Applied Economic Perspectives and Policy (2021) volume 43, number 2, pp. 531–555.
doi:10.1002/aepp.13091

Featured Article

Barriers to Using Economic Experiments in Evidence-Based Agricultural Policymaking

Stephanie Rosch*, Sharon Raszap Skorbiansky, Collin Weigel, Kent D. Messer, and Daniel Hellerstein

Stephanie Rosch is a research agricultural economist at USDA Economic Research Service. Sharon Raszap Skorbiansky is a research economist at USDA Economic Research Service. Collin Weigel is a postdoctoral fellow at the Whiting School of Engineering at Johns Hopkins University. Kent D. Messer S. Hallock du Pont Professor of Applied Economics and the director of the Center for Experimental & Applied Economics at the University of Delaware. Daniel Hellerstein is an agricultural economist at USDA Economic Research Service.
Editor in charge: Craig Gundersen.
*Correspondence may be sent to: sdrosch@alum.mit.edu

Submitted 27 September 2019; editorial decision 29 May 2020.

Abstract *Economic experiments have been used to inform evidence-based policymaking in a variety of fields but have rarely been used to address agricultural policy topics in the United States. Several barriers exist in designing and funding experimental studies with farmer participants, which limits the usefulness of this approach for informing agricultural policymaking. We review three such barriers: heterogenous treatment effects, access to participants, and aligning funding agencies' priorities. We document the extent of these barriers using original analyses of the literature. We then suggest potential methods of mitigating these barriers through changes in how experiments are designed, reviewed, and funded.*

Key words: Agricultural policy, Experimental economics, Farmer participants, Representativeness, Student participants.

JEL codes: C90, C93, Q18.

Policymakers request evidence from the research community in order to design policies and programs that are cost-effective and achieve program goals. While evidence to inform policymaking can come from a variety of sources, economics experiments can be particularly useful in three areas. First, they can be used to identify causal responses to policy changes which cannot be isolated in observational data. In a field experiment, for example, researchers can randomize the information provided to participants and/or vary the interface used to enroll in a program to observe how these factors affect program participation (Higgins et al. 2017). Moreover, laboratory and

artefactual[1] experiments can be used to pilot test changes in program parameters before the changes are implemented in the full-scale program.

The second important area in which economic experiments are particularly useful is in observing behaviors that cannot be observed from administrative records or other sources. For example, agricultural economic experiments can be used to observe specific behavioral mechanisms, such as how the demand for crop insurance changes in response to the degree of risk exposure or how land conservation auction bidding behaviors change with the number of competing bidders or auction structure. In laboratory experiments, researchers can use payments and procedures to "induce" behaviors from participants to mimic the assumptions necessary to test an economic theory.

Thirdly, economic experiments can be used to provide insight into why individuals decide to participate in government programs. Administrative records are being increasingly merged with survey data to recover information about nonparticipants, which can provide useful insights into why individuals opt into some programs. However, a key advantage of experiments is that they can allow us to recover a wider variety of outcomes of interest from individuals who choose not to participate. For example, in an experiment we can collect information on participants that refuse agricultural contracts (instead taking the "outside option") or do not participate in conservation programs due to adverse selection (Arnold, Duke, and Kent 2013).

Additionally, while there is a wealth of administrative data on many agricultural programs, they are frequently not structured to facilitate analysis. Even when program variations exist, finding comparable "control" and "treatment" subsets is not typically straightforward. Although there can be opportunities to construct "control" and "treatment" cohorts, program administrators are often reluctant to do so. Concerns with potential equity (who gets the better treatment or who gets the treatment and who stays in the control), burdens of data collection and program administration, and coordination required with cooperating entities can deter administrators from structuring programs to facilitate analysis. Results from experiments can be used to demonstrate the potential benefits from implementing program changes as well as the benefits of structuring administrative record collection to support high-powered analysis for evidence-based policy design.

Over the past two decades, economic experiments have been used in a variety of evidence-based policymaking initiatives outside of agriculture. Economic experiments have informed policymaking for the US federal government in many contexts, including allocating capacity on shuttle flights for NASA's Space Station Program Payload Office (Lei, Noussair, and Plott 2000), design of Federal Communication Commission spectrum auctions (Guala 2001; Banks et al. 2003), auctions of landing and takeoff rights at crowded airports (Ball, Donohue, and Hoffman 2006), SEC determinations about eliminating "uptick" or "short-selling" restrictions (U.S. Securities and Exchange Commission Office of Economic Analysis 2007), and the U.S. Environmental Protection Agency change of fuel economy labels on cars based on results from Larrick and Soll (2008) (Pete 2014). At the state level, Georgia used data from an economic experiment when designing auctions of

[1]*An artefactual experiment is one that samples participants from the target population (Harrison and List 2004).*

irrigation rights for the state's Environmental Protection Division (Cummings, Holt, and Laury 2004). Overseas, economic experiments have been used to inform policymaking for sales of telecom licenses in Europe (Abbink et al. 2002; Binmore and Klemperer 2002; Klemperer 2002), and for revising the EU's non-horizontal merger guidelines (Normann and Ricciuti 2009). In an experiment conducted in Brazil, Hjort et al. (2019) found that policymakers placed a high value on evidence derived from experiments and were more likely to act when a program or policy was based on experimental evidence. In fact, Ireland's Competition and Consumer Protection Commission, Commission for the Regulation of Utilities, and the Commission for Communications Regulation jointly support the Programme of Research Investigating Consumer Evaluations (PRICE) Lab, which uses laboratory and online experiments to inform government regulation. For example, PRICE studies have informed the Central Bank of Ireland on consumers' decision making for personal loans (Lunn, Bohacek, and Rybicki 2016).

Billions of dollars are allocated for agricultural programs in the United States and throughout the world. But to date, economic experiments have rarely been used to inform agricultural policymaking in high-income countries (Palm-Forster et al. 2019a) because of the difficulties involved in designing, funding, and implementing experiments with farmers, ranchers, and rural populations. In this paper, we review three major barriers that limit the application of experimental methods in agricultural policymaking: heterogenous treatment effects, access to participants, and aligning funding agencies' priorities. We document the extent of the problems that arise from these barriers through an analysis of the literature and suggest opportunities to mitigate these barriers through changes in how experiments are designed, and how experimental research proposals are reviewed and funded. To address issues of heterogenous treatment effects, we recommend creating a standard set of demographic and farm characteristics that all agricultural experimental studies could report on, and using stratified or blocked randomized designs to increase the power of designs used to test for heterogenous treatment effects. We recommend using student participants to pretest experimental protocols as a way to minimize the burdens of data collection from farmer populations. We also recommend research funding organizations focus on increasing the numbers of experimental studies comparing student and farmer behavior, to better understand the usefulness of students as a test population for agricultural policy experiments. Finally, to overcome the funding barrier, we recommend using an explicit set of criteria to fund policy-relevant experimental research, and that funders invest in management case studies of agencies' experiences implementing agri-economic experiments in government programs so researchers and policymakers can both benefit from prior institutional knowledge.

A General Concern: Balancing Internal Validity, External Validity, and Parallelism

The data used to support policy analysis can come from a variety of sources, including administrative records, surveys, focus group reports, and experiments. The usefulness of the analysis for informing policymaking depends, in part, on the quality and applicability of the underlying data. While no data source can perfectly measure all outcomes of interest for policymaking, the methods used to collect the data can affect the representativeness

and comprehensiveness of the analysis drawn from it. The fundamental design trade-offs between internal validity, external validity, and parallelism have strong implications for the representativeness and comprehensiveness of experimental data for policymaking in general and for agricultural policymaking in particular.

Any economic experiment designed to provide high-quality results that are useful for evidence-based agricultural policymaking must jointly address internal validity, external validity, and parallelism (Messer, Duke, and Lynch 2014). Internal validity requires that experiments control for potential confounding factors, or factors that could affect the treatment effect but are not the treatment themselves. If factors aside from the treatment could account for differences in behavior among the treated groups, then the results from the experiment cannot be solely attributed to the treatment. For example (Example 1), an experimenter's goal may be to estimate farmers' demand for farm operating loans when exposed to crop price risk. The experimenter could do this by comparing a treatment with a fixed price and a treatment with price randomly falling within some range. However, to ensure that the experiment does not suffer from internal validity problems, the experiment must also ensure that credit is offered at the same terms to all participants. Otherwise, the experimenter will not be able to disentangle the effect of price risk and cost of credit on demand for farm loans.

