Annual Review of Economics

Experimental Economics: Past and Future

Guillaume R. Fréchette,[1] Kim Sarnoff,[2] and Leeat Yariv[2,3,4]

[1]Department of Economics, New York University, New York, NY, USA; email: frechette@nyu.edu

[2]Department of Economics, Princeton University, Princeton, New Jersey, USA; email: ksarnoff@princeton.edu, lyariv@princeton.edu

[3]Centre for Economic Policy Research, London, United Kingdom

[4]National Bureau of Economic Research, Cambridge, Massachusetts, USA

Annu. Rev. Econ. 2022. 14:777–94

The *Annual Review of Economics* is online at economics.annualreviews.org

https://doi.org/10.1146/annurev-economics-081621-124424

JEL codes: B40, C80, C90

Keywords

experimental economics, time trends, experimental platforms, MTurk, replication

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.

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 progress in many fields of economics, from microeconomics to labor to finance to macroeconomics.[1]

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 the *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-five publications over the last decade of that period. Indeed, **Figure 1*a*** replicates this observation for 2010–2019. Top-five publications based on experimental work have significantly declined in the second half of this period.[2]

Despite this decline in the publication of experimental economics work, the impact of the published papers—as measured by the number of citations, which is often utilized in promotion and hiring decisions (see, e.g., Lehmann et al. 2006, Ellison 2013)—has consistently exceeded that of contemporary papers in other fields. **Figure 1*b*** displays the median number of citations in top-five journals during the last decade. This number is substantially higher for experimental work in nearly all years.[3]

Certainly, experimental work in the twenty-first century does not mirror work done 50 years prior. In what follows, we use data on top-five publications to identify recent trends in how

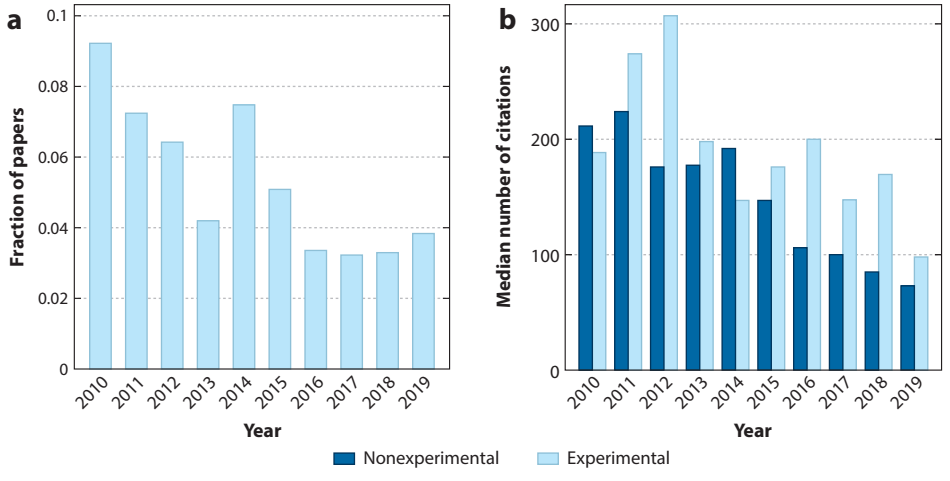

Figure 1

(*a*) Fraction of experimental economics publications in top-five journals during 2010–2019. (*b*) Median number of citations in top-five journals during 2010–2019.

[1]Kagel & Roth (1995, 2020) and Plott & Smith (2008) provide comprehensive surveys of experimental work in various fields.

[2]Throughout, we refer to the *American Economic Review*, *Econometrica*, the *Journal of Political Economy*, the *Quarterly Journal of Economics*, and the *Review of Economic Studies* as the top-five economics journals. In the data we present, we exclude nonreviewed pieces. Reuben et al. (2021) conduct a similar analysis including the *Journal of the European Economic Association* and the *Economic Journal*.

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

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 and of 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. Preregistration and pre-analysis plans, lower *p*-value thresholds for significance, and efforts to replicate existing studies have been suggested by some researchers and implemented to some extent. In the last part of this review we discuss some of the benefits and potential pitfalls that 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.

2.1. The Data

We collected detailed data on all lab experiments published in top-five journals between 2010 and 2019, both traditional experiments conducted at physical university laboratories and online or field experiments that have a lab component.[4]

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

In all the analyses we discuss below, unless noted otherwise, 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

The topics covered by published experimental work are diverse. We use the *Journal of Economic Literature* (JEL) classification to record the topics individual papers focus on. **Figure 2*a*** illustrates the five most frequent general JEL categories in the first and second half of the decade under consideration. **Figure 2*b*** provides an analogous picture for the five most frequent specific JEL categories. The breadth of the topics analyzed is substantial throughout the time period we inspect. However, we see some notable, albeit not statistically significant, shifts in focus. Comparing the

[4]The list of experimental papers comes from a database maintained by Luca Congiu and Salvatore Nunnari, from the Bocconi University in Milan, Italy. As mentioned, we exclude nonreviewed papers—comments, errata, proceedings, etc.—from all of our analyses.

