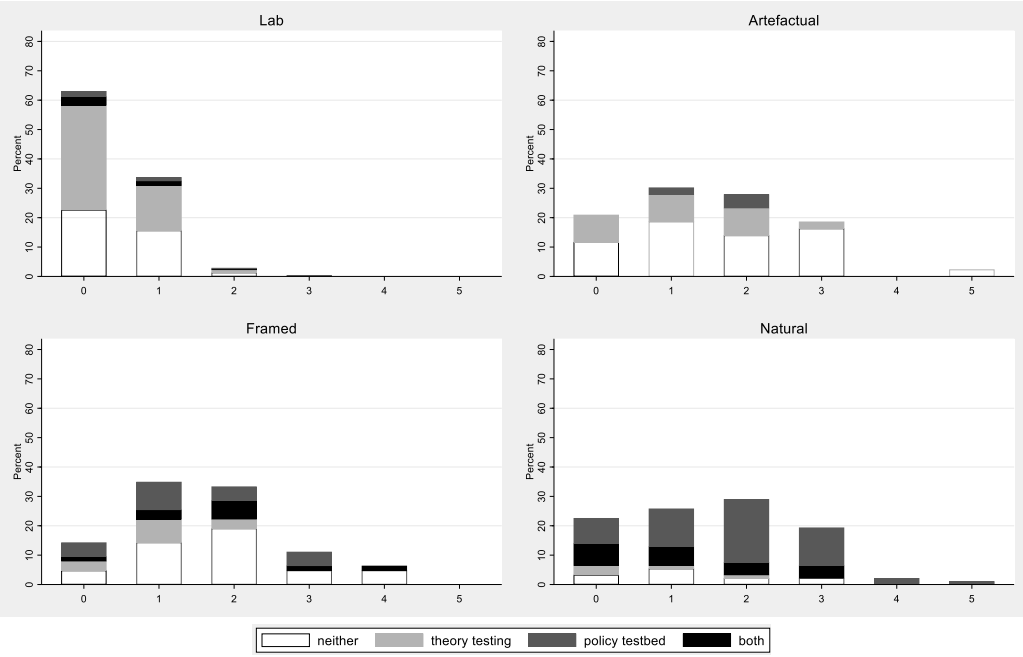
None of the field experiment categories have match rates higher than around 60% or 2/3, with country match being particularly low (bounded at 16%).

Result 3. *No category of field experiments has a match rate higher than around 2/3, with country match being no higher than around 16%.*

Figure 2 shows that a number of pure theory testing papers among the conventional laboratory, artefactual, or framed field experiments can achieve positive matches; but that, in general, many positive matches are produced by papers at least one goal of which is being a policy testbed. The next section, and the Appendix A1 with further analysis, look at this matter further.

3.2 Joint analysis on match likelihood

Figure 3. *Distribution of number of matches for each type of experiment*



Note: Each figure shows the number of matches (horizontal axis) that applies to a given proportion of papers (vertical axis) for each type of experiment.

We have so far considered the five dimensions separately. We now consider them jointly. We specifically look at which kind of papers achieves a positive match in least three dimensions. Figure 3

shows the proportion of number of matches for papers based on each type of experiments.[8] Given that age and gender have many missing observations, it is particularly interesting to focus on the proportion of papers that has at least three matches. Unsurprisingly given their nature and section 3.1, only one paper based on conventional laboratory experiments has three matches. However, field experiments do not do much better: just 23% of natural field experiments obtain a match in at least three out of five dimensions, which is not statistically different from the percentages we observe for artefactual and framed field experiments (around 20%, Mann-Whitney both $p > 0.1$). 90% of natural field experiments achieving at least three matches are policy testbeds.

Result 4. *Only around 20% of papers based on field experiments have a match rate in at least three dimensions. Papers based on natural field experiments are not statistically different from papers based on other kinds of field experiments in achieving three matches; insofar as they do achieve three matches, this is largely tied to policy testbeds.*

We present robustness checks of Result 4 in the next section. More generally, among natural field experiments, we find that the probability that a policy paper get at least three matches is twice that of a non-policy paper (25% vs. 12%). On the contrary, a theory-testing paper is 10 percent less likely to get at least three matches than a non-theory paper (25% vs. 15%).

We now present a regression analysis on which kind of papers achieves a positive match in three dimensions. Given the number of missing observations for age and gender, we restrict the regression analysis to the three other dimensions: profession, country and asset.[9] We run OLS, Logit, and Probit regressions on *sum of matches* in profession, country, and asset. Models 1-3 test our previous findings directly. The dummy variables *Artefactual FE, Framed FE*, and *Natural FE* (=1 for each respective category) compare the overall match rates in each field experiment category to the conventional laboratory experiment baseline. *Policy testbed* (=1 if policy testbed papers, 0 if otherwise) examines whether policy papers get higher matches, while *Theory testing* (=1 if theory testing papers, 0 if otherwise) tests whether theory papers achieve less matches. Models 4-6 add additional control variables and error clustering at country-level. *Top21-100* and *Top20* (=1 if the highest institution ranking is 21-100 and below 20 respectively, 0 if otherwise) control for authors' institutions. *Caucasian* (=1 if the experiment was conducted in a majority Caucasian country, 0 if otherwise) is a proxy for the W (White) of WEIRD countries. *Journal_top5* and *Jounrnal_exp* (=1 if published in world top 5 and experiment economics journals respectively – see Table 2 -, 0 if otherwise) controls

[8] Missing values in age and gender are considered not achieving a match here. An analysis excluding missing values can be found below. Given small sample size, the n = 7 mixed experimental methods papers are reviewed in the Appendix A2.
[9] The Appendix A3 includes regressions on match over each of the five dimensions separately.

for the journal the paper was published in. *Incentive* (=1 if incentivized experiments, 0 if otherwise) and *Online* (=1 if online experiments, 0 if otherwise) control for other aspects of the experiments.

Table 4 reports the results. Field experiments of all types have significantly higher match rates than conventional laboratory experiments in models 1-6, an unsurprising effect exacerbated by the fact that conventional laboratory experiments do comparatively better for gender and age, which are not represented in these regressions. The differences between the coefficients of different field experiment are not significant (all $p > .1$, except $p < .05$ for *Artefactual FE* ≠ *Natural FE* in model 1). *Policy testbed* and *Theory testing* are respectively statistically positive and negative in all models.

Result 5. *Policy testing papers are more likely to achieve a match in the three dimensions for which we do not have missing observations. Theory testing papers are less likely to achieve a match in the three dimensions instead.*

Table 4. *OLS, Logit, and Probit regressions on sum of matches (out of three)*

	Dependent Variable: ***Sum of matches = profession + country + asset***					
	OLS	Logit	Probit	OLS	Logit	Probit
	(1)	(2)	(3)	(4)	(5)	(6)
Artefactual FE	0.535***	0.396***	0.397***	0.503***	0.378***	0.374***
	(0.104)	(0.077)	(0.078)	(0.171)	(0.117)	(0.121)
Framed FE	0.733***	0.462***	0.462***	0.627***	0.414***	0.414***
	(0.093)	(0.068)	(0.068)	(0.210)	(0.132)	(0.132)
Natural FE	0.880***	0.412***	0.418***	0.741***	0.332**	0.340**
	(0.099)	(0.077)	(0.075)	(0.159)	(0.135)	(0.133)
Mixed	-0.113	-0.075	-0.079	-0.163	-0.114	-0.114
	(0.243)	(0.140)	(0.138)	(0.173)	(0.146)	(0.138)
Policy testbed	0.230***	0.093*	0.092*	0.188**	0.076*	0.076*
	(0.084)	(0.052)	(0.052)	(0.091)	(0.045)	(0.045)
Theory testing	-0.184***	-0.112***	-0.111***	-0.169***	-0.103***	-0.102***
	(0.058)	(0.037)	(0.037)	(0.051)	(0.035)	(0.036)
Incentives				-0.179	-0.096	-0.098
				(0.161)	(0.075)	(0.077)
Top21-100				-0.045	-0.028	-0.032
				(0.055)	(0.029)	(0.029)
Top20				-0.245*	-0.104**	-0.102**
				(0.128)	(0.048)	(0.046)
Online				-0.147	-0.071	-0.071
				(0.098)	(0.068)	(0.068)
Caucasian				-0.268***	-0.137***	-0.142***
				(0.101)	(0.045)	(0.048)
Journal_top5				-0.006	0.046	0.046
				(0.126)	(0.060)	(0.060)
Journal_exp				-0.138	-0.020	-0.018
				(0.093)	(0.060)	(0.060)
Clustered	no	no	no	yes	yes	yes
Observations	520	520	520	520	520	520