The bar for achieving internal validity for an agricultural policy experiment is particularly high due to the diversity and complexity of farm businesses and agricultural policies. An experiment designed to measure the effect of premium subsidies on crop insurance purchases must necessarily consider the different types of policies offered (catastrophic coverage, yield coverage, revenue coverage, rainfall indexed, etc.), coverage levels available, farm price and yield risk exposure, and alternative risk management strategies available. Most laboratory and artefactual policy experiments tend to abstract from a complicated policy environment in order to isolate the treatment effect, such as offering a single type of crop insurance policy or limiting the set of alternative risk management strategies available. This abstraction allows the experimenters to control potential confounding factors and achieve high internal validity, though at the cost of reducing the experiment's parallelism.

Parallelism (Smith 1982; Plott 1987; Levitt and List 2007a, 2007b; Camerer 2011) is the extent to which the conditions in the experiment replicate the real-world conditions of the policy setting. Experiments that have a high degree of parallelism produce behaviors observed within the experiment that can be expected to reflect behaviors in the real world. Many experiments abstract from the real-world context to better test theoretical predictions or provide more control over potential confounding factors. For example, experimental instructions may use generic terms such as "goods," "tokens," "buyers," "sellers," or "managers of a public good" to ensure that experimental participants make decisions based on the economic incentives provided in the experiment rather than any intrinsic beliefs about the goods or actions used in the actual policy context. However, factors not accounted for in the experiment can have a significant impact on choices made by the participants outside of the experiment, and without parallelism, the experimental results might not reflect the actual outcomes likely to result from implementing the policy.

Parallelism is often cited as a critical concern by agricultural policymakers and academic peer-reviewers alike when reviewing funding proposals and

evaluating experimental evidence for use in policymaking. Agricultural policies are normally tailored for specific types of crops or livestock (e.g., row crops, specialty crop, organic dairy), specific types of marketing practices (e.g., contract producer, supplier of local markets), and specific types of producers (*e.g.* beginning farmers, socially disadvantaged producers, high adjusted gross income levels). Policymakers and peer-reviewers may be skeptical of unframed experimental designs or designs that make extreme simplifications of complex policy choices and require extra justifications for why these simplifications and framing choices are necessary and reasonable.

Adding context to the experiment can increase the degree of parallelism between the experiment and the actual policy setting by giving participants "…some contextual cues about why their decision might matter in a bigger world" (Lusk and Shogren 2007). For example, Palm-Forster et al. (2019b) examine the effect of policies to reduce pollution in agricultural landscapes, telling participants that they were managers of firms sharing a watershed group. However, adding context may introduce bias if the experimental participants have strong opinions about the policy issue, and even create problems of internal validity if these opinions motivate participants in ways that contradict the financial incentives provided within the experiment.

External validity, or representativeness, is the extent to which the results generated by experimental participants represent the broader population (Muller 2014). Representativeness is also referred to as "transportability" in statistical studies and "effect homogeneity" in epidemiology (Pearl and Bareinboim 2014; Bareinboim and Pearl 2016; Lesko et al. 2017). For the results of an experiment to have external validity, the participants must respond like many or most of the target population would in that experiment. One way to achieve representativeness is to recruit participants from the target population. Another is to recruit participants who make decisions in the experiment in the same way that members of the target population would (e.g., students modeling farmers' decision-making, if it is believed that students will properly represent farmers' decisions). However, both approaches can suffer from convenience bias, which occurs when experiments do not address recruitment systematically and recruits a nonrepresentative sample of the broader population. In that case, the participants' behavior, on average, may not necessarily match those of the target group solely due to the convenience sampling rather than due to any intrinsic differences in how the model and target populations make decisions.

The problem of external validity is particularly acute in agricultural experiments because it is rarely possible in practice to recruit a large, representative sample of farmers in an artefactual experiment (see Palm-Forster et al. 2016 and Weigel et al. 2020). Power analyses indicate that large samples are often needed to detect treatment effects for some types of agri-environmental policies, and recent research has found that many of the agri-economic experiments conducted so far have been underpowered (Palm-Forster et al. 2019). As researchers seek to improve their experimental designs by increasing the number of participants, policy-relevant research will become even more expensive and time consuming to conduct, especially if external validity concerns mandate the recruitment of farmers, ranchers, and landowners.

The challenges of balancing issues of internal validity, parallelism, and external validity for agri-economic policy experiments are difficult, but these experiments can answer important policy questions not easily ascertained through other methods, such as behavior in strategic settings. For example,

the U.S. Department of Agriculture (USDA) regularly uses auctions to allocate resources. The Conservation Reserve Program (CRP) is a prominent example. In this program, auctions help to place conservation activities, like planting native grasses, in locations that are ecologically and cost efficient. Experimental auctions can be used to reveal information about the supply side (e.g., value of farming the land) that government programs do not know. Information asymmetry makes it possible for farmers and landowners to receive payments in excess of the minimum they would be willing to accept, called economic rent, which reduces the overall efficiency of the auction. Programs can change parameters, such as what information they publicly share about the selection process, the available budget, or the range of allowable bids, but theoretical models provide multiple equilibria that are often uninterpretable for policy decisions. Hellerstein, Higgins, and Roberts (2015) and Hellerstein and Higgins (2010) show how auctions parameters including asymmetric information, bid caps, and quotas could affect CRP auctions.

Strategic settings are common in competitive agricultural programs but often difficult to explore with theory or observational data. Other strategic scenarios that apply to agriculture include agglomeration bonuses (Parkhurst et al. 2002; Banerjee et al. 2014; Fooks et al. 2016; Banerjee 2018), trading markets (Cason and Gangadharan 2011; Perkis et al. 2016; Cason and de Vries 2019), market rules and public information provision (Cason and Gangadharan 2004; Duke et al. 2017; Messer et al. 2017); reduction of nonpoint source pollution (Poe et al. 2004; Spraggon 2004; Suter et al. 2008; Suter, Vossler, and Poe 2009; Suter et al. 2010; Spraggon 2013; Miao et al. 2016; Palm-Forster et al. 2019; Butler et al. 2020), groundwater extraction (Suter et al. 2012; Li et al. 2014; Suter et al. 2019), practice adoption (Hellerstein, Higgins, and Horowitz 2013; Liu 2013; Liu and Huang 2013; Brick and Visser 2015), and voluntary contributions to generic advertising (aka check-off) programs (Messer, Schmit, and Kaiser 2005; Messer, Kaiser, and Schulze 2008). Experiments allow researchers to explicitly state the costs and benefits of participating in a program, which is not known in real world programs, and allow researchers far more control over parameters that could not be easily changed or varied across farmers.

Barriers to Using Experiments for Evidence-Based Agricultural Policymaking

In addition to the ever-present general concern of balancing internal validity, external validity, and parallelism, there are other significant challenges researchers must address when designing experiments to inform evidence-based agricultural policymaking. We examine three significant barriers to using economic experiments to inform agricultural policymaking: heterogeneous treatment effects, limited access to target populations, and a lack of congruence between funding priorities and evidence-based research.

Heterogeneous Treatment Effects

Farmers in high-income countries are a diverse group of people, and their business structures, practices, and needs are equally diverse. While average treatment effects are useful for assessing the benefits of a program as a whole, policymakers are often interested in how the effects of a policy or program vary across the farm population. They might, for example, want to understand how a policy affects households with different levels of food security;

how small, medium, and corporate farm operations are affected by risk management policies; or how responses to conservation policies depend on the type of land tenure a farmer has and/or the terms of rental agreements between landowners and farm operators. When a treatment affects different groups within the population in different ways, that is called a heterogenous treatment effect. Understanding these heterogenous treatment effects is particularly useful for modeling the impacts of policies aimed at specific subpopulations such as beginning farmers, farmers producing on ecologically sensitive lands, and socially disadvantaged producers.

Economic research for policy applications often identifies treatment effects that vary across population subgroups (Heckman and Vytlacil 2001). It is now common for laboratory and field experiments to collect demographic information from participants, and popular experimental software packages such as "z-Tree" make it easy to collect this information. However, there is no standard in the experimental economics field on which sociodemographic variables of interest should be collected (Gächter 2009). This creates a two-pronged missing data problem. First, experimenters may not collect essential demographic characteristics, particularly if they are not planning to include them in the final publication. Second, even if the characteristics are collected, the experimenter may not test for interactions between characteristics and treatment effects, or not publish the interaction test results. This missing data problem makes it difficult to infer how common heterogeneous treatment effects are in agri-economic experiments and to determine the set of characteristics necessary to collect for agri-economic policy experiments.

