CHAPTER

Experimental and behavioral economics to inform agri-environmental programs and policies

80

Leah H. Palm-Forster* **and Kent D. Messer**
Department of Applied Economics and Statistics, University of Delaware,
Newark, DE, United States
**Corresponding author: e-mail address: leahhp@udel.edu*

Chapter outline

Handbook of Agricultural Economics, Volume 5, ISSN 1574-0072, https://doi.org/10.1016/bs.hesagr.2021.10.006

1 Introduction

Experimental and behavioral economics research has challenged and improved how economists think about people and their actions. By recognizing and testing myriad factors that drive behavior beyond neo-classical economic assumptions about rationality, these fields have contributed to a richer understanding of human decision-making. Furthermore, by using experimental methods that rely on randomly assigned controls and treatments, researchers have identified the marginal effects of exogenous factors such as design attributes of programs and policies aimed at improving social wellbeing and environmental conditions.

The expansion of our understanding of human behavior made possible by experimental and behavioral economics has been widely recognized. In 2019, the Nobel Memorial Prize in Economic Sciences was bestowed on three development economists—Abhijit Banerjee, Esther Duflo, and Michael Kremer—who use randomized controlled trials (RCTs, a type of field experiment) to examine programs and policies that can improve lives in poor communities, often in the context of developing countries. In 2017, the Nobel Prize went to Richard Thaler for his broad contributions to behavioral economics. These winners built on the work of several earlier behavioral and experimental economists, including 2002 Nobel Prize recipients Daniel Kahneman and Vernon Smith.

Agricultural and applied economists use experimental and behavioral economics approaches to analyze a variety of decisions of consumers and producers, including decisions that affect the environment. Economic experiments have been instrumental for examinations of consumer behavior, including consumer demand for environmentally friendly products (see, for instance, Loureiro, McCluskey, and Mittelhammer (2002), Teisl, Roe, and Hicks (2002), Lagerkvist and Hess (2011), Kecinski, Messer, and Peo (2018), and Savchenko, Kecinski, Li, Messer, and Xu (2018)), willingness to pay for green infrastructure provisions (Ellis, Fooks, Messer, & Miller, 2016), and engagement in pro-environmental behaviors (Byerly et al., 2018). Additionally, insights from behavioral economics have contributed to our nuanced understanding of consumer decision-making broadly (Foxall, 2017) and of decisions with environmental implications (Brown & Hagen, 2010). Certainly, more research is needed in this area; however, given the relatively large number of existing consumer studies of pro-environmental decision-making and existing high-quality reviews of research applying experiments to consumer behavior and food policy (Canavari, Drichoutis, Lusk, & Nayga, 2019; Just & Byrne, 2020), we focus this chapter primarily on issues related to conducting experiments on the supply side of agri-environmental programs. The literature examining producer behavior related to environmental decision-making is growing but remains small compared to the plethora of consumer research overall. Furthermore, while many insights have been applied to consumer behavior, less evidence has been developed about whether these insights will lead to the same behavioral changes in producer behavior.

Why is it critical to improve our understanding of decision-making related to agri-environmental issues? Human-engineered agricultural landscapes comprise about half of the world's habitable land area (Ritchie & Roser, 2013), and management of those landscapes has a profound impact on natural resources and provision of ecosystem services at multiple scales. Water resources are particularly affected by agriculture. For instance, while water quality in the United States has been improved over the past 50 years by reducing industrial (point source) pollution, diffuse (nonpoint source) pollution remains a major concern and agriculture is a leading contributor of it. Approximately 53% of the miles of rivers and streams and 70% of the acres of lakes, ponds, and reservoirs are listed as "impaired" and as not meeting designated use standards (U.S. Environmental Protection Agency (EPA), 2018).

Agri-environmental programs and policies are used to improve the quality of water, air, soil, and other natural resources by promoting management practices that mitigate the negative effects (externalities) of production and enhance provision of ecosystem services in agricultural landscapes. These practices are often referred to as conservation practices or best management practices (BMPs). In the United States, agri-environmental programs and policies tend to use monetary "carrots" rather than regulatory "sticks" to encourage BMP adoption by land managers (Ribaudo, 2015).

At the federal level, the U.S. Department of Agriculture (USDA) spends more than $6 billion annually on voluntary conservation programs that offer payments for ecosystem services (PES) to offset producers' costs of adopting BMPs. Well-designed agri-environmental programs can be effective in protecting and restoring environmental resources. Also important are programs that target consumer behavior and choices since consumer demand applies pressure to various points in the supply chain and can ultimately change which products are supplied and how food and fiber products are produced (Khanna, Swinton, & Messer, 2018; Waldman & Kerr, 2014).

Achieving environmental improvements with limited funding is a key challenge for agri-environmental programs (Duke, Dundas, & Messer, 2013; Messer & Allen, 2018). Consequently, many program managers are interested in applying insights from behavioral and experimental economics so they can improve how programs are designed to increase the cost-effectiveness of their efforts (Higgins, Hellerstein, Wallander, & Lynch, 2017). The evidence supporting behavioral economics insights for individual and consumer behavior is extensive, as is application of these insights to improve policy and program effectiveness (see, for instance, Chetty, 2015; Dellavigna, 2009; Madrian, 2014). However, studies that test how behavioral insights can be used to improve program performance with producers in large-scale agri-environmental programs are rare and represent a critical gap in the literature (see Palm-Forster, Ferraro, Janusch, Vossler, & Messer, 2019).

Using experimental economics to inform U.S. agri-environmental policies and programs can also be beneficial when bringing federal agri-environmental programs into compliance with the Foundations of Evidence-based Policymaking Act of 2018 (known as the Evidence Act). The Evidence Act requires federal agencies to develop agency "learning agendas" that assist them in identifying key questions related to their programs and evidence-based approaches to answer those questions (Abraham et al., 2017). To date, USDA's use of evidence from economic experiments and randomized field experiments is lagging behind efforts of other federal agencies such as the Department of Health and Human Services and the Department of Education. It is difficult to imagine the federal effort to fight COVID-19 proceeding without careful experimental protocols and testing of new vaccines. Likewise, it is nearly impossible to imagine the many benefits derived from modern agricultural seeds in terms of crop yields without careful field experiments. In many domains, the power of careful experimentation is being harnessed to improve societal outcomes, and it is vital to bring this power to agri-environmental program and policy contexts as well. Fortunately, the Evidence Act calls for all federal agencies in the United States to take this approach with their programs to ensure their effectiveness in delivering the desired objectives and their cost-effective use of taxpayer funds that support the programs.

Though agricultural and applied economists are increasingly using experimental and behavioral economics approaches to study agri-environmental issues, the amount of research conducted in this area is quite thin compared to such studies of topics in education, finance, health, and pro-environmental behaviors such as energy and water conservation. It is likely that some behavioral insights from other fields will apply to agri-environmental issues while others will not because of characteristics unique to those contexts.

In particular, agri-environmental programs and policies often aim to change long-term management decisions affecting production of impure public goods. The model of impure public goods reflects actions that produce both private and public goods (Cornes & Sandler, 1994), which is often true of agri-environmental decisions. No-till agricultural practices, for example, can reduce sediment loss, which improves water quality (a public good), and improve soil structure and infiltration, which improve crop yields (private goods). Such joint production of private *and* public benefits changes the incentives for agricultural decision-makers relative to decisions that solely affect private *or* public goods. To date, the broader behavioral science literature has mostly addressed contexts that involve purely private or purely public goods so it is not clear whether insights from that literature can be directly applied to the design of agri-environmental programs and policies. This is a critical gap in our understanding that needs to be closed.[1]

A primary appeal of using experimental approaches to evaluate and inform agri-environmental programs and policies is the ability to draw strong inferences about *causal relationships between interventions and outcomes*, making them attractive to policymakers. For example, behavioral and experimental research has suggested that agri-environmental programs can be more cost-effective by using reverse auctions to distribute funds, allowing administrators to target often-limited resources to areas where they will have the greatest impact and streamlining programs to reduce farmers' transaction costs and increase participation (Ferraro, 2008; Fooks et al., 2016; Palm-Forster, Swinton, Lupi, & Shupp, 2016; Schilizzi, 2017). Experiments have also shown that screening mechanisms can increase the cost-effectiveness of conservation programs by reducing adverse selection (Arnold, Duke, & Messer, 2013) and that greater communication improves program outcomes (Banerjee, Cason, de Vries, & Hanley, 2017). In Section 2, we provide a more-detailed summary of how experiments have been used to test economic mechanisms that can inform the design of agri-environmental policies and programs.

A second application of experimental economics research lies in using experiments to *test the effects of low-cost behavioral "nudges"* aimed at increasing participation in agri-environmental programs and enhancing ecosystem services cost-effectively. For example, Ferraro, Messer, Shukla, and Weigel (2021) found that changing the default bid level in a reverse auction could reduce the cost-share amount farmers requested from the auctioneer for adopting BMPs. Additionally, producers were more likely to participate in the program when they were given information about social norms suggesting that other farmers valued the practices being promoted. Experimental studies have also investigated the role of messengers (Butler, Fooks, Messer, & Palm-Forster, 2020), features of incentives such as the timing of payments

[1] Likewise, most behavioral economics studies have focused on individual decision-making, while decisions in agri-environmental contexts often are made by groups of decisionmakers, including non-operating landowners and multiple decisionmakers within family farms, corporate farming operations, and cooperatives.

(Duquette, Higgins, & Horowitz, 2012), norms and social comparisons (Banerjee, 2018; Wallander, Ferraro, & Higgins, 2017; Wu, Palm-Forster, & Messer, 2021), and salience, priming, and affect (Czap, Czap, Khachaturyan, Burbach, & Lynne, 2013; Wallander et al., 2017). Laboratory experiments have investigated how recognizing pro-environmental behavior (akin to recognition through agri-environmental stewardship awards) and shaming of pollution behavior affect use of pollution-reducing technologies and levels of ambient pollution (Butler et al., 2020; Palm-Forster, Griesinger, Butler, Fooks, & Messer, forthcoming).

This chapter presents a guide for designing and conducting economic experiments related to agriculture and the environment. We begin by highlighting relevant insights from behavioral economics, describing how economic experiments have been integral to testing and informing our understanding of human behavior. We then describe four key types of experiments that we think of as experiment stages (though our use of the term "stages" does not imply that all four stages of experimentation must be completed to generate valuable research contributions). For each experiment type, we discuss trade-offs required in terms of control, context, and representativeness and the key questions raised in terms of internal and external validity. These considerations assist researchers in identifying the most appropriate type of experiment (or sequence of complementary experiments) for their studies. The chapter emphasizes five contemporary issues and related best practices that researchers should consider when conducting experimental economics research: replicability, underpowered designs, publication bias, participant recruitment, and detection of heterogeneous treatment effects.[2] We also discuss important ethical considerations to consider when designing and conducting economic experiments and engaging with rural communities. In addition to providing practical guidance for researchers, we outline key recommendations for editors, reviewers, and funders to strengthen the quality of future experimental and behavioral research. We conclude the chapter by presenting a framework for prioritizing research in the face of serious time and resource constraints, and we offer advice for junior researchers who are beginning to build their research programs.

2 Behavioral insights and experimental applications

Behavioral economics is the study of why people make the decisions they do. It both challenges and extends traditional economic assumptions about rationality by examining psychological, cognitive, social, and other related factors that influence decision-making. By recognizing and studying the complexity of people's beliefs and deciphering the motivations for their actions, behavioral economists can identify

[2] Using appropriate methods to analyze and interpret data from experiments is another important topic; however, offering guidance on analyzing experimental data is beyond the scope of this chapter. For more information on statistical and econometric methods for analyzing such data, see recent resources on this topic, including a handbook chapter by Athey and Imbens (2017) and discussion about non-parametric testing approaches (Feltovich, 2003).

and explain numerous behaviors seen in practice. At times, those behaviors are consistent with the predictions of neo-classical economic theory, but in many cases they are not. Self-interest continues to be viewed as a dominant influence on human behavior, but other factors have been identified that consistently influence behavior, such as altruism, social norms, risk and time preferences, deliberate democratic processes, and biases and mental heuristics related to anchoring of beliefs, conformity, default behavior, and reference points.

Economic experiments are critical when determining whether insights from behavioral economics, often derived from theory, map to observed behaviors. In his review of Dhami's (2016) textbook, *The Foundations of Behavioral Economic Analysis*, Tyran (2017, p. 161) commented that "the general perspective of the book is that science prospers in a fruitful dialogue between theory and empirics." This sentiment perfectly describes the reciprocal value of theoretical and experimental research. Tyran (2017) went on to emphasize that this viewpoint is common among experimental economists who use controls and treatments to test theories and investigate behavioral alternatives, especially when evidence from experiments contradicts theory. This marriage of behavioral and experimental economics has provided a solid foundation for explaining inconsistent behavior and rigorously investigating alternative models of behavior.

Recognizing that existing theory sometimes provides limited policy guidance—especially in complex settings, Shogren (2004, p. 1218) suggested that "like a wind tunnel to test airplane design, lab experiments provide a testbed for what is called *economic design*—the process of constructing institutions and mechanisms to examine efficient resource allocation." As we describe in Section 3, using experiments as testbeds is particularly useful in settings in which implementation of a policy change is difficult or costly; testing alternative policies in the laboratory can provide critical insights and be highly cost-effective.

2.1 Insights from behavioral economics

Behavioral economists have been making important contributions to identifying the myriad factors that influence choices people make in various contexts. Their contributions are beginning to be applied to the intersections of agriculture, food, and the environment (Dessart, Barreiro-Hurlé, & van Bavel, 2019; Palm-Forster, Ferraro, et al., 2019; Streletskaya et al., 2020). It has become clear that neo-classical models of profit-maximization cannot adequately explain decisions made by all producers because such models are overly simplistic representations of complex thought and behavior patterns.

Another important development is the discovery that factors that influence human decisions can be used to alter the decision-making environment and influence behavior in predictable ways. This concept of "nudging" behavior was popularized by Thaler and Sunstein (2008). Nudges come in many different forms and change decision-making environments (the choice architecture) in different ways. Dolan et al. (2012) established an acronym, MINDSPACE, to categorize types of nudges used to affect behavior in various settings: messengers, incentives, norms, defaults,

Table 1 The MINDSPACE framework for behavioral change.

Cue	Behavior
Messenger	We are heavily influenced by who communicates information to us
Incentives	Our responses to incentives are shaped by predictable mental shortcuts such as strongly avoiding losses
Norms	We are strongly influenced by what others do
Defaults	We "go with the flow" of pre-set options
Salience	Our attention is drawn to what is novel and seems relevant to us
Priming	Our acts are often influenced by subconscious cues
Affect	Our emotional associations can powerfully shape our actions
Commitment	We seek to be consistent with our public promises and reciprocate acts
Ego	We act in ways that make us feel better about ourselves

Reprinted from Dolan, P., Hallsworth, M., Halpern, D., King, D., Metcalfe, R., & Vlaev, I. (2012). Influencing behaviour: The mindspace way. Journal of Economic Psychology, 33*(1), 266, Copyright (2012), with permission from Elsevier.*

salience, priming, affect, commitment, and ego (see Table 1). For more than a decade, agricultural economists have been studying the application of nudges to influence agri-environmental decisions, but this literature remains relatively thin, especially compared to the application of nudges in other contexts.

Economists are keenly interested in the role played by behavioral factors when agricultural producers make decisions, including management and input choices and whether to participate in agri-environmental programs. Several research teams have produced excellent reviews that summarize how behavioral economics has improved our understanding of farmer behavior and informed the design of agri-environmental programs and policies. Dessart et al. (2019) review findings from policy-related behavioral economics studies of voluntary adoption of sustainable farming practices. The authors develop a framework in which they identify how three types of factors affect farmer decision-making: dispositional factors (e.g., personality, values, beliefs, and preferences), social factors (e.g., social norms and signaling motives), and cognitive factors (e.g., perceptions of benefits, costs, and risks). The review by Streletskaya et al. (2020) highlights synergies between studies of agricultural technology adoption and behavioral economics. They identify three thematic areas that hold promise for cross-fertilization between those fields: behavior in the face of risk and deviations from expected utility, models of learning and social preferences, and behavioral time discounting. The authors argue that researchers can generate more-robust evidence of what does and does not work in a program by incorporating behavioral factors into their analyses. Palm-Forster, Ferraro, et al.'s (2019) review of research to test the effectiveness of nudges in influencing landowner behavior uses Dolan et al.'s (2012) MINDSPACE framework (Table 1) and examines how each nudge category can be applied to agri-environmental decisions. They also highlight gaps in the literature, outline methodological challenges, and make recommendations to promote more-robust research in this area.

Decision-making under risk has been identified as a factor that has clear implications for adoption of agricultural practices (Dessart et al., 2019; Streletskaya et al., 2020). Agricultural decisions often involve considerable risk and uncertainty. Producers must routinely choose, for example, what crops to produce and in what quantities (acreage) based on limited information about likely weather patterns and market outcomes in the upcoming season. Farmers are generally considered to be risk-averse, and that aversion can limit their willingness to adopt new, environmentally beneficial practices that deviate from conventional ones. Studies have shown that farmers, like most people, often strongly weigh small probabilities of loss (Bocquého, Jacquet, & Reynaud, 2014). Consequently, agri-environmental programs designed to reduce risk and minimize income volatility are likely to be popular with producers, motivating them to try new practices and approaches. Additionally, programs can be designed to promote incremental changes that allow for trial and error to get a farmer's "foot in the door" and acknowledge that small changes can lead to larger ones (Dessart et al., 2019). Programs can initially request behavioral changes from farmers who are relatively receptive to trying new practices, and outreach can be focused during times when farmers are more open to change and have greater bandwidth for considering new practices (e.g., after the growing season when farmers are planning for the following year).

Behavioral economic research has shown that people who already have pro-environmental values and are committed to pro-environmental behavior are more likely to engage in additional pro-environmental actions (Gosnell, 2018; Whitmarsh & O'Neill, 2010)—the additional actions align with the ethic with which they identify. Actions motivated by factors such as values and preferences can be influenced by nudges that relate to ego, commitment, and affect (as defined by Dolan et al., 2012). Humans seek to make decisions that are consistent with the way they view themselves in an effort to reinforce their self-identifies and egos. As a result, encouraging small changes in behavior that affect how people view themselves can lead them to make more-significant changes in the future. Ego nudges also can be linked to commitment nudges by encouraging people to make promises that align with their self-images and to follow through on those promises using a combination of intrinsic and extrinsic pressure. Requests for private commitments are effective with people who want to behave in a way that is consistent with their initial intentions and plans (Baca-Motes, Brown, Gneezy, Keenan, & Nelson, 2013). Public commitments observable to others, such as roadside signs that indicate enrollment in an agri-environmental program, can motivate people to act because they want to keep their promises and act pro-socially (intrinsic motivation) and, potentially, to be recognized by others for their positive actions (extrinsic motivations). Dolan et al.'s (2012) review describes links between reciprocity and commitment that arise because people are more likely to commit to something when others are also willing to do so.

Emotional (*affect*) nudges can also promote pro-environmental behavior. Experimental economics studies in the agri-environmental domain have shown that empathy nudges can positively influence conservation behavior and promote pro-environmental decisions (Czap, Czap, Banerjee, & Burbach, 2019; Czap, Czap, Lynne, & Burbach, 2015; Lynne, Czap, Czap, & Burbach, 2016). These finding have implications for outreach campaigns designed to promote voluntary adoption of agri-environmental BMPs. For example, farmers and agricultural landowners are often emotionally connected to their land and care deeply about passing it down to the next generation—they may be more willing to invest in sustainable agricultural practices when programs emphasize how their investments will benefit their children and grandchildren.

Agri-environmental behavior is also influenced by social factors such as behavioral norms and signaling motives (Dessart et al., 2019). These factors relate to how people are influenced by the actions of others and by how others perceive their actions. Others' choices and behaviors can serve as cues that guide behavior and as a benchmark by which behavior is measured. The Dolan et al. (2012) review highlights the importance of social norms as positive feedback loops that promote greater adherence as more people follow them. Norms can provide information about the actions of others (i.e., descriptive norms) and communicate behavior deemed socially acceptable (i.e., injunctive norms) (Cialdini, Reno, & Kallgren, 1990). Pairing descriptive and injunctive norms can have particularly powerful effects on behavior (Cialdini, 2003). Furthermore, heterogeneity in behavior can lead to perverse incentives when people learn that others' behaviors are worse than theirs, known as the "boomerang effect," and research has shown that the boomerang effect can be overcome by injunctive messaging that communicates social approval or disapproval (Schultz, Nolan, Cialdini, Goldstein, & Griskevicius, 2007).[3] Le Coent, Préget, and Thoyer (2021) found that interactions between descriptive and injunctive norms can lead to multiple equilibria—such as low-participation vs high-participation states—and that the design of programs and policies can influence the equilibrium outcome.

The choice of strategies by which to incorporate social factors in agri-environmental programs depends on the current level of the target effort or action in a community (Dessart et al., 2019) and the flexibility of the program to adjust payment levels and other components. When existing participation levels are high, telling farmers about other farmers who are adopting BMPs and participating in agri-environmental programs can communicate a positive social norm (Kuhfuss et al., 2016; Wu et al., 2021). However, the social-norm strategy can back-fire when the existing level of effort or participation is low, indicating that few others are contributing to the public good (Le Coent et al., 2021). Le Coent et al. (2021) suggest that, for agri-environmental programs with low rates of participation, payment rules and communication strategies can be designed to change beliefs about the behavior of others. For example, PES programs can require a minimum level of participation from agricultural producers, thus altering their beliefs about the actions of others

[3] Evidence of the "boomerang effect" in experiments addressing agri-environmental questions have been mixed with some finding boomerang effects (Fleming, Palm-Forster, & Kelley, 2021) and others not (Wu et al., 2021).

and motivating higher levels of participation (Le Coent, Thoyer, & Préget, 2014). Alternatively, initial PES payments can be greater than subsequent payments to encourage engagement early on. Strong participation changes the descriptive social norm, which, when combined with an injunctive norm, can retain participants even when payment levels decrease later in the life of the program (Le Coent et al., 2021). Communication campaigns can be used to mitigate misperceptions of social norms when the perceived norm inaccurately portrays the true level of stewardship. In those cases, enlisting influential messengers (e.g., respected farmers in the community) to communicate an injunctive norm about desired actions can be particularly effective. People are most likely to act when they receive positive information from individuals they view as similar to themselves in some way and from individuals they like and trust (Wu et al., 2021). Social networks influence behavior in complex ways that are often difficult to analyze (Maertens & Barrett, 2013). Disentangling these interactions, which are likely tied to social and physical geographies, is an important topic for future research.