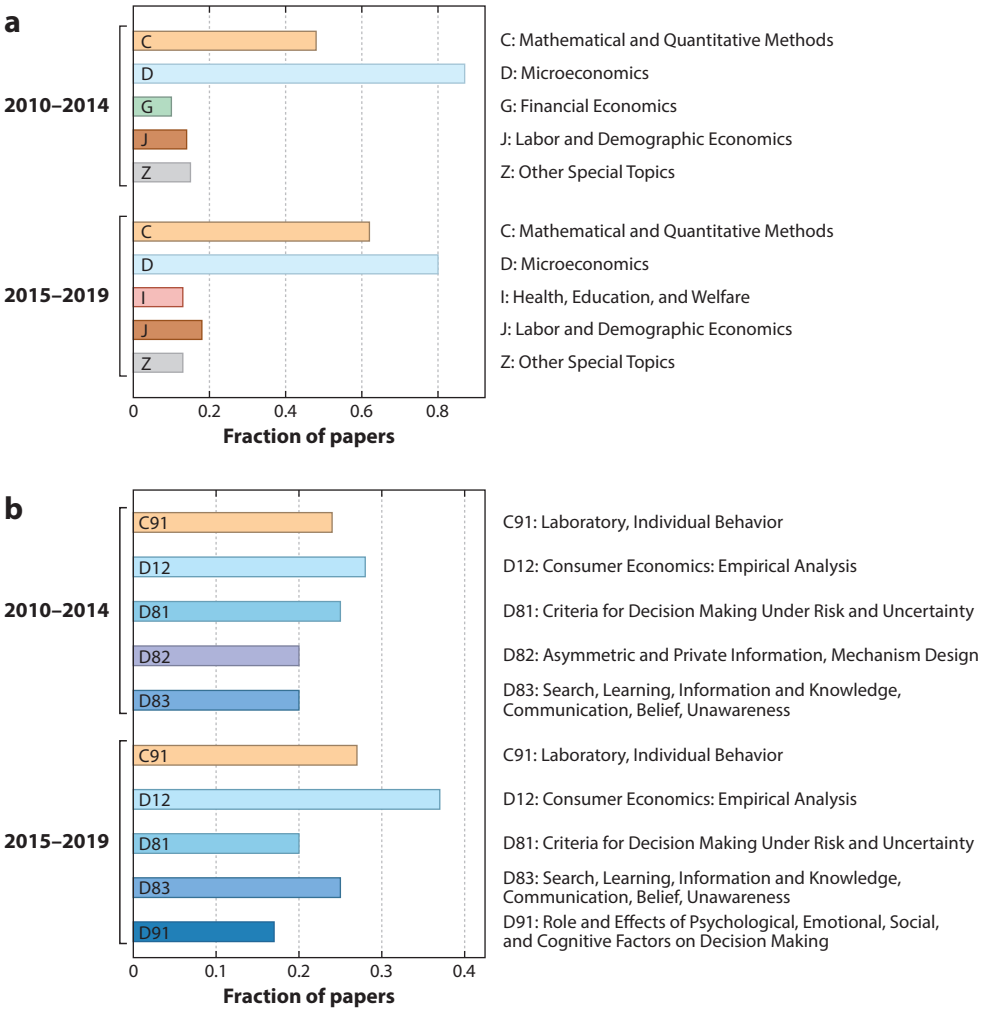

Figure 2

(*a*) Most frequent JEL general categories of papers published over time. (*b*) Most frequent JEL specific categories of papers published over time.

first and second halves of the decade, we see an increase in work focused on quantitative methods over time, a research area we touch upon in Section 4. We observe a decrease in work focusing on finance and an increase in work focusing on questions pertaining to health and education. We also see an increase in papers studying individual consumer choice and a decrease in papers studying mechanism design.