To assess how common heterogeneous treatment effects are in the context of agricultural economics research, we reviewed 83 experiments using farmers, fishermen, ranchers, and landowners as participants (see table 1 for the results, and appendixes A1 and A2 for studies reviewed). We identified the studies by searching Google Scholar for all published experiments that (i) used farmers, fishermen, ranchers, or landowners as participants; and (ii) incorporated at least one experimental treatment. We also reviewed unpublished manuscripts that seemed likely to be published in the near future. We excluded studies that were not peer-reviewed, and book chapters for which we could not determine if a peer-review had been conducted.

In our review of each study, we noted whether any tests of correlation between the treatment effects and demographic (or farm) characteristics were reported, and whether reported correlations were statistically significant. Approximately one-third of the studies (twenty-eight) did not report results of tests for correlation between any demographic or farm characteristics. The rest tested for correlation between the treatment effects and at least one demographic or farm characteristic.

We found that most of the studies reported significant correlations between treatment effects and at least one demographic/farm characteristic. Wealth, race/ethnicity, and health status were always found to be correlated with treatment effects in studies involving farmer participants. However, table 1 illustrates how inconsistently these types of correlations were reported. Though wealth, race/ethnicity, and health status were found to be correlated with the treatment effect in all of the reviewed studies, few studies reported testing for those characteristics (16%, 4%, and 1%, respectively). The most commonly tested characteristics were age, education, and gender, which were found to be correlated with treatment effects in 55%, 52%, and 40% of the studies, respectively.

Table 1 Demographic and Farm Characteristics Correlated with Treatment Effects in Studies with Farmer Participants

	Farmer, fishermen, rancher and landowner experiments	
Characteristic	**Percent of studies reporting tests for correlation with treatment effects (1)**	**Conditional on reporting, percent of studies finding significant correlation with treatment effects (2)**
Age	55%	48%
Education	52%	51%
Gender	40%	45%
Land size	30%	44%
Household size	17%	71%
Experience with farming/ fishing/ ranching	16%	46%
Wealth	16%	100%
Income	16%	62%
Marital Status	7%	17%
Race/Ethnicity	4%	100%
Health	1%	100%

Note: *Column 1 represents the studies that reported correlation tests. There were eighty-three studies in total: fifty-five studies tested for at least one correlation between the treatment effect and a demographic or farm characteristic, and twenty-eight studies included no tests of correlations between a treatment effect and any demographic or farm characteristics. Nonreporting could be due to the researchers not collecting the information or not reporting the information. Column 2 represents the percentage of studies finding a significant correlation based on having reported (Column 1).*

The results in table 1 likely mischaracterize the true extent of correlations between treatment effects and the selected demographic and farm characteristics tested. In general, studies designed to be adequately powered to detect an average treatment effect are inadequately powered to detect a differential treatment effect for subgroups within the subject pool (see Brookes et al. 2004 for simulation results and related discussion of this problem). The intuition is that when one estimates a treatment effect for two or more subgroups, not only is sample size reduced to just the subjects in each subgroup, but the goal of the experiment becomes trying to detect the difference in treatment effects between the subgroups instead of the pooled treatment effect for both groups. The difference in treatment effects will be smaller than the pooled treatment effect for both when the treatment effect for the two groups is of the same sign but different magnitudes. None of the studies reviewed indicated that the researchers had chosen the number of subjects sampled based on power requirements needed to detect differential treatment effects within the subject pools.

Underpowered studies can mislead policymakers in two ways. The first is failure to detect a true treatment effect. The second is exaggeration of detected treatment effects, be they true or false results. It is generally known that underpowered studies are unlikely to detect policy-relevant effects, but the truth of the matter is worse. Underpowered studies are also unable to detect modest but policy-relevant effects. This is because of the mechanical relationship between standard error and the effect size. If the true effect size is half the

size of the standard error, the treatment effect will necessarily be insignificant if accurately estimated. The only case in which the treatment effect is significant is when it is inaccurately estimated by a factor of four (see Gelman and Carlin 2014 for further discussion).

Testing for heterogeneous treatment effects with adequate power is challenging given the difficulty of recruiting large numbers of farmers for field experiments conducted by academic researchers not embedded within government programs (see Weigel et al. 2020). Heterogenous treatment effects can add variance and further reduce the power of a study, but this problem can be mitigated by using techniques such as stratification and block randomization (Duflo, Glennerster, and Kremer 2007). Stratification and blocking compare the treated subjects to control subjects who were grouped or stratified before assignment to treatment. Grouping or stratifying subjects based on a particular characteristic ex-ante reduces the noisiness of the estimated treatment effect within each stratum or block. For properly powered experiments, stratification also improves the power of the design to detect heterogeneous treatment effects between the strata or blocks. None of the farmer-focused experiments we reviewed used stratification or blocking, making it more likely that the studies were unable to detect true heterogeneous treatment effects and increasing the noise of their estimate if heterogenous treatment effects exist.[2]

Researchers should be aware that self-selection into experiments can reduce the number of subjects in under-represented strata, such as high-income or minority farmers. Researchers must address underrepresented strata through the recruitment process to achieve the desired level of statistical power among target subgroups. Adding more subjects to well-represented strata typically does not make up for underrepresented strata. It may not be feasible for academic researchers to recruit sufficient subjects in all underrepresented strata through conventional recruitment techniques (see Weigel et al. 2020). Options for increasing recruitment of underrepresented strata could include expanding the geographic scope of recruitment activities, partnering with private sector businesses and NGOs that serve the targeted community (e.g. credit institutions, equipment dealers), and working with state extension agents.

We offer two recommendations for researchers to assist them in characterizing the frequency of heterogeneous treatment effects when using farmers as participants and to improve the usefulness of their experimental findings for evidence-based policymaking:

> Recommendation 1: Test for correlations between a standard set of demographic and farm characteristics in all experiments in which farmers, fishermen, ranchers, and landowners are participants.

Standardizing the set of characteristics tested in each experiment will greatly expand the body of evidence available for a meta-analysis of which characteristics tend to be correlated with treatment effects. Standardization will also have the side benefit of limiting the opportunity for ex-post data mining of

[2]*Poststratification is also an option for deriving heterogeneous treatment effects after data collection is complete. However, none of the studies reported using this approach, and its usefulness in for experiments with small sample sizes or strongly imbalanced assignments to treatment is not certain (Miratrix, Sekhon, and Yu 2013).*

experimental results, a technique which has been suggested as contributing to the replicability crisis in scientific research (Duvendack, Palmer-Jones, and Reed 2017).

We propose using the characteristics listed in table 1 as an initial list of standard characteristics to test. This list could be refined over time as new studies increase the evidence available for heterogeneous treatment effects in different policy contexts. Ideally, experiments should test for all of the characteristics listed in table 1. When collecting data on the full set of characteristics is not possible (*e.g.* privacy concerns, limited time), researchers could collect data on as many characteristics as feasible and note their rationale for choosing the subset of characteristics reported in the published manuscript or online supplemental materials. Documentation of the reasons why some characteristics could not be tested will also be helpful in refining the standard list of characteristics to test based on feasibility considerations.

> Recommendation 2: Use stratified and/or block randomized designs to test for heterogeneous treatment effects whenever possible in experiments designed to inform policymaking for farmer, fishermen, rancher, and landowner populations.

While not all experiments designed to inform policymaking will need to consider heterogeneous treatment effects, the growing demand for evidence-based policymaking makes it likely that future policy-relevant experiments will be designed to look for heterogeneous treatment effects. In those cases, we recommend that researchers use stratified and/or blocked randomized designs with sufficient representation in all strata to ensure adequate power for these tests. High powered tests are essential. Under-powered studies could fail to identify heterogeneous treatment effects, leading programs relying on the experimental results to fail to address the needs of particular subsets of the population.

Our second recommendation is a big ask. As Weigel et al. 2020 points out, recruiting farmers as participants is difficult for academic researchers that are not directly collaborating within government agricultural programs. This recommendation may require agricultural researchers to change the way they think about their sample, and it is likely to require significant funding, time, and teamwork to collect data from large enough samples of farm populations. For example, researchers interested in conducting a well-powered experiment with heterogenous treatment effects that samples corn growers may have to collaborate with extension specialists from a broad region, such as Indiana, Illinois, Iowa, etc.

