Experimental Economics: Past and Future

Guillaume R. Fréchette[*], Kim Sarnoff[†], Leeat Yariv[‡]

August 2021

Abstract

Over the past several decades, lab experiments have offered economists a rich source of evidence on incentivized behavior. In this article, we use detailed data on experimental papers to describe recent trends in the literature. We also discuss various experimentation platforms and new approaches to the design and analysis of the data they generate.

Keywords: Experimental Economics, Time Trends, Experimental Platforms, MTurk, Replication

[*]Department of Economics, New York University.
[†]Department of Economics, Princeton University.
[‡]Department of Economics, Princeton University, CEPR, and NBER.

1 Introduction

Experimental economics has come of age over the past five decades. The field has assembled a large body of evidence on human behavior in the face of incentives. Its insights have affected the progress in many fields of economics, from microeconomics, to labor, to finance, to macroeconomics.

Over that period, publications based on experimental research have become commonplace in general-interest and field journals. In addition, two journals dedicated to research based on experimental work were initiated: *Experimental Economics* in 1998 and *Journal of the Economic Science Association* in 2015. Nonetheless, Nikiforakis & Slonim (2019) report several trends in experimental publications from 1975–2018 and point to a significant decline in top-5 publications over the last decade of that period. Indeed, Figure 1a replicates this observation for the period of 2010–2019. Top-5 publications based on experimental work have significantly declined in the second half.[1]

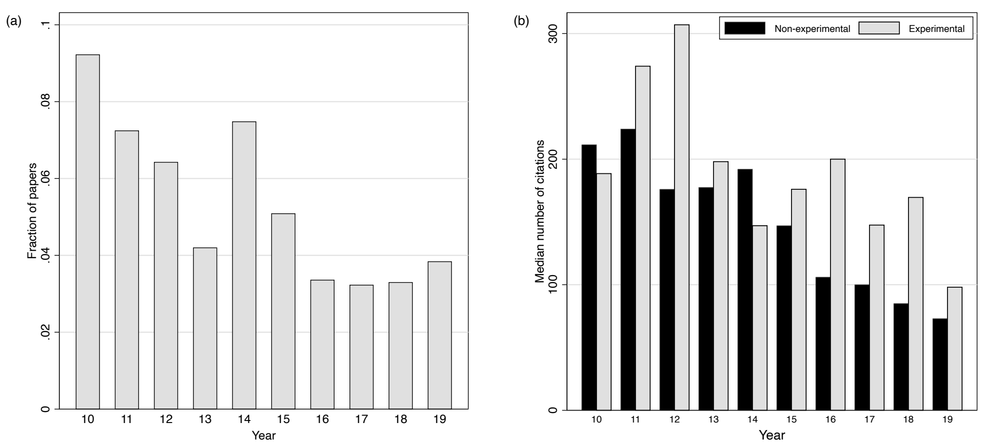

Figure 1: (a) Fraction of experimental economics publications in top-5 journals during 2010–2019, (b) Median number of citations in top-5 journals during 2010–2019

While fewer experimental papers are being published, their impact—as measured by citations, which are often utilized in promotion and hiring decision, see e.g. Lehmann, Jackson & Lautrup (2006) and Ellison (2013)—has consistently ex-

[1]Throughout, we refer to *American Economic Review*, *Econometrica*, *Journal of Political Economy*, *Quarterly Journal of Economics*, and *Review of Economic Studies* as the "top-5." In the data we present, we exclude non-reviewed pieces.

ceeded those of contemporary papers in other fields. Figure 1b displays the median number of citations in top-5 journals during the last decade. This number is substantially higher for experimental work in nearly all years.[2]

Certainly, experimental work in the 21st century does not mirror work done 50 years prior. In what follows, we use data on top-5 publications to identify recent trends in how experimental work is conducted.

The introduction of new online experimentation platforms and the transition to lab-in-the-field research allow experimental work to cover a broader pool of participants in terms of both volume and characteristics. As we document, successful experiments are becoming larger, involving an increasing number of sessions and participants.

The use of participants outside of the traditional lab has many benefits, both practical and conceptual. However, as we discuss, the online lab also has its shortcomings. Settings explored in online labs tend to be simpler, in terms of the strategic considerations participants face as well as the feedback and learning opportunities they offer. Recent research also suggests that observations collected on online platforms might be noisier than those collected in the traditional lab. We use new data comparing physical and online lab observations in a particular strategic interaction to illustrate the learning limitations entailed in online experiments.

The replication crisis in the social sciences, see Dreber & Johannesson (2019) and references therein, has led to recent attempts to develop agreed-upon best practices in empirical work. Pre-registration and pre-analysis plans, lower p-value thresholds for significance, and an effort to replicate existing studies, have been suggested by some and, to some extent, implemented. In the last part of this piece we discuss some of the benefits and potential pitfalls pursuing these directions may entail for experimental research.

2 General Time Trends

We begin by discussing trends in laboratory experiments over the past decade, in terms of the features of experimental work that has been published in leading journals and its authors' characteristics.

[2]This is in line with Anauati, Galiani & Gálvez (2020), who find that applied work tends to garner more citations than work in other fields. In that respect, experimental papers are no exception.

2.1 The Data

We collected detailed data on all lab experiments published in top-5 journals between 2010 and 2019, both traditional experiments conducted at physical university laboratories, as well as online or field experiments that have a lab component.[3]

Over this period, 164 experimental papers were published: 88 in *American Economic Review*, 28 in *Review of Economic Studies*, 21 in *Econometrica*, 16 in *Quarterly Journal of Economics*, and 11 in *Journal of Political Economy*.

In all the analyses we discuss below, the patterns we highlight are significant at the 5% level, using a t-test or a median test that compares the first and second half of the decade in aggregate.

2.2 Attributes of Experimental Papers

Published materials are becoming longer. As seen in Figure 2a, both articles and online supplementary materials have grown in length.[4] These patterns are in line with trends identified for the economics profession as a whole over the past several decades, see Ellison (2002) and work that followed. To the extent that the length of papers and supplementary materials are correlated with time spent bringing those papers to publication, researchers may be facing increased burdens in publishing their work, at least at our lead journals. This observation may be important in hiring and promotion decisions; see also Heckman & Moktan (2020).