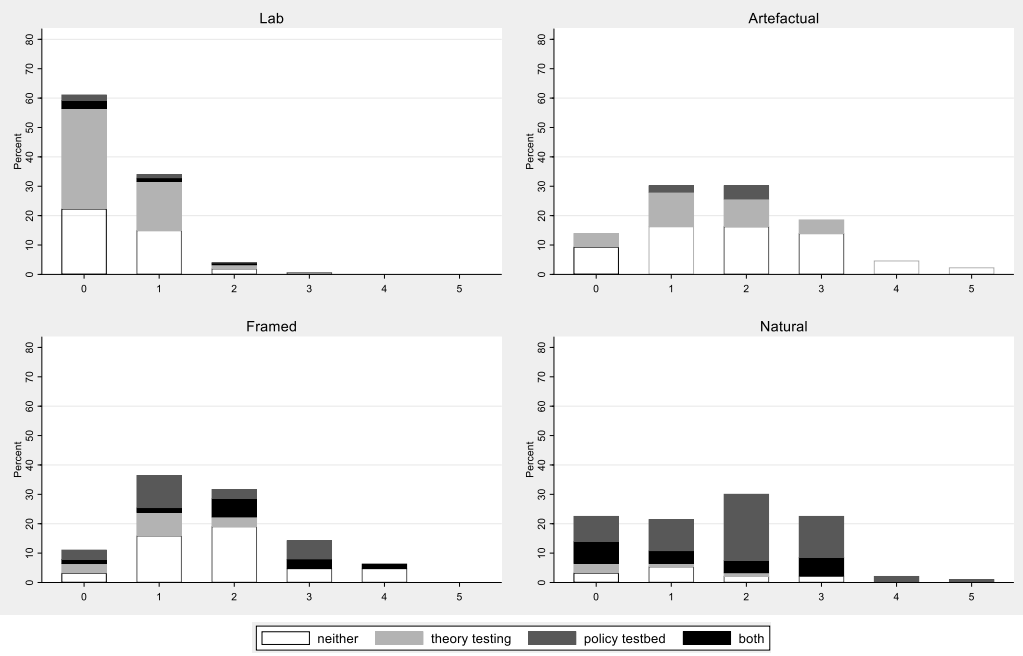
Note: Marginal effects are reported here. Robust standard errors clustered by country are in parentheses. Two-tailed $p <$ 0.1*, 0.05**, 0.01***.

Turning to the demographic covariates of authors and samples, *Top20* and *Caucasian* are negative and statistically significant in models 4-6. If at least one author is in a top 20 institution, or if the experiment was conducted in a Caucasian country, such papers can afford to claim more with fewer matches. Papers with top 20 institution co-authors can be some 10-25% less likely to achieve a match in the three dimensions and still get published in the leading journals in our sample. Similarly, if the experiment was run in a Caucasian-majority country, the authors can be some 15-25% less likely to achieve a match in the three dimensions and still get published in the leading journals in our sample. We do not, instead, find a difference in the publication strategy of world top 5 journal and of experimental economics journals.

3.3 Robustness checks

Gray area cases. As a robustness check, we exclude papers where we feel unsure to judge whether a match is positive or negative. The uncertainty mainly originates from the ambiguousness of the writing.[10] We identify 35 papers (6.7%) in the gray area. They are distributed across a variety of publications (14 lab vs. 28 field; 12 policy vs. 21 non-policy; 12 theory vs. 23 non-theory), which means this problem exists in all kinds of papers rather than a particular type. Our results are robust to excluding the gray area papers, which is unsurprising given their small number: see the Appendix B1.

Figure 4. *Percentage of each category with 0-5 matches under 'rule of 2'*

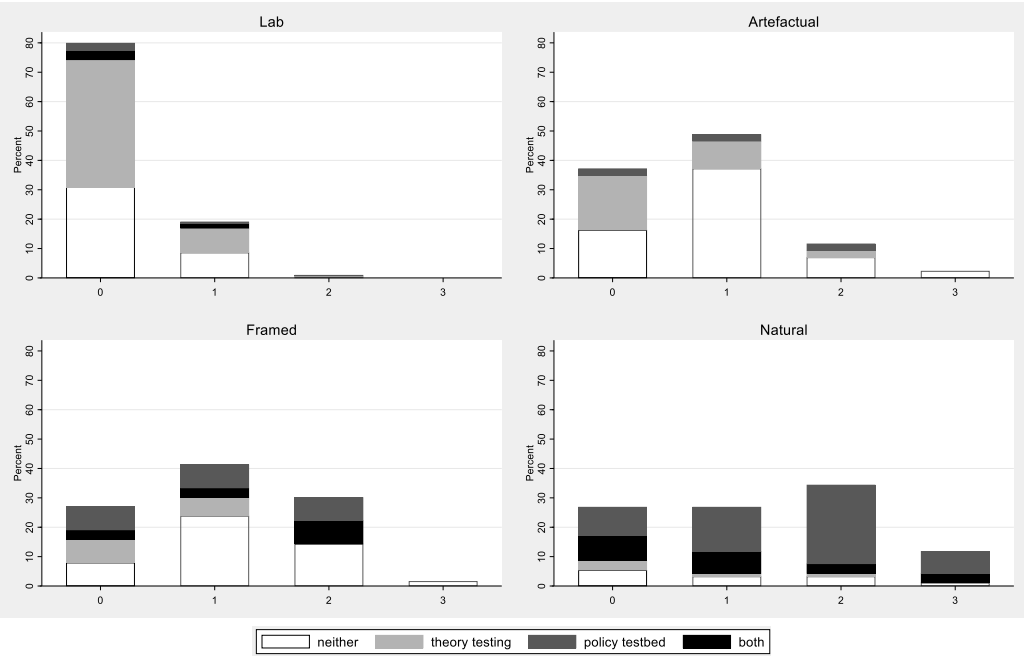


[10] For example, whether some theoretical papers are making claims about the games itself or having external implications; or whether some field policy papers specify the context where the policy is implemented or intend to extrapolate the results.

Rule of 2 and rule of 3. Section 2.2 discussed possible over-generous relaxations of the rules associated to profession, country, and asset match. We now check whether our key findings change if we apply a rule of 2 (3) where a general match is considered justified if there are at least 2 (3) professions/countries/assets being relevant to the experiment(s) in a paper. Only an additional 16 (2) profession matches, 15 (2) country matches and 15 (8) asset matches are achieved if we follow the rule of 2 (3), respectively. Figure 4 depicts the distribution of number of matches in each category under 'rule of 2'. Even under this over-generous match rule, natural field experiments still only achieve a match rate of about 24%, while artefactual and framed field experiments remain below 20%. The key findings of the regression analysis also do not change (see Appendix B2).

Missing values. Another robustness check is to verify how Figure 3 changes if we exclude age and gender when examining papers of at least three matches, given the many missing values for those two dimensions. Figure 5 shows the fraction of papers with three matches (out of three) for each type of experiment. Since artefactual and framed field experiments do comparatively well for age and gender, it is not surprising that, once we remove them, natural field experiments have a clearly higher match rate (Mann-Whitney $p = 0.12$ and 0.03 respectively). This is still driven by the natural field experiments papers on policy testing. However, only around one natural field experiment out of ten jointly achieves a positive match in profession, country and asset, with the corresponding percentage being close to 0% for artefactual and framed field experiments.

Figure 5. *Fraction of papers with three matches (out of three) for each type of experiment*



4. Discussion and conclusion

Let us emphasize once again that inference, including out-of-domain inference, plays an essential role in the scientific enterprise. We have studied inference over five dimensions: age, gender, profession, country, and asset. An inference is considered *within-domain* when, with respect to a given dimension, the experimental claim is only about the same unit or units within that dimension that characterize the experimental dataset, or the experimental dataset is otherwise representative of what is stated or implied in the experimental claim in relation to that dimension. For example, a within-domain inference with respect to country would be one where an experiment is conducted in the USA, and the claim is about a US policy (Aimone et al. 2019); or, with respect to asset, the experimental asset is health and the claim is about health (Carrera et al. 2018). Section 2 has detailed the protocol we have followed for defining within-domain inference for our five dimensions, and section 3.3 has described some robustness analysis. We have out-of-domain inference when the conditions for within-domain inference are not met.[11]

We stress again that there is nothing wrong as such with out-of-domain inference, and indeed this will be justified in many cases, for example based on a holistic reading of the available literature and what it entails for the plausibility and relevance of a confound as it may affect the validity of such inference (Zizzo 2013). Our evaluations of whether a positive or negative match has been achieved are an interpretation of fact, not an interpretation of the scientific merit of a paper. It is nonetheless interesting to verify the correspondence between claims and evidence in terms of within-domain and out-of-domain inference. There are two primary reasons for this. First, because the Economics publishing incentives arguably incentivizes over-claiming to show value added relative to existing evidence, for example to pass the desk rejection hurdle that most Economics journals now have.