To see why failing to detect true effects is important for policymaking, consider the following example (Example 2). Suppose researchers are creating an experiment to measure the effect of subsidizing the cost of crop insurance premiums on farm capital investments. Assume that beginning farmers (i.e., those with less than ten years of experience as an operator) are more likely than experienced farmers to increase capital investments in response to an increase in crop insurance premium subsidies.

Suppose the researchers in Example 2 want to use two strata: beginning farmers and experienced farmers. Assume that the standard deviation of beginning farmers' responses to the treatment is twice as large as the standard deviation of experienced farmers' responses to the treatment. If the researchers want an equal number of treated and control participants in each

stratum, then they will need to sample four times as many beginning farmers as experienced farmers to achieve equal power to reliably detect the treatment effects in both strata. This is because required sample size scales with the square of the standard deviation.[3]

To detect a difference between the two strata, researchers must recruit far more farmers than would be needed if testing only for an overall treatment effect. Assume that the beginning farmers had a treatment effect of 1.5 and the experienced farmers had a treatment effect of 1. Pooling both groups together, the experiments would need enough power to detect a treatment effect in the range of 1 to 1.5. In contrast, to detect difference in treatment effects between the two samples, the researchers would need to be able to detect a difference of 0.5, which is less than half the size of the pooled treatment effect and therefore requires 4 times as many subjects to detect as the pooled treatment effect.

It is important to note that experiments are not the only means of studying the impact of crop insurance subsidies on farm capital investment. Because of the large sample sizes involved and the volume of data collected, the combination of survey and administrative data can yield valuable insights about the distribution of farm characteristics and crop insurance purchases, the correlation between climate conditions and crop insurance purchases, and other relationships of interest to policymakers. However, experiment described in Example 2 could not be completely replicated using survey and administrative records alone. While the U.S. Department of Agriculture offers many programs that target beginning farmers, including direct and guaranteed loans and microloans, the administrative records from the crop insurance program do not identify purchases by beginning and experienced farmers. Moreover, the range of subsidies offered in the history of the program has been limited and surveys may not repeat sample farmers frequently enough to observe changes in beginning farmers' capital investment that were coincident with changes in crop insurance subsidy rates. For these reasons, the USDA Economic Research Service (ERS) has continued to fund research on the crop insurance program using experimental studies, survey data collection, administrative records, and other study designs.

Limited Availability of Target Populations (and Necessity of Using Model Populations)

The second barrier we consider is the limited access to farm populations for experimental research, and the resulting need to use experiments with model populations (such as students) to complement policy-relevant experiments involving farmer and rural populations. The previous section discussed the need for large sample sizes to ensure high-powered experimental tests of treatment effects for agri-economic policy experiments, and designing experiments to test for heterogeneous treatment effects only increases the burdens for total participants required to achieve adequate statistical power. It is often prohibitively expensive and time-consuming to recruit enough farmers to adequately power simple experiments, let alone a complex experiment with multiple treatments. The most common recruitment methods—in-person recruiting at agricultural conventions/shows and by mail—are inherently prone to generating substantial self-selection issues. Weigel et al. 2020

[3]*See Athey and Imbens (2017) for formulas and examples of power calculations for treatment effects.*

presents a detailed overview of the costs and complexities associated with this type of recruitment.

Additionally, farmers are a population of interest for a wide variety of policy-relevant questions, and as such, are in high demand from many data collectors. The demands on farmers time from competing research interests can lead to research fatigue and low response rates. McCarthy and Beckler (2000), for example, found that both limited time and concerns about data privacy influenced many farmers' decision regarding whether to participate in surveys. In fact, the farmers' response rates to surveys are declining. Every year, NASS releases the Acreage and Production surveys which provide valuable information on farmers' planting intentions and the US crop supply. The response rates for these surveys have fallen from 80%–85% in the early 1990s, to 57%–67% percent in 2016 (Johansson et al. 2017), and the rate of decline began to accelerate in 2011.

Students have long provided a cost-effective and accessible participant pool for economic experiments and can be used for initial testing prior to using farmer participants. Students have been used in economic experiments to model decision-making for a wide variety of populations, including businesses responding to new emissions trading platforms (Cason and Plott 1996), stock traders reacting to new information (Lei, Noussair, and Plott 2001), bidders in auctions for telecommunications spectrum rights (Brunner et al. 2010), organ donors (Kessler and Roth 2012), voluntary contributions of farmers to generic advertising programs (Messer, Schmit, and Kaiser 2005; Messer, Kaiser, and Schulze 2008), and public officials responding to incentives to behave corruptly (Drugov, Hamman, and Serra 2014).

Students are less expensive to recruit and incentivize than farmers. Berinsky, Huber, and Lenz (2012), for example, report per participant costs for students in political science experiments are around $5–$10, as compared to $30 for nonstudent samples recruited on academic campuses and $15–$20 for participants recruited through temporary employment agencies. Students are also easy to recruit in large numbers, are available throughout much of the year, and often have schedules that can accommodate experiments that require coordination between subjects or repeated sampling.

Another advantage of students is their relative homogeneity. Students are more similar in certain demographic characteristics, such as age and income, than the general population. This similarity reduces the number of factors that contribute to variation in the experiment, which increases statistical power and makes it more likely to detect true treatment effects (and avoid incorrect findings). By definition, students are well educated. Typically, they are also comfortable with accessing the internet and using computers and tablets, facilitating researchers' use of electronic methods in experiments that can improve the speed and accuracy of data collection.

The key question, though, is whether student participants are predictive enough of farmer decision-making to be a useful first step in policy-relevant agricultural experiments. Evidence from noneconomic studies has shown that experimental treatment effects can vary between students, professionals, and members of the general population (King and He 2006)—indicating that student results may not be predictive of all types of populations or in all types of decisions. Students typically differ from farmers in terms of age (the average age of farmers has been rising for decades), socioeconomic status, and life experiences. There are many more female students than female farmers. Consequently, using students as participants in any policy-relevant agricultural

experiments immediately raises questions about the external validity of the study. If students do not act enough like farmers when making decisions in the experiments, the results of those experiments will not provide accurate information to guide policymaking. However, as long as student results can be predictive enough of farmers' decision-making, experiments using student participants can be cost-effective and useful for pretesting experimental procedures, checking for average and heterogeneous treatment effects, and confirming that the experiment has a high degree of parallelism with the policy setting in question.

So how do we know whether students can adequately represent farmer and rural populations? To answer this question, we surveyed the literature to identify all published and unpublished experimental studies that included at least one treatment effect and had been conducted with both student and farmer participants. We limited our analysis to examine representativeness of treatment effects only (as opposed to other types of outcomes like preferences or strategies used) as this is the type of outcome most relevant for informing policy. Also, because of resource constraints, we restricted our scope to exclude choice experiments and other types of contingent valuation methods[4] that did not involve direct financial incentives.

Our search yielded 13 studies[5] that involved 64 unique treatment effects. However, only nine studies (36 unique treatment effects) provided sufficient data to compare the results between student and farmer participants. Table 2 shows the extent to which the student treatment effects matched the farmer treatment effects. The list of studies comparing farmer to student populations can be found in appendix table A3. We compared the treatment effects across samples in two ways. First, we look at whether the estimated treatment effects are significantly different from each other at the 95% confidence level. Second, we consider whether the signs of the treatment effects are different from each other.

Our analysis shows that student treatment effects were predictive of farmer treatment effects in the majority of cases. 86% of the treatment effects had the same sign for student and farmer samples, and 67% did not statistically differ at the 95% confidence level between the student and farmer samples. We also observed that treatment effects were consistent between student and farmer samples in studies in developing and developed countries, and for each type of experimental methodology sampled.

In all nine studies, the student results at least partly informed the farmer results. Three showed no statistically significant differences between any of the student and farmer treatment effects, five had some but not all treatment effects that were significantly different between the student and farmer samples, and only one study showed significantly different effects for all treatments.