The median number of sessions has increased over time as well, albeit not significantly so; see Figure 2b.[5]

[3]The list of experimental papers comes from a database maintained by Congiu & Nunnari (2021). As mentioned, we exclude non-reviewed papers—comments, errata, proceedings papers, etc.—from all of our analyses.

[4]Data publication has been common practice throughout the decade, with more than 80% of papers making their data available each year.

[5]The median number of sessions for 2010-2014 is 12; for 2015-2019, it is 15.5. The number of sessions is reported in detail for only 80 of the 164 papers. Of the remaining 84 papers, 48 have a component conducted in a laboratory without indication of the number of sessions carried out. There are 36 papers that rely on lab-in-the-field or online experiments that do not involve multiple sessions. These are equally split across the first and second half of the decade. Their consideration as papers with one session does not affect these qualitative observations.

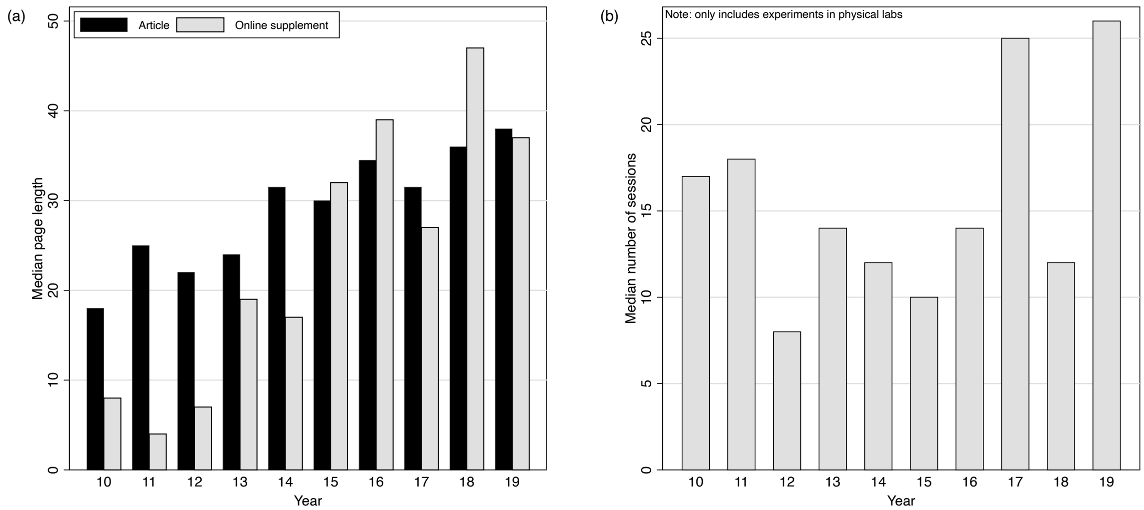

FIGURE 2: (a) Median article length, (b) Median number of sessions

2.3 Author Profiles

Next, we turn to the composition of author teams. The average size of teams hovers between 2.5 and 3 throughout the decade. Figure 3 shows changes in team demographics. The fraction of teams with at least one author from a school with a top-20 economics department has clearly grown, going from slightly under 30% in 2010 to over 60% in 2019.[6] On the other hand, representation of women and people of color has been relatively stable. The fraction of papers with at least one woman author is somewhat higher in the second half of the decade, increasing from 38% to 52%.[7] The fraction of teams with at least one non-white author is fairly constant at around 1/3.[8]

[6]We use the 2017 U.S. News and World Report rankings. We count a paper as having a top-20 co-author if someone works at a school whose economics department falls in the top 20 of that list. The trend holds if we require at least one co-author to work in the economics department of the school, rather than in any department.

[7]The increase in women authors is barely significant (just above the 10% level).

[8]Our categorization of an author's race/ethnicity is based on name, photo, undergraduate institution and, if available, citizenship or languages spoken. There are 2 papers where available information is insufficient for determining whether at least one author is non-white; these papers are categorized as having an all-white author team.

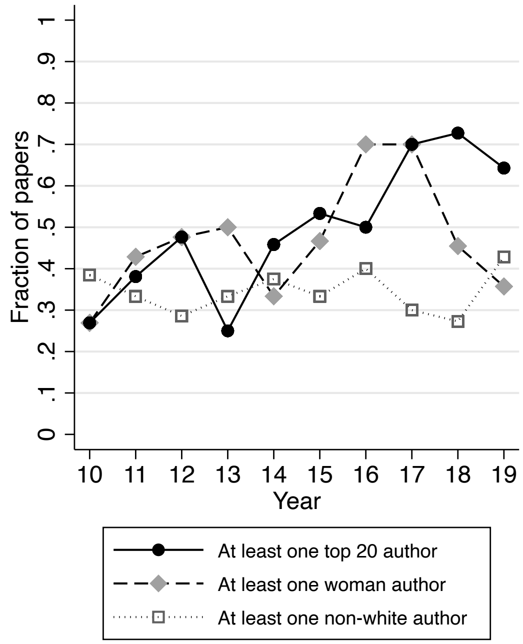

FIGURE 3: Time trends in co-author teams

2.4 Experimental Platform and Design Complexity

The number of online experiments has increased. Between 2010 and 2014, only about 5% of experiments published in the top-5 journals included an online treatment. However, between 2015 and 2019, that number had increased to 23%.

This growth in online experiments has driven a second development, an increase in sample size. Figure 4 breaks down the median number of participants by experimental platform: whether the experiment is conducted fully offline, or has at least one online treatment. When considering experimental papers conducted offline, the increase in participant volume in the second half of the decade is barely significant (the p-value is 0.13). However, when considering all experimental papers, the growth in sample size is pronounced and highly significant.

The content of experiments has also changed. Figure 5a shows an increase in the proportion of experiments that do not entail a strategic interaction. This trend is potentially due to the decrease in the share of papers relying on physical lab experiments and the growth in papers relying on online experiments, which mostly involve non-strategic tasks.