Second, it matters because of the empirical contribution that our exercise makes to the debate on conventional laboratory experiments and field experiments, and particularly natural field experiments as the gold standard of field experiments. There are aspects of this debate, such as the potential presence of experimenter demand in experiments other than natural field experiments, which our

[11] We have carefully avoided casting the within-domain versus between-domain inference in terms of the external validity versus internal validity dichotomy, given the controversies surrounding such a distinction (see section 1 for references). Readers who like to think in terms of external validity can however see the distinction as one of strength of external validity of experimental claims: within-domain inference is associated with high external validity while the external validity of out-of-domain inference may require additional assumptions or justification. An alternative framework that is closer to our own thinking is Cronbach's (1982) UTOS framework for scientific inference (see Jimenez Buedo and Russo 2021 for a summary): within-domain inference simply involves a high level conceptual interpretation over a given dimension based on the population the experimental sample is representatively drawn from; out-of-domain inference instead involves extrapolation of the experimental claim to other situations of interest.

contribution does not touch.[12] The aspect our empirical contribution is relevant for has been sometimes cast along the lines of (especially natural) field experiments having greater external validity than conventional laboratory experiments, due to them being 'closer to reality' (e.g., Cartwright 2007; List 2007, 2011; Levitt and List 2007; al-Ubaydli and List 2015a, 2015b; Samek 2019). There is a sense in which this is trivially true: it is trivially true, for example, that in moving from a conventional laboratory experiment to an artefactual field experiment, we are replacing a university student sample with a profession of interest; therefore, insofar as the profession of interest is not made of university students, there is the potential to achieve within-domain inference over this dimension. This explains the intuitive appeal of field experiments to policy makers as discussed for example by al-Ubaydli and List (2015b). What this argument however ignores is that inference is not a property of experimental datasets, but rather of claims that are made based on such experimental datasets (see Jimenez-Buedo and Russo 2021). Less than 20% or only up to around 65% of field experiments, including natural field experiments, achieve a match in any given dimension out of the five we have investigated. Around four out of five field experiments fail to match in at least three out of five of the dimensions considered. Even if we over-generously assume that running an experiment in 2 countries, or with 2 professions, or with 2 assets, is sufficient respectively to achieve country, profession or asset match, we are only able to push the percentage of papers based on natural field experiments achieving a match in three dimensions to around 25%. If we ignore gender and age because of the missing observations, only one out of ten papers based on natural field experiments achieves a match in all three of the remaining dimensions, and almost none of the artefactual and framed field experiments do.

There is no doubt that papers based on conventional laboratory experiments rely on out-of-domain inference at least as much as or more than field experiments, depending on the dimension. However, from our empirical analysis it is also clear that the large majority of papers based on field experiments, including natural field experiments, engage in out-of-domain inference. The reason is that, in our sample of leading journals, the publications based on field experiments also engage in wide and ambitious generalizations from experimental datasets. There is nothing wrong with this as such, and there is clear complementary value of both laboratory and field experiments, a point cogently made among others by Harrison (2013) and Harrison et al. (2015); but it cannot be deemed as a significant argument or element of differentiation between laboratory and field experiments as it has been made to be.

[12] See e.g. Zizzo (2010) and de Quidt et al. (2018) for a comprehensive discussion of experimenter demand effects criticisms and how they can be addressed. An example of argument instead that has been made in support of conventional laboratory experiments is that they are superiority for replicability (Roe and Just 2009), therefore helping with the systematic building up of scientific evidence.

We have additional interesting results. First, we find that, where there is a higher match rate of natural field experiments, this is driven by policy testing experiments. Persuasiveness towards policy makers may require more qualified statements. Second, and complementarily, theory testing experiments tend to have a lower match rate: this reflects a traditional methodological argument that, if an experiment's goal is to test theory, the mapping to the natural world is unimportant (e.g. Davis and Holt 1993; Camerer 2015). Third, among our set of leading journals we do not find that there are differences in publication practices across different sets of journals (world top 5 or experimental or other), although we recognize that much larger datasets would be needed before differences between journals may be picked up.

Fourth, running the experiments in Caucasian-majority countries seems to relate to more general claims than if done in, for example, Ethiopia or India. Fifth, more general claims may also be predicted if at least one of the authors is affiliated in one of the top 20 institutions in the world for Economics. The effects associated to these last two findings are quantitatively large (from 10-15% to around 25% depending on the estimation method). They may not reflect problems or biases with the publication process: for example, it may be that top 20 Economics institutions authors are more likely to do out-of-domain inference for justified and plausible reasons, accepted as such by journal editors and reviewers; similarly, a disproportionately large proportion of top Economics departments according to RePEc (or other league table indicators such as the QS World Ranking) is in Caucasian-majority countries, and a parallel unobservable quality argument may be made. There are of course alternative, less optimistic interpretations. The Caucasian-majority country finding aligns with the anecdotal view that, as an example, a dataset is more likely to be treated as having general applicability and therefore being worth serious consideration in a leading journal if comes from the United States than if the same dataset comes from say a African or Asian country. This returns us to the concerns of Henrich et al. (2010a, 2010b), who note that most people in the world are not WEIRD.

To conclude, the methodological challenge of generalizing results beyond what is within the domain of the experiments also applies to many papers based on field experiments, given their stated claims. Experimental publications on policy testing tend to limit the extent of out-of-domain inference, whereas publications based on experiments conducted in a Caucasian majority country or with at least one author in a top 20 Economics institution correlates with more out-of-domain inference.

References

Aimone, J. A., North, C., & Rentschler, L. (2019). Priming the jury by asking for Donations: An empirical and experimental study. *Journal of Economic Behavior and Organization*, *160*, 158-167.

Al-Ubaydli, O., & List, J. A. (2015a). Do natural field experiments afford researchers more or less control than laboratory experiments?. *American Economic Review, 105*(5), 462-466.

Al-Ubaydli, O., & List, J. A. (2015b). On the generalizability of experimental results in Economics. In G. Fréchette & A. Schotter (Eds.), *Handbook of Experimental Economic Methodology* (pp. 420-462). New York: Oxford University Press.

Bandiera, O., Barankay, I., & Rasul, I. (2005). Social preferences and the response to incentives: Evidence from personnel data. *The Quarterly Journal of Economics, 120*(3), 917-962.

Belot, M., Duch, R., & Miller, L. (2015). A comprehensive comparison of students and non-students in classic experimental games. *Journal of Economic Behavior and Organization*, *113*, 26-33.

Benz, M., & Meier, S. (2008). Do people behave in experiments as in the field?—evidence from donations. *Experimental economics, 11*, 268-281.

Bossuroy, T., & Delavallade, C. (2016). Experiments, policy, and theory in development economics: a response to Glenn Harrison's 'field experiments and methodological intolerance'. *Journal of Economic Methodology, 23*(2*)*, 147-156.

Brandts, J., & Charness, G. (2011). The strategy versus the direct-response method: a first survey of experimental comparisons. *Experimental Economics, 14*(3), 375–398.

Brinberg, D., & McGrath, J. E. (1985). *Validity and the research process*. Thousand Oaks: SAGE Publications.

Büyükboyacı, M., Gürdal, M. Y., Kıbrıs, A., & Kıbrıs, Ö. (2019). An experimental study of the investment implications of bankruptcy laws. *Journal of Economic Behavior and Organization*, *158*, 607-629.

Camerer, C F. (2015). The promise and success of lab–field generalizability in experimental economics: A critical reply to Levitt and List. In G. Fréchette & A. Schotter (Eds.), *Handbook of Experimental Economic Methodology* (pp. 249-296). New York: Oxford University Press.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Boston: Houghton Mifflin Company.