Moreover, published materials are becoming longer. As seen in **Figure 3*a***, both articles and online supplementary materials have grown in length.[5] These patterns are in line with trends identified for the economics profession as a whole over the past several decades (see Ellison 2002 and

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

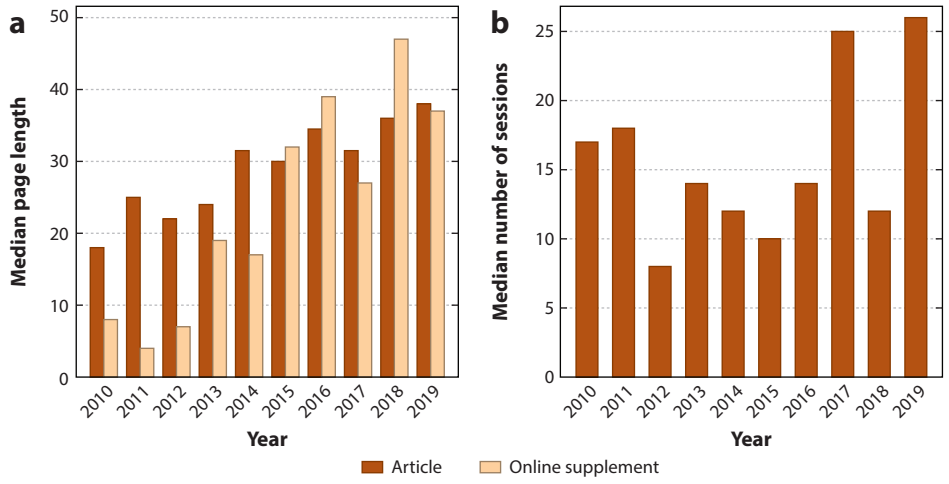

Figure 3

(*a*) Median page length of published articles. (*b*) Median number of sessions reported in published papers (only includes experiments in physical labs).

work that followed). To the extent that the length of papers and supplementary materials is correlated with time spent bringing those papers to publication, researchers may be facing increased burdens in publishing their work, at least in our lead journals. This observation may be important in hiring and promotion decisions (see also Heckman & Moktan 2020).

Figure 3*b* illustrates that the median number of sessions in laboratory experiments has increased over time as well, albeit not significantly so.[6]

2.3. Author Profiles

Next, we turn to the composition of the author teams. The average size of teams hovers between 2.5 and 3 throughout the decade. **Figure 4** 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.[7] 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%.[8] The fraction of teams with at least one nonwhite author is fairly constant, at around 33%.[9]

[6]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 we analyze. 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 our qualitative observations.

[7]We use the 2017 U.S. News & World Report rankings. We count a paper as having a top-20 coauthor 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 coauthor to work in the economics department of the school rather than in any of the school's units.

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

[9]Our categorization of authors' race/ethnicity is based on information gleaned from institutional websites and posted curriculum vitae. Two papers provide insufficient information for determining whether at least one author is nonwhite; these papers are categorized as having an all-white author team.

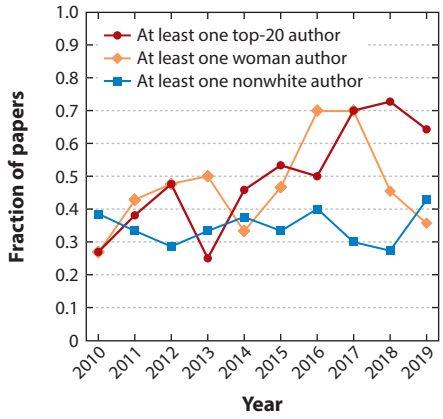

Figure 4

Time trends in coauthor teams.

2.4. Experimental Platform and Design Complexity

The number of online experiments has increased during the period under consideration. Between 2010 and 2014, only about 5% of experiments published in the top-five 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 5** breaks down the median number of participants by experimental platform, i.e., based on 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 the experiments has also changed. **Figure 6*a*** shows an increase in the proportion of experiments that do not entail a strategic interaction. This trend is potentially due to the

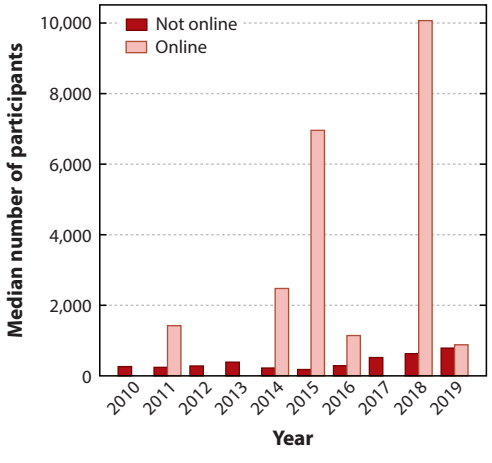

Figure 5

Time trends in number of participants.