Learning opportunities are often important for more complex tasks—participants

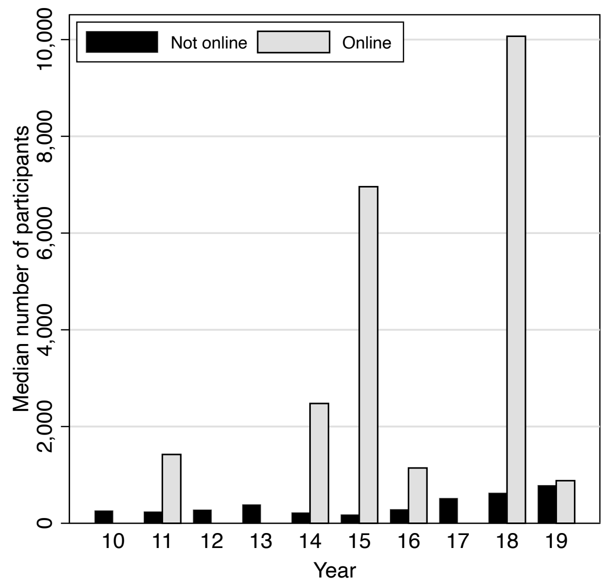

Figure 4: Time trends in number of participants

may need experience with the experimental interface and with the strategic forces in the setting they face. Naturally, more complex designs, particularly ones involving strategic interactions, may require richer learning opportunities. Related to our previous observation, Figure 5b demonstrates a decrease in the fraction of papers that are based on experiments providing feedback to participants. Online platforms are therefore more frequently utilized for non-strategic experimental designs and offer fewer explicit learning opportunities, a point we return to in the next section.

3 Online and Physical Labs

As discussed in the previous section, one of the striking trends over the past decade is an increase in the use of online experimental platforms, such as Amazon Mechanical Turk (MTurk). Virtual laboratories have the advantage of allowing researchers without access to a physical laboratory to perform experiments. They also supply a more diverse participant pool than most physical laboratories (see Fréchette (2015) and Fréchette (2016) for a survey of results pertaining to various

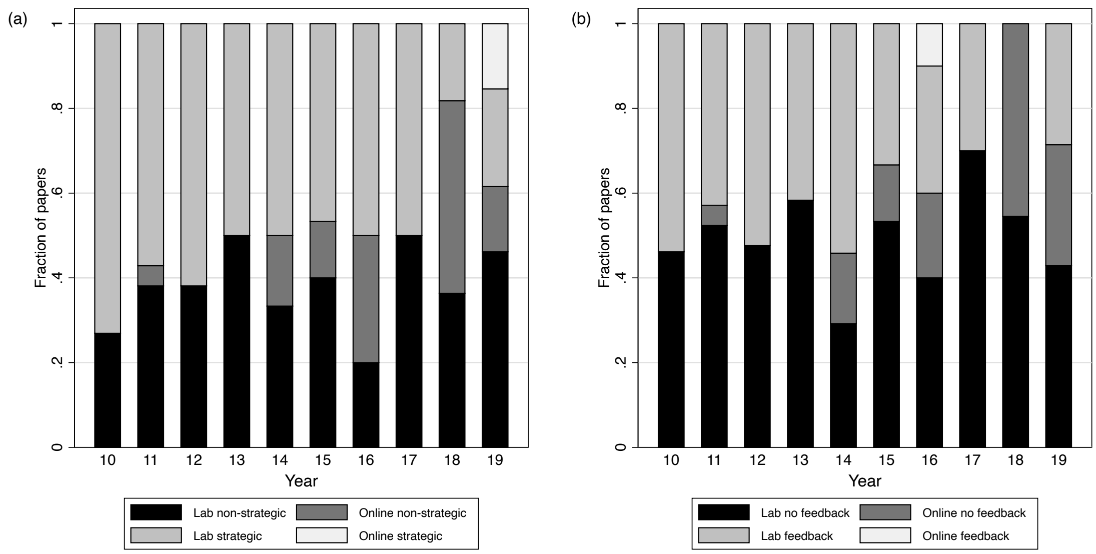

Figure 5: (a) Breakdown of experiments by presence of strategic interaction, (b) Breakdown of experiments by presence of feedback

participant pools) and offer lower per-participant costs.[9]

Given these differences, it is important to understand if and how results are affected by whether an experiment is conducted in a traditional laboratory or online. Early papers comparing experimental results using students in a physical lab and MTurk participants found encouraging results. MTurk participants behave similarly to university students on several "heuristic and biases" experiments and non-incentivized games, as well as (incentivized) repeated public goods and prisoner's dilemma games (Paolacci, Chandler & Ipeirotis (2010); Horton, Rand & Zeckhauser (2011); Berinsky, Huber & Lenz (2012); Goodman, Cryder & Cheema (2013); Arechar, Kraft-Todd & Rand (2017)). See Hauser, Paolacci & Chandler (2019) for a survey of the first results on the topic.

3.1 Noise across Platforms

Recent work depicts a somewhat more nuanced picture. Snowberg & Yariv (2021) compare university students in two universities (California Institute of Technology

[9]Lower costs are certainly an advantage: they make experimental research more economical and allow access to a broader set of researchers. Lower costs do, however, imply lower incentives to experimental participants. Since incentives are a key feature of economic experiments—see Smith (1982)— reduction in costs could be a double-edged sword.

and University of British Columbia) in the lab and online, a representative sample of the US, and MTurk. The comparison is across a wide range of incentivized, fundamental behaviors: risk and ambiguity aversion, discounting, competitiveness, cognitive sophistication, dictator giving, play in prisoner's dilemma games, guesses in beauty contests games, and many others. MTurk data were collected for low and high incentives, where low incentives matched the average payment on MTurk and high incentives were double the low ones. A variety of attention screeners were utilized in the low-stake variant.

Snowberg & Yariv (2021) observe little difference between student responses online and in the physical lab (in both of the university samples).[10] However, responses do differ significantly across the three sample types: the student sample, the representative US sample, and MTurk.[11] Nonetheless, correlations between behaviors and comparative statics are similar across the samples, with differences driven mostly by some correlations being insignificant.