Candelo, N., Croson, R. T., & Eckel, C. (2018). Transmission of information within transnational social networks: a field experiment. *Experimental Economics*, *21*, 905-923.

Cappelen, A. W., Nielsen, U. H., Tungodden, B., Tyran, J. R., & Wengström, E. (2016). Fairness is intuitive. *Experimental Economics, 19*, 727-740.

Carrera, M., Royer, H., Stehr, M., & Sydnor, J. (2018). Can financial incentives help people trying to establish new habits? Experimental evidence with new gym members. *Journal of Health Economics, 58*, 202-214.

Cartwright, N. (2007). *Hunting causes and using them: Approaches in philosophy and economics*. Cambridge: Cambridge University Press.

Cason, T. N., & Plott, C. R. (2014). Misconceptions and game form recognition: Challenges to theories of revealed preference and framing. *Journal of Political Economy, 122*(6), 1235-1270.

Central Intelligence Agency. (2021). Country comparisons -- median age. *The World Factbook*. Retrieved from https://www.cia.gov/the-world-factbook/field/median-age/country-comparison

Charness, G., & Fehr, E. (2015). From the lab to the real world. *Science, 350*(6260), 512-513.

Chen, Y., Jiang, M., & Krupka, E. L. (2019). Hunger and the gender gap. *Experimental Economics, 22*, 885-917.

Cilliers, J., Kasirye, I., Leaver, C., Serneels, P., & Zeitlin, A. (2018). Pay for locally monitored performance? A welfare analysis for teacher attendance in Ugandan primary schools. *Journal of Public Economics*, *167*, 69-90.

Crockett, E., & Crockett, S. (2019). Endowments and risky choice. *Journal of Economic Behavior and Organization*, *159*, 344-354.

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. San Francisco, CA: Jossey-Bass.

Croson, R., & Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature, 47*(2), 448-474.

Davis, D. D., & Holt, C. A. (1993). *Experimental economics*. Princeton: Princeton University Press.

de Quidt, J., Haushofer, J., & Roth, C. (2018). Measuring and bounding experimenter demand. *American Economic Review, 108*(11), 3266-3302.

Falk, A., & Heckman, J. J. (2009). Lab experiments are a major source of knowledge in the social sciences. *Science, 326*(5952), 535-538.

Festré, A., & Garrouste, P. (2015). Theory and evidence in psychology and economics about motivation crowding out: A possible convergence?. *Journal of Economic Surveys, 29*(2), 339-356.

Frechette, G. (2015). Laboratory experiments: professionals versus students. In G. Fréchette & A. Schotter (Eds.), *Handbook of Experimental Economic Methodology* (pp. 360-390). New York: Oxford University Press.

Frey, B. S., & Oberholzer-Gee, F. (1997). The cost of price incentives: An empirical analysis of motivation crowding-out. *American Economic Review*, *87*(4), 746-755.

Gneezy, U., Haruvy, E., & Yafe, H. (2004). The inefficiency of splitting the bill. *The Economic Journal, 114*(495), 265-280.

Hackethal, A., Kirchler, M., Laudenbach, C., Razen, M., & Weber, A. (2022). On the role of monetary incentives in risk preference elicitation experiments. *Journal of Risk and Uncertainty*, 1-25.

Haggag, K., Pope, D. G., Bryant-Lees, K. B., & Bos, M. W. (2019). Attribution bias in consumer choice. *The Review of Economic Studies*, *86*(5), 2136-2183.

Harrison, G. W. (2013). Field experiments and methodological intolerance. *Journal of Economic Methodology*, *20*(2), 103-117.

Harrison, G. W., Lau, M. I. & Rutström, E. I. (2015). Theory, experimental design, and econometrics are complementary (and so are lab and field experiments). In G. Fréchette & A. Schotter (Eds.), *Handbook of Experimental Economic Methodology* (pp. 296-338). New York: Oxford University Press.

Harrison, G. W., & List, J. A. (2004). Field experiments. *Journal of Economic Literature, 42*(4), 1009-1055.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010a). Most people are not WEIRD. *Nature, 466*(7302), 29-29.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010b). The weirdest people in the world?. *Behavioral and Brain Sciences, 33*(2-3), 61-83.

Herbst, D., & Mas, A. (2015). Peer effects on worker output in the laboratory generalize to the field. *Science, 350*(6260), 545-549.

Jiménez-Buedo, M., & Russo, F. (2021). Experimental practices and objectivity in the social sciences: re-embedding construct validity in the internal–external validity distinction. *Synthese, 199*(3-4), 9549-9579.

Kessler, J., & Vesterlund, L. (2015). The external validity of laboratory experiments: The misleading emphasis on quantitative effects. In G. Fréchette & A. Schotter (Eds.), *Handbook of Experimental Economic Methodology* (pp. 391-406). New York: Oxford University Press.

Khan, A. Q., Khwaja, A. I., & Olken, B. A. (2019). Making moves matter: Experimental evidence on incentivizing bureaucrats through performance-based postings. *American Economic Review*, *109*(1), 237-270.

Laury, S. K., & Taylor, L. O. (2008). Altruism spillovers: Are behaviors in context-free experiments predictive of altruism toward a naturally occurring public good?. *Journal of Economic Behavior and Organization, 65*(1), 9-29.

Levitt, S. D., & List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world?. *Journal of Economic Perspectives, 21*(2), 153-174.

Levitt, S. D., & List, J. A. (2009). Field experiments in economics: The past, the present, and the future. *European Economic Review, 53*(1), 1-18.

Lin, H., Werner, K. M., & Inzlicht, M. (2021). Promises and perils of experimentation: The mutual-internal-validity problem. *Perspectives on Psychological Science*, *16*(4), 854-863.

List, J. A. (2003). Does market experience eliminate market anomalies?. *The Quarterly Journal of Economics, 118*(1), 41-71.

List, J. A. (2006). The behavioralist meets the market: Measuring social preferences and reputation effects in actual transactions. *Journal of political Economy, 114*(1), 1-37.

List, J. A. (2007a). Field experiments: a bridge between lab and naturally occurring data. *The BE Journal of Economic Analysis & Policy, 6*(2).

List, J. A. (2007b). On the interpretation of giving in dictator games. *Journal of Political Economy, 115*(3), 482-493.

List, J. A. (2011). Why economists should conduct field experiments and 14 tips for pulling one off. *Journal of Economic Perspectives , 25*(3), 3-16.

Manski, C. F. (2013). *Public policy in an uncertain world: analysis and decisions*. Cambridge, MA: Harvard University Press.

Jimenez-Buedo, M., & Miller, L. M. (2010). Why a trade-off? The relationship between the external and internal validity of experiments. *Theoria. Revista de Teoría, Historia y Fundamentos de la Ciencia*, *25*(3), 301-321.

Muralidharan, K., Singh, A., & Ganimian, A. J. (2019). Disrupting education? Experimental evidence on technology-aided instruction in India. *American Economic Review*, *109*(4), 1426-1460.

Reuben, E., Li, S. X., Suetens, S., Svorenčík, A., Turocy, T., & Kotsidis, V. (2022). Trends in the publication of experimental economics articles. *Journal of the Economic Science Association*, *8*(1-2), 1-15.

Roe, B. E., & Just, D. R. (2009). Internal and external validity in economics research: Tradeoffs between experiments, field experiments, natural experiments, and field data. *American Journal of Agricultural Economics, 91*(5), 1266-1271.

Samek, A. (2019). Advantages and disadvantages of field experiments. In *Handbook of Research Methods and Applications in Experimental Economics* (pp. 104-120). Cheltenham: Edward Elgar Publishing.

Schram, A. (2005). Artificiality: The tension between internal and external validity in economic experiments. *Journal of Economic Methodology*, *12*(2), 225-237.

Trafimow, D. (2022). A new way to think about internal and external validity. *Perspectives on Psychological Science*, 17456916221136117.

World Population Review. (2022). *Caucasian Countries*. https://worldpopulationreview.com/country-rankings/caucasian-countries

Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Experimental Economics, 13*, 75-98.