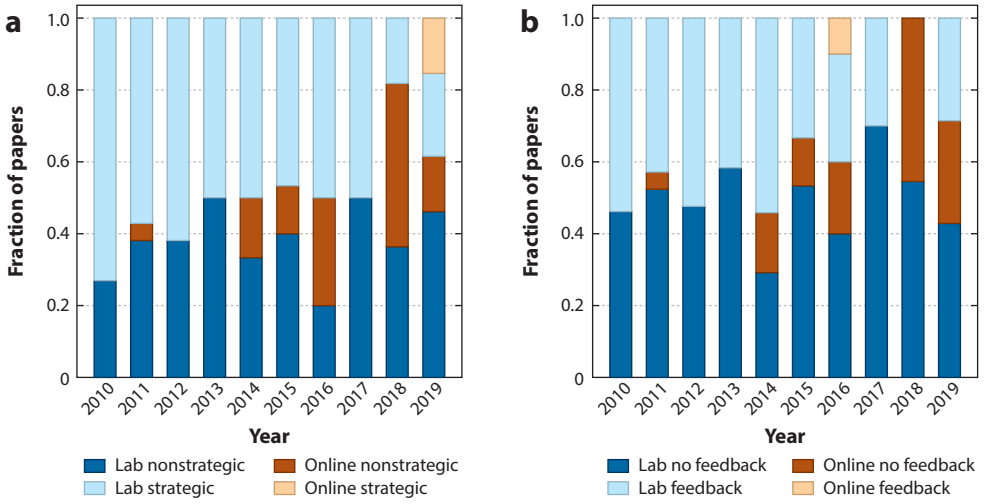

Figure 6

(*a*) Breakdown of experiments by presence of strategic interaction. (*b*) Breakdown of experiments by presence of feedback.

decrease in the share of papers relying on physical lab experiments and the growth in the number of papers relying on online experiments, which mostly involve nonstrategic tasks.

Learning opportunities are often important for more complex tasks: 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 the increase in nonstrategic experimental designs, **Figure 6*b*** demonstrates a decrease in the fraction of papers that are based on experiments providing feedback to participants. Online platforms are more frequently utilized for nonstrategic 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 has been 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, 2016 for a survey of results pertaining to various participant pools) and offer lower per-participant costs.[10]

Given these differences, it is important to understand if and how results are affected by the setting (traditional laboratory or online) in which an experiment is conducted. 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

[10]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. Because incentives are a key feature of economic experiments (see Smith 1982), reduction in costs could be a double-edged sword.

biases" experiments and nonincentivized games, as well as (incentivized) repeated public goods and prisoner's dilemma games (Paolacci et al. 2010, Horton et al. 2011, Berinsky et al. 2012, Goodman et al. 2013, Arechar et al. 2017). Hauser et al. (2019) provide a survey of the first results on the topic.

3.1. Noise Across Platforms

Recent work depicts a somewhat nuanced picture. Snowberg & Yariv (2021) address concerns about the external validity of experiments with student participants. In order to assess whether experimental insights can be generalized beyond experimental labs and university student populations, they compare university students in two universities (California Institute of Technology and University of British Columbia), both in the lab and online, with a representative sample of the US population and a sample of MTurk participants. 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 contest games, and many others. MTurk data are collected with low and high incentives, where low incentives match the average payment on MTurk and high incentives are double the low ones. A variety of attention screeners are 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.[11] However, responses do differ significantly across the three sample types: the student sample, the representative US sample, and MTurk.[12] Nonetheless, correlations between behaviors and comparative statics are similar across the samples, with differences mostly driven by some correlations being insignificant.

One important insight of Snowberg & Yariv's (2021) analysis relates to noise. The authors use duplicate elicitations in their survey. Consider a parameter of interest X^* with standard deviation σ_{X^*}. Suppose we have two elicitations of X^*, denoted $X^a = X^* + \nu_X^a$ and $X^b = X^* + \nu_X^b$, where ν_X^a and ν_X^b are independent and identically distributed random variables with mean zero and standard deviation σ_{ν_X}.[13] Then, we have that

$$1 - \widehat{\text{Corr}}[X^a, X^b] \to_p 1 - \text{Corr}[X^a, X^b] = \frac{\sigma_{\nu_X}^2}{\sigma_{X^*}^2 + \sigma_{\nu_X}^2}.$$

Thus, $1 - \widehat{\text{Corr}}[X^a, X^b]$ is an estimate of the proportion of variation of an elicitation that is due 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 the MTurk sample.[14]

Gupta et al. (2021) suggest one channel through which noise might be generated: inattention. They compare behavior in a standard physical laboratory, on MTurk, and on the Prolific experimental platform. 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