One important insight of Snowberg & Yariv (2021) relates to noise. For the eight measures for which noise is quantified, university students exhibit the lowest noise level among the three samples for all but one elicitation. Furthermore, for all elicitations, the student sample is less noisy than MTurk.[12]

Gupta, Rigotti & Wilson (2021) suggest one channel through which noise might be generated: inattention. They compare behavior in a standard physical laboratory, MTurk, and on Prolific. Their focus is on games with a tension between individual rationality and social efficiency. Their main treatments compare behavior in four one-shot games: two prisoner's dilemma games and two symmetric games with a dominant strategy, where the unique equilibrium is socially efficient. They use the latter to evaluate noise, which they interpret as inattention—in a game with a dominant strategy and no apparent motives for selecting other actions, inattention seems a sensible rationalization for any out-of-equilibrium behavior.[13] Noisy behavior accounts for 60% of choices on MTurk, 19% on Prolific, and 14% in the standard laboratory. In addition, Gupta et al. (2021) also report that lab partici-

[10]Using data on historical participation in lab experiments, they also note little selection into the physical lab.

[11]Interestingly, they observe virtually no differences between behaviors of MTurk participants with high incentives and with those incentives halved.

[12]Since correlations between noisy variables are attenuated, this helps explain some of the correlation differences that are due to insignificance.

[13]The data analyzed are only from participants who successfully passed a comprehension quiz.

pants are far more sensitive to treatment variations across the different prisoner's dilemma games, in line with the insensitivity of MTurk participants to the magnitude of incentives that Snowberg & Yariv (2021) document.

The above two studies suggest that, while comparative statics are by and large similar across experimental platforms, virtual laboratories with a more diverse participant pool may exhibit greater noise, particularly when participants are only lightly vetted, as on MTurk. Certainly, researchers can generate sensitive filters themselves, a practice promoted within the field, see e.g. Berinsky, Margolis & Sances (2014). As the two studies discussed above indicate, this may not eliminate the increased noise. In fact, one may worry that matters become more complicated for experiments in which experience and attention to feedback are important for behavior. We provide novel data that speaks to this point.

3.2 Feedback and Experience across Platforms

Fudenberg & Peysakhovich (2016) study a version of the Acquire a Company game (Samuelson & Bazerman, 1984), also known as the Additive Lemons Problem, on MTurk. In their experiment, there are two players, a (human) buyer and a (computerized) seller. At the outset, the seller "owns" an item of value v, drawn from a uniform distribution between 0 and 10. The value of the item to the buyer is $v+k$, where $k > 0$ is a pre-specified constant. Thus, there are always gains from trade. However, only the seller knows the realized value of v. The buyer does not. The buyer can make a single take-it-or-leave-it offer b to the seller. If the seller accepts this offer, the buyer receives the item and pays b to the seller.

This game has a unique Nash equilibrium in weakly undominated strategies. It is weakly dominant for the seller to accept all offers above v and reject all offers below v. Solving the buyer's maximization shows that the optimal bid is k. Thus, in the unique equilibrium in weakly undominated strategies, the buyer offers k. The seller accepts the offer when $v < k$ and rejects it when $v > k$.

In the experiment, the constant k was either 3 or 6, and participants, acting as buyers, played the game for 30 rounds.

In the baseline treatment, participants know the support of possible item values, but are not informed of the distribution. They receive feedback about the realized v at the end of each round only if their offer is accepted. There are two additional treatments. In one, participants are informed at the outset that each item value is equally likely (treatment Info), with feedback as in the baseline treatment. This treatment therefore mimics the theoretical game sketched above. In another treatment, participants are not informed of the distribution of item values, but receive richer feedback: they learn the realized value v regardless of whether their offer is accepted (treatment CF—for counterfactual).

As Figure 6 suggests, when run on MTurk, there are no significant differences between behaviors in the three treatments, either initially or after 30 rounds of play (those are labeled MTurk Baseline, Info, and CF in the figure).[14] As in Gupta et al. (2021), who find very small reactions to treatment variations on MTurk, the absence of a response to changing feedback (treatment CF) could possibly indicate participants' lack of attention.[15]

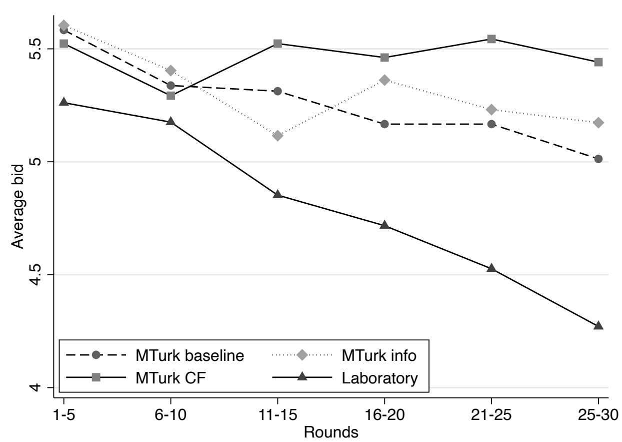

Figure 6: Bids in the Additive Lemons Game on MTurk and in a student laboratory

Potentially more revealing in that figure are the "Laboratory" data recently collected by Drew Fudenberg and Guillaume Fréchette. These data are generated from student participants in a physical laboratory playing the same additive lemons game as Fudenberg & Peysakhovich (2016) with $k = 3$. Participants were informed of the underlying uniform distribution of item values and were given

[14]The figure depicts average bids for $k = 3$ sessions. Significance results derive from linear regressions with participant-level clustering. P-values are above 0.1 for any pairwise or joint comparison.

[15]Here too, the data analyzed pertain only to participants who successfully passed a comprehension quiz.

feedback regardless of whether their offer was accepted.[16]

As can be seen, the laboratory data display significant movement toward equilibrium as participants gain experience. While the figure shows that bids are slightly lower in the laboratory even in the first five rounds ($p < 0.05$), this is likely due to the reaction to feedback early on: in the first round, the average bids are approximately 5.2 in both platforms ($p > 0.1$, no participant-specific clustering in this case). In fact, even when considering the first five rounds, the modal bid is five on both MTurk and in the laboratory (25% and 28% of bids respectively). By round 30, however, the average laboratory bid has dropped by more than 20% from its original value. For rounds 25 to 30, the modal bid is 3 in the laboratory (exactly the equilibrium and represents 26% of bids), while it is still 5 on MTurk (28% of bids with only 11% of bids at 3).