Note that only a third of the studies reviewed were conducted in a high-income country (see appendixes A1 and A2). On the other hand, out of the total studies reviewed, those that could be included for treatment analysis

[4]*There is a large literature on hypothetical bias in choice experiments, which is typically measured as the difference in willingness-to-pay or willingness-to-accept for a good or service under both real and hypothetical payment conditions. Meta-analyses of these literatures have reported mixed conclusions. Murphy et al. (2005) found that students were more likely to exhibit a hypothetical bias than nonstudent samples, while Horowitz and McConnell (2002) and Schläpfer and Fischhoff (2012) both finds no statistical difference in hypothetical bias between student and nonstudent samples.*

[5]*Akay* et al *(2012) also sampled students and farmers but their design did not include any treatment effects.*

Table 2 Summary Statistics and Results for Tests of a Strongly Representative Model Population for Experiments with Farmer Participants

Study	Location	Experiment type	Number of farmers	Number of students	Same country for students and farmers?	Number of treatment effects suitable for analysis	Number of treatment effects statistically different between students and farmers	Number of treatment effects with different signs between students and farmers
Carpenter & Seki (2011)	Japan	Multi-player game	27	26	Y	2	0	1
Castillo et al (2011)	Colombia and Thailand	Multi-player game	120	40	Y	0	NA	NA
Ferré et al (2017)	Switzerland	Multi-player game	282	912	Y	4	0	0
Fooks et al. (2016)	United States	Auction	24	96	Y	0	NA	NA
Herberich & List (2012)	United States	Risk	41	27	Y	3	0	1
Hermann & Mussoff (2016)	Germany	Risk	111	178	Y	4	3	0
Janssen et al (2011)	Colombia and Thailand	Multi-player game	120	99	N	0	NA	NA
Nagler et al (2013)	United States	Market	52	72	Y	4	2	1

(Continues)

Table 2 Continued

Study	Location	Experiment type	Number of farmers	Number of students	Same country for students and farmers?	Number of treatment effects suitable for analysis	Number of treatment effects statistically different between students and farmers	Number of treatment effects with different signs between students and farmers
Palm-Forster et al. (2016)	United States	Auction	51	72	Y	0	NA	NA
Peth & Mussoff (2018)	Germany	Nudge	163	144	Y	6	2	2
Suter & Vossler (2014)	United States	Multi-player game	48	48	Y	2	1	0
Tellez Foster et al (2016)	Mexico	Multi-player game	84	120	N	3	3	0
Waichmann & Ness (2012)	Germany	Market	45	45	Y	8	1	0

were for the most part from a high-income country, with five of them from the United States, three from Germany, and one from Switzerland and Japan each. For those in low- or middle-income countries, not all used student populations from the same country. For example, Janssen et al. (2011) compared irrigation decisions between US students and rural villagers in Colombia and Thailand, and Tellez Foster et al. (2016) compared water management decision-making between US students and Mexican farmers. Farmers in low-income countries face some of the same issues as farmers in the US, although they also face very different issues than farmers in the US. Yet, the small sample of studies collected show that students may have potential to be useful in modeling decision-making for farmers both in the US and abroad.

These findings suggest that pretesting with students could be informative when developing policy-relevant experiments using farmer and rural populations. However, the results draw from a relatively small number of studies, and some of those studies involved relatively small sample sizes. Additionally, many of these studies had mixed results as far as the consistency of treatment effects between student and farmer participants. Thus, it is reasonable to expect that the degree to which student results will be informative of farmer results may also depend on the type of outcome measured.

We offer two recommendations for improving the usefulness of the experimental findings for researchers who integrate student participants into their research protocols:

> Recommendation 3: Use student participants to pretest experimental protocols designed to be used later with farmer and rural populations.

Pretesting can be used to check for heterogeneous treatment effects and to generate estimates of the sampling variability of treatment effects, which can be used later to inform the power calculations for the subsequent experiments conducted with farmer participants. Using our earlier example of an experiment to test crop insurance subsidies on farm capital investment (Example 2), suppose the researchers believed ex-ante that beginning and experienced farmers would make different capital investment decisions in response to a change in premium subsidy. Performing this experiment as a randomized controlled trial would require a very large budget given the large number of farmers required and the expense involved in subsidizing real crop insurance purchases.

A less expensive alternative would be to have farmers play a stylized game that mimicked the decision-making involved in purchasing crop insurance and investing in farm capital investments. In that case, experiments with student participants could be used to (i) test various parameterizations of the game to identify the parameters that best replicate the real world context of the crop insurance/ farm capital investment decision; (ii) induce different levels of operating capital in the game and test for differential treatment effects caused by differences in operating capital; and (iii) test for consistencies between participants' willingness to purchase insurance and/or make business investments within the game and participants' willingness to purchase actual insurance products and/or make actual financial investments. All these tests conducted on student experiments would contribute to establishing the internal validity and degree of parallelism achieved by the stylized game before asking farmers to invest their valuable time to play it. Moreover, the results from the student experiments could provide a starting point

estimate of the likely sampling variability of the outcome variable in power calculations for farmer participants in the event that actual data is not otherwise available.

> Recommendation 4: Plan to use both student and farmer participants in experiments that investigate the behavioral mechanisms that underpin decision-making in policy-relevant experiments.

Policy-relevant experiments typically answer a research question along the lines of "What happens to outcome X when policy Y changes to policy Y′?" In addition, policymakers often want to know what the impacts of changing policy Y will be on other outcomes, and/or the impact of shifting from policy Y to policy Y′. Because it is often prohibitively expensive to study all of these variations in experiments using farmer participants, experiments designed to shed light on the behavioral mechanisms that connect outcome X to policies Y and Y′ can be particularly useful for developing models to predict the effects of other policy changes and on other outcomes of interest.

Returning again to our hypothetical experiment about the effect of crop insurance premium subsidies on farm capital investment (Example 2), suppose that policymakers are also interested in the effect of a changing insurance policies to provide higher premium subsidies on acreage where farmers plant cover crops. In this situation, researchers could benefit from a behavioral model that connects revenue variability to farmers' incentives to purchase crop insurance and make capital investments. The model could then be modified to account for changes in revenue risk associated with use of cover crops and estimate the resulting impact on crop insurance purchases and capital investments.

Experiments with farmer participants would be useful to validate the model's predictions. However, using farmers to validate every change to the behavioral model would likely be costly and take a long time to recruit enough participants. Instead, experiments with student participants could be used to validate the predictions of the behavioral model, with periodic replications conducted using farmer participants. This approach would reduce the burden on farmers for research to improve the behavioral model and build a body of evidence by which to characterize the situations when student experimental results do and do not correspond well with farmer results.

Research Funding Priorities

The final barrier we consider is the difficulty associated with funding a research agenda that supports evidence-based agricultural policymaking. Data from experimental research is part of the toolbox for evidence-based policy that is highly valued by policymakers. Experimental data is particularly useful in instances where other data sources may not capture a full range of responses or when other data sources are not structured to identify causal relationships. However, experiments also require substantial funding to maximize the potential for the results to be useful for policymaking. Research funding is always scarce, and grant reviewers and program officers may seek to maximize the value of available research dollars by prioritizing experiments that address multiple research questions. However, as noted earlier, studies with multiple experimental treatments and involving farmer participants are often underpowered because of the innate heterogeneity of the US

farming population and the small sample sizes typically used in experiments. Proposals to investigate a small number of research questions using methods with strong internal validity can also provide good value for the research dollar, particularly when replicated with multiple samples.

When assessing the merits of a proposed study, reviewers often weigh several criteria including the internal validity of the design, the external validity of the research protocol, and the study's potential to generate novel findings. While all of these factors are important for proposals to conduct agri-economic policy experiments, there are additional considerations specific to policy-relevant experimental research that matter for establishing connections between the experimental results and the policy context the study is meant to inform.

Hallmarks of a research proposal that has good potential to yield useful findings agricultural policymaking include:

1. Evidence of how the experiment's design mimics and deviates from the actual policy setting. Most experimental designs require some deviations from the policy setting in order to control for potential confounding effects. These deviations are likely to be smaller in field experiments than for laboratory experiments, according to the typology of experiments described by Harrison and List (2004). Proposals that articulate the rationale for these deviations and describe a plan for verifying the concordance of the experimental findings with actions taken in the real-world policy setting are more likely to generate policy-relevant findings than proposals that do not consider these issues in the design phase.
2. Evidence of how the experiment's design corresponds to an underlying behavioral model of decision-making in the policy context. Experimental treatments may not necessarily generate statistically significant differences in outcomes. However, an experimental design that provides evidence that the behavioral model has some predictive power in the context of the policy in question can still be useful for informing policymaking—even when the treatments do not change outcomes.
3. Evidence of how the experiment's participants, sample size(s), and recruitment strategies are informed by the policy setting of interest. Practical and budget considerations can preclude researchers from planning to recruit a representative sample of the policy-relevant population for every experiment. A research proposal that relies on a non-representative sample of the target population can still be useful for informing policymaking if the researchers can collect data about heterogeneity of treatment effects within sample. Furthermore, research that relies on a nonrepresentative sample of an alternative model population (such as students) can inform policymaking as part of a broader agenda that builds up to experiments involving the policy-relevant population.