[11] Using data on historical participation in lab experiments, they also note little selection into the physical lab.
[12] Interestingly, the authors observe virtually no differences between the behaviors of MTurk participants with high incentives and those of participants with those incentives halved.
[13] This implies that $\mathbb{E}[\nu_X^a \nu_X^b] = 0$ and $\frac{\text{Var}[\nu_X^a]}{\text{Var}[X^a]} = \frac{\text{Var}[\nu_X^b]}{\text{Var}[X^b]} := \frac{\text{Var}[\nu_X]}{\text{Var}[X]}$.
[14] Because correlations between noisy variables are attenuated, this helps explain some of the correlation differences that are due to insignificance.

seems a sensible rationalization for any out-of-equilibrium behavior.[15] Noisy behavior accounts for 60% of the choices on MTurk, 19% on Prolific, and 14% in the standard laboratory. In addition, Gupta et al. (2021) also report that lab participants 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, although 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 et al. 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. Next, we provide novel data that speak 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 problem yields 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 the participants acting as buyers played the game for 30 rounds.

In the baseline treatment, participants are not informed of the distribution of possible item values but know its support. They receive feedback about the realized value 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 Information, or 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 Counterfactual, or CF).

As **Figure 7** suggests, when the experiment is run on MTurk, there are no significant differences between behaviors in the three treatments (labeled MTurk baseline, Info, and CF in the figure), either initially or after 30 rounds of play.[16] As observed by 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.[17]

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

[16]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.

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

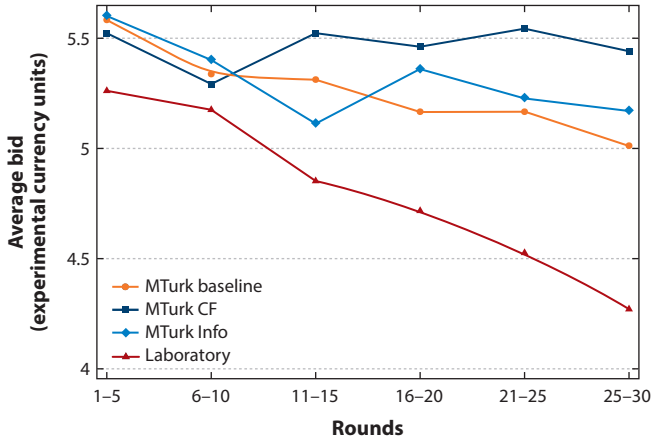

Figure 7

Bids in the additive lemons game on MTurk and in a student laboratory. CF stands for the Counterfactual treatment; Info stands for the Information treatment.

Potentially more revealing in **Figure 7** are the laboratory data recently collected by Drew Fudenberg and Guillaume Fréchette (unpublished). These data are generated from student participants in a physical laboratory playing the same additive lemons game as described by Fudenberg & Peysakhovich (2016), with $k = 3$. Participants were informed of the underlying uniform distribution of item values and were given feedback regardless of whether their offer was accepted.[18]

As can be seen, the laboratory data display significant movement toward equilibrium as participants gain experience. The figure shows that bids are slightly lower in the laboratory in the first 5 rounds ($p < 0.05$), which is likely due to the reaction to feedback early on: In the first round, the average bids are approximately 5.2 on both platforms ($p > 0.1$; no participant-specific clustering in this case). When considering the first 5 rounds, the modal bid is 5 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 representing 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 needed 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

[18] Data were originally collected as part of a separate experimental setting. As such, this study is not a replication per se.

experimental economics is not immune to such problems, the results of Camerer et al. (2016) seem to suggest that replicability issues may not be as acute as in other fields, most notably psychology (see Camerer at al. 2016, figure 4).

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 an easy evaluation of the sensitivity of reported results to data selection (particularly if the 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 et al. (2003) experimentally investigate Baron & Ferejohn's (1989) 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.[19] 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, and so on. Furthermore, subsequent studies also explore other parameter constellations for which the same qualitative predictions apply.[20] 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 et al.'s (2000) work 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. Importantly, as this list suggests, quasi-replications appear in well-published papers. They do not hinder or diminish publication prospects.

[19]In the closed-rule version, in each period, an agent is selected at random to propose a division of resources within the group. Members of the group vote on the proposed division. If there is a majority of votes, the proposal is implemented; otherwise, the process starts over. Agents discount time, that is, the number of proposals before agreement. In the open-rule version, after a proposal is made, another randomly selected agent has the possibility to amend it. However, doing so implies delays and, consequently, discounting.