Another way to tap into social networks and norms is via programs that allow farmers to send credible signals about their stewardship actions, such as certification and verification programs (Dessart et al., 2019; Palm-Forster, Griesinger, Butler, Fooks, & Messer, forthcoming). The programs can offer a variety of benefits to participating farmers, including allowing some to differentiate their products and thereby access niche markets and/or garner price premiums. The economic benefits of verification programs can be limited for many farmers, especially those producing commodity crops (Waldman & Kerr, 2014). However, the social and community aspects of such programs can encourage pro-environmental behavior. Programs that recognize agri-environmental stewardship can, for example, influence dispositional and social factors since these programs involve commitment to an action and publicly recognize actions that tap into elements of ego and the power of social norms. Consequently, signaling programs enable social comparisons and thus can contribute to long-run changes in social norms.

Agri-environmental programs and policies can also benefit from designs that consider the influence of cognitive factors that influence farmers, including how farmers learn, discount time and money, and perceive costs, benefits, and risks (Dessart et al., 2019; Streletskaya et al., 2020). For example, programs must be salient to farmers. Dolan et al.'s (2012) review notes that stimuli that most effectively attract attention are often novel, accessible, and simple. The authors emphasize that simplicity is key because people tend to pay attention to things they can readily understand and relate. Salience also involves raising farmers' awareness of desirable practices and programs, which can be accomplished through information campaigns conducted by extension and advisory/consultancy services (Dessart et al., 2019) and by sending letters reminding them to enroll (Higgins et al., 2017). Studies have emphasized the importance of reducing perceived costs and risks associated with participation by, for example, limiting transaction costs (McCann & Claassen, 2016; Palm-Forster et al., 2016).

Though economists have long recognized that incentives influence behavior in meaningful ways, behavioral economists have identified detailed information about how various features and presentations (referred to as framing) affect responses. Dolan et al. (2012) summarizes how incentive attributes change behavior based on insights from the behavioral economics literature: (1) reference points determine what people view as losses and as gains; (2) we dislike losses more than we like equivalent gains (i.e., we are loss-averse); (3) we overweight small probabilities, especially for risks and events that are hazardous and easy to imagine; (4) we engage in mental accounting in which money is allocated to discrete accounts; and (5) we exhibit present bias by making choices that reflect living for today at the expense of planning for tomorrow.

Program administrators can make programs more attractive to farmers and rural landowners by reducing the perceived risk of participating, establishing regular payment schedules farmers can count on and making some payments upfront (Duquette et al., 2012), offering insurance options, and promoting cost-free trial periods (Pannell et al., 2006). Outreach efforts should focus on actions that lead to clear, tangible environmental benefits and emphasize a program's benefits rather than its cost (Dessart et al., 2019). Changing the default enrollment option is another important tool that can increase producers' investments in agri-environmental programs and mitigate a lack of attention to the options presented to them.

Farmer heterogeneity must also be kept in mind when designing programs to change their behavior. For example, some farmers manage their land primarily for agricultural profit while others manage it for non-pecuniary factors such as preserving the family legacy and managing land resources to support recreation such as hunting and fishing. Thus, there are likely to be no one-size-fits-all policy approaches. Instead, in the context of agri-environmental programs, administrators generally need to use a mix of strategies, including both voluntary programs (carrots) and mandatory requirements (sticks), to achieve their desired outcomes (Dessart et al., 2019; Ferraro, Messer, & Wu, 2017; Ribaudo, 2015).

2.2 Using economic experiments for evidence-based policymaking

Over the past two decades, economic experiments have been used in a variety of evidence-based policymaking initiatives outside of agriculture. For example, the U.S. federal government has used economic experiments to inform policymaking in contexts such as design of auctions used by the Federal Communications Commission (Banks, Olson, Porter, Rassenti, & Smith, 2003) and use of package labeling to indicate whether food products contain bioengineered (genetically modified) ingredients (Just & Kaiser, 2016). Outside the United States, economic experiments have informed policymaking related to sales of telecom licenses in Europe (Abbink et al., 2005; Binmore & Klemperer, 2002; Klemperer, 2002) and revisions of European Union (EU) guidelines on non-horizontal mergers (Normann & Ricciuti, 2009). In Ireland, several government commissions have jointly supported the Programme of Research Investigating Consumer Evaluations (PRICE) Lab, which uses experiments to inform government policies. For example, PRICE studies have analyzed consumers' decisions regarding personal loans, and the resulting information has informed actions by the Central Bank of Ireland (Lunn et al., 2016).

In 2014, to promote use of behavioral and experimental economics when designing agri-environmental programs, USDA established the Center for Behavioral and Experimental Agri-Environmental Research (CBEAR). CBEAR is co-headquartered at the University of Delaware and Johns Hopkins University. In a similar effort, researchers in the EU established the Research Network on Economic Experiments for the Common Agricultural Policy (REECAP) in 2017. Discussions are ongoing to establish similar programs in Canada, Australia, and other developed countries.

The Evidence Act in the United States suggests that evidence can come from a variety of sources, but economic experiments involving non-hypothetical decisions are particularly useful in informing agri-environmental programs and policies for three reasons. First, such experiments can allow researchers to identify causal responses to policy changes that otherwise cannot be isolated from administrative or observational data (Rosch, Skorbiansky, Weigel, Messer, & Hellerstein, 2021). For example, policymakers may want to know how different contract lengths affect enrollment in voluntary cost-share programs and whether these different contract lengths impact long-term persistence of practices when the cost-share ends. Experiments can be used to identify the causal effect of changing the length of a contract by randomly assigning participants to control and treatment groups with different contract lengths and then measuring outcomes of interest.

One of the goals of this chapter is to identify different progressive stages of economic experiments and illustrate how each stage best informs evidence-based policymaking. For example, laboratory and artefactual experiments are valuable for doing initial tests of potential changes in a program. They can be a wonderful way to test changes suggested by theory and observational research. Promising results in the laboratory can be further refined for robustness via field experiments that recruit the target population (e.g., farmers and rural landowners). Field settings allow researchers to randomize how various programs, institutions, and information are presented to individuals, including new interfaces, and test whether the changes affect key outcomes, like participation. If the changes perform well in the field with targeted participants, they are likely to be well suited for implementation. Ideally, the treatments are embedded in government and non-governmental programs using principles of experiment design such as randomization so additional information can be gathered to see how much the change improves outcomes compared to the status quo control treatment.

The second key area in which economic experiments have shown to be particularly useful is to observe behaviors that would not be observable using classic data collection methods and administrative records (Rosch et al., 2021). Furthermore, policymakers and program administrators are often interested in how unobservable factors (e.g., environmental attitudes, risk perceptions) influence decision making. Lab and field experiments can be designed to investigate behaviors and behavioral drivers that are typically unobservable. We emphasize the importance of thinking carefully about how experiments are designed and parameterized to reflect the policy settings in which these unobservable factors are hypothesized to be important.

Appropriate selection of experiment parameters depends on the purpose of the experiment and what it is designed to test. For example, if the experiment is designed to test economic theory, parameters are selected that support underlying theoretical assumptions. When the experiment is intended to reflect actual conditions, such as pollution of water from nonpoint sources, the parameterization process should mimic those conditions to ensure that participant behavior in the control treatment leads to the typical observed collective outcome. This design element, described in greater detail in Section 3, affects how well the study represents or parallels the relevant policy context. Once a laboratory experiment has been parameterized to mimic field experiments successfully (e.g., it captures the general scale and functional form of producer profits, social benefits, and environmental impacts), treatments can be introduced to test the extent (if any) to which they influence participant behavior and collective outcomes.

Different types of field experiments also can be used to observe specific behavioral mechanisms related to specific program contexts. For example, in a conservation auction setting, experiments can examine how bidding behavior changes when a program is modified (e.g., changing the number of competing bidders or the auction structure). An advantage of field experiments is that they tend to be appealing to policymakers because of the additional context provided in the decision environment and potential to recruit the target population as participants. However, from a research perspective, much of the participants' underlying value structures are likely to remain hidden.

The third area in which economic experiments are particularly useful in agri-environmental settings relates to the ability to derive insights into landowner decisions to participate in or opt out of voluntary programs. Researchers in this area are increasingly merging administrative records with survey data to recover information about the motives and drivers of the decisions. Agri-environmental programs often produce a wealth of data that are not structured to facilitate analysis of causality (Rosch et al., 2021). It can be difficult, for example, to find comparable control and treatment subsets in a program. And when opportunities exist to construct control and treatment cohorts as a new program is rolled-out or marketed to potential participants, program administrators generally have been reluctant to do so because of concerns about potential burdens associated with required data collection, program administration, and coordination with other entities and about equity (e.g., who is assigned the potentially better treatment vs the control group).

Reluctance to carefully test new agri-environmental programs and record key outcomes has a potentially significant cost to society and the environment; without strong supporting evidence, an agri-environmental program offered to farmers and landowners is essentially a large and costly "uncontrolled" experiment involving billions of dollars in which the funders (taxpayers) cannot determine whether their investments are cost-effective. This reluctance has seriously hampered efforts to assess the actual degree of benefits provided by programs as implemented and potential gains associated with modifying them (Ferraro et al., 2017). Experiments can demonstrate the potential benefits of modifying incentives and structuring administrative records to support high-powered analyses needed for evidence-based policy design.

2.3 Experimental tests of economic mechanisms to improve agri-environmental outcomes

In this section, we summarize key findings from the experimental economics literature regarding designing mechanisms for agri-environmental programs. We pay particular attention to two key areas of research: (1) the design of voluntary PES programs and the use of reverse auctions to allocate scarce PES funds; and (2) the design of policy mechanisms to control ambient water pollution and improve groundwater management. We do not attempt to provide an exhaustive review of experimental economics research in these areas; rather, we highlight how different types (and progressive stages) of economic experiments have been used to investigate the performance of various economic mechanisms designed to improve agri-environmental outcomes.

2.3.1 Reverse auctions and payments for ecosystem services programs

Design of PES programs has been a primary focus of experimental economics research in the agri-environmental domain. Much of this work has examined approaches for increasing cost-effectiveness of voluntary programs, including the use of reverse auctions as tools for allocating scarce resources. Reverse auctions can provide guidance when allocating limited conservation funds provided by buyers (e.g., the government) to sellers of ecosystem services (e.g., farmers). In such auctions, sellers submit offers that declare the minimum payment required for them to take a specific action that would generate ecosystem services (e.g., adoption of one or more BMPs). The process for evaluating, ranking, and selecting offers depends on the program's priorities and limitations. In general, however, a reverse auction mechanism can allocate limited funds cost-effectively by awarding payments to low-cost high-value offers. This benefit is particularly relevant since one of the world's largest land conservation programs—USDA's Conservation Reserve Program—uses a type of reverse auction to allocate a majority of its funds.

A myriad of design, implementation, and contracting considerations can affect the performance of agri-environmental auctions, making the laboratory an excellent testbed. Schilizzi (2017) provides a comprehensive overview of research on conservation auctions using laboratory experiments. Studies have evaluated various pricing mechanisms by comparing uniform and discriminatory pricing in one-shot and repeated auctions. The results generally suggest that one-shot, discriminatory, first-price auctions achieve the greatest budgetary cost-effectiveness (Cason & Gangadharan, 2005; Iftekhar & Latacz-Lohmann, 2017; Schilizzi & Latacz-Lohmann, 2007). However, only uniform-price auctions are incentive-compatible and thus are capable of revealing information about private costs (Schilizzi, 2017). Consequently, researchers need to consider how important private cost information is. Results by Liu (2021) suggest that, in multiple award settings, generalized second-price reverse auctions can have advantages over first (discriminatory)-price and uniform-price auctions because they simultaneously offer strong performance in cost-revelation and cost-effectiveness.

Laboratory experiments used to analyze auctions have identified a number of other factors that reduce budgetary cost-effectiveness, including adverse selection (Arnold et al., 2013), noncompliance (Kawasaki, Fujie, Koito, Inoue, & Sasaki, 2012), transaction costs (Li, Palm-Forster, & Bhuiyanmishu, in review), rent-seeking linked to bid-cap discovery (Hellerstein & Higgins, 2010), low levels of competition (Boxall, Perger, Packman, & Weber, 2017; Conte & Griffin, 2019), large private benefits (Conte & Griffin, 2019), and provision of information about the quality of offers (Banerjee & Conte, 2018; Cason & Gangadharan, 2004; Cason, Gangadharan, & Duke, 2003; Conte & Griffin, 2017; Messer, Duke, & Lynch, 2014) and past market outcomes (Messer, Duke, Lynch, & Li, 2017). Laboratory experiments have also been used to investigate the role of learning (Schilizzi & Latacz-Lohmann, 2007), performance of single-round vs multiple-round formats (Boxall, Perger, & Weber, 2013; Reeson et al., 2011; Rolfe, Windle, & McCosker, 2009), impacts of various selection criteria (Iftekhar & Latacz-Lohmann, 2017; Iftekhar, Tisdell, & Sprod, 2018), timing of entry in repeated auctions (Fooks, Messer, & Duke, 2015), performance of target-constrained vs budget-constrained auctions (Boxall et al., 2017), spatial coordination (Banerjee, Kwasnica, & Shortle, 2012; Fooks et al., 2016; Krawczyk, Bartczak, Hanley, & Stenger, 2016), and group bidding (Banerjee & Cason, 2020).

Research on using reverse auctions for PES programs highlights the value of pairing laboratory and field experiments. Results from laboratory experiments have revealed conditions under which particular auction formats perform better. For example, these studies have found that performance of reverse auctions and the relative budgetary cost-effectiveness of various auction designs depends on factors such as cost function heterogeneity and risk preferences (Boxall et al., 2013; Wichmann, Boxall, Wilson, & Pergery, 2017), and these factors are often unknown to program administrators. Framed field experiments can supplement knowledge gained from laboratory settings by testing how participant characteristics and preferences affect bidding behavior and auction performance. For example, Palm-Forster, Swinton, and Shupp (2017) examined farmer preferences for different types of incentives for BMP adoption. Farmers were wary of novel incentives, like BMP insurance, due to high transaction costs, and they demonstrated these preferences by requesting higher BMP payments thus reducing auction cost-effectiveness. Knowledge from framed field experiments complements findings from laboratory experiments and can be used to further refine recommendations for designing reverse auctions that will be successful in agri-environmental programs.

In addition to using laboratory experiments, researchers have tested alternative reverse auction designs in field experiments (see Rolfe et al., 2018 for a review). Their studies have compared behaviors observed in the field to results obtained from laboratory experiments and, in addition, revealed on-the-ground challenges not anticipated when conducting controlled laboratory experiments. An important distinction between laboratory and field experiments relates to expectations about participation. In lab settings, student participants arrive at sessions generally ready to participate and their actions and decisions typically are driven by salient costs and benefits. This is important in the context of reverse auctions. When auction entry is essentially costless, we expect that all subjects will choose to participate. In the field, on the other hand, participation in agri-environmental reverse auctions can be costly in terms of time, and low participation rates are common (Rolfe et al., 2018). Participant transaction costs and uncertainty about bid acceptance have been identified as significant barriers that limit participation and reduce cost-effectiveness of auctions in practice (Comerford, 2013; Palm-Forster et al., 2016; Whitten, Wünscher, & Shogren, 2017).

Strides made in advancing the design of reverse auctions and addressing challenges associated with conducting them in the field highlight the value of pairing laboratory and field experiments. Such pairing allows researchers to push the frontier of mechanism design while remaining grounded regarding challenges that arise when moving out of controlled settings. Research has also offered examples of how economic experiments can be used to inform policy and program design directly. Cummings, Holt, and Laury (2004) described a case in which policymakers used an experimental auction as a tool to guide development of the implemented auction for irrigation permits. Hellerstein (2017) highlighted key findings from auction experiments that could be used to improve the cost-effectiveness of USDA's Conservation Reserve Program.

Other studies have used laboratory and field experiments to investigate how PES mechanisms can be designed to improve spatial coordination of participants. These works have tested use of agglomeration bonuses (Liu et al., 2019; Parkhurst & Shogren, 2007; Parkhurst et al., 2002), agglomeration preferences in buyers' value functions (Fooks et al., 2016), non-pecuniary nudges (Banerjee, 2018), and voluntary conservation agreements with assurances (VCAAs) (Reeling, Palm-Forster, & Melstrom, 2019). Other studies have analyzed the impacts of network effects (Banerjee et al., 2012), transaction costs and communication (Banerjee et al., 2017), and heterogeneous land profitability (Jones Ritten et al., 2017) on achieving spatial coordination.

Laboratory experiments have also been used to investigate factors contributing to thin credit markets and low rates of participation in PES programs, including the role of risk and uncertainty (Jones & Vossler, 2014). In a study of habitat exchanges, Lamb, Hansen, Bastian, Nagler, and Jones Ritten (2019) investigated the role of three types of risk: not identifying a buyer or seller (matching risk), not being able to sell generated credits (inventory loss risk), and credit failure (post-production risk). Their findings emphasized the importance of risk mitigation strategies in habitat exchange markets, such as reimbursing sellers for credits that, despite good faith efforts, fail to generate sufficient habitat protection. Reducing risks for sellers increases market participation and improves welfare and market efficiency, which also benefit credit buyers.

2.3.2 Regulatory and market mechanisms to improve water quality

Economic experiments have served as valuable testbeds for policies not currently used in practice, such as the use of ambient pollution taxes and subsidies for nonpoint source pollution originally proposed by Segerson (1988). Most of these experiments have been limited to the laboratory, and we know of at least one framed field experiment study that involved dairy farmers (Suter & Vossler, 2014). Laboratory experiments have been used to test the performance of various ambient policy structures, including linear and non-linear ambient taxes (Suter, Vossler, Poe, & Segerson, 2008) and dynamic (Vossler, Suter, & Poe, 2013), input (Cochard, Willinger, & Xepapadeas, 2005), and average Pigouvian taxes (Sarr, Bchir, Cochard, & Rozan, 2019), as well as taxes linked to conservation compliance (Palm-Forster, Suter, & Messer, 2019). Studies have also investigated how the performance of ambient policies is affected by factors such as firm heterogeneity (Spraggon, 2004, 2013; Wu et al., 2021), communication (Vossler, Poe, Schulze, & Segerson, 2006), information about and monitoring of water quality with sensors (Miao et al., 2016), non-pecuniary nudges (Boun My & Ouvrard, 2019), informal peer sanctions (Cason & Gangadharan, 2013), social pressure (Palm-Forster, Griesinger, Butler, Fooks, & Messer, forthcoming), and the structure of damages (Willinger, Ammar, & Ennasri, 2014).

Furthermore, studies have used economic experiments to analyze water-quality trading markets and offset programs designed to reduce ambient pollution (Jones & Vossler, 2014; Liu & Swallow, 2016; Suter, Spraggon, & Poe, 2013), protect wildlife habitats (Lamb et al., 2019), and improve water availability (Bayer & Loch, 2017; Tisdell, 2010). Some water-quality trading and offset programs have been established (see Stephenson and Shabman (2017), table 3), but their success in improving water quality has been limited, due largely to limited participation. The laboratory can be a fruitful place to investigate challenges observed in those programs to determine which elements of the market design are contributing to the lack of success. For example, experiments have shown that emission dischargers (municipal stormwater and wastewater systems) tend to overinvest in upfront capital-intensive abatement technologies, which limits demand for credits (Suter et al., 2013). Studies also have highlighted how uncertainty about credit demand reduces firms' willingness to adopt technologies that could generate water quality credits (Jones & Vossler, 2014). This work highlights the market impediments generated by institutional rules that require binding pre-commitments when high fixed costs are associated with abatement technologies. In addition to identifying strategies to reduce barriers to participation and increase trading activity, experimental research has investigated the role credit valuation plays in water-quality trading markets. Liu and Swallow (2016) demonstrated how accounting for public values for co-benefits in water quality credit transactions can improve market efficiency by creating additional incentives for credit suppliers.

2.3.3 Policies and institutions for sustainable water withdrawals

Examining decisions on groundwater withdrawals has been a fruitful area of study using economic experiments to explore behavior, policies, and water use efficiency. This area of research is especially well suited for experiments because of the difficulty of observing actual groundwater extractions and their impacts on aquifers and other users of the water resources. In most places in the world, few groundwater wells have meters that can accurately measure the amount of water withdrawn from underground aquifers. And when such meters are present, the resulting information is often unavailable to the public and to resource managers. Groundwater users typically are reluctant to voluntarily reveal information about their withdrawals because there is a strategic incentive not to reveal their true levels of pumping. They tend to be concerned about potential regulations and/or liability for affecting the availability of groundwater to others who want to use the same aquifer. Researchers can observe behaviors that are likely to occur by developing experimental frameworks that parallel many of the key economic, psychological, and social aspects of groundwater pumping. To date, most experiments in this area have induced valuations in laboratory experiments involving student subjects.[4]

Broadly speaking, economic experiments investigating aquifer use are related to experiments that have studied behavior in common pool resources, especially ones involving spatial variability (e.g., Janssen, Anderies, & Cardenas, 2011; Janssen & Ostrom, 2008). This line of research often parameterizes the experiments such that private benefits to potential users increase upon entrance into the market. The experiments frequently incorporate a negative externality so that entrance and subsequent use negatively affect the private benefits of other resource users. Using these setups, a variety of policies and institutions can be exogenously or endogenously imposed, and studies can measure the impacts of the interventions (see, for instance, Casari & Plott, 2003; Gardner, Moore, & Walker, 1997; Mason & Phillips, 1997; Rodriguez-Sickert, Guzmán, & Cárdenas, 2008).

In aquifer use, groundwater extraction by one user increases the private costs of other users of the resource. This increase is often assumed to be associated with pumping water from a great distance, which requires additional energy. Much of the experimental research conducted in this area has sought to understand the coupled relationship between human behavior and a model of water flow through an aquifer. The research arose, in part, because some studies, such as Gisser and Sanchez (1980), suggested that there would be few social inefficiencies even if the aquifers were not managed. The estimates often involved simplistic "bathtub" models of aquifers that did not capture true spatial and temporal movements of groundwater. When more-realistic models were included in calculations in non-experimental studies, it was easier to identify when social inefficiencies arose (i.e., Brozović, Sunding, & Zilberman, 2010; Guilfoos, Pape, Khanna, & Salvage, 2013).