Additionally, replication studies are particularly useful for informing policymaking. In an ideal world, policymakers would be able to compare results from initial experiments with replication studies using multiple samples to provide a robust estimate of treatment effects before they need to make a decision about implementing a policy change in a real program. Even if multiple replications aren't available to inform a policy change, policy environments do evolve continuously over time. Replicating an experiment after a new policy has taken effect can provide evidence of the stability of treatment effects to changes in the policy environment and/or target population.

We offer a recommendation for program officers and grant reviewers to facilitate their reviews of proposals to fund experiments that are designed to inform evidence-based policymaking:

> Recommendation 5: Use an explicit set of criteria to assess the quality of proposals to fund policy-relevant experimental research.

It is a rare experimental study that jointly delivers high internal validity, external validity, and potential for generating novel policy-relevant results at a low cost. Grant reviewers are often placed in the unenviable position of having to rank proposals that offer different levels of each of these factors. Their jobs are made easier when the funding agency assigns relative weights for each of these factors in accordance with the agency's priorities.

Returning to our Example 2, imagine two proposals for this experiment. Proposal 1 has participants play a stylized game that mimics the decision-making involved in the real-world problem. Proposal 2 measures actual capital investments in response to actual changes in crop insurance premium subsidies. Both proposals call for using farmers as participants, but proposal 1 recruits a larger number of farmers and plans to replicate the game in multiple locations across the country. Both proposals request the same level of funding, use methods that are likely to generate strong internal validity, recruit from the same participant pool, provide a novel contribution to the literature, and have the potential to generate policy-relevant results. Without explicit criteria from the funding agency as a guide, reviewers will likely find it difficult to rank the proposals in terms of their potential utility for policymaking. Some funding agencies may prefer the proposal that relies on a stylized game and does not affect the actual program, while others might prefer the proposal that more closely replicates the real-world context of the program. By clearly communicating its priorities, the agency makes it possible for the reviewers to compare diverse proposals.

Like research funds, program funding available for experimentation and evaluation may also be scarce and split among multiple priorities. To help policymakers evaluate the relative merit of incorporating experiments into program activities, we also make a final recommendation for research funding organizations:

> Recommendation 6: Consider funding the development and dissemination of management case studies to document how federal agencies have used agri-economic policy experiments to create, modify, and/or evaluate the impact of federal programs.

Policymakers considering incorporating an agri-economic experiment within a government program may not be aware of prior experience implementing similar experiments in other government agencies. However, sharing journal articles and program reports may not be the most effective way to communicate to policymakers the merits and challenges associated with incorporating agri-economic experiments in programs. Management case studies could be used to document the benefits and costs associated with agri-economic policy experiments, as well as lessons learned, from the perspective of administrators who have already implemented experiments within their programs. Case studies could also give perspective on the time frames, skills, and collaborations required to implement experiments within programs, and discuss any

factors that impacted the extent to which the project achieved the desired policy goals for the experiment. By funding the creation and dissemination of such case studies, research organizations could help support demand for evidence-based policymaking in general, and increase opportunities for government agencies to share their institutional knowledge in ways that benefit researchers and policymakers alike.

Discussion

At times, experimental economics may be the most practical method to test theoretical predictions and estimate agricultural policies before putting them into effect. Risky or costly policies can be tested in the lab first and avoided in the field if these policies fail in a laboratory setting. Alternatively, given success in the lab, the policy may be amended given treatment results. Since experiments can provide otherwise hard to obtain evidence, policymakers can benefit from the application of experimental methods when seeking to achieve evidence-based policy. High-quality and policy-relevant experimental evidence is difficult and costly to obtain. Small changes in the way experiments are designed and funded can provide better insight into the heterogeneity of treatment effects and the underlying behaviors that drive decision-making.

We provide several recommendations. Since heterogeneity in response is common, it is helpful to collect data from respondents on a standard set of demographic and farm characteristics. In addition, the use of stratified or block randomization can better leverage the ability of limited sample sizes to detect differences attributable to respondent characteristics. Pretesting with student participants can be useful, particularly for investigations of behavioral mechanisms underpinning decision-making. The review of proposals is facilitated by having a predetermined explicit set of criteria. Finally, research organizations could consider funding the creation and dissemination of management case studies about government agencies' experiences implementing agri-economic experiments within their programs so that program administrators and researchers can learn from past experiences and share institutional knowledge across agencies.

Disclaimer

The findings and conclusions in this publication are those of the authors and should not be construed to represent any official USDA or US government determination or policy.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Acknowledgments

This research was supported by the U.S. Department of Agriculture, Economic Research Service.

References

Abbink, Klaus, Bernd Irlenbusch, Bettina Rockenbach, Abdolkarim Sadrieh, and Reinhard Selten. 2002. The Behavioural Approach to the Strategic Analysis of Spectrum Auctions: The Case of the German DCS-1800 Auction. *Ifo Studien* 48(3): 457–480.

Arnold, Michael A., Joshua M. Duke, and D. Kent. 2013. Adverse Selection in Reverse Auctions for Environmental Services. *Land Economics* 89(3): 387–412.

Athey, Susan, and Guido W. Imbens. 2017. The econometrics of randomized experiments. In *Handbook of Economic Field Experiments*, Vol 1, ed. Esther Duflo and Abhijit Banerjee, 73–140. Amsterdam: Elsevier.

Ball, Michael, George Donohue, and Karla Hoffman. 2006. Auctions for the Safe, Efficient, and Equitable Allocation of Airspace System Resources. In *Combinatorial Auctions*, ed. Peter Cramton, Yoav Shoham, and Richard Steinberg, 507–538. Cambridge, MA: MIT Press.

Banerjee, Simanti. 2018. Improving Spatial Coordination Rates under the Agglomeration Bonus Scheme: A Laboratory Experiment with a Pecuniary and a Non-pecuniary Mechanism (Nudge). *American Journal of Agricultural Economics* 100(1): 172–197.

Banerjee, Simanti, Frans P. De Vries, Nick Hanley, and Daan P. van Soest. 2014. The Impact of Information Provision on Agglomeration Bonus Performance: An Experimental Study on Local Networks. *American Journal of Agricultural Economics* 96(4): 1009–1029.

Banks, Jeffrey, Mark Olson, David Porter, Stephen Rassenti, and Vernon Smith. 2003. Theory, Experiment and the Federal Communications Commission Spectrum Auctions. *Journal of Economic Behavior & Organization* 51(3): 303–350.

Bareinboim, Elias, and Judea Pearl. 2016. Causal Inference and the Data-Fusion Problem. *Proceedings of the National Academy of Sciences of the United States of America* 113 (27): 7345–7352.

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

Binmore, Ken, and Paul Klemperer. 2002. The Biggest Auction Ever: The Sale of the British 3G Telecom Licences. *The Economic Journal* 112(478): C74–C96.

Brick, Kerri, and Martine Visser. 2015. Risk Preferences, Technology Adoption and Insurance Uptake: A Framed Experiment. *Journal of Economic Behavior & Organization* 118: 383–396.

Brookes, Sara T., Elise Whitely, Matthias Egger, George Davey Smith, Paul A. Mulheran, and Tim J. Peters. 2004. Subgroup Analyses in Randomized Trials: Risks of Subgroup-Specific Analyses: Power and Sample Size for the Interaction Test. *Journal of Clinical Epidemiology* 57(3): 229–236.

Brunner, Christoph, Jacob K. Goeree, Charles A. Holt, and John O. Ledyard. 2010. An Experimental Test of Flexible Combinatorial Spectrum Auction Formats. *American Economic Journal: Microeconomics* 2(1): 39–57.

Butler, Julianna M., Jacob R. Fooks, Kent D. Messer, and Leah H. Palm-Forster. 2020. Addressing Social Dilemmas with Mascots, Information, and Graphics. *Economic Inquiry* 58(1): 150–168.