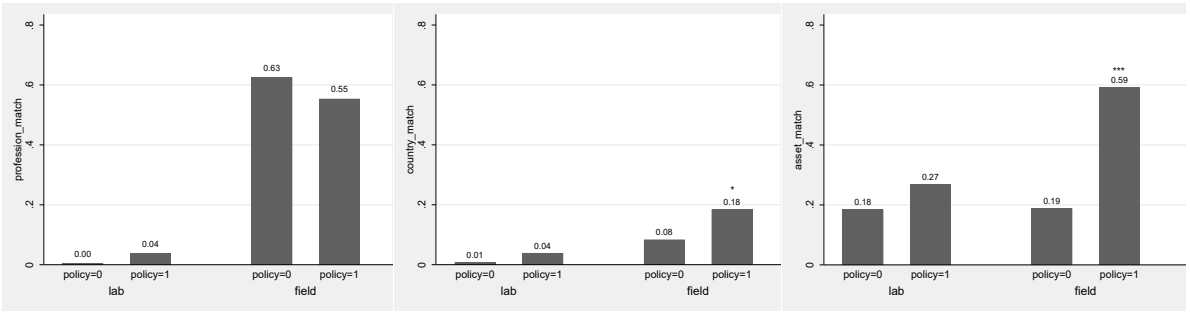
Zizzo, D. J. (2013). Claims and confounds in economic experiments. *Journal of Economic Behavior and Organization*, *93*, 186-195.

Online Appendix

A. Additional analysis

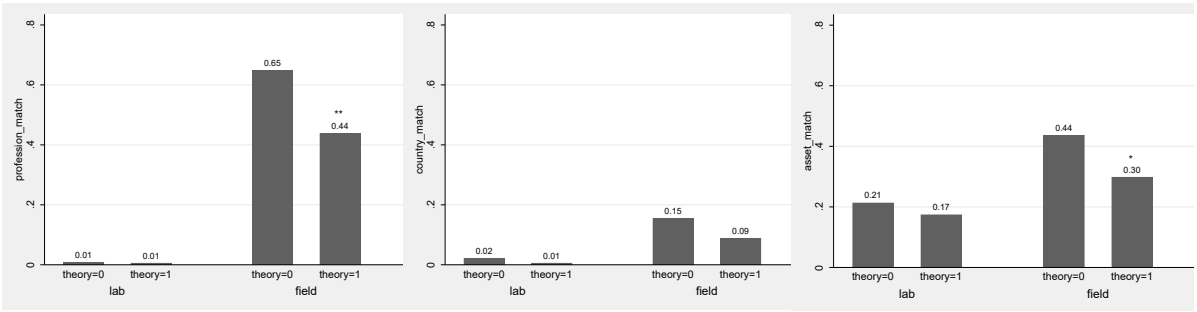
A1: Matches over policy testbed and theory testing papers

Figure A1.1*. Match rates of policy testbed vs. non-policy papers by lab and field experiments*



profession country asset

Figure A1.2. *Match rates of theory testing vs. non-theory papers by lab and field experiments.*



profession country asset

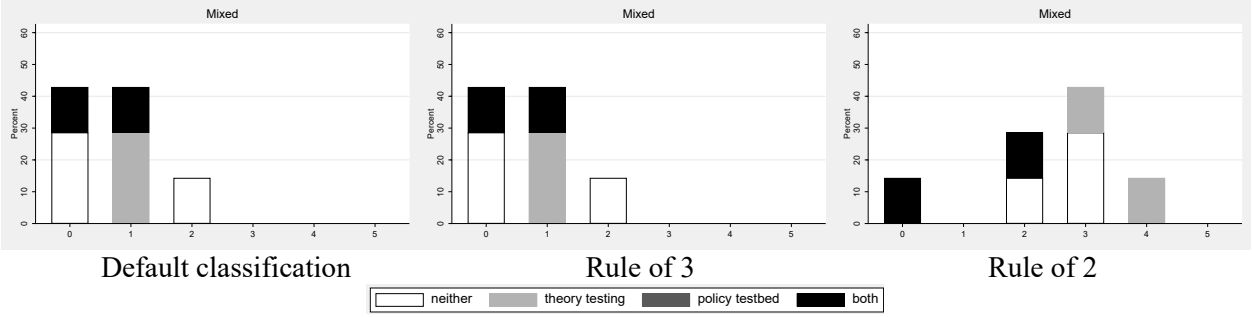
Note: MWU tests compare policy (theory) papers to non-policy (non-theory) papers within each partition, with 2-tailed significance value of $p < 0.1$ (0.05) [0.01] is denoted by * (**) [***].

Figure A1.1 (A1.2) compares the match rates between policy testbed (theory testing) papers and others in lab and field experiments. Laboratory experiments do not make a difference on matches between policy (theory) and non-policy (non-theory) papers. As for field experiments, policy testbed papers tend to have higher match rates in country and asset, compared to non-policy papers. Theory testing papers tend to have lower match rates, especially in the profession and asset dimensions.[1]

[1] Policy testbed papers usually specify the context of country where the policy is implemented, and the asset being employed in the experiment. Theory papers tend to make more general claims.

A2: Papers with mixed methods

Figure A2.1. *Distribution of number of matches for the mixed methods category*



Note: Each figure shows the number of matches (horizontal axis) that applies to a given proportion of papers (vertical axis) for mixed type, out of seven.

Figure A2.1 shows the proportion of the number of matches for mixed methods papers based on different matching rules. Given that it is based on only 7 studies, interpretations based on this figure should be made with caution.

A3: Regressions on match over each dimension

We ran OLS, Logit, and Probit regressions on profession, country, asset, age, and gender *match*. The control variables are as same as in Table 4 of the main text.

Table A3.1 reports the results on *Profession match*. All types of field experiments have significantly higher match rates than conventional laboratory experiments in all models. The coefficients of artefactual and framed field experiments are slightly higher than natural field experiments ($p < .05$ for *Framed FE ≠ Natural FE* in all models). *Theory testing* is negative in all models.

Table A3.1. *OLS, Logit, and Probit regressions on profession match*

	Dependent Variable: *Profession match* = 0 or 1					
	OLS	Logit	Probit	OLS	Logit	Probit
	(1)	(2)	(3)	(4)	(5)	(6)
Artefactual FE	0.554***	0.538***	0.542***	0.584***	0.569***	0.564***
	(0.051)	(0.077)	(0.077)	(0.113)	(0.093)	(0.092)
Framed FE	0.657***	0.642***	0.642***	0.644***	0.597***	0.593***
	(0.045)	(0.063)	(0.063)	(0.099)	(0.086)	(0.088)
Natural FE	0.504***	0.487***	0.483***	0.478***	0.410***	0.408***
	(0.048)	(0.073)	(0.071)	(0.089)	(0.084)	(0.089)
Mixed	0.138	0.148	0.141	0.187	0.131	0.134
	(0.119)	(0.140)	(0.137)	(0.137)	(0.134)	(0.129)
Policy testbed	-0.008	-0.004	-0.000	-0.029	-0.025	-0.017
	(0.041)	(0.034)	(0.034)	(0.046)	(0.030)	(0.031)
Theory testing	-0.077***	-0.082***	-0.080***	-0.068***	-0.074***	-0.072***
	(0.028)	(0.027)	(0.028)	(0.020)	(0.023)	(0.022)
Incentives				-0.081	-0.042	-0.043
				(0.069)	(0.030)	(0.031)
Top21-100				0.022	0.030	0.026
				(0.028)	(0.023)	(0.023)
Top20				-0.082	-0.024	-0.026
				(0.061)	(0.038)	(0.037)
Online				-0.143***	-0.062***	-0.061***
				(0.045)	(0.022)	(0.022)
Caucasian				-0.068	-0.052	-0.051
				(0.052)	(0.032)	(0.032)
Journal_top5				-0.123*	-0.065*	-0.067*
				(0.068)	(0.034)	(0.035)
Journal_exp				-0.092***	-0.077**	-0.077***
				(0.033)	(0.030)	(0.029)
Clustered	no	no	no	yes	yes	yes
Observations	520	520	520	520	520	520

Note: Marginal effects are reported here. Robust standard errors clustered by country are in parentheses. Two-tailed $p <$ 0.1*, 0.05**, 0.01***.