This body of experimental research has uncovered several key findings. First, the behavior of groundwater users varies depending on the spatial and temporal parameterization of the aquifer (Suter, Duke, Messer, & Michael, 2012). More-complex modeling of spatial characteristics affects the behavior observed in groundwater

[4] We are aware of several large-scale field experiments involving groundwater extraction. To date, those studies have looked at how best to encourage users to voluntarily report their monthly water use (Meiselman et al., in development), how accurate voluntary reporting is compared to government-required annual reports (Savchenko et al. *in development*), and the impacts of knowing how much water is being extracted by neighboring landowners (Suter et al. *in development*). To date, no results have been published.

users, and, ultimately, the additional complexity influences the degree to which unregulated groundwater extraction has a negative impact on efficiency.

Several studies have investigated how the number of groundwater users affects the efficiency of aquifer use. For instance, Gardner et al. (1997), using experiments, found that social efficiency increased when the number of users was limited. They also showed that social efficiency could increase with use of individual quotas.

Another key finding from experimental economic studies of groundwater is that having greater information about the condition of the aquifer can change the behavior of aquifer users (Li, Michael, Duke, Messer, & Suter, 2014; Saak & Peterson, 2007). Though some earlier research (e.g., Tisdell, Ward, & Capon, 2004) found that the impact of that type of information was relatively small, others, such as Li et al. (2014) found that both the existence of the information and how it is displayed affect behavior. They further showed that groundwater users were more responsive to easy-to-understand depictions of the severity of the risk of overpumping relative to greater quantities of information and more-complex scientific descriptions of this risk.

Other experiment-based research in this area has examined the importance of entry and regulations using internal and external models to determine whether the effects on the social efficiency of groundwater management vary (Suter et al., n.d.; Liu, Suter, Messer, Duke, & Michael, 2014). Since many groundwater policies involve imposition of taxes on water users, studies also have tested the efficiency and behavioral impacts of different means of redistributing the tax revenue (Duke, Liu, Suter, Messer, & Michael, 2020).

3 Designing experiments to inform agri-environmental programs and policies

Researchers have written extensively about the proper design of economic experiments in the laboratory and field (Davis & Holt, 1993; Friedman & Sunder, 1994; Lusk & Shogren, 2007), including two handbooks filled with perspectives and recommended best practices from leading experimental economists (see Banerjee & Duflo, 2017; Kagel & Roth, 2016). Summarizing this rich body of knowledge exceeds the scope of this chapter so we recommend consulting those resources when undertaking experimental research. With the limited space available here, we concentrate on several issues that are particularly important when designing experiments to inform agri-environmental policies and programs. We address concerns related to internal and external validity and interactions of experimental control, context, and representativeness.

We present what we think of as four stages of experimentation for research designed to inform agri-environmental programs and policies. The stages consist of (i) laboratory experiments with students, (ii) artefactual and framed field experiments with the target population (often farmers and rural landowners), (iii) field experiments that can result in real changes in agri-environmental management, and (iv) randomized controlled trials (RCTs). We certainly acknowledge the stand-alone value of research findings that can be generated by a single type of experiment (e.g., findings from a laboratory experiment), and we are not suggesting that a researcher or research team must always use all experiment types progressively to be informative on a particular issue. That said, we believe that there is a logical progression of research designs and want to highlight important benefits of building upon findings from laboratory experiments with students by conducting context-specific field experiments and RCTs with farmers. We also briefly address considerations when conducting experiments with partner organizations and when designing experiments to analyze long-term behavioral changes, though we note that more work is needed to develop best practices for analyzing long-term agri-environmental behavior. In Section 4, we discuss five contemporary issues associated with designing and implementing experimental studies.

3.1 Internal and external validity

Rigorously testing new agri-environmental interventions is critical to ensuring internal and external validity before recommending an intervention for use. *Internal validity* refers to the ability to isolate a treatment effect by controlling for potential confounding factors (covariates) that could influence the outcome of interest. That is, internal validity refers to the ability to argue that observed correlations between treatments and outcomes are causal (Roe & Just, 2009). In many research endeavors, the goal is to determine the precise effect of a treatment on an outcome, but many other factors can influence the outcome. Hence the saying among students of economics and other social science fields: "correlation does not imply causation." Economists frequently have a limited ability to identify causality because of the nature of the data used, which are often observational and present various challenges. Much of researchers' empirical training is focused on analyzing observational data and using sophisticated econometric approaches to identify causal effects when warranted.

Through the random assignment of participants to control and treatment groups, experiments have been touted as one of the most reliable tools economists have to uncover credible evidence of causal effects of a treatment (Banerjee & Duflo, 2017). In experiments, researchers can control treatment assignments. In observational settings, assignment of treatments is beyond the control of the researcher and often depends on confounding factors. An example in an agri-environmental context is farmers who voluntarily enroll in a new agri-environmental program—in part because of the program attributes and in part because of the farmers' underlying stewardship ethics, characteristics of their farm operations, and unobservable social pressures from neighboring farmers. The program attributes are observable, but the other factors influencing their decisions usually are not, making it impossible to attribute their participation to the program features. A carefully designed experiment, on the other hand, can randomly assign farmers to a treatment group presenting the new agri-environmental program or to a control group, which can only enroll in the current program. Random assignment eliminates concerns about confounding factors and permits researchers to determine causal effects on the outcome of interest (e.g., participation).

External validity refers to the ability to extend causal relationships identified in a study to settings, contexts, and individuals with different characteristics. That is, external validity relates to the generalizability of the findings of a study (Roe & Just, 2009). The role and importance of external validity is more nuanced—and, consequently, more subject to debate—than internal validity, which is widely acknowledged as a fundamentally important issue for economic experiments. Conversations about the importance (or lack of importance) of external validity often occur when applying lessons from a laboratory experiment to the field (Camerer, 2011). In field experiments and RCTs, researchers and policymakers are frequently concerned about the external validity between field settings. It is not always known whether outcomes observed in one field setting will be replicated in field settings involving different populations and/or other points in time (Athey & Imbens, 2017). Furthermore, it is typically uncertain how well the results and lessons from field experiments will "scale up."

Do concerns about external validity make experiments less valuable? Our short answer is "No." The longer answer is found in a statement loved and loathed (and frequently used) by economists: "It depends…" It depends on the broader research goal and how the experiment will support that goal. Is the goal to learn something general about behavior by people in decision-making frameworks or is it to say something about behavior in a specific decision-making context? Camerer (2011) presented this dichotomy by juxtaposing what he called the "scientific view" and the "policy view" and posited reasons for concerns about lab-to-field generalizability often being exaggerated. From a scientific view, external validity is not a pre-condition of a well-designed experiment to show how behavior is generally influenced by one or more factors. Many people believe experiments must have external validity to be policy-relevant, but we view that argument as only partially valid. On one hand, if the goal of the research is to characterize behavior that will likely be observed in the field, it can be important for the experiment to reflect conditions in the field (see the following discussion of *parallelism*) and for the results to be generalizable. On the other hand, knowledge about general behavioral responses to changes in decision-making environments can be policy-relevant even when the observations are made in lab settings. We argue that the research goal is what should ultimately dictate how much weight should be placed on external validity.

Concerns about internal and external validity are not unique to experimental economics, and additional considerations are required when designing and implementing an experiment and analyzing experiment-derived data. Throughout this section, we highlight how the choice of experiment type and design relates to internal and external validity. In Section 4 (Issues 4 and 5), we discuss how efforts to improve participant recruitment and identify heterogeneous treatment effects can improve the external validity of agri-environmental economics experiments.

3.2 Design considerations: Control, context, and representativeness

An experiment's structure is critical for its ability to provide accurate and useful insights. Researchers make dozens of decisions when designing any particular experiment. These decisions are typically guided by a desire to develop the experiment that is best suited to answer the research question(s) at hand. It would be impractical of us to attempt to provide guidance for all the possible design decisions researchers might ponder when planning an economic experiment. Instead, we focus on three dimensions that are integral in designing experiments that can inform policy: control, context, and representativeness. We define each dimension and describe how certain design decisions influence fundamental tradeoffs among control, context, and representativeness. In the following section, we revisit these tradeoffs in our discussion of four stages of experimentation that can be used to generate credible evidence related to agri-environmental programs and policies.

Lusk and Shogren (2007, p. 6) defined control as limiting the environment of the experiment "such that no unmeasured external force drives choices. That is, confounding of cause and effect is eliminated." Experiments in general provide the researcher with a level of control through the exogenous assignment of participants to treatment and control groups. The highest levels of control are typically achieved in laboratory experiments, in part because the administrator can monitor all participants simultaneously and prohibit or limit communication among participants through rules established at the beginning of the experiment. Laboratory experiments can also be designed to allow for free communication among participants and for records of that communication to be saved and used in data analysis. In field experiments, on the other hand, researchers have far less (and sometimes no) ability to control communication among participants or to record it directly and often must rely on self-reported post-experiment surveys to gather information about the nature and amount of communication among participants.

The source of values driving decision making can also influence the level of control in an experiment. Most laboratory experiments rely on *induced values,* which are set within the experiment and thus are known to both the researcher and participant. For example, in a reverse auction experiment, participants are assigned costs that they can consider when deciding whether they will participate and how much to bid. Additionally, when payoff functions are only influenced by induced values, the researcher knows clearly which incentives are available to drive participant behavior. *Endogenous values* arise entirely from the participants and may or may not be revealed to the researcher through measuring participants' behaviors. If endogenous values are unobservable, but influence observed behavior, the researcher lacks some level of control especially if these external forces end up being key drivers of participant decisions.

Lusk and Shogren (2007, p. 15) defined *context* as participants having "some contextual cues about why their decision[s] might matter in a bigger world." Among other things, context affects the tone of language used in experiments. An important decision is whether instructions should be written using context-specific language (e.g., landowner, profit, taxes, pollution, conservation, government program) or neutral language (e.g., agent, earnings, deduction, external effects, project). Context-specific language can make instructions and the experiment setting easier for participants to understand. However, it also can invoke emotional responses and negative associations (e.g., terms such as taxes and pollution). The language, therefore, must make the incentives adequately salient to participants. Whether these changes truly affect participants' behaviors is subject to debate and has not yet been thoroughly tested.

A related concept is experiment *parallelism*, typically defined as the extent to which the conditions and parameters in the experiment replicate actual policy conditions (Camerer, 2011; Levitt & List, 2007a, 2007b; Messer, Kaiser, & Poe, 2007; Plott, 1987; Smith, 1982). A high degree of parallelism in an experiment is desirable as it induces behaviors in an experiment that reflect participants' behaviors in actual settings.

While a generically framed experiment can produce individual and aggregate results that closely parallel behavior in an agri-environmental program, providing participants with additional context increases the degree of parallelism by associating their choices with opportunities to "do good" for the environment and others. For example, in a laboratory experiment conducted by Palm-Forster, Suter, and Messer (2019) regarding policies to reduce ambient pollution in a primarily agricultural watershed, participants were asked to act as firm managers operating in a shared watershed. The downside of context arises from potential bias introduced when experiment participants have strong opinions about the issue. For instance, the terms "tax" and "transfers" can have identical functions in an experiment setting, but "tax" can invoke strong reactions in participants who view any tax as undesirable—even if, in a particular context, the tax would increase social welfare. Such negative reactions create problems of internal validity when they give participants a perverse incentive to reject a program's financial incentive. However, if the negative reactions are similar to what would be observed in actual programs, use of the additional context can be quite beneficial.

Neutral, abstract language is typically used in the instructions for experiments designed to test theoretical predictions thus providing greater control over potential confounding factors. For example, to ensure that participants make their decisions based on the economic incentive offered in the experiment and not on their feelings about the study topic, researchers can use generic terms such as goods, tokens, buyers, sellers, firms, and managers in experiment instructions.

Parallelism has become a critical concern among agricultural policymakers, academic peer-reviewers, and funders trying to evaluate the results of experimental studies. Policymakers and peer-reviewers are often skeptical of results from unframed experiment designs and designs that grossly simplify complex policy choices even when there is little evidence that the framing affects participant behavior. Their skepticism has led them to require additional justification for the need for and reasonableness of simplifications and framing choices.

Messer et al. (2014) built on the description of parallelism by adding a third dimension, *representativeness,* related to how similar behaviors of sampled participants are to behaviors of people making actual economic decisions. Representativeness is particularly critical for agri-environmental experiments because researchers can rarely sample large groups of agricultural producers and landowners (Weigel, Paul, Ferraro, & Messer, 2021). Several studies employing experiments in agri-environmental contexts have indicated that landowner behaviors are similar to behaviors observed in laboratory experiments with undergraduate students (see a summary in Rosch et al., 2021). However, government agencies so far remain reluctant to use experimental results to establish policies, in part because of questions about whether the results are broadly applicable to producers targeted by their specific programs and focus areas. Consequently, researchers should be aware of the inherent trade-offs associated with various types of participant pools when defining what constitutes causality, achieving external (face) validity, and testing theories and policies related to land economics.

In our following description of different types of experiments that can be used to inform evidence-based policymaking in the agri-environmental domain, we discuss how design choices influence tradeoffs among control, context, and representativeness.

3.3 Experiment stages

A single experiment does not need to "do it all." As researchers, we contribute incrementally to the stock of knowledge in a particular field, and experiments are but one tool we use to contribute to that knowledge. We learn something from every experiment so the goal is not to design one experiment from which we can learn everything about a particular topic. (And we would certainly fail if we tried.) The goal is to implement a well-designed experiment that clearly tests one or two hypotheses and leads to the next research question and next experiment. With this incremental approach to research in mind, we suggest that investigators think carefully about their overarching research questions and the level of experiment testing that will allow them to answer those questions. We further note that it can be helpful to conduct surveys, focus groups, and/or interviews prior to designing experiments since those tools can provide important insights into needed elements of the design.

Different types of laboratory and field experiments offer distinct advantages and pose unique challenges; used together, they represent a strong versatile toolkit for experimental economists. Cason and Wu (2019) argued that, in general, tests of economic theory are best accomplished using laboratory experiments involving student subjects and that field experiments with the target population are preferrable for questions about specific policies and programs and when the role of heterogenous preferences and characteristics of the target population is important.

We propose four stages of experimentation that can inform the design and implementation of agri-environmental programs and policies. Experimental approaches in these stages range from using induced-value, context-neutral laboratory experiments involving student participants to RCTs in collaboration with agency partners in which observations come directly from the population of interest. In Table 2, we present seven key attributes of experiments in each stage that we feel are particularly important to consider when designing experiments in agri-environmental contexts. These attributes include, (1) the location of the experiment, (2) source of values, (3) participants' awareness of the research, (4) experiment framing, (5) whether experiment decisions link to real world behavioral changes, (6) the participant pool, and (7) experiment incentives. These attributes influence the strengths and weaknesses of using a specific type of experiment to answer a relevant research question, depending on the nature of the question. In Fig. 1 we present a graphical depiction of the tradeoffs among control, context, and representativeness for experiments designed to investigate questions related to agri-environmental issues.[5] In describing the types of experiments within each stage, we will reference Table 2 and Fig. 1 to discuss how attributes of those experiments connect to fundamental trade-offs among control, context, and representativeness that strongly influence the balance between internal and external validity and the level of parallelism.

Using multiple complementary stages of experimentation provides a comprehensive research approach for generating credible research findings that can inform evidence-based policymaking in agri-environmental contexts. Naturally, there are contexts and questions for which a stage is not appropriate, and it is not necessarily feasible to conduct all of the stages because of budget, time, and other constraints.[6] Note also that these stages, ideally, are pursued sequentially, developing knowledge in Stage I that is rigorously tested in subsequent stages; however, we recognize that there are situations in which findings in later stages require additional work using Stage I or Stage II approaches before the underlying behavioral mechanism can be understood. We use the experiment stage framework as a guide to show how different types of experiments can build upon each other to inform program and policy design, but we are not suggesting that researchers must always work through each stage to generate meaningful findings. We next discuss the value of each experiment stage to assist readers in determining the stages and sequence of experiments best suited to particular research questions.

[5] Figure 1 is a modification of one developed in Messer et al. (2014) and is an extension of Lusk and Shogren's (2007) framing methodology, which discussed trade-offs between *control* and *context* in experimental designs for auctions.

[6] Constraints beyond those imposed by budgets and timelines may limit a researcher's ability to conduct a particular type of experiment. For example, constraints related to participant recruitment are often faced by researchers trying to engage farmers and rural landowners in economic experiments (and other types of research, for that matter). We discuss recruitment challenges in Section 4 (Issue 4).

Table 2 Stages of experiment testing to inform agri-environmental policy and program design.

Stage	Type of experiment	Location	Source of values	Experiment framing	Participant pool	Awareness of research participation	Do experiment decisions link to real behavioral changes?	Experiment incentives
I	(IA) Context-neutral lab experiment (IB) Context-specific lab experiment	University lab	(IA) Induced (IB) Induced/ endogenous[a]	(IA) Context-neutral (IB) Context-specific	Students	Yes	No	Experiment dollars[b]
II	(IIA) Artefactual field experiment (IIB) Framed field experiment	University or mobile lab	(IIA) Induced (IIB) Induced/ endogenous[a]	(IIA) Context-neutral (IIB) Context-specific	Target population[c]	Yes	No	Experiment dollars[b]
III	Field experiments with potential on-farm implications	In the field	Endogenous	Context-specific	Target population[c]	Yes[d]	Yes	Actual currency
IV	Randomized controlled trials (RCTs)	Natural decision-making environment[e]	Endogenous	Context-specific	Target population[c]	No	Yes	Actual currency

[a] *In laboratory experiments related to agri-environmental topics, values are typically induced in the sense that incentives and payoffs functions are assigned by the researcher. There may be some cases in which endogenous values enter an experiment, particularly when context-specific language is used.*

[b] *Experiment dollars are traditionally converted to real currency at a specified conversion rate, which means that typically the stakes are much lower in Stage I and II experiments relative to Stages III and IV.*

[c] *Farmers and rural landowners are often target populations for experiments on agri-environmental decision-making.*

[d] *In Stage III, we consider field experiments in which participants are aware of the research study. These types of field experiments are often designed to test agri-environmental research questions. Awareness of research participation is not a defining feature of all field experiments in general.*

[e] *Information about decisions can often be obtained via administrative data.*

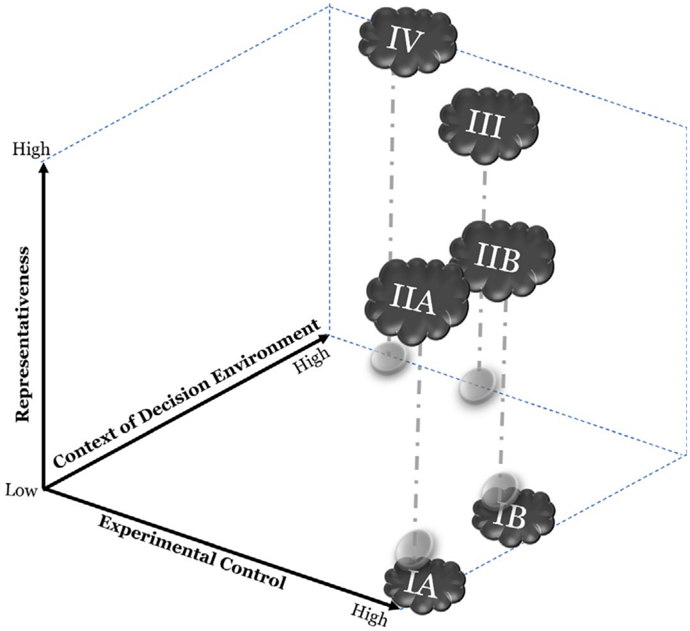

FIG. 1

Stages of experiment testing involving trade-offs in control over the experiment versus the context of the decision environment and representativeness of the sample and incentives.

3.3.1 Stage I: Laboratory experiments with student participants

As shown in Table 2 and Fig. 1, Stage I experiments provide strong control because they are generally conducted in university laboratories but often have weak representativeness by recruiting student participants and using researcher-induced values. We subdivide Stage I experiments into Stage IA experiments, which use context-neutral (abstract) language, and Stage IB experiments, which use context-specific language.

The ability to control participant valuations and, therefore, the incentives via induced values is particularly important (1) when using settings and examining behaviors for which direct data are not available and (2) when participants form their valuations using information generally hidden from outside observers, as occurs with rent-seeking and adverse selection in conservation auctions, and nonpoint source pollution behavior. Stage I experiments are also valuable because they are the simplest and least expensive to conduct since researchers can readily recruit participants from large pools of undergraduate students and generally can offer students relatively small financial incentives (approximately the regular hourly wage for student workers on campus).

Laboratory experiments involving student subjects allow researchers to "induce" participants with pre-set monetary values known to the researchers and use financial rewards that are salient to students, creating incentives that can mimic assumptions from theory. Tests of theoretical questions generally can be considered valid in a variety of settings, including simple, well-controlled university experimental economics laboratories with relatively homogeneous student participants. Therefore, Stage I experiments are particularly useful for researchers primarily interested in testing theory, and they also allow experimenters to place greater stress on theories by testing them under a variety of assumptions that would likely be relevant for programs implemented in real world settings.

In practical terms, laboratory experiments offer some clear advantages over field experiments, including being significantly less costly and time-consuming, providing the greatest control over the experiment environment, and limiting irrelevant variations in participants that can affect results (Cason & Wu, 2019). The relatively low cost allows researchers to enroll larger numbers of participants. With larger samples, researchers can generally test a greater number of hypotheses and obtain greater statistical power. Lower-cost experiments also mean that researchers can spend more time focusing on research and less time writing grant proposals, which can be time-consuming to develop and have success rates of only about 10–20% at leading funding agencies such as USDA's National Institute for Food and Agriculture and the National Science Foundation. Although we typically do not consider external validity as a strength of laboratory experiments, Cason and Wu (2019, p. 746) noted that laboratory experiments involving students "can enhance external validity because the experimenter can manipulate numerous variables and factors to put stress on the theory and determine how sensitive the predictions of the theory are to context."

Another key advantage of using students is their relative homogeneity in terms of demographic characteristics. Students tend to be more similar to each other than members of the general public are in terms of income, age, and educational background. This similarity can reduce the number of factors potentially contributing to variations in behavior in an experiment. Greater homogeneity can also improve statistical power and increase the likelihood of identifying true treatment effects (i.e., avoiding incorrect findings). By definition, students are generally well-educated and tend to be comfortable using technologies such as computers and tablets, which are often used in experiments. Furthermore, Fréchette (2015) found that student and non-student subjects behaved similarly when making decisions framed in the same way.