Camerer, Colin 2011. *The Promise and Success of Lab-Field Generalizability in Experimental Economics: A Critical Reply to Levitt and List.* Available at SSRN 1977749.

Cason, Timothy N., and Frans P. de Vries. 2019. Dynamic Efficiency in Experimental Emissions Trading Markets with Investment Uncertainty. *Environmental and Resource Economics* 73(1): 1–31.

Cason, Timothy N., and Lata Gangadharan. 2004. Auction Design for Voluntary Conservation Programs. *American Journal of Agricultural Economics* 86(5): 1211–1217.

——. 2011. Price Discovery and Intermediation in Linked Emissions Trading Markets: A Laboratory Study. *Ecological Economics* 70(7): 1424–1433.

Cason, Timothy N., and Charles R. Plott. 1996. EPA's New Emissions Trading Mechanism: A Laboratory Evaluation. *Journal of Environmental Economics and Management* 30(2): 133–160.

Carpenter, J., & Seki, E. (2011). Do social preferences increase productivity? Field experimental evidence from fishermen in Toyama Bay. *Economic Inquiry*, 49(2), 612–630.

Cummings, Ronald G., Charles A. Holt, and Susan K. Laury. 2004. Using Laboratory Experiments for Policymaking: An Example from the Georgia Irrigation Reduction Auction. *Journal of Policy Analysis and Management* 23(2): 341–363.

Drugov, Mikhail, John Hamman, and Danila Serra. 2014. Intermediaries in Corruption: An Experiment. *Experimental Economics* 17(1): 78–99.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2007. Using Randomization in Development Economics Research: A Toolkit. In *Handbook of Development Economics*, Vol 4, ed. T. Paul Schultz and John A. Strauss, 3895–3962.

Duke, Joshua M., Kent D. Messer, Lori Lynch, and Tongzhe Li. 2017. The Effect of Information on Discriminatory-Price and Uniform-Price Reverse Auction Efficiency: An Experimental Economics Study of the Purchase of Ecosystem Services. *Strategic Behavior and the Environment* 7(1–2): 41–71.

Duvendack, Maren, Richard Palmer-Jones, and W. Robert Reed. 2017. What Is Meant by" Replication" and why Does it Encounter Resistance in Economics? *American Economic Review* 107(5): 46–51.

Fooks, Jacob R., Nathaniel Higgins, Kent D. Messer, Joshua M. Duke, Daniel Hellerstein, and Lori Lynch. 2016. Conserving Spatially Explicit Benefits in Ecosystem Service Markets: Experimental Tests of Network Bonuses and Spatial Targeting. *American Journal of Agricultural Economics* 98(2): 468–488.

Gächter, Simon. 2009. Improvements and Future Challenges for the Research infrastructure in the Field 'Experimental Economics'. Working Paper No. 56, German Data Forum (RatSWD).

Gelman, Andrew, and John Carlin. 2014. Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science* 9(6): 641–651.

Guala, Francesco. 2001. Building Economic Machines: The FCC Auctions. *Studies in History and Philosophy of Science Part A* 32(3): 453–477.

Harrison, Glenn W., and John A. List. 2004. Field Experiments. *Journal of Economic Literature* 42(4): 1009–1055.

Heckman, James J., and Edward Vytlacil. 2001. Policy-Relevant Treatment Effects. *American Economic Review* 91(2): 107–111.

Herberich, D. H., & List, J. A. (2012). Digging into background risk: experiments with farmers and students. *American Journal of Agricultural Economics*, 94(2), 457–463.

Hellerstein, Daniel, and Nathaniel Alan Higgins. 2010. The Effective Use of Limited Information: Do Bid Maximums Reduce Procurement Cost in Asymmetric Auctions? *Agricultural and Resource Economics Review* 39(2): 288–304.

Hellerstein, Daniel, Nathaniel Alan Higgins, and Michael Roberts 2015. Options for Improving Conservation Programs: Insights from Auction Theory and Economic Experiments. Amber Waves, February.

Hellerstein, Daniel, Nathaniel Higgins, and John Horowitz. 2013. The Predictive Power of Risk Preference Measures for Farming Decisions. *European Review of Agricultural Economics* 40(5): 807–833.

Hermann, D., & Musshoff, O. (2016). Measuring time preferences: Comparing methods and evaluating the magnitude effect. *Journal of Behavioral and Experimental Economics*, 65, 16–26.

Horowitz, J. K., & McConnell, K. E. (2002). A review of WTA/WTP studies. *Journal of environmental economics and Management*, 44(3), 426–447.

Higgins, Nathaniel, Daniel Hellerstein, Steven Wallander, and Lori Lynch. 2017. *Economic Experiments for Policy Analysis and Program Design: A Guide for Agricultural Decisionmakers*. No. 1477–2017-4104.

Hjort, Jonas, Diana Moreira, Gautam Rao, and Juan Francisco Santini. 2019. How Research Affects Policy: Experimental Evidence from 2,150 Brazilian Municipalities. NBER Working Paper No. 25941.

Janssen, M. A., Anderies, J. M., & Cardenas, J. C. (2011). Head-enders as stationary bandits in asymmetric commons: Comparing irrigation experiments in the laboratory and the field. *Ecological Economics*, 70(9), 1590–1598.

Johansson, R., Effland, A., & Coble, K. (2017). Falling response rates to USDA crop surveys: Why it matters. University of Illinois Farmdoc Daily, 7.

Kessler, Judd B., and Alvin E. Roth. 2012. Organ Allocation Policy and the Decision to Donate. *American Economic Review* 102(5): 2018–2047.

King, William R., and Jun He. 2006. A Meta-Analysis of the Technology Acceptance Model. *Information & Management* 43(6): 740–755.

Klemperer, Paul. 2002. What Really Matters in Auction Design. *Journal of Economic Perspectives* 16(1): 169–189.

Larrick, Richard P., and Jack B. Soll. 2008. The MPG Illusion. *Science.* 320(5883): 1593–1594.

Lei, Vivian, Charles N. Noussair, and Charles R. Plott. 2000. A Market-Based Mechanism for Allocating Space Shuttle Secondary Payload Priority. *Experimental Economics* 2(3): 173–195.

——. 2001. Nonspeculative Bubbles in Experimental Asset Markets: Lack of Common Knowledge of Rationality Vs. Actual Irrationality. *Econometrica* 69(4): 831–859.

Lesko, Catherine R., Ashley L. Buchanan, Daniel Westreich, Jessie K. Edwards, Michael G. Hudgens, and Stephen R. Cole. 2017. Generalizing Study Results: A Potential Outcomes Perspective. *Epidemiology (Cambridge, Mass.)* 28(4): 553.

Levitt, Steven D., and John A. List. 2007a. On the Generalizability of Lab Behaviour to the Field. *Canadian Journal of Economics/Revue Canadienne d'économique* 40(2): 347–370.

——. 2007b. What Do Laboratory Experiments Measuring Social Preferences Reveal about the Real World? *Journal of Economic Perspectives* 21(2): 153–174.

Li, Jingyuan, Holly A. Michael, Joshua M. Duke, Kent D. Messer, and Jordan F. Suter. 2014. Behavioral Response to Contamination Risk Information in a Spatially Explicit Groundwater Environment: Experimental Evidence. *Water Resources Research* 50(8): 6390–6405.

Liu, Elaine M. 2013. Time to Change What to Sow: Risk Preferences and Technology Adoption Decisions of Cotton Farmers in China. *Review of Economics and Statistics* 95(4): 1386–1403.

Liu, Elaine M., and JiKun Huang. 2013. Risk Preferences and Pesticide Use by Cotton Farmers in China. *Journal of Development Economics* 103: 202–215.

Lunn, Pete, Marek Bohacek, and Alicia Rybicki. 2016. An Experimental Investigation of Personal Loan Choices. Economic and Social Research Institute (ESRI), https://www.esri.ie/system/files?file=media/file-uploads/2016-07/BKMNEXT314.pdf (accessed September 21, 2020).

Lusk, Jayson L., and Jason F. Shogren. 2007. *Experimental Auctions: Methods and Applications in Economic and Marketing Research*. New York: Cambridge University Press.

McCarthy, Jaki, and Daniel G. Beckler. 2000. *Survey Burden and its Impact on Attitudes Toward the Survey Sponsor*. No. 1496–2016-130653.