This evidence is only suggestive and, indeed, we hope more data comparing various experimental platforms will be collected in years to come, particularly as experimental designs evolve to speak to a broader set of participants. Nonetheless, the results are in line with prior observations indicating that MTurk participants are perhaps less attentive and noisier than their student counterparts. Furthermore, these data highlight that even when one-shot play is unaffected by the experimental platform, longer-run behaviors may differ substantially. Thus, even if online platforms can offer a reasonable trade-off between noise and costs for certain simple tasks, caution may be wise when studying more complex strategic interactions in which participants require experience and attention to feedback in order to appreciate the strategic forces at play.

4 Robustness of Experimental Results

In recent years, concerns have been raised regarding the replicability of results in the social sciences (Dreber & Johannesson, 2019) as well as the natural sciences (Baker, 2016). Although experimental economics is not immune to such problems, the results from Camerer, Dreber, Forsell, Ho, Huber et al. (2016) seem to suggest replicability issues may not be as acute as in other fields, most notably psychology (see Figure 4 of that paper).

[16] Data was originally collected as part of a separate experimental setting. As such, it is not a replication per se.

In this section, we start by describing standard practices in experimental economics that have potentially shielded the field from severe reproducibility issues. We then turn to some of the strategies suggested within the economics profession to alleviate replication concerns and discuss their potential implications for experimental work.

4.1 Standard Practices in Experimental Economics

Experimental economics has developed a set of standard practices and traditions that we believe are helpful, which we now discuss.

It has long been the norm for experimental economists to share their data after publication, even before the practice became a common requirement at journals. In fact, early papers sometimes presented the entire data set in appendices. Access to data allows easy evaluation of the sensitivity of reported results to data selection (particularly if results are reported for a subset of observations) and the assessment of alternative econometric specifications.

Also, experimental papers building on prior work frequently replicate variants of the original design. More often than not, these are *quasi-replications*, in that the design parameters and interface details may differ from those initially used. Standard replication exercises commonly mimic the original design and sample as much as possible. In turn, quasi-replications are particularly important for assessing the robustness of results to particular details of the experimental protocol, including the utilized sample, the interface participants interact through, and the precise parameters implemented.

For example, Fréchette, Kagel & Lehrer (2003) experimentally investigate the Baron and Ferejohn bargaining model, comparing two procedures referred to as closed and open rule in bargaining groups of five subjects with a discount rate of 0.8. Their experiment was conducted using pen and paper. Agranov & Tergiman (2014) study communication in this environment, and conduct as a baseline the closed rule treatment of Fréchette et al. (2003). However, they do so using computers, different instructions, etc. Furthermore, subsequent studies also explore other parameter constellations for which the same qualitative predictions apply.[17]

[17] Fréchette, Kagel & Morelli (2005a), Fréchette, Kagel & Morelli (2005b), Kagel, Sung & Winter (2010), Bradfield & Kagel (2015), and Fréchette & Vespa (2017) all include one closed rule treatment with either a different number of subjects, a different discount factor, or both. They all

Similar examples abound: Kübler & Weizsäcker (2004) quasi-replicate the social-learning experiment of Anderson & Holt (1997), Healy (2006) quasi-replicates the implementation of the Groves-Ledyard mechanism of Chen & Plott (1996), Goeree & Yariv (2011) quasi-replicate a version of Guarnaschelli, McKelvey & Palfrey (2000) in a strategic voting setup, Agranov & Yariv (2018) quasi-replicate independent, private-value first- and second-price auctions à la Kagel & Levin (1993), and so on and so forth. Importantly, as this list suggests, quasi-replications appear in well-published papers. They do not hinder or diminish publication prospects.

Quasi-replications can lead researchers to stumble on discoveries that *pure* replications would fail to uncover. When Smith (1994) discussed what we term quasi-replications, he used an analogy from Franklin & Allan (1990): "if you want to know the correct time, it is more informative to compare your watch with another's than for either of you to look at your own watch twice." For instance, Charness, Fréchette & Kagel (2004) intended to study the impact of team production on gift exchange. To do so, their baseline aimed to reproduce the standard gift exchange result of Fehr, Kirchsteiger & Riedl (1993). The result failed to replicate—they observed very little gift exchange—despite implementing the same parameters as the original study and many subsequent ones that followed. As it turned out, their inclusion of a payoff table summarizing payoffs for combinations of wages and efforts was responsible for the difference.[18] Such "accidental" discoveries can be important. In this particular example, the results highlight that canonical observations about gift exchange are sensitive to implementation details—namely, that some amount of confusion can alter results dramatically. Pure replications are of great use, but are not designed to assess the robustness of results to various design details.

Quasi-replications are sufficiently prevalent that meta-studies are often used to describe the patterns and robustness of results on a topic, and re-analyze results from multiple sources. For instance, Baranski & Morton (2021) provides a meta-study on experimental studies of Baron and Ferejohn bargaining with a closed rule. The literature has seen meta-studies on many topics: public goods (Zelmer, 2003), dictator games (Engel, 2011), ultimatum games (Cooper & Dutcher, 2011),

confirm the original qualitative results.

[18] The original study informed participants of how payoffs related to wages and efforts but did not provide a summary table computing those payoffs.

discrimination (Lane, 2016), finitely repeated prisoner's dilemma games (Embrey, Fréchette & Yuksel, 2018), indefinitely repeated prisoner's dilemma (Dal Bó & Fréchette, 2018), social dilemmas (Mengel, 2018), stag-hunt games (Dal Bó, Fréchette & Kim, 2021), and others.

Finally, many experiments in economics are constrained by the underlying model they test. This often provides clear guidance as to the dependent variables of interest and the main comparative statics to be inspected. As such, authors' ability to present as a finding that was not of interest from the start is limited.

In what follows, we discuss two recent approaches that have been suggested and pursued in the hopes of alleviating replication issues in economics: increased transparency of study designs and analysis, as well as tightening criteria for what qualifies as a statistically significant result.

4.2 Pre-registration and Pre-analysis Plans

One approach advocated widely within the empirical fields of economics has been to increase transparency, and hopefully reproducibility, through pre-registration of studies and pre-analysis plans, see e.g. Christensen & Miguel (2018) and references therein. Indeed, one potential channel that may contribute to the reproducibility problem is publication bias in favor of significant findings. Even well-intentioned researchers are induced to report specifications yielding significant results, so-called *p-hacking*, see Simmons, Nelson & Simonsohn (2011). The natural solution is to constrain researchers at the outset. If they pre-specify the analyses that will be carried out, there is limited scope for dredging the data later on.