Obviously, however, homogeneous subjects are not ideal when seeking to understand behavior by non-student stakeholders who are heterogeneous on multiple dimensions and when examining how stakeholder characteristics affect behavior. However, even in those cases, studies can benefit from first conducting a laboratory experiment and then moving on to more-heterogeneous environments. Initial laboratory experiments can provide solid baselines for observations and opportunities to test potential experiment designs.[7]

The framing used in experiment instructions is also an important design consideration. Stage IA describes experiments in which the instructions use context-neutral language. Participants are asked, for example, to sell "units" to a "buyer" under various market rules or to make "production" decisions that lead to different "payoffs" and "costs" for them and others in a group. As discussed earlier, some researchers worry that these generic terms are difficult for participants to understand and apply and therefore prefer language that is more context-specific (represented by Stage IB).

An important question in terms of agri-environmental policies and programs is whether the behaviors of student participants in a laboratory are adequately *predictive* of farmers' actual behaviors. Students generally are much younger and have less work experience than agricultural producers and generally are less wealthy. The majority of U.S. farmers are men while the majority of college students are women, and students generally are more ethnically diverse than agricultural landowners. And, most students have no experience with agricultural production or land management. Thus, at first glance, questions about the use of student participants for policy-relevant agricultural experiments make sense. If a sample of students does not behave similarly enough to producers, experiments designed for agri-environmental contexts cannot provide meaningful, accurate (and therefore useful) information for policymakers. However, given researchers' limited resources and difficulty recruiting agricultural producers, student results viewed as "predictive enough" can be highly cost-effective as a means of pre-testing procedures to be used with producers. Just as medical studies use animal models to study how cancer can affect humans, economic researchers can benefit from identifying conditions under which student participants respond similarly and differently from target populations, allowing them to better calibrate results and determine when extrapolating from student results is appropriate.

[7] Another low-cost option is to recruit participants through online platforms such as MTurk and Qualtrics. A key advantage of these platforms is that they tend to involve participants who vary more in age than can be found at most universities. However, whether these low-cost participant pools are better than student pools has been the subject of significant investigation (see, for instance, Berinsky, Huber, & Lenz, 2012; Huff & Tingley, 2015; Mullinix, Leeper, Druckman, & Freese, 2015; Snowberg & Yariv, 2018; Goldberg, van der Linden, Ballew, Rosenthal, & Leiserowitz, 2019; Boas, Christenson, & Glick, 2020). Snowberg and Yariv (2018) found that MTurk participants tended to be younger and more educated and to have lower incomes than the general population. As a result, they argued that researchers gain little from using MTurk respondents over student samples. Berinsky et al. (2012) found that MTurk participants tended to be more representative of the general U.S. population than samples obtained using in-person convenience, and several other studies found that results drawn from MTurk participants were similar to ones obtained from national samples (Savchenko et al., n.d.; Mullinix et al., 2015). Regarding samples from Qualtrics, Boas et al. (2020) found that Qualtrics samples were more representative than MTurk samples but also costed more.

Unfortunately, there is not yet enough data to determine how important these differences are. Evidence from some experiments related to agri-environmental programs has suggested that student and farmer behavior is not all that different (Fooks et al., 2016; Suter & Vossler, 2014). Other non-economic studies have shown that the behavior of students, the public, and professionals in response to experimental treatments can vary (King & He, 2006).

3.3.2 Stage II: Artefactual and framed field experiments with non-student participants

Stage II experiments are similar to lab experiments, but they are conducted with the target population, typically agricultural producers and rural landowners for agri-environmental studies. These experiments are often conducted in locations that are convenient for farmers, but farmers may also be invited to a university laboratory. These experiments can be designed using context-neutral (abstract) or context-specific language. When language is context-neutral, Harrison and List (2004) referred to these experiments as "artefactual field experiments" and we adopt their terminology for experiments in Stage IIA. Gneezy and Imas (2017) used the term "lab-in-the-field" to refer to experiments conducted in the field but designed with a an induced-value laboratory structure. Context-specific language is used in Stage IIB experiments, and we call these "framed field experiments" (also following Harrison & List, 2004).

Recruiting experiment participants from the group of interest improves the representativeness of the study, and retaining an induced-value (laboratory-based) structure provides a relatively high degree of control. These experiments can be conducted in university laboratories. However, it is usually difficult and significantly more expensive in terms of compensation to recruit agricultural landowners and producers when they must travel to the laboratory. Consequently, Stage II experiments are ideally conducted using mobile laboratories at sites where farmers tend to gather, such as state fairs and agricultural extension conferences. Other options include university facilities in rural areas that specialize in extension activities and other convenient gathering places in rural communities. Regardless of where the experiments are conducted, using agricultural producers and rural landowners as subjects significantly enhances the representativeness of the studies and allows researchers to assess the effects of context-specific instructions.

Like most design decisions, the choice between using context-neutral and context-specific language is guided by the goals of the experiment and by the research questions the experiment is expected to answer. Researchers must also weigh other considerations when making this choice, like how to most efficiently use their limited time and scarce research funds. If a researcher is investing the additional effort and expense to recruit farmers, they are likely interested in how farmers make agri-environmental decisions in specific contexts. Designing an experiment with context-specific language may also be most salient for farmers participating in the study. Some level of researcher control is forfeited when participants consider factors external to the experiment in making their decisions, but this is a tradeoff that is often accepted to capture behavior that better reflects decision-making outside of the experiment.

Agricultural producers naturally bring heterogeneous preferences, experiences, social attitudes, and norms to experiments that can affect their behavior, particularly when context-specific language is used. Therefore, some degree of control over participants making true valuations can be lost. A significant concern is landowners worrying that the study results will influence the policies and regulatory environments they face in the future, which gives them an incentive to attempt to skew the results. This is particularly problematic when experiments involve treatments representing policies that potentially have large financial impacts. For instance, consider experiments that have tested ambient taxes to improve water quality (see Section 2). Such tax policies would likely go into effect only if voluntary efforts to reduce pollution are not effective. To demonstrate that voluntary actions are sufficient (and, thus, that taxes are unnecessary), participants could choose to act strategically in the experiment by making different choices (e.g., larger reductions in fertilizer use and thus production) than they would for their own fields. When some participants act strategically in response to suspected policy implications, their decisions are not motivated by the salient reward offered in the experiment. Consequently, control of the research setting is lost, and the responses do not necessarily reflect the effectiveness of the incentive being tested. After all, the financial stakes in an economic experiment, even when paying hundreds or thousands of dollars to participants, likely pale compared to potential costs that new regulations and taxes could impose on farmers. Note, also, that neutral framing does not guarantee that private motivations will not affect participants' choices. Future taxes and/or implications of a study can still be apparent to the participants, especially when an experiment is conducted solely with farmers.

Experimentalists often believe that the data generated through an experiment tell them everything they need to know. While that may be mostly true in context-neutral laboratory experiments with high levels of control, it is less true in the field. As discussed, participants from a group of interest (e.g., farmers) bring rich knowledge, experiences, beliefs, and values into an experiment and these factors, which are typically unobservable to the researcher, can influence their decisions. Finding ways to capture these behavioral drivers can lead to valuable insights. We have found that coupling experiments with debriefing sessions, focus groups, and surveys can enrich the researchers' understanding of decision making in different contexts, and this information can be particularly informative when applying insights from experimental studies to program and policy contexts.

3.3.3 Stage III: Field experiments with potential on-farm implications

Researchers studying questions related to agri-environmental decision-making can construct experiments in which participants make actual decisions that can influence their agricultural operations and/or the land they manage. We broadly refer to these studies as field experiments, and they can take a variety of different forms. The behavioral changes often measured in field experiments can involve significant costs for a participant (e.g., purchasing an agricultural input or deciding to use a new BMP). In our classification of experiment stages, the real-world nature of decisions and the associated costs in Stage III field experiments are what differentiate them from Stage II framed field experiments that are typically set up more like context-specific lab experiments with lower stakes. Sometimes a Stage III field experiment may be implemented as a pilot project with external partners, other times it might be feasible through a large grant. We differentiate Stage III field experiments from "natural field experiments" and "RCTs" based on an assumption that in the latter cases, participants are not aware that they are involved in an experiment. In the Stage III field experiments described in this section, we assume people know they are engaging in research.

Field experiments with the potential to influence real land management decisions are relatively expensive to conduct but clearly provide greater context and representativeness, which are attractive attributes for stakeholders who may be interested in using the findings to inform their work. The amount of control a researcher has over the environment in a field experiment tends to be limited by factors such as the natural policy environment involved and the amount of available research funding. Given their cost, researchers are likely to use field experiments to build on previous promising findings obtained in simpler, less-expensive settings.

Field experiments are most likely to provide convincing external validity. Note, however, that the external validity of field experiments is limited by attributes of the research setting and characteristics of the participants. For example, one cannot necessarily generalize the results of a field experiment conducted with northeastern U.S. vegetable crop farmers to row crop farmers in the Midwest or to vegetable farmers in other developed countries. Furthermore, field experiments can be conducted in a variety of settings that may influence the generalizability of observed behavior. For example, some settings may convey more of a research focus (e.g., a university-sponsored booth at an agricultural expo) vs having participants make decisions via an online platform or using a paper form that can be mailed or delivered to a community partner, like a soil and water conservation district.

As discussed in Weigel et al. (2021), a major challenge associated with Stage III research is recruitment of rural landowners, and we discuss this issue in Section 4 (Issue 4). Numerous studies by leading researchers have never been fully implemented (and have not been published) because of recruiting challenges. And this problem has only become more difficult as researchers have started relying more on power analyses (correctly so) to determine the necessary sample size prior to launching studies and incorporating the power analyses in pre-analysis plans. As we discuss hereafter, a well-powered study generally requires a dramatic increase in the number of participants—from the less than one hundred participants used in many published studies to thousands—to test treatments. In Section 4, we further describe issues related to statistical power and discuss why statistical power is critical for deriving robust findings.

3.3.4 Stage IV: Randomized controlled trials

Stage IV experiments are RCTs which are often implemented in coordination with the governmental or non-governmental organization that administers the agri-environmental program in question. RCTs represent the extreme opposite of the induced-value, neutrally-framed, university laboratory experiments involving student participants of Stage IA. After researchers have identified promising treatments using framed field experiments, RCTs are the logical "next step" to test the treatments in the field with individuals who will be affected by the policies (Behaghel, Macours, & Subervie, 2019). RCTs allow for strong context-specific framing and representativeness but often sacrifice experimenter control. The degree of control over the experiment environment varies because the researcher is often constrained by cooperation with the partner organization.

Few papers describe the benefits and drawbacks of using RCTs for agricultural policymaking in developed countries. Researchers helping to fill this gap include Behaghel et al. (2019) and Colen et al. (2016)—they describe benefits of using RCTs to inform the design of the European Union's Common Agricultural Policy (CAP). There is, however, a much richer literature on using RCTs in the context of development economics, and we point readers to this literature for a more comprehensive discussion about using RCTs for program and policy research (see Barrett & Carter, 2010, 2020; Gueron 2017).

Program administrators frequently consider new ways to implement agri-environmental programs without implementing careful controls that would allow them to establish a causal link between a program change and outcomes such as reductions in pollution. RCTs are an excellent method for identifying causal relationships. For example, a program could be interested in testing the impact of providing information to landowners using engagement strategies that are meant to increase the salience of stewardship. By randomly assigning the new-information treatment to some landowners and not to others and measuring outcomes such as sign-up rates, researchers can estimate the causal effect of the program change on desired outcomes.

We see conducting RCTs at the launch of new programs, and as part of significant revisions to programs, as promising opportunities, and we encourage researchers to pursue partnership opportunities. RCTs have not been widely used to evaluate agri-environmental programs and policies in the United States and Europe (Behaghel et al., 2019) but offer important benefits. A well-designed RCT with sufficient controls can accomplish a relatively high degree of internal and external validity. The current lack of agri-environmental RCTs reflects, in part, the challenges associated with their use, including political questions about the fairness of randomly exposing only a subset of individuals to the treatment.

We also view RCTs as an experimental tool that can support the measurement of long-term behavioral changes. Most economic experiments measure behavioral changes in the short run. However, agri-environmental questions in general and questions about producer behavior in particular often relate to long-term and/or repeated behaviors. To truly understand these behaviors, researchers need to make repeated observations over time. Long-run impacts are particularly relevant when evaluating the kind of large-scale interventions used in government programs (Colen et al., 2016; Czibor, Jimenez-Gomez, & List, 2019) in which the goal is some type of sustainability, which implicitly means motivating consistent pro-environmental behavior over time.

Evidence suggests that applying lessons learned from short-run experiments can be short-sighted (Czibor et al., 2019). While some studies have shown long-term impacts, other studies suggest that the effects of experimental treatments persist for short periods. The reasons for this relatively brief impact are unclear. Perhaps, in fact, the treatment effect does not persist. However, the lack of an ongoing effect could be related to participants' long-term exposure (or lack thereof) to the treatment being measured. This type of effect is related to attenuation bias and the ability of participants to self-select into treatments over time. Careful experimentation over long time horizons is necessary to analyze the persistence of program interventions in specific contexts.

3.4 Collaborating with government agencies and other partner organizations

Some of the most well-known field experiments in environmental economics have been conducted in collaboration with partner organizations such as public utilities (see, for example, Allcott, 2011). Working with government agencies and non-governmental organization (NGO) partners allows researchers to test new approaches to program delivery and determine how participant behavior is likely to change in specific decision spaces. Partnerships are also advantageous because the agencies and organizations likely possess baseline data on participants' behavior under the status quo that allow researchers to test new policy approaches using carefully designed controls and observe how their behaviors change in response to the treatments. In addition, as we discuss in Section 4 (Issue 4), it is time-consuming, difficult, and expensive for academic researchers to recruit agricultural producers for experiments. By partnering with organizations in contact with hundreds (or thousands) of farmers, academic researchers can use those pools to quickly and easily recruit participants, potentially eliminating the need to do any recruiting of their own.

Glennerster (2017) describes several attributes of an ideal partner for conducting field experiments. These attributes include having the capacity to operate at the scale necessary for the program, the technical competence and expertise required to successfully implement the program, a strong reputation, low staff turnover, flexibility, and a desire to know the truth. This is a lofty list of attributes, and it is often difficult to find a partner organization that checks all of these boxes. Our experience suggests that low staff turnover, flexibility, and having a strong desire to improve their program, even if research reveals harsh truths, are key for conducting experiments with partners in the agri-environmental context.

Partnering with organizations introduces factors that are beyond researchers' control and can threaten the viability of a study and the generalizability of results. For example, a partner organization's priorities and/or leadership can change, leading to a reduction in or even abandonment of projects that were once a high priority for the organization. For instance, in the United States, CBEAR was originally established during the Obama administration, was renewed during the Trump administration, and has continued to operate during the Biden administration. Throughout that period, priorities and even the structure of the federal government underwent large changes with interest in environmental protection waxing and waning. Changes in staff and leadership also can affect how much active support there is for a project in the organization. And those problems can be amplified when approval processes involve individuals who are political appointees; appointments change with the administration and program priorities tend to change as well.

One frequently noted disadvantage of partnering with agencies and organizations relates to oversight of the research. Collaboration often requires allowing program staff members to review and approve research plans and budgets, which can improve the research but also invite critiques, introduce perverse incentives from outside organizations, and (usually) extend project timelines, often adding years to the research process. Likewise, some organizations will want to develop legal documents such as memorandums of understanding to formalize the relationship. Such documents make sense in terms of protecting the privacy of the data but also tend to be time-consuming to develop and require various reviews, which can complicate the process for researchers when the documents also require authorization from university legal officials.

When working with a government agency, such as USDA's Economic Research Service, researchers typically must obtain approval from agency managers through a "gateway" process that includes a full review by the Office of Management and Budget. The approval processes tend to be quite time-consuming, often requiring more than a year to complete and potentially creating conflicts with other timelines (e.g., grant deadlines and job performance evaluations) that require researchers to demonstrate progress. Furthermore, as the number and difficulty of required approvals increases, it becomes increasingly difficult to satisfy everyone and the time required to obtain so many approvals can grow and grow. However, with the right project partners, the rewards for overcoming these difficulties can certainly be worth the extra effort.

4 Contemporary issues, best practices, and recommendations

Conducting quality research is always demanding. In this section, we highlight contemporary challenges researchers must tackle to produce credible findings and emphasize how those challenges are particularly salient for experimental economists. We also highlight issues that are critical for agri-environmental research, including recruitment challenges and detecting heterogeneous treatment effects. After introducing each challenge, we suggest best practices (drawn from the literature and our own experiences) for overcoming the challenge to produce high-quality experimental and behavioral economic research. The five issues we highlight do not comprise an exhaustive list of the challenges researchers will face when designing and implementing economic experiments on agri-environmental topics. That said, we believe these issues are sufficiently important to warrant focused attention so that researchers can use best practices to carefully plan experimental studies that can provide credible information to further advancing the evidence on how best to design agri-environmental programs and policies.

4.1 Issue 1: Replicability crisis in the social sciences

Credible scientific knowledge serves as the foundation of evidence-based policies and programs. Questions about the integrity of this knowledge base damage researchers' reputations and leave policymakers to search for reliable information. Several studies have demonstrated that results of social science studies could not be replicated, leading researchers to question the validity of the published findings (Camerer et al., 2018; Open Science Collaboration, 2015), including experimental economics (Camerer et al., 2016) and environmental and resource economics (Ferraro & Shukla, 2020). Camerer et al. (2018) attempted to replicate 21 social science experiments published in two journals—*Nature* and *Science*—and found that false positives and inflated effect sizes contributed to failed replication for 8 of the studies (38%) and that replicated effect sizes were 75% of the original effect size in 13 replicated studies. As in other scientific disciplines, the ability to replicate the results of agri-environmental experiments provides confidence that the original findings are robust, and therefore valuable, sources of information that can be used to inform policy and program designs. Ferraro and Shukla (2020) emphasized that a replicability crisis in environmental and resource economics will damage the reputation of the field as a source of credible, unbiased research and raise questions about the value of the research for evidence-based policymaking.

The ongoing replicability crisis has been fueled by numerous factors, including so-called questionable research practices, which include underpowered designs, strategic sampling, selective exclusion of data points, multiple comparisons, and selective reporting of results (Ferraro & Shukla, 2020), and overtly deceptive practices such as p-hacking.[8] Publication bias also plays an important role by censoring the kinds of results that are published and creating an incentive for researchers to take actions that will increase the likelihood of their studies being published. We further discuss the issues of underpowered designs and publication bias later in this section (these are Issues 2 and 3, respectively), and we suggest best practices for addressing these issues specifically. Here, we focus first on two best practices that address the replicability crisis by reducing the use of questionable research practices and making studies easier to replicate: (1) writing pre-analysis plans and pre-registering experiment designs and (2) designing and presenting economic experiments with replication in mind.

[8] p-Hacking generally refers to misleading efforts by some researchers who repeatedly analyze data and/or subsets of the data to obtain results that meet traditional thresholds for statistical significance (*P* values less than 0.05), thus dramatically increasing the potential for false positives.

4.1.1 Best Practice A: Pre-analysis plans and pre-registration of experiment designs

Pre-analysis plans are documentation generated before beginning to collect or analyze data that describe how a study will be conducted, including the plan for empirical analysis. For experimental economics studies, pre-analysis plans describe the research hypotheses; experiment design, including all treatments and controls; power calculations; characteristics of people who will be sampled; sample sizes; and methods for analyzing the data, including construction of the key variables, procedures for cleaning the data, and strategies for estimations (Janzen & Michler, 2021; Olken, 2015).

Creation of a pre-analysis plan can be inherently valuable as a research planning tool. It is also important as a public commitment device when it is posted on a public registry before the start of data collection. Published pre-analysis plans bind researchers to the established protocols and data analysis strategy regardless of the results, reducing the incentive to use questionable research practices that would increase the likelihood of finding significant results. Published plans also increase research transparency by clearly stipulating the original aims of experiments and requiring researchers to be forthright about results that were discovered during exploratory analyses of experiment data. Additionally, registries help address the file drawer problem by creating a record of tested hypotheses and generating an incentive for researchers to report the results regardless of their statistical significance.

To be clear, we (along with many other researchers) are not suggesting that pre-analysis plans should restrict exploratory analyses. Such analyses often generate interesting insights and new hypotheses that can be tested in future work. Pre-analysis plans require that researchers be transparent about whether results came from planned analyses testing their original hypotheses or were generated during exploratory analyses. To make these distinctions clear to readers, we encourage authors to clearly label each type of results in the result section of their papers.

Registration of pre-analysis plans is not universally required of economic researchers, but journals are becoming increasingly strict about registration requirements for certain types of studies, including RCTs.[9] The American Economic Association (AEA) operates a registry of RCTs (www.socialscienceregistry.org), and researchers must have registered a pre-analysis plan to submit papers to journals under the AEA umbrella. Currently, the AEA registry is set up primarily for RCTs. We encourage AEA to update its registry to better accommodate pre-analysis plans for laboratory and field experiments and even non-experimental studies, as the value of doing pre-analysis plans certainly extends to research other than RCTs. Other public registries for pre-analysis plans include Open Science Framework (https://osf.io/registries) and AsPredicted (https://aspredicted.org/).

[9] Registering pre-analysis plans is required in other fields. For example, clinical trials regulated by the U.S. Food and Drug Administration must be pre-registered at a site managed by the National Institutes of Health (NIH) National Library of Medicine (*ClinicalTrials.gov*).

In addition to their "public" benefits, pre-analysis plans can provide private benefits to researchers. Olken (2015) suggested that registered plans can streamline the analysis and presentation of data by requiring researchers to commit to specific steps in advance. Janzen and Michler (2021) proposed that the process of developing a pre-analysis plan is inherently valuable to researchers because it requires them to proactively think about their approaches and methods and about timelines and constraints associated with the design and to document the research design. The study noted that developing a pre-analysis plan is particularly valuable when working as part of a team in which members are responsible for portions of the project and all must acknowledge and agree to the plan before finalizing the research design and beginning data collection. Registered plans also provide opportunities for feedback prior to data collection that can prevent costly mistakes and oversights that would be difficult or impossible to correct later.