Table A3.2*. OLS, Logit, and Probit regressions on country match*

	Dependent Variable: *Country match* = 0 or 1					
	OLS	Logit	Probit	OLS	Logit	Probit
	(1)	(2)	(3)	(4)	(5)	(6)
Artefactual FE	0.049	0.060	0.059	0.032	0.039	0.034
	(0.037)	(0.044)	(0.042)	(0.047)	(0.055)	(0.053)
Framed FE	0.100***	0.100***	0.098**	0.020	0.033	0.019
	(0.033)	(0.039)	(0.040)	(0.044)	(0.026)	(0.027)
Natural FE	0.089**	0.076**	0.077**	0.021	0.024	0.016
	(0.035)	(0.034)	(0.036)	(0.052)	(0.034)	(0.034)
Mixed	-0.026			-0.058		
	(0.087)			(0.040)		
Policy testbed	0.068**	0.050*	0.050*	0.064***	0.048***	0.049***
	(0.030)	(0.026)	(0.026)	(0.024)	(0.014)	(0.013)
Theory testing	-0.034	-0.042*	-0.040*	-0.028	-0.036	-0.031
	(0.021)	(0.025)	(0.023)	(0.022)	(0.033)	(0.026)
Incentives				-0.132**	-0.076***	-0.081***
				(0.054)	(0.026)	(0.027)
Top21-100				-0.029	-0.030	-0.030
				(0.022)	(0.024)	(0.023)
Top20				-0.097**	-0.057**	-0.059**
				(0.045)	(0.025)	(0.025)
Online				-0.026	-0.006	-0.010
				(0.040)	(0.027)	(0.028)
Caucasian				-0.076**	-0.066**	-0.064***
				(0.033)	(0.026)	(0.023)
Journal_top5				0.073	0.043	0.044
				(0.059)	(0.036)	(0.035)
Journal_exp				-0.033	-0.034	-0.038**
				(0.021)	(0.021)	(0.019)
Clustered	no	no	no	yes	yes	yes
Observations	520	513	513	520	513	513

Note: Marginal effects are reported here. Robust standard errors clustered by country are in parentheses. Two-tailed $p <$ 0.1*, 0.05**, 0.01***. Mixed methods papers are dropped in Logit and Probit regressions as dependent variable equals to 0 for all Mixed methods papers.

Table A3.2 reports the results on *Country match*. Artefactual and natural field experiments have significantly higher match rates than conventional laboratory experiments ($p < .01$ in models 1-3). The differences among the coefficients for different categories of field experiments are not significant (all $p > .1$). *Policy testbed* is positive in all models.

Table A3.3. *OLS, Logit, and Probit regressions on asset match*

	Dependent Variable: *Asset match* = 0 or 1					
	OLS	Logit	Probit	OLS	Logit	Probit
	(1)	(2)	(3)	(4)	(5)	(6)
Artefactual FE	-0.069	-0.072	-0.075	-0.114*	-0.112	-0.122*
	(0.067)	(0.062)	(0.061)	(0.064)	(0.070)	(0.069)
Framed FE	-0.024	-0.020	-0.024	-0.038	-0.041	-0.044
	(0.059)	(0.057)	(0.057)	(0.110)	(0.101)	(0.101)
Natural FE	0.287***	0.258***	0.261***	0.242***	0.197***	0.202***
	(0.063)	(0.077)	(0.077)	(0.063)	(0.074)	(0.073)
Mixed	-0.225			-0.293***		
	(0.156)			(0.055)		
Policy testbed	0.170***	0.145***	0.152***	0.153**	0.130**	0.136**
	(0.054)	(0.047)	(0.049)	(0.074)	(0.053)	(0.055)
Theory testing	-0.073*	-0.073*	-0.074**	-0.073**	-0.075**	-0.077***
	(0.037)	(0.037)	(0.037)	(0.029)	(0.030)	(0.029)
Incentives				0.035	0.020	0.021
				(0.097)	(0.082)	(0.083)
Top21-100				-0.038	-0.037	-0.040
				(0.035)	(0.035)	(0.034)
Top20				-0.066	-0.070	-0.074
				(0.072)	(0.069)	(0.067)
Online				0.022	0.017	0.018
				(0.066)	(0.067)	(0.066)
Caucasian				-0.124**	-0.113***	-0.118***
				(0.052)	(0.043)	(0.045)
Journal_top5				0.044	0.045	0.050
				(0.069)	(0.069)	(0.066)
Journal_exp				-0.014	-0.016	-0.012
				(0.068)	(0.069)	(0.069)
Clustered	no	no	no	yes	yes	yes
Observations	520	513	513	520	513	513

Note: Marginal effects are reported here. Robust standard errors clustered by country are in parentheses. Two-tailed $p <$ 0.1*, 0.05**, 0.01***. Mixed methods papers are dropped in Logit and Probit regressions as dependent variable equals to 0 for all Mixed methods papers.

Table A3.3 reports the results on *Asset match*. Only natural field experiments have significantly higher match rates than conventional laboratory experiments ($p < .05$ in all models). Natural field experiments have higher match rates than 'less pure' categories of field experiments ($p < .01$ in all models). *Policy testbed* and *Theory testing* are respectively statistically positive and negative in all models.

Table A3.4. *OLS regressions on age match*

Dependent Variable: *Age match* = 0 or 1						
	OLS (1)	Logit (2)	Probit (3)	OLS (4)	Logit (5)	Probit (6)
Artefactual FE	0.473***			0.537***		
	(0.0792)			(0.112)		
Framed FE	0.334***	-0.148	-0.148	0.278***	-0.314***	-0.314***
	(0.0758)	(0.133)	(0.133)	(0.100)	(0.105)	(0.102)
Natural FE	0.280***	-0.205	-0.211	0.242**	-0.351***	-0.369***
	(0.0953)	(0.158)	(0.161)	(0.104)	(0.121)	(0.120)
Mixed	0.245	-0.243	-0.246	0.357	-0.191	-0.208
	(0.177)	(0.235)	(0.236)	(0.231)	(0.207)	(0.200)
Policy testbed	0.0443	0.0655	0.0712	0.0936	0.150	0.166
	(0.0765)	(0.132)	(0.133)	(0.112)	(0.194)	(0.193)
Theory testing	0.0378	0.0794	0.0808	0.0222	0.0642	0.0643
	(0.0565)	(0.112)	(0.113)	(0.0344)	(0.0707)	(0.0719)
Incentives				-0.158	-0.176	-0.176
				(0.122)	(0.119)	(0.116)
Top21-100				0.0768	0.148	0.152
				(0.0983)	(0.152)	(0.149)
Top20				0.0830	0.140	0.156
				(0.153)	(0.157)	(0.159)
Online				-0.0689	-0.0755	-0.0801
				(0.0995)	(0.0979)	(0.101)
Caucasian				0.191**	0.264***	0.262***
				(0.0726)	(0.0965)	(0.0916)
Journal_top5				-0.0623	-0.134	-0.144
				(0.117)	(0.133)	(0.133)
Journal_exp				-0.0451	-0.103	-0.107
				(0.0639)	(0.123)	(0.125)
Clustered	no	no	no	yes	yes	yes
Observations	170	87	87	170	87	87

Note: Marginal effects are reported here. Robust standard errors clustered by country are in parentheses. Two-tailed $p <$ 0.1*, 0.05**, 0.01***. Conventional laboratory papers will be dropped in Logit and Probit regressions as dependent variable equals to 0 for all conventional laboratory papers.

Table A3.4 reports the results on *Age match*. All categories of field experiments have significantly higher match rates than conventional laboratory experiments in models 1 and 4. Due to conventional laboratory papers being dropped in the estimation process in models 2-3 and 5-6, the coefficients of each experiment category show the comparison with artefactual field experiments as baseline. Artefactual field experiments have higher matches rates in age dimension than framed and natural field experiments in models 5 and 6.[2]

[2] In model 1, $p < .1$ for *Artefactual FE ≠ Framed FE*, and *Artefactual FE ≠ Natural FE.*