While forcing increased transparency of research protocols appears to resolve some important pitfalls that may generate irreproducible results to begin with, its precise form is still taking shape. We suspect that feasible implementations might not be a panacea when it comes to experimental economics.

Certainly, there are logistical constraints. Pre-registration is useful only insofar that it is monitored. A researcher could pre-register multiple studies and analysis plans, or simply specify a broad umbrella of specifications that will be considered. Furthermore, particularly with the emergence of online experimental platforms, there is a risk that unmonitored pilots grow rampant and guide researchers in new ways to experimental designs and analyses that generate a significant set of results:

a substitution of p-hacking with *design-hacking*. In the words of Simmons et al. (2011), absent careful monitoring, pre-registration still leaves many "researcher degrees of freedom."

The second limitation pertains to the costs pre-registration imposes on the discovery process. Frequently, approaches to data analysis evolve as results shine through. An attempt to understand the mechanism generating the pattern of results observed in an experiment often leads to new analyses that would be difficult to pre-conceive at the outset. Certainly, one could report such results and admit their unplanned nature. It is still unclear how such caveats would be understood. In many ways, if they are accepted at face value, the commitment embedded in the pre-analysis plan could come undone. Alternatively, one could run a new study inspired by the original one and submit a more informed pre-analysis plan. This comes at a monetary cost that many scholars would not be able to afford on a regular basis. Furthermore, again, it runs the risk of undoing the benefits of pre-registration by converting preliminary studies into effective pilots.

4.3 Increasing Significance Requirements

Another approach for combating the replication crisis is to require smaller p-values for results to be considered significant, see Benjamin, Berger, Johannesson, Nosek, Wagenmakers et al. (2018). This can potentially mitigate type-I errors in research, making "false positives" more challenging to achieve.

This is in line with trends we observe. Figure 7 depicts the significance levels of main results published in the top-5 journals over the 2010-2019 decade.[19] As can be seen, the most significant results in papers, corresponding to the figure's left panel, exhibit somewhat smaller p-values in recent years. At the same time, the least significant results, corresponding to the figure's right panel, are more likely to be deemed null within this one-decade horizon.

Could this be a solution? Certainly, results associated with lower p-values entail greater statistical confidence. At the same time, a reduction in type-I error may come at the cost of an increase in type-II errors, whereby authors may be quick to dismiss a relationship that, with a larger data set, or more refined measurement,

[19] We classified a result as a "main result" if it was mentioned in the abstract, specified in the paper's introduction as a main result, or enumerated in the paper as a main result.

would emerge.[20] Furthermore, in view of the tendency of journals to publish only highly-significant results, some experimental designs and results may be lost to the literature due to lower significance.

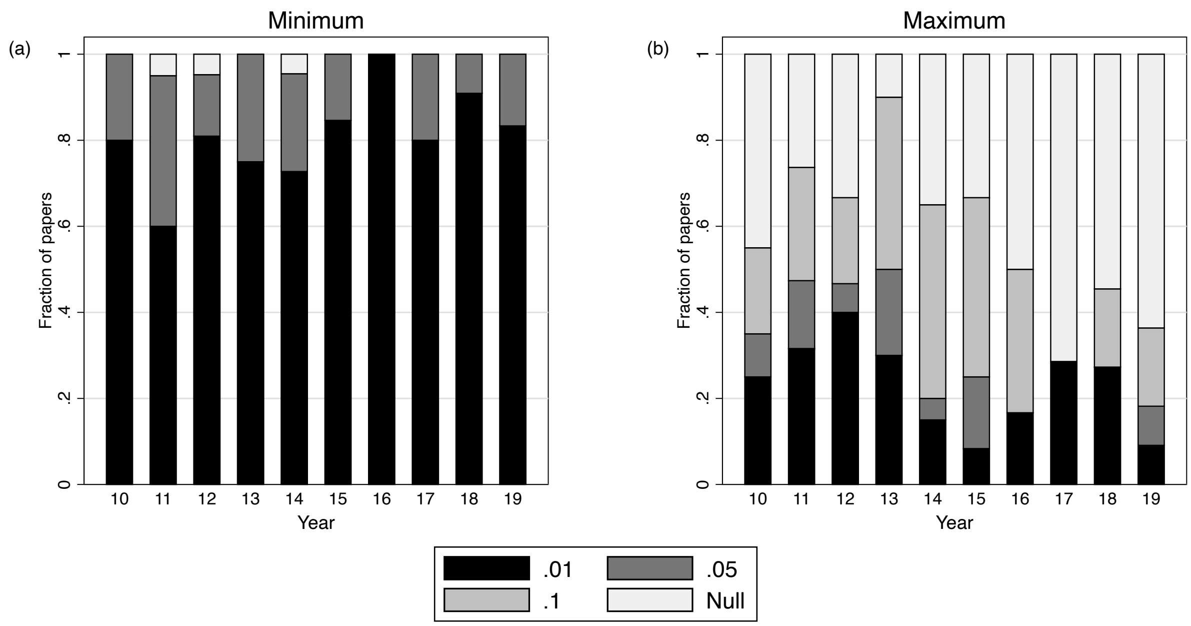

Figure 7: Results' significance levels over time

Achieving low p-values often entails larger data sets. With logistical and financial costs of physical labs, a more demanding significance level may drive researchers to other platforms, which exhibit potential shortcomings discussed in the previous section. In addition, the higher costs of collecting large experimental data sets could preclude researchers with limited access to funds from producing publishable experimental work. Indeed, the third panel of Figure 3 suggests a recent increase in top-university authors among those of top-5 papers. Certainly, this pattern in and of itself could emerge for a variety of reasons, but it might be useful for the profession to take stock of the implications of any change in journal acceptance requirements.

What more can be done? Assessing reproducibilicity is of great value and several institutional efforts in this direction might prove valuable for both our understanding of which results hold up to further testing and as a taming mechanism

[20] A related discussion, looking at the impacts of measurement error in experimental work, appears in Gillen, Snowberg & Yariv (2019).

for new research.[21] Further encouraging authors to replicate and quasi-replicate experiments their new designs are based on could prove useful as well. Finally, it seems important to reward (through publication) well done meta-analyses, as this provides incentives for authors to invest in such projects.