[20]Fréchette et al. (2005a,b), Kagel et al. (2010), Bradfield & Kagel (2015), and Fréchette & Vespa (2017) all include one closed-rule treatment with either a different number of participants, a different discount factor, or both. They all confirm the original qualitative results: positive but less than predicted proposal power, a majority of immediate bargaining agreements, and a majority of minimal winning coalitions.

Quasi-replications can lead researchers to stumble on discoveries that pure replications would fail to uncover. When (Smith 1994, p. 128) 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 et al. (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 et al. (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 the payoffs for combinations of wages and efforts was responsible for the difference.[21] 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 to reanalyze results from multiple sources. For instance, Baranski & Morton (2021) provide a meta-study on the experimental studies of bargaining à la Baron & Ferejohn (1989) 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), discrimination (Lane 2016), finitely repeated prisoner's dilemma games (Embrey et al. 2018), indefinitely repeated prisoner's dilemma games (Dal Bó & Fréchette 2018), social dilemmas (Mengel 2018), stag-hunt games (Dal Bó et al. 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 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, and tightened 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—what is often called *p*-hacking (see Simmons et al. 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.

Although 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.

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

Certainly, there are logistical constraints. Pre-registration is useful only insofar as 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 will grow rampant and guide researchers in new ways toward 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, p. 1359), absent careful monitoring, preregistration still leaves many "researcher degrees of freedom."

The second limitation pertains to the costs that 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 preconceive at the outset. Certainly, one could report such results and admit their unplanned nature. It is still unclear, however, 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, and 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 et al. 2018). This can potentially mitigate type 1 errors in research, making false positives more challenging to achieve.

This is in line with trends we observe. **Figure 8** depicts the significance levels of main results published in the top-five journals over the 2010–2019 decade.[22] As can be seen, the most significant

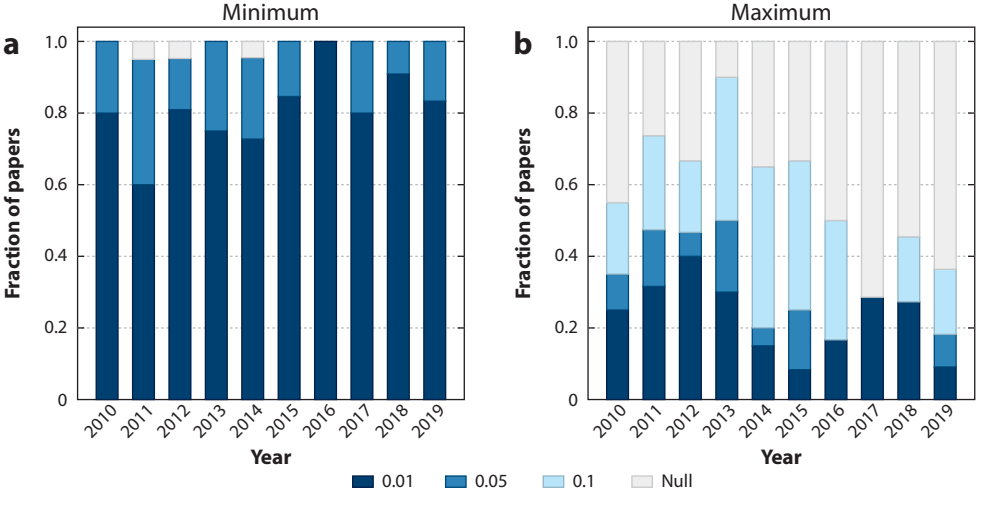

Figure 8

Results' significance levels over time.

[22] 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.

results in papers, shown in **Figure 8*a***, exhibit somewhat smaller *p*-values in recent years. At the same time, the least significant results, shown in **Figure 8*b***, 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 1 errors may come at the cost of an increase in type 2 errors, whereby authors may be quick to dismiss a relationship that would emerge with a larger data set or a more refined measurement.[23] 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.

Achieving low *p*-values often entails larger data sets. With the 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, **Figure 4** suggests a recent increase in top-university researchers among the authors of top-five 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 reproducibility is of great value, and several institutional efforts in this direction might prove valuable both for our understanding of which results hold up to further testing and as a taming mechanism for new research.[24] Further encouraging authors to replicate and quasi-replicate experiments on which their new designs are based 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 the samples collected—undoing some of the platforms' cost benefits—and to the complexity of the experimental designs. 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

[23] A related discussion, looking at the impacts of measurement error in experimental work, is provided by Gillen et al. (2019).

[24] Readers may consult the Experimental Economics Replication Project website (at **https://experimentaleconreplications.com/**) for an example of such an enterprise.

in greater detail. They, too, suggest the 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.