Messer, Kent D., Joshua M. Duke, and Lori Lynch. 2014. Applying Experimental Economics to Land Economics: Public Information and Auction Efficiency in Land Preservation Markets. In *Oxford Handbook of Land Economics*, ed. Joshua M. Duke and JunJie Wu, 481–546. Oxford, UK: Oxford University Press.

Messer, Kent D., Joshua M. Duke, Lori Lynch, and Tongzhe Li. 2017. When Does Public Information Undermine the Effectiveness of Reverse Auctions for the Purchase of Ecosystem Services? *Ecological Economics* 134: 212–226.

Messer, Kent D., Harry M. Kaiser, and William D. Schulze. 2008. The Problem of Free Riding in Voluntary Generic Advertising: Parallelism and Possible Solutions from the Lab. *American Journal of Agricultural Economics* 90(2): 540–552.

Messer, Kent D., Todd M. Schmit, and Harry M. Kaiser. 2005. Optimal Institution Designs for Generic Advertising: An Experimental Analysis. *American Journal of Agricultural Economics* 87(4): 1046–1060.

Miao, Haoran, Jacob R. Fooks, Todd Guilfoos, Kent D. Messer, Soni M. Pradhanang, Jordan F. Suter, Simona Trandafir, and Emi Uchida. 2016. The Impact of Information on Behavior under an Ambient-Based Policy for Regulating Nonpoint Source Pollution. *Water Resources Research* 52: 3294–3308.

Miratrix, Luke W., Jasjeet S. Sekhon, and Bin Yu. 2013. Adjusting Treatment Effect Estimates by Post-Stratification in Randomized Experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75(2): 369–396.

Muller, Sean (2014). Randomised Trials for Policy: A Review of the External Validity of Treatment Effects.

Murphy, James J., P. Geoffrey Allen, Thomas H. Stevens, and Darryl Weatherhead. 2005. A Meta-Analysis of Hypothetical Bias in Stated Preference Valuation. *Environmental and Resource Economics* 30(3): 313–325.

Nagler, A. M., Menkhaus, D. J., Bastian, C. T., Ehmke, M. D., & Coatney, K. T. (2013). Subsidy incidence in factor markets: An experimental approach. *Journal of Agricultural and Applied Economics*, 45(1), 17–33.

Normann, Hans-Theo, and Roberto Ricciuti. 2009. Laboratory Experiments for Economic Policy Making. *Journal of Economic Surveys* 23(3): 407–432.

Palm-Forster, L. H., Ferraro, P. J., Janusch, N., Vossler, C. A., & Messer, K. D. (2019). Behavioral and experimental agri-environmental research: methodological challenges, literature gaps, and recommendations. *Environmental and resource economics*, 73(3), 719–742.

Palm-Forster, Leah H., Paul J. Ferraro, Nicholas Janusch, Christian A. Vossler, and Kent D. Messer. 2019a. Behavioral and Experimental Agri-Environmental Research: Methodological Challenges, Literature Gaps, and Recommendations. *Environmental and Resource Economics* 73(3): 719–742.

Palm-Forster, Leah H., Jordan F. Suter, and Kent D. Messer. 2019b. Experimental Evidence on Policy Approaches that Link Agricultural Subsidies to Water Quality Outcomes. *American Journal of Agricultural Economics* 101(1): 109–133.

Palm-Forster, Leah H., Scott M. Swinton, Todd M. Redder, Joseph V. DePinto, and Chelsie M.W. Boles. 2016. Using Conservation Auctions Informed by Environmental Performance Models to Reduce Agricultural Nutrient Flows into Lake Erie. *Journal of Great Lakes Research* 42(6): 1357–1371.

Parkhurst, Gregory M., Jason F. Shogren, Chris Bastian, Paul Kivi, Jennifer Donner, and Rodney B.W. Smith. 2002. Agglomeration Bonus: An Incentive Mechanism to Reunite Fragmented Habitat for Biodiversity Conservation. *Ecological Economics* 41 (2): 305–328.

Perkis, D. F., Cason, T. N., & Tyner, W. E. (2016). An experimental investigation of hard and soft price ceilings in emissions permit markets. *Environmental and Resource Economics*, 63(4), 703–718.

Pearl, Judea, and Elias Bareinboim. 2014. External Validity: From Do-Calculus to Transportability across Populations. *Statistical Science*: 579–595.

Pete, Lunn. 2014. *Regulatory Policy and Behavioural Economics*. Paris: OeCD Publishing.

Peth, D. and O. Mussoff (2018). *Comparing Compliance Behavior of Students and Farmers: Implications for Agricultural Policy Impact Analysis. Diskussionsbeitrag, No. 1809*, Georg-August-Universität Göttingen, Department für Agrarökonomie und Rurale Entwicklung (DARE), Göttingen.

Plott, Charles R. 1987. Dimensions of Parallelism: Some Policy Applications of Experimental Methods. In *Laboratory Experimentation in Economics: Six Points of View*, ed. Alvin Roth, 193–219. New York, NY: Cambridge University Press.

Poe, Gregory L., William D. Schulze, Kathleen Segerson, Jordan F. Suter, and Christian A. Vossler. 2004. Exploring the Performance of Ambient-Based Policy Instruments when Nonpoint Source Polluters Can Cooperate. *American Journal of Agricultural Economics* 86(5): 1203–1210.

Schläpfer, Felix, and Baruch Fischhoff. 2012. Task Familiarity and Contextual Cues Predict Hypothetical Bias in a Meta-Analysis of Stated Preference Studies. *Ecological Economics* 81: 44–47.

Smith, Vernon L. 1982. Microeconomic Systems as an Experimental Science. *The American Economic Review* 72(5): 923–955.

Spraggon, John. 2004. Testing Ambient Pollution Instruments with Heterogeneous Agents. *Journal of Environmental Economics and Management* 48(2): 837–856.

Spraggon, John M. 2013. The Impact of Information and Cost Heterogeneity on Firm Behaviour under an Ambient Tax/Subsidy Instrument. *Journal of Environmental Management* 122: 137–143.

Suter, Jordan F., Sam Collie, Kent D. Messer, Joshua M. Duke, and Holly A. Michael. 2019. Common Pool Resource Management at the Extensive and Intensive Margins: Experimental Evidence. *Environmental and Resource Economics* 73(4): 973–993.

Suter, Jordan F., Joshua M. Duke, Kent D. Messer, and Holly A. Michael. 2012. Behavior in a Spatially Explicit Groundwater Resource: Evidence from the Lab. *American Journal of Agricultural Economics* 94(5): 1094–1112.

Suter, Jordan F., Kathleen Segerson, Christian A. Vossler, and Gregory L. Poe. 2010. Voluntary-Threat Approaches to Reduce Ambient Water Pollution. *American Journal of Agricultural Economics* 92(4): 1195–1213.

Suter, Jordan F., Christian A. Vossler, and Gregory L. Poe. 2009. Ambient-Based Pollution Mechanisms: A Comparison of Homogeneous and Heterogeneous Groups of Emitters. *Ecological Economics* 68(6): 1883–1892.

Suter, Jordan F., Christian A. Vossler, Gregory L. Poe, and Kathleen Segerson. 2008. Experiments on Damage-Based Ambient Taxes for Nonpoint Source Polluters. *American Journal of Agricultural Economics* 90(1): 86–102.

Suter, J.F., and C.A. Vossler. 2014. Towards an Understanding of the Performance of Ambient Tax Mechanisms in the Field: Evidence from Upstate New York Dairy Farmers. *American Journal of Agricultural Economics* 96: 92–107.

Tellez Foster, E., Rapoport, A., and Dinar, A., 2016. Alternative policies to manage electricity subsidies for groundwater extraction: a field study in Mexico. UCR SPP Working Paper Series, July 2016, WP# 16-05Torero

U.S. Securities and Exchange Commission Office of Economic Analysis. 2007. Economic Analysis of the Short Sale Price Restrictions under the Regulation SHO Pilot. Securities and Exchange Commission Working Paper.

Waichman, I., Ness, C., 2012. Farmers' performance and subject pool effect in decentralized bargaining markets. *Economics Letters*. 115, 366–368.

Weigel, Collin, Laura Paul, Paul Ferraro, and Kent Messer. 2020. Challenges in Recruiting U.S. Farmers for Policy-Relevant Economic Field Experiments. *Applied Economics Policy & Perspectives* (this issue).