5 Conclusions

This article presents recent trends in experimental economics papers. We highlight two in particular: the increased use of virtual labs and the recent concerns with the replicability of studies.

Online experimental platforms may be another instance of a case in which there is no free lunch. Platforms that offer little screening of participants, such as MTurk, may provide cheaper access to a large sample. However, behavior can be noisier and participants may exhibit a shorter attention span than that observed in traditional labs. As a consequence, adjustments may be required, both to the size of samples collected—undoing some of the platforms' cost benefits—and to experimental designs' complexity. Some of the shortcomings of virtual platforms that we note may be due to the participant pools they access, rather than the virtual technology itself: on a variety of tasks, student participants appear to behave similarly online and in a physical lab. As virtual technologies evolve, benchmarking experimental platforms on a set of commonly used elicitations may prove useful and allow data-based comparisons of venues.

Recent replicability concerns have triggered attempts to modify perceived best practices. We highlight some of the more subtle effects several of the suggested approaches could yield. We direct the interested reader to Coffman & Niederle (2015), who discuss some of these issues in greater detail. They too suggest potential limitations of pre-analysis plans and argue for the value of replications via simulations. They also propose ways to stimulate more replications. We highlight complementary instruments for assessing results' robustness that are commonplace within the experimental economics field: *quasi*-replications and meta-studies. We hope their importance is recognized and their existence encouraged further.

[21] See https://experimentaleconreplications.com/ for an example of such an enterprise.

The discussion of best practices often ignores the cost implications on publishable research. We believe research costs should play a role in these debates. Keeping the costs of experiments low fosters discovery. This is important not only for the generation of new results, but also for the use of quasi-replications that identify robustness of prior insights. Increasing experimental costs also amplifies incentives to pilot designs. Tailoring design parameters to generate significant results raises a wide host of concerns. Even well-intentioned scholars may face non-trivial dilemmas in identifying the design aspects pilots help "get right."[22] Finally, increased costs can affect researchers differentially: established scholars with access to large research funds are less sensitive than their junior and less-established counterparts.

One trend that merits more analysis is the seeming increase in designs that include a battery of elicitations at the end of experimental sessions, with no a-priori justifications. Those may include gender, race, college major, risk attitudes, etc. The practice could be an artifact of the publication process: if review teams frequently ask for arbitrary associations, researchers are better off preempting such requests by generating data for responses. Requirements for pre-registration and pre-analysis plans may only increase the prevalence of these practices since producing such data at a later stage may come at higher costs. These are seemingly free data that allow for richer insights: why would there be reason for concern? Naturally, the risk of finding spurious linkages is high when many such elicitations are run. Combined with a tendency to publish significant results, exploring many correlates could paint a misleading picture about the relationship between a phenomenon of interest and other characteristics. Careful statistics and publication of insignificant linkages, not just significant ones, would alleviate, perhaps resolve, such concerns. Our impression is that these may not yet be standard. We hope more thought is given to how large elicitation menus are handled and reported. Particularly as data sets become bigger and richer, these issues are more and more germane.

Ultimately, many approaches for combating reproducibility concerns aim at correcting two types of behavior: consciously nefarious manipulations, and well-

[22]Certainly, pilots can be useful in instructing scholars how long a task takes, the clarity of the interface's instructions, etc. As a simple heuristic, pilots that do not entail data analysis may be less prone to issues of design-hacking.

intentioned practices that accidentally fall prey to degrees of freedom. Consciously nefarious intentions are difficult to alleviate. Indeed, many of the suggested approaches could easily be circumvented. The hope is that well-intentioned researchers' degrees of freedom can be limited in productive ways. Suggested solutions effectively restrain scholars' scope for discretion. However, without clear criteria for what proper discretion is, researchers may respond to new restrictions in unexpected ways, generating unintended consequences. We hope a healthy research culture, in which replications, quasi-replications, and meta-studies are standard practice, will be promoted and encouraged.

References

Agranov M, Tergiman C. 2014. Communication in multilateral bargaining. *Journal of Public Economics* 118:75–85

Agranov M, Yariv L. 2018. Collusion through communication in auctions. *Games and Economic Behavior* 107:93–108

Anauati MV, Galiani S, Gálvez RH. 2020. Differences in citation patterns across journal tiers: The case of economics. *Economic Inquiry* 58(3):1217–1232

Anderson LR, Holt CA. 1997. Information cascades in the laboratory. *American Economic Review* :847–862

Arechar AA, Kraft-Todd GT, Rand DG. 2017. Turking overtime: how participant characteristics and behavior vary over time and day on amazon mechanical turk. *Journal of the Economic Science Association* 3(1):1–11

Baker M. 2016. 1,500 scientists lift the lid on reproducibility. *Nature News* 533(7604):452

Baranski A, Morton R. 2021. The determinants of multilateral bargaining: A comprehensive analysis of baron and ferejohn majoritarian bargaining experiments. Working Paper

Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, et al. 2018. Redefine statistical significance. *Nature Human Behaviour* 2(1):6–10

Berinsky AJ, Huber GA, Lenz GS. 2012. Evaluating online labor markets for experimental research: Amazon. com's mechanical turk. *Political Analysis* 20(3):351–368

Berinsky AJ, Margolis MF, Sances MW. 2014. Separating the shirkers from the workers? making sure respondents pay attention on self-administered surveys. *American Journal of Political Science* 58(3):739–753

Bradfield AJ, Kagel JH. 2015. Legislative bargaining with teams. *Games and Economic Behavior* 93:117–127

Camerer CF, Dreber A, Forsell E, Ho TH, Huber J, et al. 2016. Evaluating replicability of laboratory experiments in economics. *Science* 351(6280):1433–1436

Charness G, Fréchette GR, Kagel JH. 2004. How robust is laboratory gift exchange? *Experimental Economics* 7(2):189–205

Chen Y, Plott CR. 1996. The groves-ledyard mechanism: An experimental study of institutional design. *Journal of Public Economics* 59(3):335–364

Christensen G, Miguel E. 2018. Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature* 56(3):920–80

Coffman LC, Niederle M. 2015. Pre-analysis plans have limited upside, especially where replications are feasible. *Journal of Economic Perspectives* 29(3):81–98