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 the incentives to pilot designs. Tailoring design parameters to generate significant results raises a wide host of concerns. Even well-intentioned scholars may face nontrivial dilemmas in identifying the design aspects that pilots help get right.[25] Finally, increased costs can affect researchers differentially: Established scholars with access to large research funds are less sensitive to them 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. These often 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 with many such elicitations. 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 of significant ones, would alleviate, and perhaps resolve, such concerns. Our impression is that these practices 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-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.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

[25] Certainly, pilots can be useful in instructing scholars on 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.

ACKNOWLEDGMENTS

The authors thank Alessandro Lizzeri and Salvo Nunnari for helpful conversations. L.Y. gratefully acknowledges financial support from the National Science Foundation through grants SES-1629613 and SES-1949381.

LITERATURE CITED

Agranov M, Tergiman C. 2014. Communication in multilateral bargaining. *J. Public Econ.* 118:75–85

Agranov M, Yariv L. 2018. Collusion through communication in auctions. *Games Econ. Behav.* 107:93–108

Anauati MV, Galiani S, Gálvez RH. 2020. Differences in citation patterns across journal tiers: the case of economics. *Econ. Inq.* 58(3):1217–32

Anderson LR, Holt CA. 1997. Information cascades in the laboratory. *Am. Econ. Rev.* 87(5):847–62

Arechar AA, Kraft-Todd GT, Rand DG. 2017. Turking overtime: how participant characteristics and behavior vary over time and day on Amazon Mechanical Turk. *J. Econ. Sci. Assoc.* 3(1):1–11

Baker M. 2016. 1,500 scientists lift the lid on reproducibility. *Nat. News* 553(7604):452–54

Baranski A, Morton R. 2021. The determinants of multilateral bargaining: a comprehensive analysis of Baron and Ferejohn majoritarian bargaining experiments. *Exp. Econ.* 2021. **https://doi.org/10.1007/s10683-021-09734-7**

Baron DP, Ferejohn JA. 1989. Bargaining in legislatures. *Am. Political Sci. Rev.* 83(4):1181–206

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

Berinsky AJ, Huber GA, Lenz GS. 2012. Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Anal.* 20(3):351–68

Berinsky AJ, Margolis MF, Sances MW. 2014. Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *Am. J. Political Sci.* 58(3):739–53

Bradfield AJ, Kagel JH. 2015. Legislative bargaining with teams. *Games Econ. Behav.* 93:117–27

Camerer CF, Dreber A, Forsell E, Ho TH, Huber J, et al. 2016. Evaluating replicability of laboratory experiments in economics. *Science* 351(6280):1433–36

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

Chen Y, Plott CR. 1996. The Groves-Ledyard mechanism: an experimental study of institutional design. *J. Public Econ.* 59(3):335–64

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

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

Cooper DJ, Dutcher EG. 2011. The dynamics of responder behavior in ultimatum games: a meta-study. *Exp. Econ.* 14(4):519–46

Dal Bó P, Fréchette GR. 2018. On the determinants of cooperation in infinitely repeated games: a survey. *J. Econ. Lit.* 56(1):60–114

Dal Bó P, Fréchette GR, Kim J. 2021. *The determinants of efficient behavior in coordination games*. Work. Pap., Brown Univ., Providence, RI

Dreber A, Johannesson M. 2019. Statistical significance and the replication crisis in the social sciences. In *Oxford Research Encyclopedia of Economics and Finance*. Oxford, UK: Oxford Univ. Press. **https://doi.org/10.1093/acrefore/9780190625979.013.461**

Ellison G. 2002. The slowdown of the economics publishing process. *J. Political Econ.* 110(5):947–93

Ellison G. 2013. How does the market use citation data? The Hirsch index in economics. *Am. Econ. J. Appl. Econ.* 5(3):63–90

Embrey M, Fréchette GR, Yuksel S. 2018. Cooperation in the finitely repeated prisoner's dilemma. *Q. J. Econ.* 133(1):509–51

Engel C. 2011. Dictator games: a meta study. *Exp. Econ.* 14(4):583–610

Fehr E, Kirchsteiger G, Riedl A. 1993. Does fairness prevent market clearing? An experimental investigation. *Q. J. Econ.* 108(2):437–59