Table A3.5. *OLS, Logit, and Probit regressions on gender match*

Dependent Variable: *Gender match* = 0 or 1						
	OLS	Logit	Probit	OLS	Logit	Probit
	(1)	(2)	(3)	(4)	(5)	(6)
Artefactual FE	-0.141	-0.142	-0.141	-0.132	-0.133	-0.132
	(0.104)	(0.102)	(0.101)	(0.0952)	(0.0938)	(0.0930)
Framed FE	-0.0823	-0.0840	-0.0822	-0.128	-0.127	-0.127
	(0.121)	(0.119)	(0.118)	(0.0990)	(0.0971)	(0.0963)
Natural FE	-0.194	-0.196	-0.194	-0.203	-0.204	-0.203
	(0.143)	(0.140)	(0.139)	(0.143)	(0.140)	(0.139)
Mixed	-0.191	-0.193	-0.191	-0.199	-0.200	-0.200
	(0.212)	(0.210)	(0.208)	(0.153)	(0.152)	(0.155)
Policy testbed	-0.0426	-0.0424	-0.0423	-0.0176	-0.0178	-0.0160
	(0.0947)	(0.0929)	(0.0929)	(0.0887)	(0.0855)	(0.0856)
Theory testing	-0.105	-0.103	-0.103	-0.0780	-0.0764	-0.0768
	(0.0690)	(0.0664)	(0.0670)	(0.0643)	(0.0607)	(0.0606)
Incentives				-0.115	-0.115	-0.116
				(0.0999)	(0.101)	(0.0976)
Top21-100				0.108	0.108	0.109
				(0.0707)	(0.0690)	(0.0691)
Top20				0.155	0.154	0.154
				(0.159)	(0.155)	(0.154)
Online				0.00939	0.00952	0.0134
				(0.0622)	(0.0624)	(0.0612)
Caucasian				-0.0150	-0.0151	-0.0157
				(0.0799)	(0.0784)	(0.0773)
Journal_top5				-0.176	-0.179	-0.182
				(0.153)	(0.156)	(0.155)
Journal_exp				0.0269	0.0278	0.0285
				(0.0795)	(0.0774)	(0.0772)
Clustered	no	no	no	yes	yes	yes
Observations	238	238	238	238	238	238

Note: Marginal effects are reported here. Robust standard errors clustered by country are in parentheses. Two-tailed $p <$ 0.1*, 0.05**, 0.01***.

Table A3.5 reports the results on *Gender match*. No significant difference is found among all experiment categories in all models (all $p >.1$).

B. Robustness checks

B1: Gray area cases

Table B1.1. *Descriptive statistics of grey area papers*

Type of Experiment	Conventional Lab: 14 Artefactual FE: 4, Framed FE: 6, Natural FE: 11
Characteristics	Policy testbeds: 14, Theory testing: 12, Online: 5, Incentivized: 30
Institution Ranking	Top 20: 9, Top 21-100: 9
Country	Caucasian countries:25
Experimental Economics Journals	Experimental Economics: 3, Games and Economic Behavior: 4, Journal of Economic Behavior & Organization: 8, Journal of Risk and Uncertainty: 1
Top 5 Journals	American Economic Review: 1, Journal of Political Economy: 1, Review of Economic Studies: 1, The Quarterly Journal of Economics: 1
Other Journals	International Economic Review: 1, Journal of Agricultural Economics: 1, Journal of Development Economics: 3, Journal of Economic Theory: 1, Journal of Environmental Economics and Management: 4, Journal of Health Economics: 1, Journal of Public Economics: 3, Journal of the European Economic Association: 1

Note: Numbers refer to the number of papers.

Table B1.2. *Descriptive statistics of positive match variables*

Variables	Dataset excluding grey area papers		Grey area papers	
	Frequency	**Percentage**	**Frequency**	**Percentage**
Profession match (=1)	110	22.7%	10	28.6%
Country match (=1)	28	5.8%	3	8.6%
Asset match (=1)	129	26.6%	10	28.6%
Age match (=1)	28	17.7%	3	25%
Missing value	327	67.4%	23	65.7%
Gender match (=1)	114	51.1%	9	60%
Missing value	262	54.0%	20	57.1%

Note: Where there are missing values, percentages are computed as a proportion of the papers for which values are available.

Table B1.1 presents descriptive statistics of grey area papers. It shows they are distributed across a variety of journals. Table B1.2 shows the match rates of grey area papers and of the dataset excluding them. The features of the dataset remain broadly unchanged after excluding grey area papers. There is no obvious difference between grey area papers and non grey area papers, except that the latter have slightly higher match rates.

Table B1.3. *OLS, Logit, and Probit regressions on sum of matches (excluding grey area papers)*

	Dependent Variable: *Sum of matches = profession + country + asset*					
	OLS	Logit	Probit	OLS	Logit	Probit
	(1)	(2)	(3)	(4)	(5)	(6)
Artefactual FE	0.516***	0.384***	0.384***	0.470***	0.353***	0.347***
	(0.107)	(0.080)	(0.081)	(0.157)	(0.107)	(0.112)
Framed FE	0.765***	0.480***	0.482***	0.671***	0.438***	0.441***
	(0.094)	(0.070)	(0.069)	(0.233)	(0.150)	(0.148)
Natural FE	0.878***	0.394***	0.402***	0.728***	0.304**	0.312**
	(0.102)	(0.082)	(0.080)	(0.157)	(0.136)	(0.134)
Mixed	-0.120	-0.076	-0.080	-0.209	-0.126	-0.130
	(0.238)	(0.140)	(0.137)	(0.174)	(0.137)	(0.129)
Policy testbed	0.283***	0.121**	0.121**	0.246**	0.106**	0.107**
	(0.086)	(0.052)	(0.053)	(0.108)	(0.043)	(0.043)
Theory testing	-0.215***	-0.129***	-0.127***	-0.200***	-0.121***	-0.120***
	(0.059)	(0.037)	(0.037)	(0.054)	(0.036)	(0.036)
Incentives				-0.157	-0.079	-0.082
				(0.160)	(0.077)	(0.078)
Top21-100				-0.050	-0.034	-0.038
				(0.056)	(0.033)	(0.032)
Top20				-0.181	-0.077	-0.074
				(0.152)	(0.071)	(0.067)
Online				-0.130	-0.059	-0.059
				(0.110)	(0.073)	(0.074)
Caucasian				-0.257**	-0.133***	-0.139***
				(0.103)	(0.047)	(0.050)
Journal_top5				0.030	0.064	0.065
				(0.134)	(0.074)	(0.072)
Journal_exp				-0.120	-0.010	-0.009
				(0.105)	(0.066)	(0.065)
Clustered	no	no	No	yes	yes	yes
Observations	485	485	485	485	485	485

Note: Marginal effects are reported here. Robust standard errors clustered by country are in parentheses. Two-tailed $p <$ 0.1*, 0.05**, 0.01***.

Table B1.3 reports the regressions on *sum of matches* in profession, country, and asset excluding grey area papers. The significance of *Artefactual FE, Framed FE, Natural FE, Policy testbed, Theory testing,* and *Caucasian* remains the same: our findings are robust to excluding grey area papers. *Top 20* is still negative although it loses statistical significance in models 4-6.

B2: Rule of 2 (3)

Overall, there are 37 (9) papers that examine at least 2 (3) professions/countries/assets in the experiments. Table B2.1 shows the match rates when implementing the rule of 2 and 3. Unsurprisingly, match rates increase slightly with over-generous relaxations of the rules. Compared to our default classification methodology, the profession match rate rises from 23.1% to 26.2% (23.5%), the country

match rate rises from 6% to 8.9% (6.4%), and the asset match rate rises from 26.7% to 29.6% (28.3%) with a rule of 2 (3). Additional 12 (5) papers obtain at least three matches under a rule of 2 (3).

Table B1.1. *Descriptive statistics of positive match variables*

Variables	Rule of 2		Rule of 3	
	Frequency	**Percentage**	**Frequency**	**Percentage**
Profession match (=1)	136	26.2%	122	23.5%
Country match (=1)	46	8.9%	33	6.4%
Asset match (=1)	154	29.6%	147	28.3%
Three matches papers (out of five)	54	10.4%	47	9.0%

Note: Rule of 2 (3) will not affect age and gender match, which are thus not reported here.