Some economic researchers have pushed back against requiring registered pre-analysis plans. They typically do not believe that the use of inappropriate research practices is a significant problem and view pre-analysis plans as time-consuming activities that overly restrict their freedom as researchers (Coffman & Niederle, 2015). However, several leading experimental researchers have argued that the benefits of pre-analysis plans outweigh the costs and they have emphasized that the plans do not have to be overly burdensome. Banerjee et al. (2020) suggested that the plans can be short documents that outline the general research plan, thus allowing for flexibility. Additionally, Janzen and Michler (2021) pointed out that many components of pre-analysis plans are developed when drafting grant proposals to fund the research thus the time required to develop and post the pre-analysis plan may be less than expected. We further note that, for many experiments, the design of the instructions and protocols is already a very deliberative process as many decisions are being made about how to set up the situation to test treatments of interest. Thus, documenting these decisions and how the data will be analyzed in a pre-analysis plan does not need to be too time consuming.

Since most experiments involve treatments set by the researchers, it should not be burdensome to develop pre-analysis plans that describe why the treatments were selected, the research questions they were designed to answer, and the statistical approaches that will be used to test the hypotheses. Registering a concise plan generates value by motivating research transparency and creating an incentive to communicate statistically significant and insignificant results, which could counter publication bias (discussed in Issue 3). Additionally, if editors and reviewers begin to view a pre-analysis plan as a critical element of a well-designed and properly powered experiment, registered plans should facilitate ongoing publication of studies in influential journals even when results from key treatments are not statistically significant. Clearly, registered pre-analysis plans involve upfront costs for researchers, but we and other proponents of the plans believe that the benefits greatly outweigh the costs.[10]

[10] See Janzen and Michler (2021) for a comprehensive review of the history, format, and debate regarding pre-analysis plans.

4.1.2 Best Practice B: Design and present studies with replication in mind

Replicability increases the value and reliability of results from experimental economics studies. More than 30 years ago, Smith (1982) suggested that progress in the discipline depended on replicability of experiments, and in the wake of scrutiny recently about the credibility of social science discoveries, researchers are increasingly calling for more replication studies (Coffman, Niederle, & Wilson, 2017; Czibor et al., 2019). To support these efforts, researchers can design and present studies with replication in mind.

Replication studies can take several forms, and researchers can design and present research to support each form (see Hamermesh, 2007 for a detailed overview of the replication types highlighted in this section). The most basic form, *pure replication*, involves reanalyzing the original data using the original models to confirm the results. Pure replication can uncover errors associated with data cleaning, calculations, and coding. *Statistical replication* reruns experiments with a second sample of the original population and analyzes the new data using the original model. It typically is applied to address questions about sampling errors and to studies that had weak statistical power. The broadest form of replication, *scientific replication*, involves running the same experiment in terms of overall goals and design with a sample from a different population and using the same or different models to analyze the results. It is used to test whether the original study's findings are robust and generalizable.

Furthermore, experimental economists can replicate their own studies by rerunning the experiments with new participant pools. Within-study replications can make valuable contributions by testing whether generalized findings from a study hold in different contexts and with participants from different populations.

Friedman and Sunder (1994) outlined four types of records experimental researchers should keep to ensure replicability: (1) written instructions for participants and details of the recruitment process; (2) copies of the software and hardware used, when applicable, to allow them to be made available to others to replicate the experiment; (3) documentation of the laboratory activities in a log that includes dates, times, and particulars of the activities (including copies of the instructions) and copies of the raw data; and (4) a record and copies of statistical programs and code used to analyze the data. Consequently, researchers can proactively promote transparency and facilitate replication by publicly archiving the data and the coding used to construct the variables and clean and analyze the data.

Top economics and applied economics journals already require publishing of the data and code with papers, but many agricultural, resource, and environmental economics journals do not require the data to be published or do not enforce the requirement (Lybbert & Buccola, 2021). We expect that publication of data and coding with papers will become more common in the near future. In addition, an increasing number of economics journals submit results presented in papers to data editors who conduct pure replications before the papers are published. Some universities have begun to offer free replication services to support replicability and transparency.[11]

Another step researchers can take to facilitate replicability is to conduct sufficiently powered studies, which we discuss in greater detail in the next section. Adequate statistical power is essential for replication. When an original study and its replication rely on small samples, variations in the samples can be amplified, leading to different results, and a lack of adequate statistical power always raises questions about the reliability of findings.

4.2 Issue 2: Challenges presented by underpowered studies

Statistical power refers to the probability that a false null hypothesis will be rejected at a given significance threshold (Ellis, 2010). Researchers who fail to reject a false null hypothesis commit Type II errors known as false negatives. The probability of a Type II error typically is denoted as β and power is denoted as 1–β. In experiments, statistical power is influenced by the design of the study, how strongly treatments influence participant behavior, the number of participants, and how the data will be analyzed, including the type and number of hypotheses tested. It is important to remember that power is not a data analysis tool and is not related to causal inference (Vasilaky & Brock, 2020).

Lack of statistical power has plagued numerous types of research and a wide variety of methods, including economic experiments (Smaldino & McElreath, 2016; Zhang & Ortmann, 2013). In terms of statistical power, the rule of thumb is to design studies to achieve 80% power or better (i.e., the probability of not making a Type II error is 80% or greater). However, a survey of the economics literature (including experimental and non-experimental studies) suggests that the median statistical power of published studies is less than 20% (Ioannidis, Stanley, & Doucouliagos, 2017).

A lack of statistical power can produce misleading experiment results, a particularly significant drawback when experiments are used to guide policymaking and program design. The danger associated with using underpowered experiments is twofold. First, low power increases the probability that the researcher will fail to identify the effect of a tested treatment (a Type II error). Second, low power increases the likelihood that a statistically significant effect will be exaggerated (a Type M error) or even have the wrong sign (a Type S error) (Button et al., 2013; Gelman & Carlin, 2014). Low power is also a problem when testing multiple hypotheses (multiple treatments), treatment effects for multiple subgroups, and multiple outcomes (Vasilaky & Brock, 2020).

It is important to understand that concerns about insufficiently powered studies are closely related to the replicability crisis (Issue 1) and publication bias (Issue 3).

[11] For example, Cornell University's Cornell Institute for Social and Economic Research (CISER) offers a service called Results Reproduction (R-Squared) that is described as *enhanced proofreading for your Data and Code* (https://socialsciences.cornell.edu/research-support/R-squared).

Specifically, Type M and S errors are less problematic when the results of all studies, including replications, are published so researchers can draw conclusions about the "true" results. When journals are likely to publish only statistically significant results and experiments are rarely replicated, it is difficult to identify exaggerated estimates and incorrect null effects in prior studies.

In terms of agri-environmental decision-making, underpowered designs are particularly prevalent in field experiments investigating behavioral nudges because true effect sizes tend to be relatively small and can only be detected by using large samples. When a study is underpowered, the most likely inference is that the intervention was either unsuccessful or had an exaggerated effect even when the true effect is positive. The incorrect inferences can thwart future refinements of interventions, encourage adoption of interventions that will have little or no effect, and, sadly, prevent widespread adoption of interventions that are actually successful (especially ones that have relatively small effects but are extremely cost-effective, such as nudges, which often cost essentially nothing to implement). The consequence of incorrect inferences combined with publication bias and lack of replication is potential misallocation of scarce agri-environmental resources. To identify underpowered designs in advance, researchers should conduct power analyses before implementing experiments. See Bellemare, Bissonnette, and Kröger (2016), Ellis (2010), List, Sadoff, and Wagner (2011), and Vasilaky and Brock (2020) for guidance on conducting power analyses for experimental studies.[12]

Some researchers have argued that reporting results from underpowered studies provides helpful (but imperfect) results that can inform policymaking via their inclusion in meta-analyses. We recommend caution in this regard. Meta-analyses generally draw from the published literature and thus generally can only estimate true effects by assuming that all underpowered studies are equally likely to be published. However, editors and reviewers currently tend to favor underpowered studies that show statistically significant results to ones that show null results or results that are counter to the current wisdom. This publication bias is likely leading to overestimates of treatment effects and this impact can also bias the results of meta-analyses.

4.2.1 Best Practice C: Conduct statistical power analyses

Conducting an *ex ante* statistical power analysis informs the design of an experiment and the sampling strategy. By considering statistical power, researchers can increase the efficiency of their experimental designs and avoid using samples that are too small to detect effects (plus avoid costly oversampling) (Vasilaky & Brock, 2020). Using a power analysis to guide sample size reduces use of misguided practices such as endogenously choosing sample sizes based on initial results from a handful of experiment sessions and continuing to increase the sample size until a statistically significant result is found. Both practices are bad science and increase the

[12] Janzen and Michler (2021) emphasized that researchers should justify their alpha levels (statistical significance levels) along with other decisions they make when designing a study.

likelihood that the study will generate results that are not credible. Stopping or continuing studies based on the presence or absence of statistical significance has long been recognized as a source of bias (Armitage, McPherson, & Rowe, 1969) and can introduce Type I errors by rejecting true null hypotheses. Consequently, adequate sample sizes should always be determined before conducting experiments.

Sufficiently powered studies also make null results more convincing because researchers can demonstrate that the study was adequately powered to detect an effect of a certain size if it existed. For agri-environmental policies, one can argue that *economically significant* results are needed—that the effect is sufficiently large to justify new actions and policies. Researchers who conduct power analyses, which are part of publicly available pre-analysis plans, can point to the power analyses and the magnitudes of the effects the studies were designed to detect. In addition, rather than evaluating papers based on the magnitude and significance of the results, editors and reviewers should reward authors for reporting *ex ante* power analyses in their pre-analysis plans and evaluate the quality of the experimental designs and the importance of the questions asked.

Statistical power depends on many factors, including sample sizes, significance levels (α), distributions of outcome variables, minimum detectable effect sizes desired by the researchers, and testing procedures employed in the experiments (e.g., econometric models vs simple parametric and non-parametric tests), including the types and numbers of hypotheses to be tested. Other facets of the design, such as measures of repeat decisions from the same observational units (individuals or groups) and choices about treatment randomization (within-subject or between-subject) also need to be considered. Bellemare et al. (2016) and Vasilaky and Brock (2020) provide useful resources and examples of power analyses specifically designed for economic experiments.

The primary challenge associated with power analysis is specifying the minimum detectable effect size and expected outcome distributions associated with treatment and control conditions. One approach to identifying that information is to collect data from the population of interest via a pilot study. Pilot studies are useful for other reasons as well, such as refining an experiment's parameters, procedures, software, and information materials. In addition to pilot studies, researchers can seek related studies and meta-analyses on the study topic to guide selection of standardized effect sizes to use in power calculations. However, we caution researchers to interpret published effect sizes with caution because challenges with underpowered designs and publication bias can lead to publication of exaggerated effect sizes. Field studies of agri-environmental issues have often found effect sizes of 0.10 standard deviations or less (Palm-Forster, Ferraro, et al., 2019); therefore, we suggest that researchers use conservative effect-size estimates when conducting *ex ante* power analyses.

Conducting power analyses and reporting the power of studies are two concrete steps researchers can take to increase the credibility and enable replication of their findings. That said, power calculations are valuable only if researchers follow their guidance, which is not without challenges. First, a power analysis can indicate that the experiment design or recruitment plan needs to be modified, which requires time and effort. Second, larger sample sizes typically involve higher recruitment costs and a significantly larger budget for participation fees, raising the cost of the study. If resources are not available to recruit a larger sample, increasing the power of the study can be accomplished by using a simpler experiment design, which could limit the number of treatments tested. Third, as previously noted, researchers need to be transparent in subsequent write-ups about which of their hypotheses were part of the original design for which a power analysis was calculated and which hypotheses were generated later. As previously noted in the section on pre-analysis plans (Best Practice A), our suggestions are not meant to devalue research discretion or exploratory analyses. Instead, such results should be reported transparently and conclusions from exploratory analysis should come with the caveat that they are more *suggestive* than definitive.

4.2.2 Best Practice D: Report standardized effect sizes

We recommend that authors publish standardized effect sizes when describing results of a study, which will allow for comparison of the magnitudes of estimated treatment effects across treatments and outcomes in papers and to treatment effects estimated in related studies. Published standardized effect sizes also provide valuable information for researchers seeking strong priors when selecting the minimum detectable effect size needed for an *ex ante* power analysis. However, we caution researchers to be wary of selecting experiment designs based solely on large anticipated effect sizes since subsequent power analyses could indicate that the sample sizes are too small to detect the true (smaller) effects.

Ellis (2010) and Thalheimer and Cook (2002) provide guidance on how to calculate standardized effect sizes. Cohen's *d* (Cohen, 1988) is commonly used. It is calculated by computing the difference between the mean outcome of the control group and the mean outcome of the treatment group (i.e., the treatment effect) and dividing that difference by the pooled standard deviation for the two groups.[13] For example, if the outcome variables for the control and treatment groups are 4.5 and 6.5, respectively, and the pooled standard deviation is 5.0, the standardized effect size is 0.40—equivalent to a change in the outcome variable of 0.40 standard deviations in magnitude.

4.2.3 Best Practice E: Correct for multiple hypothesis testing

Economic studies, including experiments, commonly test multiple hypotheses. Multiple hypothesis testing (MHT) can arise from testing the effects of multiple treatments on a specific outcome, the effects of one or more treatments on multiple outcomes, and heterogeneous subgroup treatment effects. In a review of 34 studies involving field experiments published in top economic journals,

[13] Other common formulas for calculating standardized effect sizes include Glass's delta and Hedge's *g* (Ellis, 2010).

Fink, McConnell, and Vollmer (2014) found that 76% of the studies included subgroup analyses and 29% reported estimated treatment effects for ten or more subgroups.

Testing multiple hypotheses is a common practice, and it is not inherently wrong. However, if problems associated with it are not acknowledged and addressed, MHT can generate findings that are not robust. Two primary concerns arise with MHT, one related to statistics and the other to researcher behavior. Based solely on the statistical properties of hypothesis testing, the more hypotheses we test, the more likely we are to spuriously reject a null hypothesis (Type I error) (List, Shaikh, & Xu, 2019). In other words, with MHT, we worry about the probability that a P value will fall below the critical threshold (e.g., $\alpha = 0.05$) and a null hypothesis will be rejected even though it is true. The second related concern, about researcher behavior, arises when there is an incentive to search for statistically significant results and researchers test hypotheses until they obtain significant results. It also can occur when researchers attempt to address reviewer requests to test for heterogeneous treatment effects, as sometimes occurs for papers on experimental studies.

Correcting for MHT is common in other disciplines but has not been widely used in economics. That said, economic researchers are recognizing the problems associated with MHT, and many are calling for solutions to address the problems (Christensen & Miguel, 2018; List et al., 2019; Olken, 2015). Correcting MHT essentially consists of adjusting the critical values (P values) used for inference. One strategy is to control the false discovery rate—the expected proportion of false positives among rejected hypotheses—often with Bonferroni-type corrections. Procedures for controlling false discovery rates are described by Benjamini and Hochberg (1995), Benjamini and Yekutieli (2001), Yekutieli (2008), and Xie, Cai, Maris, and Li (2011). In short, the approaches revise the rejection rules for hypothesis testing (e.g., they use alternatives to "reject if $P < 0.05$") based on the desired false discovery rate, the number of hypotheses to be tested, and, typically, a set of P values from prior individual tests. One can also account for MHT by controlling the family-wise error rate: the probability of incorrectly rejecting even one null hypothesis (of committing at least one Type I error). The Romano-Wolf procedure (Clarke, Romano, & Wolf, 2020; Romano & Wolf, 2010) is an approach for controlling family-wise error rates that may offer greater statistical power than Bonferroni-type procedures. Applications of the Romano-Wolf procedure to experimental economics were demonstrated by List et al. (2019).

Plans to conduct MHT should be taken into account when conducting *ex ante* power analyses (Czibor et al., 2019). To adjust power analyses for MHT, researchers can specify lower significance levels (α) for the planned hypothesis tests. Remember that, by specifying lower α levels, the sample size required for a study to be sufficiently powered will increase for all elements of the experiment. As previously noted, larger sample sizes increase the cost of the study, highlighting what Czibor et al. (2019) called the "hidden costs" of testing additional hypotheses related to an outcome, treatment, or subgroup.

4.3 Issue 3: Publication bias

Publication bias occurs when the statistical significance of results obtained in a study determines how likely the research is to be published (Brodeur, Cook, & Heyes, 2020; Doucouliagos & Stanley, 2013). Researchers are generally aware of this bias, and most have likely made some judgment calls in their own research regarding how to respond and whether to frame papers strategically around statistically significant results discovered when analyzing the data. The desire for statistically significant results can tempt researchers to engage in unethical practices such as p-hacking, prioritizing specifications that generate statistically significant results, and not attempting to publish studies with null results (i.e., the file drawer problem). In a review of 50,000 tests presented in studies in top economics journals, Brodeur, Lé, Sangnier, and Zylberberg (2016) found strange patterns in the *P* value distributions around 0.05 and 0.10, suggesting that presence of practices such as p-hacking and selecting specifications for statistical significance.

Publication bias is not perpetuated solely by editors and reviewers; researchers and readers contribute to it as well. In general, researchers want to publish, review, and read about new, exciting, statistically significant results in high-ranking academic journals. This desire creates little space for publishing null results and results of replication studies. Thus, often these studies are never published or are published in lower-ranked journals. Furthermore, researchers in many economic departments and some top-ranked agricultural economics programs have been reluctant to publish in lower-ranked journals because some misguided colleagues and department chairs view it as a signal of a lower quality research program. Consequently, many papers that could usefully inform policymaking and contribute to general academic knowledge are stuck in desk drawers or buried on hard drives, eroding the quality of the science. And since much of the research was funded by public agencies, it is unethical not to communicate those findings. The public deserves to have access to all of the results of its scientific investments, assuming that the studies were well designed and executed.

4.3.1 Best Practice F: Publish replication studies

Currently, the incentive to publish replication studies is small while the cost of replication studies can be large (Coffman et al., 2017) so it can be difficult to find published replication studies. Coffman et al. (2017) further point out that replication attempts often are hidden in the original studies, in which replication served as the baseline or was reported as a secondary result.

To provide a greater incentive to publish replications, scholars are calling on journals to publish the results of replication studies, create special replication sections, and consider establishing new journals dedicated to replications (Coffman et al., 2017; Czibor et al., 2019). Coffman et al. (2017) has further called on researchers to cite replication attempts with citations to the original studies. In a world in which citations signal research credibility and build scholars' academic reputations, citing replicative studies would provide a tangible payoff for researchers who attempt to replicate results presented in influential papers. Replicative studies would be easier to find if journals explicitly published them as replications instead of requiring the replications to be embedded in the original studies. Additionally, journals specializing in replications could encourage greater research transparency and accountability. Scholars would know that their studies are likely to be replicated and that the replication attempts would be published and cited alongside the original studies. Other novel ideas to enhance replication incentives are being discussed, including journal editors commissioning replication studies with publication guaranteed subject to peer review (Hamermesh, 2007) and having lead authors offer co-authorship to researchers who replicate their studies before they publish their results (see Butera & List, 2017).

4.3.2 Best Practice G: Publish null results from well-designed studies

Historically, researchers have been rewarded for statistically significant results (e.g., acceptance for publication) so they can be tempted to seek out and highlight statistically significant results. As noted in the section on replicability, professional rewards for statistical significance create perverse incentives that can undermine the scientific process when researchers use questionable practices, including overt manipulation of results and more benign but still problematic practices that bias their estimates. One research practice that is common but problematic is testing multiple hypotheses but reporting *only* statistically significant results. Even well-meaning researchers can fall into the trap of continuing to test hypotheses until they find “interesting” (significant) results. However, those actions contribute to publication bias because valuable information is withheld from the literature when researchers do not report results that do not reject null hypotheses.

Null results from a properly designed study can convey critically important, policy-relevant information about interventions and mechanism designs that likely will not work in certain contexts (see Brown, Lambert, & Wojan, 2019 for a discussion about statistical approaches for interpreting null results). As long as a study is well-designed, it is important to communicate null results, even when they conflict with results in the literature (or hopes of researchers and/or funders). Our scientific knowledge base is not generated by individual papers; it is a product of the entire literature built incrementally by individual papers. Publishing null results makes the literature more complete and allows results to be synthesized through thoughtful reviews and meta-analyses to identify robust findings that can inform the design of improved policies and publicly financed programs in support of a variety of outcomes, including improved agri-environmental performance.

When seeking to publish null results, a proper power analysis and pre-analysis plan become particularly important. As part of the development of a power analysis, we encourage researchers to identify the size of treatment effects that would be “economically significant” to detect. By including this in the pre-analysis plan, it makes the researchers’ assumptions known to the public prior to conducting the study. In an agricultural context, for example, a 3% change in producer behavior (e.g., participation in a BMP cost-share program) may not be economically significant if the program is costly and the policy change is burdensome to farmers and program managers. However, a 3% change in producer behavior could be quite economically significant if the treatment is relatively inexpensive and easy to implement, such as changes in the messenger or messages promoting enrollment in a program. Once the researcher explains how the study had sufficient power to detect an economically significant effect, it is easier to demonstrate the value of null results: though researchers cannot rule out the possibility that a smaller statistically significant effect may be found in a sample large enough to detect it, they can argue that the impact would be small enough to not be economically meaningful in a policy context.

4.4 Issue 4: Participant recruitment

Studies based on revealed-preference experiments raise some unique issues related to recruitment, including convenience and non-response biases. Convenience bias arises from correlation between a recruited participant being eager or willing to participate in an experiment and other characteristics, such as educational background. The effort required to participate in an experiment can result in selection of a subset of the targeted group consisting of people who share certain unobserved characteristics (e.g., strong interest in promotion of a certain policy, intense disdain for pollution of the environment, a belief that the types of programs and regulations in question need to be changed, a strong interest in agriculture and environmental issues, even just a strong desire to support research) and exclusion of people who share different characteristics (e.g., do not feel knowledgeable enough to contribute to the research question, less inclined toward higher education). In addition, people who decline the opportunity to participate in a lab experiment can differ systematically from people who choose to participate, the equivalent to non-response bias in survey-based studies.

4.4.1 Best Practice H: Offer appropriate payments for participation

A fundamental principal in economic theory is that people can be motivated to exert effort by financial and other types of incentives. Economic experiments generally require subjects to complete tasks that are somewhat difficult cognitively (and occasionally physically) and need them to invest sufficiently in the experiment to provide thoughtful, truthful responses. To encourage people to participate in experiments and take the experiments tasks seriously, researchers provide them with some base compensation (a show-up fee) and an opportunity to increase their compensation via their individual performance and/or their group's performance of the task. According to incentive theory, the amount of effort a person will invest in a difficult task increases with the magnitude of the potential personal reward. Thus, the question is how much a researcher needs to pay participants to ensure that they are motivated by the incentive offered in the experiment. That is, how salient is the incentive to the participants?