Franklin A, Allan F. 1990. *Experiment, Right or Wrong*. Cambridge, UK: Cambridge Univ. Press
Fréchette GR. 2015. Laboratory experiments: professionals versus students. In *Handbook of Experimental Economic Methodology*, ed. A Fréchette, A Schotter, pp. 360–90. Oxford, UK: Oxford Univ. Press
Fréchette GR. 2016. Experimental economics across subject populations. In *The Handbook of Experimental Economics*, Vol. 2, ed. JH Kagel, AE Roth, pp. 435–80. Princeton, NJ: Princeton Univ. Press
Fréchette GR, Kagel JH, Lehrer SF. 2003. Bargaining in legislatures: an experimental investigation of open versus closed amendment rules. *Am. Political Sci. Rev.* 97(2):221–32
Fréchette GR, Kagel JH, Morelli M. 2005a. Behavioral identification in coalitional bargaining: an experimental analysis of demand bargaining and alternating offers. *Econometrica* 73(6):1893–937
Fréchette GR, Kagel JH, Morelli M. 2005b. Nominal bargaining power, selection protocol, and discounting in legislative bargaining. *J. Public Econ.* 89(8):1497–517
Fréchette GR, Vespa E. 2017. The determinants of voting in multilateral bargaining games. *J. Econ. Sci. Assoc.* 3(1):26–43
Fudenberg D, Peysakhovich A. 2016. Recency, records, and recaps: learning and nonequilibrium behavior in a simple decision problem. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, pp. 971–86. New York: ACM
Gillen B, Snowberg E, Yariv L. 2019. Experimenting with measurement error: techniques with applications to the Caltech cohort study. *J. Political Econ.* 127(4):1826–63
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. *J. Behav. Decis. Mak.* 26(3):213–24
Guarnaschelli S, McKelvey RD, Palfrey TR. 2000. An experimental study of jury decision rules. *Am. Political Sci. Rev.* 94(2):407–23
Gupta N, Rigotti L, Wilson AJ. 2021. The experimenters' dilemma: inferential preferences over populations. arXiv:2107.05064 [econ.GN]
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*, ed. FR Kardes, PM Herr, N Schwarz, pp. 319–37. New York: Routledge
Healy PJ. 2006. Learning dynamics for mechanism design: an experimental comparison of public goods mechanisms. *J. Econ. Theory* 129(1):114–49
Heckman JJ, Moktan S. 2020. Publishing and promotion in economics: the tyranny of the top five. *J. Econ. Lit.* 58(2):419–70
Horton JJ, Rand DG, Zeckhauser RJ. 2011. The online laboratory: conducting experiments in a real labor market. *Exp. Econ.* 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. *Econ. J.* 103(419):868–79
Kagel JH, Roth AE, eds. 1995. *The Handbook of Experimental Economics*. Princeton, NJ: Princeton Univ. Press
Kagel JH, Roth AE, eds. 2020. *The Handbook of Experimental Economics*, Vol. 2. Princeton, NJ: Princeton Univ. Press
Kagel JH, Sung H, Winter E. 2010. Veto power in committees: an experimental study. *Exp. Econ.* 13(2):167–88
Kübler D, Weizsäcker G. 2004. Limited depth of reasoning and failure of cascade formation in the laboratory. *Rev. Econ. Stud.* 71(2):425–41
Lane T. 2016. Discrimination in the laboratory: a meta-analysis of economics experiments. *Eur. Econ. Rev.* 90:375–402
Lehmann S, Jackson AD, Lautrup BE. 2006. Measures for measures. *Nature* 444(7122):1003–4
Mengel F. 2018. Risk and temptation: a meta-study on prisoner's dilemma games. *Econ. J.* 128(616):3182–209
Nikiforakis N, Slonim R. 2019. Editors' preface: trends in experimental economics (1975–2018). *J. Econ. Sci. Assoc.* 5:143–48
Paolacci G, Chandler J, Ipeirotis PG. 2010. Running experiments on Amazon Mechanical Turk. *Judgm. Decis. Mak.* 5(5):411–19
Plott CR, Smith VL, eds. 2008. *Handbook of Experimental Economics Results*, Vol. 1. Amsterdam: Elsevier
Reuben E, Li SX, Suetens S, Svorenčík A, Turocy T, Kotsidis V. 2021. *Trends in the publication of experimental economics articles.* Work. Pap., New York Univ., New York

Samuelson W, Bazerman MH. 1984. *The winner's curse in bilateral negotiations*. Work. Pap. 1513-84, Sloan Sch. Manag., Mass. Inst. Technol., Cambridge

Simmons JP, Nelson LD, Simonsohn U. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* 22(11):1359–66

Smith VL. 1982. Microeconomic systems as an experimental science. *Am. Econ. Rev.* 72(5):923–55

Smith VL. 1994. Economics in the laboratory. *J. Econ. Perspect.* 8(1):113–31

Snowberg E, Yariv L. 2021. Testing the waters: behavior across participant pools. *Am. Econ. Rev.* 111(2):687–719

Zelmer J. 2003. Linear public goods experiments: a meta-analysis. *Exp. Econ.* 6(3):299–310