Congiu L, Nunnari S. 2021. Database of laboratory experiments in top economics journals. https://docs.google.com/spreadsheets/d/1434ApdJsvdtRNIpDR_O3h-kmUffcL24_cJj7BXyCezs/

Cooper DJ, Dutcher EG. 2011. The dynamics of responder behavior in ultimatum games: a meta-study. *Experimental Economics* 14(4):519–546

Dal Bó P, Fréchette GR. 2018. On the determinants of cooperation in infinitely repeated games: A survey. *Journal of Economic Literature* 56(1):60–114

Dal Bó P, Fréchette GR, Kim J. 2021. The determinants of efficient behavior in coordination games. Working Paper

Dreber A, Johannesson M. 2019. Statistical significance and the replication crisis in the social sciences. In *Oxford Research Encyclopedia of Economics and Finance*

Ellison G. 2002. The slowdown of the economics publishing process. *Journal of Political Economy* 110(5):947–993

Ellison G. 2013. How does the market use citation data? the hirsch index in economics. *American Economic Journal: Applied Economics* 5(3):63–90

Embrey M, Fréchette GR, Yuksel S. 2018. Cooperation in the finitely repeated prisoner's dilemma. *Quarterly Journal of Economics* 133(1):509–551

Engel C. 2011. Dictator games: A meta study. *Experimental Economics* 14(4):583–610

Fehr E, Kirchsteiger G, Riedl A. 1993. Does fairness prevent market clearing? an experimental investigation. *Quarterly Journal of Economics* 108(2):437–459

Franklin A, Allan F. 1990. Experiment, right or wrong. Cambridge University Press

Fréchette GR. 2015. Laboratory experiments: Professionals versus students. In *Handbook of Experimental Economic Methodology*. Oxford University Press, 360–390

Fréchette GR. 2016. Experimental economics across subject populations. In *The Handbook of Experimental Economics, Volume 2*. Princeton University Press, 435–480

Fréchette GR, Kagel JH, Lehrer SF. 2003. Bargaining in legislatures: An experimental investigation of open versus closed amendment rules. *American Political Science Review* 97(2):221–232

Fréchette GR, Kagel JH, Morelli M. 2005a. Behavioral identification in coalitional bargaining: An experimental analysis of demand bargaining and alternating offers. *Econometrica* 73(6):1893–1937

Fréchette GR, Kagel JH, Morelli M. 2005b. Nominal bargaining power, selection protocol, and discounting in legislative bargaining. *Journal of Public Economics* 89(8):1497–1517

Fréchette GR, Vespa E. 2017. The determinants of voting in multilateral bargaining games. *Journal of the Economic Science Association* 3(1):26–43

Fudenberg D, Peysakhovich A. 2016. Recency, records, and recaps: Learning and nonequilibrium behavior in a simple decision problem. *ACM Transactions on Economics and Computation (TEAC)* 4(4):1–18

Gillen B, Snowberg E, Yariv L. 2019. Experimenting with measurement error: Techniques with applications to the caltech cohort study. *Journal of Political Economy* 127(4):1826–1863

Goeree JK, Yariv L. 2011. An experimental study of collective deliberation. *Econometrica* 79(3):893–921

Goodman JK, Cryder CE, Cheema A. 2013. Data collection in a flat world: The strengths and weaknesses of mechanical turk samples. *Journal of Behavioral Decision Making* 26(3):213–224

Guarnaschelli S, McKelvey RD, Palfrey TR. 2000. An experimental study of jury decision rules. *American Political Science Review* 94(2):407–423

Gupta N, Rigotti L, Wilson AJ. 2021. The experimenters' dilemma: Inferential preferences over populations. Working Paper

Hauser D, Paolacci G, Chandler J. 2019. Common concerns with mturk as a participant pool: Evidence and solutions. In *Handbook of Research Methods in Consumer Psychology*. Routledge/Taylor & Francis Group, 319–337

Healy PJ. 2006. Learning dynamics for mechanism design: An experimental comparison of public goods mechanisms. *Journal of Economic Theory* 129(1):114–149

Heckman JJ, Moktan S. 2020. Publishing and promotion in economics: the tyranny of the top five. *Journal of Economic Literature* 58(2):419–70

Horton JJ, Rand DG, Zeckhauser RJ. 2011. The online laboratory: Conducting experiments in a real labor market. *Experimental Economics* 14(3):399–425

Kagel JH, Levin D. 1993. Independent private value auctions: Bidder behaviour in first-, second-and third-price auctions with varying numbers of bidders. *The Economic Journal* 103(419):868–879

Kagel JH, Sung H, Winter E. 2010. Veto power in committees: an experimental study. *Experimental Economics* 13(2):167–188

Kübler D, Weizsäcker G. 2004. Limited depth of reasoning and failure of cascade formation in the laboratory. *Review of Economic Studies* 71(2):425–441

Lane T. 2016. Discrimination in the laboratory: A meta-analysis of economics experiments. *European Economic Review* 90:375–402

Lehmann S, Jackson AD, Lautrup BE. 2006. Measures for measures. *Nature* 444(7122):1003–1004

Mengel F. 2018. Risk and temptation: A meta-study on prisoner's dilemma games. *The Economic Journal* 128(616):3182–3209

Nikiforakis N, Slonim R. 2019. Editors' preface: Trends in experimental economics (1975–2018). *Journal of the Economic Science Association* 5:143–148

Paolacci G, Chandler J, Ipeirotis PG. 2010. Running experiments on amazon mechanical turk. *Judgment and Decision Making* 5(5):411–419

Samuelson W, Bazerman MH. 1984. The winner's curse in bilateral negotiations. MIT Working Paper

Simmons JP, Nelson LD, Simonsohn U. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22(11):1359–1366

Smith VL. 1982. Microeconomic systems as an experimental science. *American Economic Review* 72(5):923–955

Smith VL. 1994. Economics in the laboratory. *Journal of Economic Perspectives*

8(1):113–131

Snowberg E, Yariv L. 2021. Testing the waters: Behavior across participant pools. *American Economic Review* 111(2):687–719

Zelmer J. 2003. Linear public goods experiments: A meta-analysis. *Experimental Economics* 6(3):299–310