First, experiments and associated incentives must be salient enough to participants to override their other interests, such as trying to please the experiment administrator, boredom, and/or trying to outcompete other participants. These interests represent a focus on “winning the experiment” rather than on revealing their true valuations or maximizing their own payments regardless of what others earn. The participants essentially ignore the fundamental financial incentive tested in the experiment, jeopardizing the internal validity of the results.

As a rule, the greater the economic reward, the more attention to detail one can expect from participants. By setting a marginal incentive such that participants gain significantly more money from “optimal” and “near optimal” choices than from random choices, researchers can expect that participants will dedicate greater cognitive effort to determining the true optimal choice.

It is important to take participants’ opportunity costs in terms of time and proximity to laboratory and field settings into account when establishing compensation. Generally, undergraduate students will invest significant effort for relatively small rewards. In our experiments, we provide compensation to undergraduate students in the United States of $15 per hour on average.[14] Busy agricultural producers in developed countries, however, require more-substantial compensation for their time and inconvenience. In our experiments involving farmers, the compensation typically is $50 to $100 per hour (sometimes more).

Experimental economists have long focused on the level of a reward required to make it salient, but the academic literature suggests that the link between the level of payment and the observed precision of behavior is less direct than previously thought (Poe, 2016). For example, farmers can be motivated wholly or partly by the opportunity to participate when the subject of the experiment particularly interests them (e.g., cost-share auctions, eliminating invasive species) or when they have personal and/or altruistic desires to improve agri-environmental programs that affect them and/or their peers.

4.4.2 Best Practice I: Successful recruitment of farmers and rural landowners

Recruiting farmers and rural landowners as participants in experiments can be particularly valuable when the target audience for the research is policymakers, regulators, and program managers. Though these professionals often have graduate degrees, they tend to have little training in economics and almost none have training in experimental economics. Thus, when they learn that a study involves undergraduate students as subjects, they are almost immediately concerned about the external validity (generalizability) of the results (Levitt & List, 2009). Federal agricultural policymakers in the United States, on the other hand, are supposed to give greater weight to results of experiment-based economic studies that use farmers as

[14] Of course, in our experiments, a participant’s choices and the general design and outcome of the experiment (through group behavior or random assignment) determines the actual payoff for the participant.

subjects since passage of the Evidence Act (Rosch et al., 2021). The appeal of recruiting farmers and landowners to participate in experiments is obvious but numerous recruitment challenges remain.

One of the most serious challenges for economic field experiments is limited access to farmers. In the United States, agricultural producers have been declining in number for decades, their average age has been increasing (meaning they could be uncomfortable with the advanced computer equipment used in experiments), and the number of farms operated by non-owners has been growing. So, at a time when the number of "farmers" is decreasing, experiments designed to have significant statistical power are requiring larger samples in general and much larger samples when testing heterogeneous treatment effects, which appeal to policymakers and funders interested in how farmer behavior differs across socio-economic characteristics. In other words, to have sufficiently powered studies, researchers find it can be prohibitively expensive in terms of time and funds to recruit the number of farmers needed even for basic single-treatment experiments. A further complication is that the most common methods of recruiting farmers, such as in-person at agricultural trade shows and conventions and by mail, are likely to suffer from significant self-selection.

Further complicating recruitment of farmers is the relatively small number of them being recruited by a relatively large number of academic and government research projects. There is growing demand for their time and attention, and multiple competing research queries can reduce response rates as farmers begin to experience research fatigue. McCarthy and Beckler (2000), for instance, found that concerns about data privacy and time limits had significant impacts on whether farmers would agree to participate in a research survey. Such research fatigue is evident in the rate of responses to the annual Acreage and Production Survey conducted by USDA's National Agricultural Statistics Service. The survey is widely followed; it provides key information on the planting intentions of farmers and provides data used to estimate the U.S. crop supply, which is closely followed by financial markets, the agricultural sector, and government agencies. Despite the importance of the survey, response rates have fallen from 80% to 85% in the early 1990s to just 57% to 67% in 2016 (Johansson, Effland, & Coble, 2017).

A summary of recruiting strategies for large field experiments involving farmers by Weigel et al. (2021) provided sobering results. They evaluated 10 strategies used to recruit U.S. farmers for two large-scale field experiments that offered relatively large financial incentives. Despite the numerous strategies tested, the recruitment rates were quite low. Of 25,616 farmers contacted, the overall response rate for the two participant pools collectively was just 2% (4% of the 9960 farmers in the first experiment and just 0.7% of the 15,656 farmers in the second). The authors also report that *none* of the farmers who were contacted by email responded and that the response rate increased to only 2.4% when that group was excluded from the calculations. Their study showed that larger monetary incentives and sending messages reminding farmers to participate were effective but expensive recruiting strategies. Costless strategies, such as prominently citing a well-known institution as a sponsor of the research, also had small positive effects on recruitment.

The good news from Weigel et al. (2021) is that using the correct mix of recruitment strategies can lead to substantial cost savings. They found, for example, that the total cost of successfully recruiting a farmer using a mix of cost-effective strategies was 31–32% lower than the total cost of the least successful strategies. Thus, for a given population of farmers, the most cost-effective strategies recruited 133–208% more farmers as research participants than the least cost-effective strategies. Also encouraging is their finding that several costless strategies such as revised messaging increased rates of response relative to the status quo language. To assist researchers in determining the best strategies for recruiting farmers, Weigel et al. (2021) provide an Excel workbook as part of the paper's supplemental materials. Researchers can enter various key assumptions about the desired population and sample size and compare the cost and efficacy of each recruitment approach. The workbook can also be used to identify the potential cost-effectiveness of a recruitment strategy.

In their closing, Weigel et al. (2021) noted that economic field experiments involving U.S. farmers are likely to struggle to recruit adequate sample sizes, at least in situations in which the recruitment solicitations come from academics. To improve those response rates, they suggest embedding the experiments in new and existing federal programs—in other words, by partnering and collaborating with governmental and non-governmental agencies as described in Stage IV experiments.

This approach is likely to be particularly effective since federal programs generally reach far greater numbers of farmers, allowing experiments to generate externally valid insights cost-effectively as long as the experiments use proper methods (e.g., randomization) to establish meaningful comparison groups. However, embedding research into agri-environmental programs requires a cooperative partnership with a governmental or non-governmental organization in which staff members are truly interested in how to make their programs as effective as possible. These partnerships can be difficult for relatively inexperienced researchers to develop since they require a great deal of trust between the researcher and leaders of the organization.

Given these recruitment challenges, researchers (and funders) likely will have to change their expectations regarding recruiting farmer samples for experiments and recognize that field experiments are likely to require larger monetary incentives and greater effort to recruit adequate samples. For instance, to conduct a well-powered experiment involving growers of a specific crop that could involve heterogeneous treatment effects, researchers will need to collaborate with professionals who have good relationships with producers across a broad region, such as extension specialists and industry representatives.

4.5 Issue 5: Detecting heterogeneous treatment effects

Agricultural producers and rural landowners are diverse, as are their business structures, practices, household compositions, needs, and preferences. While measurements of average treatment effects can be quite useful for assessing a program's overall benefits, policymakers (and, frankly, reviewers and editors of academic journals and research funders) are frequently interested in whether and how policy effects are heterogeneous and whether there is variation across individual farmers. For example, questions inevitably arise about how an agri-environmental policy or program affects farms based on the farms' sizes, corporate/family structures, and whether they are owner-operated or rented. Identifying heterogeneous treatment effects can be particularly important when modeling the impacts of policies on specific sub-populations such as family farming operations, women landowners, beginning farmers, owners of ecologically sensitive lands, and socially disadvantaged farmers.

When testing the effects of a potential program, researchers frequently use the results of economic experiments to identify heterogeneous treatment effects in subgroups of targeted populations (Heckman & Vytlacil, 2001). Thus, participants in laboratory and field experiments complete simple post-experiment surveys that collect their demographic information. However, experimental economics has not established any standards regarding which socio-demographic characteristics to collect and which characteristics to compare in academic papers, creating two data challenges. First, researchers do not always collect important demographic characteristics as part of experiments. This is particularly true when pre-analysis plans do not specify the characteristics to be analyzed. The second challenge is that researchers do not always conduct statistical tests of interactions between the characteristics for which they collect data and the treatment effects. Though this problem can be a natural consequence of low-power studies, it limits subsequent researchers' ability to draw broad conclusions from the literature. These missing-data problems make it difficult, if not impossible, to infer how commonly heterogeneous treatment effects are found in agri-environmental experiments. They also affect future research because the lack of information makes it difficult for future researchers to identify participant characteristics that are vital to collect.

Rosch et al. (2021) reviewed 83 economic experiments in which farmers, landowners, fishers, and ranchers were recruited as participants to assess the extent to which heterogeneous treatment effects were used in experiments related to agricultural and natural resources. For each study, the authors recorded whether tests of correlation between the treatment effects and demographic (or farm) characteristics were reported and whether any reported correlations were statistically significant. They found that one-third of the papers did not include reports of correlation. Of the two-thirds of the papers that reported at least one correlation, a majority found that some of the correlations were statistically significant. The characteristics most frequently tested—age, education, and gender—were found to be correlated with treatment effects 55%, 52%, and 40% of the time, respectively.

Rosch et al.'s (2021) review further notes that the study results likely mischaracterize the actual extent of the correlations. In general, experiments designed to have enough statistical power to detect an average treatment effect are not adequately powered to detect potential differential treatment effects for subgroups of a sample pool.[15] In the review, they explain that the general intuition associated with this problem is that the process of estimating treatment effects for multiple subgroups decreases the size of the samples making up the subgroups, sometimes dramatically, thereby reducing the statistical power of the analyses.

4.5.1 Best Practice J: Stratified and blocked randomized designs

Using stratification and block randomization in the experimental protocol can mitigate heterogeneous treatment effects that otherwise would add variance to the data and thus further reduce statistical power (Duflo, Glennerster, & Kremer, 2007). Rosch et al. (2021) note that:

> *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. (p. 10)*

To better understand why detecting true treatment effects is especially important for policymaking, consider an example stylized from Rosch et al. (2021) in which researchers develop a field experiment to measure the extent to which subsidizing the cost of various conservation practices affects farmers' willingness to adopt the practices and persist in using them. The researchers could hypothesize that inexperienced farmers would be more likely than experienced farmers to adopt a conservation practice initially and then persist in using it in response to increases in the proportion of the cost reimbursed by the program. In this case, two strata are needed in the experiment design: inexperienced farmers and experienced farmers.

Next, consider the situation in which the standard deviation of responses to the treatment by inexperienced farmers is twice the magnitude of the standard deviation of responses by experienced farmers. In this case, to have equally powered samples for treated and control participants in each stratum and reliably detect treatment effects for the strata, the researchers must sample four times as many inexperienced farmers as experienced farmers because the sample size must scale with the square of the standard deviation.[16]

So, to identify differences between two strata, the experiment design requires recruitment of significantly more participants than are required to test only the overall treatment effects. Returning to the first example, assume that the standardized treatment effects are 1.5 for inexperienced farmers and 1.0 for experienced farmers.

[15] Brookes et al. (2004) provide a useful simulation and discussion of this problem.

[16] Athey and Imbens (2017) provide helpful examples and formulas that can be used to calculate the power needed for various treatment effects.

In this case, to pool the groups, the experiment must have sufficient power to identify a treatment effect in the magnitude of 1.0 to 1.5. However, to detect differences in treatment effects between the subgroups, the experiment must be sufficiently powered to detect a difference of just 0.5–less than half the magnitude required for the pooled treatment effect. This extra dimensionality requires recruitment of four times as many farmers!

It is important to take challenges arising from participant self-selection into experiments into account. Self-selection can decrease the representativeness of subjects in strata that are already under-represented; in an agricultural context, this could be minority and women farmers. To achieve the needed statistical power for various subgroups, recruitment efforts need to pay attention to those characteristics. Simply adding subjects to a well-represented strata and thereby increasing the total number of participants in the study generally will not fully compensate for the lack of subjects in under-represented strata. At times, it will not be possible to recruit enough participants from under-represented strata using conventional techniques.

So when a field experiment is sufficiently powered to test the effects of treatments on the overall sample but not on subgroups, should researchers report subgroup results at all? Some have advocated not reporting those results under any circumstances, but we believe it is acceptable to report subgroup results as long as the authors are clear about the lack of power in the design and that the results should be seen more as suggestive than definitive. We recommend labeling these types of results as suggestive even when statistical significance of the effects of the treatments for subgroups appears to be quite large since small samples can yield large measured treatment effects that are not robust to replication with larger samples.

4.5.2 Best Practice K: Standardize collection and reporting of demographic data

Rosch et al. (2021) recommend collecting a standard set of demographic and farm characteristics in all experiments designed to inform agricultural and environmental policies and testing them for correlation:

(1) Age
(2) Education
(3) Gender
(4) Land size
(5) Experience with farming/fishing/ranching
(6) Wealth
(7) Income
(8) Marital status
(9) Race/Ethnicity
(10) Health

Having a standard set of characteristics like this collected and tested in every experiment would greatly expand the amount of available evidence that could be used for meta-analysis of correlations between treatment effects and the characteristics. Inevitably, the list would be further refined for new studies and when considering various policy contexts.

At times, of course, especially when working with administrative data, collecting information on the entire set of characteristics will not be possible because of issues such as privacy concerns and limited resources. However, we encourage economists to collect information on as many of the characteristics as feasible and, in manuscripts, to note the rationale for selecting the characteristics reported in paper and associated materials. Explaining the reason for omitting some characteristics would also be valuable for subsequent investigations.

4.6 Recommendations for researchers, editors, reviewers, funding agencies, and partner organizations

Based on the preceding discussion, in Table 3 we offer 12 recommendations that summarize our advice for researchers, editors, reviewers, funding agencies, and partner organizations.

5 Research ethics and community engagement

> *As researchers we need to be constantly aware that the issues we study deal either directly or indirectly with people's livelihoods and well-being. We cannot take this responsibility lightly.*
>
> **Prokopy (2008), on collaborative natural resource management research**

5.1 Research ethics and policy-focused research

As applied economists, we often ask stakeholder-driven questions that have important implications for the livelihoods of the people involved. Therefore, we have a responsibility to think carefully about our conduct as researchers and ensure that we are ethical in how we choose research ideas and design, implement, analyze, and disseminate research (Josephson & Michler, 2018; Michler, Masters, & Josephson, 2021). Scrutiny of the credibility of economics and replicability of research in the social sciences, including in agricultural and applied economics (Lybbert & Buccola, 2021), reinforces the need for experimental economists to strengthen our commitment to making ethical decisions in our research activities.

Though we spend years learning about microeconomic principles and econometric methods, we rarely receive formal training in research ethics and few papers have provided guidance on the topic. However, ethical concerns are receiving ever greater attention, including a special issue on ethics topics in *Applied Economic Perspectives and Policy* (2021).

Table 3 Twelve recommendations for producing high-quality behavioral and experimental agri-environmental research.

<table>
<tr><th colspan="4">Recommendations for Researchers</th></tr>
<tr>
<td>1. Spend sufficient time planning new experiments
a. Conduct a power analysis to inform the sample size
b. Write a pre-analysis plan
c. Pre-register experiments
d. Plan ahead to consider heterogeneous treatment effects
i. Use a stratified and/or block randomized design
ii. Collect data on a standardized set of participant demographics and characteristics</td>
<td>2. Use conservative assumptions in experimental design
a. Design experiments to detect effect sizes that are smaller than you expect
b. Simplify experimental designs, when possible</td>
<td>3. Be thoughtful in how you recruit participants and engage with agricultural stakeholders
a. Appropriately compensate participants
b. Build relationships with stakeholders and community partners that can assist with participant recruitment
c. Respect the knowledge and perspectives agricultural stakeholders bring to the table as research participants
d. Avoid deception in experiments, especially when it can damage relationships with the communities you sample from</td>
<td>4. Report results responsibly
a. Be transparent when reporting results
i. Report all results outlined in the pre-analysis plan—even null results
ii. Indicate which reported results came from exploratory analysis
b. Account for multiple hypothesis testing
c. Report standardized effect sizes
d. Publish experiment instructions, data, and code alongside the paper</td>
</tr>
<tr><th colspan="4">Recommendations for Editors and Reviewers</th></tr>
<tr>
<td>5. Place higher emphasis on the quality of the research design and importance of research questions</td>
<td>6. Avoid putting undue emphasis on statistical significance—null findings are also valuable</td>
<td>7. Encourage replication papers</td>
<td>8. Require researchers to publish data and code with their papers</td>
</tr>
<tr><th colspan="4">Recommendations for Funding Agencies</th></tr>
<tr>
<td>9. Require power analyses and pre-analysis plans</td>
<td colspan="3">10. Have realistic expectations about what can be accomplished within budget limitations, especially regarding testing of multiple hypothesis or sub-group analysis</td>
</tr>
<tr><th colspan="4">Recommendations for Partner Organizations</th></tr>
<tr>
<td colspan="2">11. Appreciate the value of results from a well-designed experiment to improve program performance and to help justify continued support for a successful program</td>
<td colspan="2">12. Work with researchers to embed experiments in programs, especially during the launch of a new program or a major program revision</td>
</tr>
</table>

Much of the research on ethical issues in applied economics has focused on considerations related either to (1) working with human subjects in terms of actions overseen by Institutional Review Boards (IRBs) or (2) research design and analysis in terms of replicability, transparency, and use of questionable practices such as p-hacking (Lybbert & Buccola, 2021). While these are incredibly important topics, we believe that concern about conducting ethical research should start much earlier than collecting and analyzing data and stretch well beyond communicating research findings. Michler et al. (2021) made a similar point in their rich overview of ethical issues related to generation of new research ideas. Furthermore, Josephson and Michler (2018) highlighted additional ethical issues related to media attention, Josephson and Smale (2021) discussed practices for obtaining informed consent, and Barrett and Carter (2020) describe ethical issues related to RCTs.

We believe that additional ethical questions must be considered when conducting economic field experiments related to agricultural producers' livelihoods and that researchers' actions can either plant seeds for future researchers or limit future opportunities by damaging relationships. These considerations are especially important when working in close-knit agricultural communities and designing incentives to motivate participants to change how they manage their land.

In Section 4 (Issue 1), we discussed ethical issues related to replicability, so we do not rehash those challenges now. Instead, we turn to ethical topics that have received less attention in the literature, focusing on researchers' interactions with participants and the communities in which they live and work.

We first review basic guidelines set forth for working with human subjects and highlight how current processes can fail to recognize the full suite of ethical research considerations. More than 30 years ago, in its 1978 *Belmont Report*, the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (1979) defined three guiding principles for protecting human research subjects—respect for individuals, beneficence, and justice—and noted that researchers can apply the principles using three concrete actions: obtaining informed consent, assessing the scope of risks and benefits to participants, and selecting subjects fairly. The World Bank's Development Impact Evaluation (DIME) group (2020) listed four related components of research ethics: protecting human subjects, informed consent, confidentiality, and ethics approvals.

Economists are accustomed to considering these aspects of research and obtaining approval to work with human subjects. At universities and other institutions in Western countries, the process is managed by IRBs and Research Ethics Boards (REBs).[17] Researchers sometimes have been tempted to regard the process of obtaining IRB approval as a time-consuming headache that brings little to the quality of the research. However, we view the approval as a critical element of the research process

[17] Josephson and Michler (2018) highlight that, though review boards oversee human subjects for research conducted by Western universities and institutions, research conducted outside of those regions often is subject to limited oversight regarding treatment of human subjects. For example, they note that only 8 of the 15 CGIAR centers require approval from an ethical review board.

since IRBs review, approve, and monitor studies to ensure that the rights of people participating in them are protected during all stages of the research.

IRB oversight is limited, however, to specific topics and cannot prevent poor research practices or unethical behavior such as use of misleading data collection methods (Josephson & Michler, 2018) and misrepresenting results. Furthermore, IRB applications are sometimes submitted shortly before the experiment is to be conducted, and approval (or disapproval) is decided *ex ante* (Barrett & Carter, 2020). The process relies on researchers updating the review board regarding new protocols when the submitted design is modified, but there are few negative or positive incentives to do so.

It is important to note that experimental economists rarely have significant IRB issues for experiments conducted in traditional laboratory settings (Stages IA and IB). Traditions of voluntary recruitment of participants, paying participants "show-up" fees and additional payments based on the explicit rules of the experiment, and lack of deception all make approval of laboratory experiments relatively easy to obtain in traditional IRB reviews.[18] However, ethical considerations do not end with an IRB approval. In fact, university IRB reviews do not cover many of the ethical concerns we view as most important.

Another challenge is that researchers cannot always fully anticipate ethical issues before a project begins. For example, they sometimes discover how contentious a topic is only after working closely with the community and could not fully inform the IRB regarding the relative risks and benefits for participants prior to conducting the research (Prokopy, 2008).

Agricultural and applied economics researchers such as Josephson and Michler (2018) have contemplated ways for the discipline to take ethical considerations more seriously. They recommend that journals require authors to submit the ethics approvals with their papers. And though Josephson and Michler (2018) do not explicitly state that the approved protocols must match the protocol used in the study, one can imagine that could be a criterion for publication.

The potential presence of deception in economic experiments is an ethical issue that agricultural and applied economics researchers must carefully consider when planning research. To be clear, deception is generally discouraged in economics, and some journals refuse to publish studies that used deception. Often in economics, the discussion is not whether or not deception should be acceptable, but instead revolves around agreeing on the definition of what is deception. Among experimental

[18] An IRB's definition of deception often is different than the definition typically accepted in the experimental economics community. IRBs may raise concerns about deception if a researcher does not provide participants with full information, including information about the different treatments being tested. This type of information omission is typically not considered deception by experimental economists, which defines deception as an act that actively misleads participants by providing false information or implying something that is not true (Cason & Wu, 2019). Researchers may need to include in their IRB protocols additional justification about certain features of an experiment and their potential impacts on the wellbeing of human subjects.

economists, deception is typically defined as "a situation where subjects are *actively* misled by the experimenter rather than a situation where subjects are only provided with incomplete information," (Cason & Wu, 2019, p. 755). Misleading participants could involve providing false information or implying something that is not true about any element of the experiment. For example, providing false information about the decisions of other participants or misleading participants about the underlying payoff structures in the experiment would both be blatant forms of deception. On the other hand, omitting information about all of the treatments being tested in a study is generally not considered deceptive.