Figure B2.1. *Match rates of claims based on each experiment type (rule of 2)*

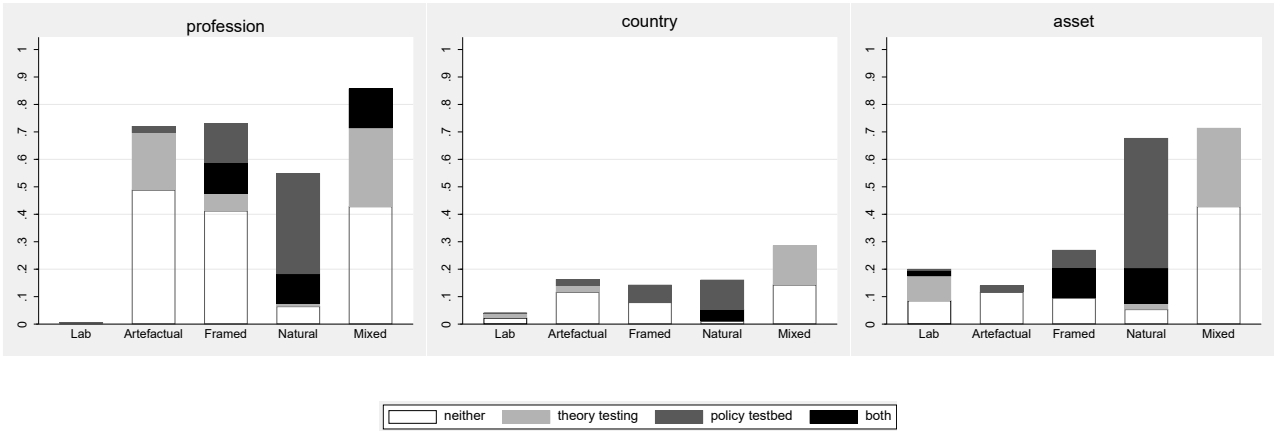


Figure B2.2. *Match rates of claims based on each experiment type (rule of 3)*

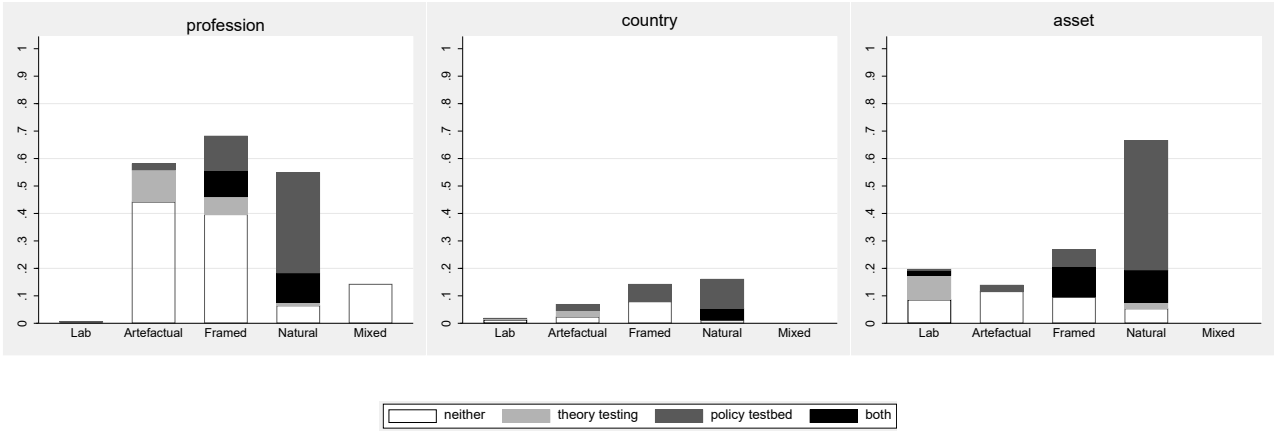


Figure B2.1 (B2.2) depicts the match rates of each experiment category in three dimensions that will be affected with a rule of 2 (3). As expected, the match rates in mixed methods papers that incorporate several professions, countries, or assets rise substantially. However, the match rates in other categories do not change much. Natural field experiments have a higher asset match rates than other categories

of field experiments in asset match. The difference is still driven by policy testbed experiments. Match rates are still similar across different categories of field experiments for the other dimensions.

Table B2.1. *OLS, Logit, and Probit regressions on sum of matches (rule of 2)*

	Dependent Variable: Sum of matches = profession + country + asset					
	OLS	Logit	Probit	OLS	Logit	Probit
	(1)	(2)	(3)	(4)	(5)	(6)
Artefactual FE	0.736***	0.488***	0.487***	0.625***	0.462***	0.453***
	(0.111)	(0.071)	(0.071)	(0.169)	(0.094)	(0.096)
Framed FE	0.780***	0.474***	0.475***	0.621***	0.407***	0.411***
	(0.099)	(0.067)	(0.066)	(0.227)	(0.155)	(0.151)
Natural FE	0.922***	0.404***	0.411***	0.718***	0.304**	0.320**
	(0.105)	(0.077)	(0.074)	(0.190)	(0.153)	(0.147)
Mixed	1.562***	0.602***	0.599***	1.397***	0.563***	0.534***
	(0.259)	(0.133)	(0.133)	(0.393)	(0.162)	(0.165)
Policy testbed	0.233***	0.093*	0.090*	0.188*	0.076	0.072
	(0.090)	(0.053)	(0.053)	(0.096)	(0.054)	(0.052)
Theory testing	-0.160**	-0.090**	-0.090**	-0.143***	-0.081***	-0.082***
	(0.062)	(0.037)	(0.038)	(0.051)	(0.026)	(0.027)
Incentives				-0.114	-0.095	-0.094
				(0.166)	(0.083)	(0.082)
Top21-100				-0.080	-0.026	-0.030
				(0.057)	(0.030)	(0.030)
Top20				-0.215*	-0.105**	-0.099**
				(0.129)	(0.047)	(0.044)
Online				-0.031	-0.059	-0.052
				(0.118)	(0.073)	(0.071)
Caucasian				-0.400***	-0.160***	-0.161***
				(0.116)	(0.049)	(0.049)
Journal_top5				-0.076	0.022	0.018
				(0.181)	(0.063)	(0.061)
Journal_exp				-0.168	-0.037	-0.034
				(0.107)	(0.072)	(0.071)
Clustered	no	no	no	yes	yes	yes
Observations	520	520	520	520	520	520

Note: Marginal effects are reported here. Robust standard errors clustered by country are in parentheses. Two-tailed $p <$ 0.1*, 0.05**, 0.01***.

Table B2.1 (B2.2) reports the regressions on *Sum of matches* in profession, country, and asset with a rule of 2 (3). The significance of *Artefactual FE, Framed FE, Natural FE, Theory testing, Top20,* and *Caucasian* remains the same: the results are robust to the rules of 2 and 3. *Policy testbed* is still positive although it loses statistical significance in models 5-6.[3] *Mixed* is significantly positive in all models of Table B2.1, indicating more mixed methods papers obtain a match in profession, country, and asset under the rule of 2 (but not under the rule of 3: see Table B2.2).

[3] It is significant when errors are not clustered.

Table B2.2. *OLS, Logit, and Probit regressions on sum of matches (rule of 3)*

	Dependent Variable: Sum of matches = profession + country + asset					
	OLS	Logit	Probit	OLS	Logit	Probit
	(1)	(2)	(3)	(4)	(5)	(6)
Artefactual FE	0.526***	0.386***	0.387***	0.474***	0.370***	0.365***
	(0.107)	(0.077)	(0.078)	(0.173)	(0.116)	(0.121)
Framed FE	0.745***	0.453***	0.453***	0.621***	0.403***	0.403***
	(0.095)	(0.068)	(0.068)	(0.210)	(0.136)	(0.136)
Natural FE	0.910***	0.416***	0.422***	0.755***	0.335**	0.344**
	(0.101)	(0.077)	(0.074)	(0.179)	(0.141)	(0.139)
Mixed	-0.133	-0.085	-0.089	-0.204	-0.123	-0.122
	(0.249)	(0.140)	(0.138)	(0.189)	(0.149)	(0.142)
Policy testbed	0.264***	0.093*	0.092*	0.220**	0.074	0.074
	(0.086)	(0.052)	(0.053)	(0.093)	(0.046)	(0.046)
Theory testing	-0.172***	-0.112***	-0.110***	-0.158***	-0.103***	-0.102***
	(0.059)	(0.037)	(0.037)	(0.052)	(0.032)	(0.033)
Incentives				-0.154	-0.093	-0.095
				(0.159)	(0.078)	(0.079)
Top21-100				-0.068	-0.026	-0.030
				(0.055)	(0.031)	(0.031)
Top20				-0.277**	-0.109**	-0.107**
				(0.114)	(0.048)	(0.046)
Online				-0.111	-0.075	-0.074
				(0.102)	(0.069)	(0.069)
Caucasian				-0.300***	-0.139***	-0.143***
				(0.105)	(0.047)	(0.049)
Journal_top5				-0.010	0.034	0.034
				(0.145)	(0.061)	(0.060)
Journal_exp				-0.160*	-0.030	-0.029
				(0.095)	(0.068)	(0.067)
Clustered	no	no	no	yes	yes	yes
Observations	520	520	520	520	520	520

Note: Marginal effects are reported here. Robust standard errors clustered by country are in parentheses. Two-tailed $p <$ 0.1*, 0.05**, 0.01***.