We emphasize two issues related to using deception that are concerning for economic experiments in general and particularly so for experiments that aim to inform policy and research involving nonstudent participants. First, deception in an experiment can reduce experimental control and reduce the internal and external validity of a study by introducing confounding factors and leading to selection bias (Cason & Wu, 2019). Second, deception can erode trust in researcher within a potential participant pool and reduce participation in future studies. This concern is particularly relevant for experiments on agri-environmental issues, which rely on participation by farmers and landowner from typically small, tightknit agricultural communities. Damaging relationships with members of these communities generates external costs on other researchers and on society if future research opportunities are jeopardized. To avoid these detrimental research outcomes, we discourage the use of deception in economic experiments, and we direct researchers to Cason and Wu (2019) for a more in-depth analysis of the role of deception in agricultural and applied economics.

Ethical challenges are more pronounced when experiments involve landowners and producers as participants (Stages II, III, and IV). While their participation is voluntary, the experiments and results can have dramatic implications by affecting their time and finances and modifying programs and policies that affect them regardless of the intent of the researcher. These concerns are particularly resonant for Stage IV RCTs in which agricultural land managers may be affected by the treatments being tested within a program or policy setting. Discussions about ethical concerns related to medical RCTs have been ongoing, but these conversations have lagged in fields that use RCTs to evaluate public policy interventions (MacKay, 2018). In her discussion of the ethics of RCTs in political science, Phillips (2021, p. 279) notes that "At this point, there seems to be no common terminology or taxonomy in discussions about the ethical issues associated with field experiments." She proceeds to present a taxonomy that includes six commonly cited factors that must be considered when assessing the ethics of social science RCTs: harms, benefits, risk/benefit ratio, autonomy, partnerships, and professionalism. Economists are also engaging in the conversation about ethics related to experiments and RCTs in particular (Abramowicz & Szafarz, 2020); however, much more discussion is needed to fully consider the ethical implications of our work and to establish robust professional norms that respect and protect the communities that we study.

Researchers in general and relatively inexperienced researchers in particular benefit from planning studies with the trajectory of the discipline in mind in terms of advancements in rigorous research methods and other research best practices. Spending time considering and documenting ethical considerations and actions that uphold high ethical standards will be time well spent—both for experiment participants and for subsequent researchers.

5.2 Considerations when working in agricultural communities

Obviously, ethical considerations are easier to manage when conducting experiments in a laboratory with student subjects (Stages IA and IB). The decisions students make in the experiments do not have direct impacts on their livelihoods. Nor do the experiments ask them to disclose critical private information about their finances and profits. Furthermore, faculty members and students are accustomed to engaging one another daily. Students can be easily and inexpensively recruited and incentivized and generally consider their participation as a task by which they earn extra money rather than as affecting public policies.

Experiments involving stakeholder participants (Stages II, III, and IV) are inherently more complex and require additional time and effort. First, researchers often must build relationships with key members of the community. Without relationships, agricultural producers generally will not be willing to participate in open, frank discussions about challenges they face. Strong relationships also provide sounding boards regarding how to approach producers and design incentives and experiments. Sometimes the conversations are humbling as stakeholders provide researchers with reality checks about their plans, requiring them to re-evaluate portions of proposed projects. Overall, the conversations are insightful and rewarding and improve the resulting research.

Building the necessary relationships takes considerable time. Researchers must first determine who to engage and how. And researchers must make time to maintain those relationships once initiated and seek to ensure that their research provides partner organizations with meaningful value-added information. Researchers at land-grant universities benefit from their ability to work with extension-focused colleagues and Cooperative Extension county agents. Extension faculty and staff members typically have established relationships in agricultural communities and can introduce researchers to leaders and other influential members who are “gatekeepers” within the community. In fact, Herberich, Levitt, and List (2009) suggested that connections with extension programs provide agricultural economists with a comparative advantage in conducting field experiments. Still, it is critical to respect Extension staff members and their relationships with and responsibilities to their clients. Sadly, some Extension faculty and staff members can point to examples of unproductive and even damaging project outcomes, usually because of a rush to “get things done.” With time and respect, collaborations can be fruitful between researchers who do not have extension appointments and colleagues working within Cooperative Extension programs.

Whenever researchers engage with communities and especially when they engage with the tight-knit communities that often characterize rural communities, there are additional concerns related to building trust and strong foundations for long-lasting, mutually-beneficial relationships. As researchers, it can be easy to wade so deep into the small details of a research study that we forget to consider how our decisions affect the experiences of our participants. We can also fail to recognize a simple reality that many people are unfamiliar with the research process, and unfamiliarity can raise concerns that researchers may need to address. Indeed, we have encountered participants who express their skepticism about elements of a research study. Additionally, participating in research takes time and often participants engage with researchers in the hopes of learning something. Therefore, when researchers fail to connect with the community to share their results, participants can feel like their efforts were wasted. Further complicating the matter is that the participants may expect definitive results to be delivered back to them in a much quicker timetable than occurs with most economics research. Negative experiences participating in research can damage relationships between researchers and members of the community and can limit the ability of subsequent researchers to recruit participants and conduct robust experimental studies. Researchers should be sensitive to these potential challenges and exert effort to conducting research in a way that creates trust and mutual respect between scientists and members of the community.

6 Conclusion and framework for prioritizing research projects

Agri-environmental programs and policies aim to improve the sustainability of agricultural landscapes by mitigating externalities, enhancing provision of public goods, and improving management of common pool resources. Though experimental and behavioral economics tools have been applied broadly to environmental issues such as energy consumption, application of those tools to agri-environmental issues has lagged. The intersection of agriculture and the environment is ripe with questions for which behavioral and experimental economics tools are well-suited. But what kind of questions and projects should researchers tackle?

To prioritize behavioral and experimental agri-environmental research, we recommend weighing four factors: (1) the monetary and time costs of conducting a well-designed, sufficiently powered experiment to test an intervention; (2) the expected social net benefit of the intervention being tested; (3) the degree of uncertainty regarding the expected net benefit of the intervention; and (4) the expected benefit to the researchers in terms of their careers. Low-hanging research fruit exists when study costs are low, the expected social net benefit of the intervention is large, the research appears to be promising for professional growth and academic promotion, and enough uncertainty about the intervention's benefits remains to warrant additional research. Costly studies of interventions that potentially offer significant social benefits should also be a high priority. Conversely, costly studies of interventions that are not expected to be strongly beneficial are not good candidates, and researchers can make wise use of project funds by analyzing interventions that have a greater potential net benefit to society.

Experiments are costly to conduct, particularly in the field with farmers and rural landowners. Some laboratory experiments with student subjects can be conducted for less than \$5000 (not including the cost of the researchers' time) while the direct cost of a field experiment involving farmers can quickly exceed $50,000, $100,000, or more depending on scale and scope.

What if you do not have the budget to run a planned experiment with sufficient power? Is it not worth doing? We understand this common frustration. It can be tempting to say "It's okay—we'll still learn a lot from this project. Let's run the underpowered design anyway and be transparent about its limitations." In fact, we both have been in this very situation in the past and made the same proclamations. Therefore, it is with humility that we encourage researchers to be more rigorous when deciding how to approach these cases. Transparency is excellent but will not prevent the statistical errors discussed in Section 4 (Issue 2). Your budget can potentially permit a simpler experiment involving fewer treatments. Alternatively, you sometimes can test your ideas in the laboratory or via a framed field experiment. Deviating from the initial plan is frustrating but not as frustrating as spending thousands (or hundreds of thousands) running an underpowered study that yields results that are too weak to be useful.

Time costs are another important consideration, and include time required to build necessary relationships with other researchers and stakeholders in the community, obtain funding, plan and design the experiment, implement the experiment, analyze data, communicate findings, and shepherd your paper through the peer-review process until it finds its forever home. It is common to underestimate the amount of time required for a particular project, so a conservative approach is to double or triple your initial estimates, particularly if you have less experience or perpetually underestimate the amount of time required to accomplish tasks (like we almost always do). Time constraints are especially important considerations for relatively inexperienced researchers simultaneously pursuing graduate degrees or endeavoring to obtain tenure or similar promotions, processes that involve separate, strict schedules. Additionally, a grant deadline imposes a ticking clock regardless of the researchers' experience.

Expansive studies typically require large, multi-collaborator and even multi-institutional grants, and the proposals for such grants typically require considerable time and effort to build needed collaborations, develop strong ideas for the proposal, construct proposal narratives, create the supporting documentation, develop budgets, and work with the institution to obtain all necessary approvals. The process of building collegial relationships and thinking critically about new research ideas is rewarding, and the seeds you sow in that process can lead to bountiful harvests in the future. However, do not ignore the trade-offs involved. The time dedicated just to the proposal is time that can be spent conducting research, writing papers, preparing classes, and advising students. While we certainly encourage junior faculty members to participate in these large efforts, we caution them against being the principal investigator since the administrative burdens associated with accounting and reporting for these grants usually are intensive.

The researcher must also consider how the project will support their short- and long-term career goals. We urge researchers to consider the expectations of their current employer as well as the broader profession. Keeping in mind the expectations of the broader profession and trying to achieve them, enables the researcher to be in a

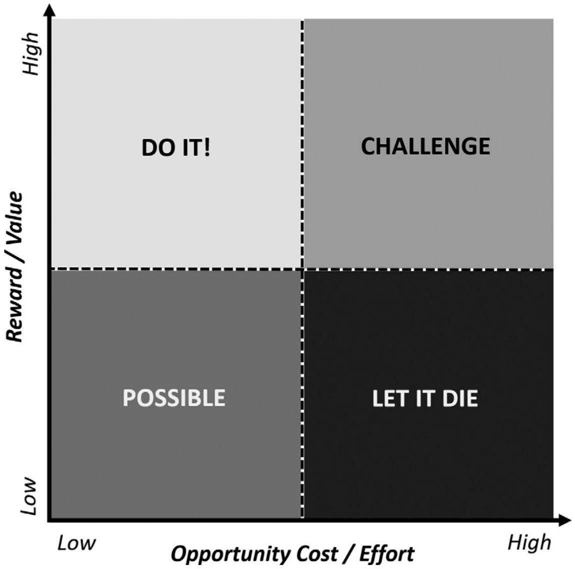

FIG. 2

A framework for assessing the benefits and costs of projects.

stronger position to change employers, if needed. This can be particularly important for a researcher's sense of professional satisfaction as it can avoid a researcher being 'stuck' at a university and feeling like they have few, if any alternatives.

To assist researchers in prioritizing projects, Fig. 2 presents a framework we have found useful when prioritizing our efforts and considering whether to pursue a new project.

There is an ongoing need for experimental and behavioral economics research to improve our understanding of decision-making at the nexus of agriculture and the environment. We hope this chapter motivates additional research in this area and provides researchers with the insights, recommendations, tools, and resources needed to conduct high-quality economic experiments that inform agri-environmental programs and policies.

Acknowledgments

We thank David Just, Chris Barrett, Stephen Swallow, Sarah Jacobson, and an anonymous reviewer for their insightful comments, which have greatly improved this chapter. We also appreciate feedback from participants in the 2020 AAEA Online Workshop for the Handbook of Agricultural Economics. Finally, we would like to thank the many researchers who have helped advance the use of behavioral and experimental economics to inform agri-environmental program and policy design. We appreciate the supportive community of researchers working on these topics, and we look forward to continuing to learn from our collective contributions. We are particularly grateful for all of the colleagues we have worked with in the Center for

Experimental and Applied Economics. Thank you to Lindsey Kelley for helping us cross reference a long list of citations. This publication was made possible by the National Science Foundation EPSCoR Grant No. 1757353 and the State of Delaware. Support was also provided from the Center for Experimental and Behavioral Agri-Environmental Research (USDA NIFA #2019-67023-29854) and NSF EPSCoR Project WICCED.

References

Abbink, K., Irlenbusch, B., Pezanis-Christou, P., Rockenbach, B., Sadrieh, A., & Selten, R. (2005). An experimental test of design alternatives for the British 3G/UMTS auction. *European Economic Review*, *49*(2), 505–530.

Abraham, K., Haskins, R., Glied, S., Groves, R., Hahn, R., Hoynes, H., et al. (2017). *The promise of evidence-based policymaking: Report of the commission on evidence-based policymaking*. https://www.cep.gov/news/sept6news.html.

Abramowicz, M., & Szafarz, A. (2020). Ethics of RCTs: Should economists care about equipoise? In F. Bédécarrats, I. Guérin, & F. Roubaud (Eds.), *Randomized control trials in the field of development: A critical perspective* Oxford University Press.

Allcott, H. (2011). Social norms and energy conservation. *Journal of Public Economics*, *95*(9–10), 1082–1095.

Armitage, P., McPherson, C. K., & Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society Series A*, *132*, 235–244. https://doi.org/10.2307/2343787.

Arnold, M., Duke, J. M., & Messer, K. D. (2013). Adverse selection in reverse auctions for environmental services. *Land Economics*, *89*(3), 387–412.

Athey, S., & Imbens, G. W. (2017). The econometrics of randomized experiments. In *Vol. 1. Handbook of field experiments* (pp. 73–140). North-Holland.

Baca-Motes, K., Brown, A., Gneezy, A., Keenan, E. A., & Nelson, L. D. (2013). Commitment and behavior change: Evidence from the field. *Journal of Consumer Research*, *39*(5), 1070–1084.

Banerjee, S. (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, S., & Cason, T. N. (2020). *Spatial coordination and joint bidding in conservation auctions**. Working Paper. https://www.krannert.purdue.edu/faculty/cason/papers/joint_bid_auction.pdf.

Banerjee, S., Cason, T. N., de Vries, F. P., & Hanley, N. (2017). Transaction costs, communication and spatial coordination in payment for ecosystem services schemes. *Journal of Environmental Economics and Management*, *83*, 68–89.

Banerjee, S., & Conte, M. N. (2018). Information access, conservation practice choice, and rent seeking in conservation procurement auctions: Evidence from a laboratory experiment. *American Journal of Agricultural Economics*, *100*(5), 1407–1426.

Banerjee, A., & Duflo, E. (Eds.). (2017). *Handbook of field experiments* North-Holland.

Banerjee, A., Duflo, E., Finkelstein, A., Katz, L. F., Olken, B. A., & Sautmann, A. (2020). *In praise of moderation: Suggestions for the scope and use of pre-analysis plans for RCTs in economics*. Working Paper 26993 National Bureau of Economic: National Bureau of Economic Research.

Banerjee, S., Kwasnica, A. M., & Shortle, J. S. (2012). Agglomeration bonus in small and large local networks: A laboratory examination of spatial coordination. *Ecological Economics*, *84*, 142–152.

Banks, J., Olson, M., Porter, D., Rassenti, S., & Smith, V. (2003). Theory, experiment and the federal communications commission spectrum auctions. *Journal of Economic Behavior and Organization*, *51*(3), 303–350.

Barrett, C. B., & Carter, M. R. (2010). The power and pitfalls of experiments in development economics: Some non-random reflections. *Applied Economic Perspectives and Policy*, *32*(4), 515–548.

Barrett, C. B., & Carter, M. R. (2020). Finding our balance? Revisiting the randomization revolution in development economics ten years further on. *World Development*, *127*, 104789.

Bayer, R. C., & Loch, A. (2017). Experimental evidence on the relative efficiency of forward contracting and tradable entitlements in water markets. *Water Resources and Economics*, *20*, 1–15.

Behaghel, L., Macours, K., & Subervie, J. (2019). How can randomised controlled trials help improve the design of the common agricultural policy? *European Review of Agricultural Economics*, *46*(3), 473–493.

Bellemare, C., Bissonnette, L., & Kröger, S. (2016). Simulating power of economic experiments: The powerBBK package. *Journal of the Economic Science Association*, *2*, 157–168. https://doi.org/10.1007/s40881-016-0028-4.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, *29*(4), 1165–1188.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, *20*(3), 351–368.

Binmore, K., & Klemperer, P. (2002). The biggest auction ever: The sale of the British 3G Telecom licences. *The Economic Journal*, *112*(478), C74–C96.

Boas, T. C., Christenson, D. P., & Glick, D. M. (2020). Recruiting large online samples in the United States and India: Facebook, Mechanical Turk, and Qualtrics. *Political Science Research and Methods*, *8*(2), 232–250. Available at https://doi.org/10.1017/psrm.2018.28.

Bocquého, G., Jacquet, F., & Reynaud, A. (2014). Expected utility or prospect theory maximisers? Assessing farmers' risk behaviour from field-experiment data. *European Review of Agricultural Economics*, *41*(1), 135–172.

Boun My, K., & Ouvrard, B. (2019). Nudge and tax in an environmental public goods experiment: Does environmental sensitivity matter? *Resource and Energy Economics*, *55*, 24–48.

Boxall, P. C., Perger, O., Packman, K., & Weber, M. (2017). An experimental examination of target based conservation auctions. *Land Use Policy*, *63*, 592–600. Available at https://doi.org/10.1016/j.landusepol.2015.03.016.

Boxall, P. C., Perger, O., & Weber, M. (2013). Reverse auctions for agri-environmental improvements: Bid-selection rules and pricing for beneficial management practice adoption. *Canadian Public Policy*, *39*(Suppl. 2), S23–S36.

Brodeur, A., Cook, N., & Heyes, A. (2020). Methods matter: p-Hacking and publication bias in causal analysis in economics. *American Economic Review*, *110*(11), 3634–3660.

Brodeur, A., Lé, M., Sangnier, M., & Zylberberg, Y. (2016). Star wars: The empirics strike back. *American Economic Journal: Applied Economics*, *8*(1), 1–32.

Brookes, S. T., Whitely, E., Egger, M., Smith, G. D., Mulheran, P. A., & Peters, T. J. (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.

Brown, G., & Hagen, D. A. (2010). Behavioral economics and the environment. *Environmental and Resource Economics*, *46*(2), 139–146.

Brown, J. P., Lambert, D. M., & Wojan, T. R. (2019). The effect of the conservation reserve program on rural economies: Deriving a statistical verdict from a null finding. *American Journal of Agricultural Economics*, *101*(2), 528–540.

Brozović, N., Sunding, D. L., & Zilberman, D. (2010). On the spatial nature of the groundwater pumping externality. *Resource and Energy Economics*, *32*(2), 154–164.

Butera, L., & List, J. A. (2017). *An economic approach to alleviate the crises of confidence in science with an application to the public goods game*. No. w23335 National Bureau of Economic Research. Available at https://www.nber.org/papers/w23335.

Butler, J. M., Fooks, J. R., Messer, K. D., & Palm-Forster, L. H. (2020). Addressing social dilemmas with mascots, information, and graphics. *Economic Inquiry*, *58*(1), 150–168.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., et al. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376.

Byerly, H., Balmford, A., Ferraro, P. J., Wagner, C. H., Palchak, E., Polasky, S., et al. (2018). Nudging pro-environmental behavior: Evidence and opportunities. *Frontiers in Ecology and the Environment*, *16*(3), 159–168.

Camerer, C. (2011). *The promise and success of lab-field generalizability in experimental economics: A critical reply to Levitt and List*. Available at SSRN 1977749.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., et al. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644.

Canavari, M., Drichoutis, A. C., Lusk, J. L., & Nayga, R. M., Jr. (2019). How to run an experimental auction: A review of recent advances. *European Review of Agricultural Economics*, *46*(5), 862–922.

Casari, M., & Plott, C. (2003). Decentralized management of common property resources: Experiments with a centuries-old institution. *Journal of Economic Behavior and Organization*, *51*(2), 217–247.

Cason, T. N., & Gangadharan, L. (2004). Auction design for voluntary conservation programs. *American Journal of Agricultural Economics*, *86*(5), 1211–1217.

Cason, T. N., & Gangadharan, L. (2005). A laboratory comparison of uniform and discriminative price auctions for reducing non-point source pollution. *Land Economics*, *81*(1), 51–70.

Cason, T. N., & Gangadharan, L. (2013). Empowering neighbors versus imposing regulations: An experimental analysis of pollution reduction schemes. *Journal of Environmental Economics and Management*, *65*(3), 469–484.

Cason, T. N., Gangadharan, L., & Duke, C. (2003). A laboratory study of auctions for reducing non-point source pollution. *Journal of Environmental Economics and Management*, *46*(3), 446–471.

Cason, T. N., & Wu, S. Y. (2019). Subject pools and deception in agricultural and resource economics experiments. *Environmental and Resource Economics*, *73*(3), 743–758.

Chetty, R. (2015). Behavioral economics and public policy: A pragmatic perspective. *American Economic Review*, *105*(5), 1–33.

Christensen, G., & Miguel, E. (2018). Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*, *56*(3), 920–980.

Cialdini, R. B. (2003). Crafting normative messages to protect the environment. *Current Directions in Psychological Science*, *12*(4), 105–109.

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, *58*(6), 1015.

Clarke, D., Romano, J. P., & Wolf, M. (2020). The Romano–Wolf multiple-hypothesis correction in Stata. *The Stata Journal*, *20*(4), 812–843.

Cochard, F., Willinger, M., & Xepapadeas, A. (2005). Efficiency of nonpoint source pollution instruments: An experimental study. *Environmental and Resource Economics*, *30*(4), 393–422.

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

Coffman, L. C., Niederle, M., & Wilson, A. J. (2017). A proposal to organize and promote replications. *American Economic Review*, *107*(5), 41–45.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Colen, L., Gomez y Paloma, S., Latacz-Lohmann, U., Lefebvre, M., Préget, R., & Thoyer, S. (2016). Economic experiments as a tool for agricultural policy evaluation: Insights from the European CAP. *Canadian Journal of Agricultural Economics/Revue canadienne d'agroeconomie*, *64*(4), 667–694.

Comerford, E. (2013). The impact of permanent protection on cost and participation in a conservation programme: A case study from Queensland. *Land Use Policy*, *34*, 176–182.

Conte, M. N., & Griffin, R. M. (2017). Quality information and procurement auction outcomes: Evidence from a payment for ecosystem services laboratory experiment. *American Journal of Agricultural Economics*, *99*(3), 571–591.

Conte, M. N., & Griffin, R. (2019). Private benefits of conservation and procurement auction performance. *Environmental and Resource Economics*, *73*(3), 759–790.

Cornes, R., & Sandler, T. (1994). The comparative static properties of the impure public good model. *Journal of Public Economics*, *54*(3), 403–421.

Cummings, R. G., Holt, C. A., & Laury, S. K. (2004). Using laboratory experiments for policymaking: An example from the Georgia irrigation reduction auction. *Journal of Policy Analysis and Management*, *23*(2), 341–363.

Czap, N. V., Czap, H. J., Banerjee, S., & Burbach, M. E. (2019). Encouraging farmers' participation in the Conservation Stewardship Program: A field experiment. *Ecological Economics*, *161*, 130–143.

Czap, N. V., Czap, H. J., Khachaturyan, M., Burbach, M. E., & Lynne, G. D. (2013). Smiley or frowney: The effect of emotions and empathy framing in a downstream water pollution game. *International Journal of Economics and Finance*, *5*(3), 9.

Czap, N. V., Czap, H. J., Lynne, G. D., & Burbach, M. E. (2015). Walk in my shoes: Nudging for empathy conservation. *Ecological Economics*, *118*, 147–158.

Czibor, E., Jimenez-Gomez, D., & List, J. A. (2019). The dozen things experimental economists should do (more of). *Southern Economic Journal*, *86*(2), 371–432.

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

Dellavigna, S. (2009). Psychology and economics: Evidence from the field. *Journal of Economic Literature*, *47*(2), 315–372.

Dessart, F. J., Barreiro-Hurlé, J., & van Bavel, R. (2019). Behavioural factors affecting the adoption of sustainable farming practices: A policy-oriented review. *European Review of Agricultural Economics*, *46*(3), 417–471.

Dhami, S. (2016). *The foundations of behavioral economic analysis*. Oxford University Press.

Dolan, P., Hallsworth, M., Halpern, D., King, D., Metcalfe, R., & Vlaev, I. (2012). Influencing behaviour: The mindspace way. *Journal of Economic Psychology*, *33*(1), 264–277.

Doucouliagos, C., & Stanley, T. D. (2013). Are all economic facts greatly exaggerated? Theory competition and selectivity. *Journal of Economic Surveys*, *27*(2), 316–339.

Duflo, E., Glennerster, R., & Kremer, M. (2007). Using randomization in development economics research: A toolkit. *Handbook of Development Economics*, *4*, 3895–3962.

Duke, J. M., Dundas, S. J., & Messer, K. D. (2013). Cost-Effective conservation planning: Lessons from economics. *Journal of Environmental Management*, *125*, 126–133.

Duke, J. M., Liu, Z., Suter, J. F., Messer, K. D., & Michael, H. A. (2020). Some taxes are better than others: An economic experiment analyzing groundwater management in a spatially explicit aquifer. *Water Resources Research*, *56*(7). https://doi.org/10.1029/2019WR026426, e2019WR026426.

Duquette, E., Higgins, N., & Horowitz, J. (2012). Farmer discount rates: Experimental evidence. *American Journal of Agricultural Economics*, *94*(2), 451–456.

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. New York: Cambridge University Press.

Ellis, S. F., Fooks, J. R., Messer, K. D., & Miller, M. J. (2016). The effects of climate change information on charitable giving for water quality protection: A field experiment. *Agricultural and Resource Economics Review*, *45*(2), 319–337.

Ferraro, P. J. (2008). Asymmetric information and contract design for payments for environmental services. *Ecological Economics*, *65*(4), 810–821.

Feltovich, N. (2003). Nonparametric tests of differences in medians: Comparison of the Wilcoxon–Mann–Whitney and robust rank-order tests. *Experimental Economics*, *6*(3), 273–297.

Ferraro, P., Messer, K. D., Shukla, P., & Weigel, C. (2021). *Behavioral biases among producers: Experimental evidence of anchoring in procurement auctions*. Working Paper.

Ferraro, P., Messer, K. D., & Wu, S. (2017). Applying behavioral insights to improve water security. *Choices*, *32*(4), 1–6.

Ferraro, P. J., & Shukla, P. (2020). Feature—Is a replicability crisis on the horizon for environmental and resource economics? *Review of Environmental Economics and Policy*, *14*(2), 339–351.

Fink, G., McConnell, M., & Vollmer, S. (2014). Testing for heterogeneous treatment effects in experimental data: False discovery risks and correction procedures. *Journal of Development Effectiveness*, *6*(1), 44–57.

Fleming, P.M., Palm-Forster, L.H., and Kelley, L.E. 2021. The effect of legacy pollution information on landowner investments in water quality: Lessons from economic experiments in the field and the lab. *Environmental Research Letters*. Available at: http://iopscience.iop.org/article/10.1088/1748-9326/abea33. Accessed March 9, 2021.

Fooks, J. R., Higgins, N., Messer, K. D., Duke, J. M., Hellerstein, D., & Lynch, L. (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.

Fooks, J., Messer, K. D., & Duke, J. (2015). Dynamic entry, reverse auctions, and the purchase of environmental services. *Land Economics*, *91*(1), 57–75.

Foxall, G. R. (2017). Behavioral economics in consumer behavior analysis. *The Behavior Analyst*, *40*(2), 309–313.

Fréchette, G. (2015). Experimental economics across subject populations. In *Handbook of experimental economics* (pp. 435–480). Princeton: Princeton University Press.

Friedman, D., & Sunder, S. (1994). *Experimental methods: A primer for economists*. Cambridge University Press.

Gardner, R., Moore, M., & Walker, J. (1997). Governing a groundwater commons: A strategic and laboratory analysis of western water law. *Economic Inquiry*, *35*, 218–234.

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651.

Gisser, M., & Sanchez, D. A. (1980). Competition versus optimal control in groundwater pumping. *Water Resources Research*, *16*, 638–642. Available at https://doi.org/10.1029/WR016i004p00638.

Glennerster, R. (2017). The practicalities of running randomized evaluations: partnerships, measurement, ethics, and transparency. In A. V. Banerjee, & E. Duflo (Eds.), *Handbook of field experiments* (pp. 175–243). North-Holland. Available at: https://www.sciencedirect.com/science/article/pii/S2214658X16300150.

Gneezy, U., & Imas, A. (2017). Lab in the field: Measuring preferences in the wild. In A. V. Banerjee, & E. Duflo (Eds.), *Handbook of field experiments* (pp. 439–464). North-Holland. (chapter 10). Available at: https://www.sciencedirect.com/science/article/pii/S2214658X16300058.

Goldberg, M., van der Linden, S., Ballew, M. T., Rosenthal, S. A., & Leiserowitz, A. (2019). Convenient but biased? The reliability of convenience samples in research about attitudes toward climate change. OSF Preprints. *Leiserowitz*. https://doi.org/10.31219/osf.io/2h7as.

Gosnell, G. K. (2018). Communicating resourcefully: A natural field experiment on environmental framing and cognitive dissonance in going paperless. *Ecological Economics*, *154*, 128–144.

Gueron, J. M. (2017). The politics and practice of social experiments: Seeds of a revolution. In A. V. Banerjee, & E. Duflo (Eds.), *Handbook of Field Experiments* (pp. 27–69). North-Holland. (chapter 2). Available at: https://www.sciencedirect.com/science/article/pii/S2214658X16300198.

Guilfoos, T., Pape, A. D., Khanna, N., & Salvage, K. (2013). Groundwater management: The effect of water flows on welfare gains. *Ecological Economics*, *95*, 31–40. Available at https://doi.org/10.1016/j.ecolecon.2013.07.013.

Hamermesh, D. S. (2007). Viewpoint: Replication in economics. *Canadian Journal of Economics/Revue canadienne d'économique*, *40*(3), 715–733.

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

Heckman, J. J., & Vytlacil, E. (2001). Policy-relevant treatment effects. *American Economic Review*, *91*(2), 107–111.

Hellerstein, D. M. (2017). The US Conservation Reserve Program: The evolution of an enrollment mechanism. *Land Use Policy*, *63*, 601–610.

Hellerstein, D., & Higgins, N. (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.

Herberich, D. H., Levitt, S. D., & List, J. A. (2009). Can field experiments return agricultural economics to the glory days? *American Journal of Agricultural Economics*, *91*(5), 1259–1265.

Higgins, N., Hellerstein, D. M., Wallander, S., & Lynch, L. (2017). *Economic experiments for policy analysis and program design: A guide for agricultural decisionmakers*. U.S. Department of Agriculture Economic Research Service Economic Research Report 236.

Huff, C., & Tingley, D. (2015). Who are these people? Evaluating the demographic characteristics and political preferences of MTurk survey respondents. *Research and Politics*, *2*(3). 2053168015604648. Available at https://doi.org/10.1177/2053168015604648.

Iftekhar, M. S., & Latacz-Lohmann, U. (2017). How well do conservation auctions perform in achieving landscape-level outcomes? A comparison of auction formats and bid selection criteria. *Australian Journal of Agricultural and Resource Economics*, *61*(4), 557–575.

Iftekhar, M. S., Tisdell, J. G., & Sprod, D. (2018). Can partial project selection improve conservation auction performances? *Australasian Journal of Environmental Management*, *25*(2), 212–232.

Ioannidis, J. P. A., Stanley, T. D., & Doucouliagos, H. (2017). The power of bias in economics research. *The Economic Journal*, *127*(605), F236–F265.

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.

Janssen, M. A., & Ostrom, E. (2008). TURFS in the lab: Institutional innovation in real-time dynamic spatial commons. *Rationality and Society*, *20*(4), 371–397.

Janzen, S. A., & Michler, J. D. (2021). 'Ulysses' pact or Ulysses' raft: Using pre-analysis plans in experimental and nonexperimental research. *Applied Economic Perspectives and Policy*, *43*(4), 1286–1304.

Johansson, R., Effland, A., & Coble, K. (2017). Falling response rates to USDA crop surveys: Why it matters. *University of Illinois Farmdoc Daily*, *7*(9), 1–9.

Jones Ritten, C., Bastian, C., Shogren, J., Panchalingam, T., Ehmke, M., & Parkhurst, G. (2017). Understanding pollinator habitat conservation under current policy using economic experiments. *Land*, *6*(3), 57.

Jones, L. R., & Vossler, C. A. (2014). Experimental tests of water quality trading markets. *Journal of Environmental Economics and Management*, *68*(3), 449–462.

Josephson, A., & Michler, J. D. (2018). Viewpoint: Beasts of the field? Ethics in agricultural and applied economics. *Food Policy*, *79*, 1–11.

Just, D. R., & Byrne, A. T. (2020). Evidence-based policy and food consumer behaviour: How empirical challenges shape the evidence. *European Review of Agricultural Economics*, *47*(1), 348–370.

Josephson, A., & Smale, M. (2021). What do you mean by 'informed consent?' Ethics in economic development research. *Applied Economic Perspectives and Policy*, *43*(4), 1305–1329.

Just, D., & Kaiser, H. (2016). GMO labeling bill good for both environment and the poor. *The Hill*, (14 July 2016). Accessed on February 28, 2021 https://thehill.com/blogs/pundits-blog/energy-environment/287699-gmo-labeling-bill-good-for-both-environment-and-the.

Kagel, J. H., & Roth, A. E. (Eds.). (2016). *Vol. 2. The handbook of experimental economics*. Princeton: Princeton University Press. Available at: https://press.princeton.edu/books/hardcover/9780691139999/the-handbook-of-experimental-economics-volume-2.

Kawasaki, K., Fujie, T., Koito, K., Inoue, N., & Sasaki, H. (2012). Conservation auctions and compliance: Theory and evidence from laboratory experiments. *Environmental and Resource Economics*, *52*(2), 157–179.

Kecinski, M., Messer, K. D., & Peo, A. J. (2018). When cleaning too much pollution can be a bad thing: A field experiment of consumer demand for oysters. *Ecological Economics*, *146*, 686–695.

Khanna, M., Swinton, S. M., & Messer, K. D. (2018). Sustaining our natural resources in the face of increasing societal demands on agriculture: Directions for future research. *Applied Economics Policy and Perception*, *40*(1), 38–59.

King, W. R., & He, J. (2006). A meta-analysis of the technology acceptance model. *Information and Management*, *43*(6), 740–755.

Klemperer, P. (2002). What really matters in auction design. *Journal of Economic Perspectives*, *16*(1), 169–189.

Krawczyk, M., Bartczak, A., Hanley, N., & Stenger, A. (2016). Buying spatially coordinated ecosystem services: An experiment on the role of auction format and communication. *Ecological Economics*, *124*, 36–48.

Kuhfuss, L., Préget, R., Thoyer, S., Hanley, N., Coent, P. L., & Désolé, M. (2016). Nudges, social norms, and permanence in agri-environmental schemes. *Land Economics*, *92*(4), 641–655.

Lagerkvist, C. J., & Hess, S. (2011). A meta-analysis of consumer willingness to pay for farm animal welfare. *European Review of Agricultural Economics*, *38*(1), 55–78.

Lamb, K., Hansen, K., Bastian, C., Nagler, A., & Jones Ritten, C. (2019). Investigating potential impacts of credit failure risk mitigation on habitat exchange outcomes. *Environmental and Resource Economics*, *73*(3), 815–842.

Le Coent, P., Préget, R., & Thoyer, S. (2021). Farmers follow the herd: A theoretical model on social norms and payments for environmental services. *Environmental and Resource Economics*, *78*(2), 287–306.

Le Coent, P., Thoyer, S., & Préget, R. (2014). Why pay for nothing? An experiment on a conditional subsidy scheme in a threshold public good game. In *1. Conférence annuelle de la FAERE. Montpellier, France*. 13 p. Available at https://hal.inrae.fr/hal-02739425.

Levitt, S. D., & List, J. A. (2007a). On the generalizability of lab behaviour to the field. *Canadian Journal of Economics/Revue canadienne d' économique*, *40*(2), 347–370.

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

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

Li, J., Michael, H. A., Duke, J. M., Messer, K. D., & Suter, J. F. (2014). Behavioral response to contamination risk information in a spatially explicit groundwater environment: Experimental evidence. *Water Resources Research*, *50*(8), 6390–6405.

Li, T., Palm-Forster, L. H., & Bhuiyanmishu, S. (in review). Transaction costs, competitiveness, and participation in reverse auctions: Evidence from a laboratory experiment. APEC Research Reports.

List, J. A., Sadoff, S., & Wagner, M. (2011). So you want to run an experiment? Now what? Some simple rules of thumb for optimal experimental design. *Experimental Economics*, *14*(4), 439–457.

List, J. A., Shaikh, A. M., & Xu, Y. (2019). Multiple hypothesis testing in experimental economics. *Experimental Economics*, *22*(4), 773–793.

Liu, P. (2021). Balancing cost effectiveness and incentive properties in conservation auctions: Experimental evidence from three multi-award reverse auction mechanisms. *Environmental and Resource Economics*, *78*(3), 417–451.

Liu, Z., Suter, J. F., Messer, K. D., Duke, J. M., & Michael, H. A. (2014). Strategic entry and externalities in groundwater resources: Evidence from the lab. *Resource and Energy Economics*, *38*, 181–197.

Liu, P., & Swallow, S. K. (2016). Integrating cobenefits produced with water quality BMPs into credits markets: Conceptualization and experimental illustration for EPRI's Ohio River Basin Trading. *Water Resources Research*, *52*(5), 3387–3407.

Liu, Z., Xu, J., Yang, X., Tu, Q., Hanley, N., & Kontoleon, A. (2019). Performance of agglomeration bonuses in conservation auctions: Lessons from a framed field experiment. *Environmental and Resource Economics*, *73*(3), 843–869.

Loureiro, M. L., McCluskey, J. J., & Mittelhammer, R. C. (2002). Will consumers pay a premium for eco-labeled apples? *Journal of Consumer Affairs*, *36*(2), 203–219.

Lunn, P., Bohacek, M., Somerville, J., Ní Choisdealbha, Á., McGowan, F., & Economic and Social Research Institute. (2016). Price lab: An investigation of consumers' capabilities with complex products. In *Report BKMNEXT306, Research Series* Economic and Social Research Institute.

Lusk, J. L., & Shogren, J. F. (2007). *Experimental auctions: Methods and applications in economic and marketing research*. Cambridge University Press.

Lybbert, T. J., & Buccola, S. T. (2021). The evolving ethics of analysis, publication, and transparency in applied economics. *Applied Economic Perspectives and Policy*, *43*(4), 1330–1351.

Lynne, G. D., Czap, N. V., Czap, H. J., & Burbach, M. E. (2016). A theoretical foundation for empathy conservation: Toward avoiding the tragedy of the commons. *Review of Behavioral Economics*, *3*(3/4), 243–279.

MacKay, D. (2018). The ethics of public policy RCTs: The principle of policy equipoise. *Bioethics*, *32*(1), 59–67.

Madrian, B. C. (2014). Applying insights from behavioral economics to policy design. *Annual Review of Economics*, *6*, 663–688.

Maertens, A., & Barrett, C. B. (2013). Measuring social networks' effects on agricultural technology adoption. *American Journal of Agricultural Economics*, *95*(2), 353–359.

Mason, C. F., & Phillips, O. R. (1997). Mitigating the tragedy of the commons through cooperation: An experimental evaluation. *Journal of Environmental Economics and Management*, *34*(2), 148–172.

McCann, L., & Claassen, R. (2016). Farmer transaction costs of participating in federal conservation programs: Magnitudes and determinants. *Land Economics*, *92*(2), 256–272.

McCarthy, J., & Beckler, D. G. (2000). Survey burden and its impact on attitudes toward the survey sponsor. In *235077. NASS Research Reports* United States Department of Agriculture, National Agricultural Statistics Service. https://doi.org/10.22004/ag.econ.235077.

Meiselman, B.S., C. Weigel, P.J. Ferraro, M. Masters, K.D. Messer, O. Savchenko et al. in development. Lottery incentives and resource management: Evidence from the Agricultural Data Reporting Incentive Program (AgDRIP).

Messer, K. D., & Allen, W. A. (2018). *The science of strategic conservation: Protecting more with less*. Cambridge, England: Cambridge University Press.

Messer, K. D., Duke, J., & Lynch, L. (2014). Applying experimental economics to land economics: Public information and auction efficiency in land preservation markets. In J. Duke, & J. Wu (Eds.), *Oxford handbook of land economics* Oxford Press.

Messer, K. D., Duke, J., Lynch, L., & Li, T. (2017). When does public information undermine the effectiveness of reverse auctions for the purchase of ecosystem services? *Ecological Economics*, *134*, 212–226.

Messer, K. D., Kaiser, H. M., & Poe, G. L. (2007). Voluntary funding for generic advertising using a provision point mechanism: An experimental analysis of option assurance. *Review of Agricultural Economics*, *29*(3), 612–631.

Miao, H., Fooks, J., Guilfoos, T., Messer, K. D., Pradhanang, S. M., Suter, J., et al. (2016). The impact of information on behavior under an ambient-based policy for regulating nonpoint source pollution. *Water Resources Research*, *52*, 3294–3308.

Michler, J. D., Masters, W. A., & Josephson, A. (2021). Research ethics beyond the IRB: Selection bias and the direction of innovation in applied economics. *Applied Economic Perspectives and Policy*, *43*(4), 1352–1365.

Mullinix, K. J., Leeper, T. J., Druckman, J. N., & Freese, J. (2015). The generalizability of survey experiments. *Journal of Experimental Political Science*, *2*(2), 109–138.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research*. Department of Health, Education, and Welfare. https://www.hhs.gov/ohrp/sites/default/files/the-belmont-report-508c_FINAL.pdf.

Normann, H.-T., & Ricciuti, R. (2009). Laboratory experiments for economic policy making. *Journal of Economic Surveys*, *23*(3), 407–432.

Olken, B. A. (2015). Promises and perils of pre-analysis plans. *Journal of Economic Perspectives*, *29*, 61–80. https://doi.org/10.1257/jep.29.3.61.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), 943. Available at https://science.sciencemag.org/content/349/6251/aac4716.

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, L.H., M. Griesinger, J.M. Butler, J.R. Fooks, and K.D. Messer. forthcoming. Stewardship signaling and use of social pressure to reduce nonpoint source pollution. Land Economics.

Palm-Forster, L. H., Suter, J. F., & Messer, K. D. (2019). Experimental evidence on policy approaches that link agricultural subsidies to water quality outcomes. *American Journal of Agricultural Economics*, *101*(1), 109–133.

Palm-Forster, L. H., Swinton, S. M., Lupi, F., & Shupp, R. S. (2016). Too burdensome to bid: Transaction costs and pay-for-performance conservation. *American Journal of Agricultural Economics*, *98*(5), 1314–1333.

Palm-Forster, L. H., Swinton, S. M., & Shupp, R. S. (2017). Farmer preferences for conservation incentives that promote voluntary phosphorus abatement in agricultural watersheds. *Journal of Soil and Water Conservation*, *72*(5), 493–505.

Pannell, D. J., Marshall, G. R., Barr, N., Curtis, A., Vanclay, F., & Wilkinson, R. (2006). Understanding and promoting adoption of conservation practices by rural landholders. *Australian Journal of Experimental Agriculture*, *46*(11), 1407–1424.

Parkhurst, G. M., & Shogren, J. F. (2007). Spatial incentives to coordinate contiguous habitat. *Ecological Economics*, *64*(2), 344–355.

Parkhurst, G., Shogren, J., Bastian, C., Kivi, P., Donner, J., & Smith, R. (2002). Agglomeration bonus: An incentive mechanism to reunite fragmented habitat for biodiversity conservation. *Ecological Economics*, *41*(2), 305–328.

Phillips, T. (2021). Ethics of field experiments. *Annual Review of Political Science*, *24*(1), 277–300.

Plott, C. R. (1987). Dimensions of parallelism: Some policy applications of experimental methods. In A. E. Roth (Ed.), *Laboratory experimentation in economics: Six points of view* (pp. 193–219). Cambridge University Press.

Poe, G. L. (2016). Behavioral anomalies in contingent values and actual choices. *Agricultural and Resource Economics Review*, *45*(2), 246–269.

Prokopy, L. S. (2008). Ethical concerns in researching collaborative natural resource management. *Society and Natural Resources*, *21*(3), 258–265.

Reeling, C., Palm-Forster, L. H., & Melstrom, R. T. (2019). Policy instruments and incentives for coordinated habitat conservation. *Environmental and Resource Economics*, *73*(3), 791–813.

Reeson, A. F., Rodriguez, L. C., Whitten, S. M., Williams, K., Nolles, K., Windle, J., et al. (2011). Adapting auctions for the provision of ecosystem services at the landscape scale. *Ecological Economics*, *70*(9), 1621–1627.

Ribaudo, M. (2015). The limits of voluntary conservation programs. *Choices*, *30*(2), 1–5.

Ritchie, H., & Roser, M. (2013). Land use. In *Our world in data*. Available at: https://ourworldindata.org/land-use. (Accessed 23 June 2021).

Rodriguez-Sickert, C., Guzmán, R. A., & Cárdenas, J. C. (2008). Institutions influence preferences: Evidence from a common pool resource experiment. *Journal of Economic Behavior and Organization*, *67*(1), 215–227.

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

Rolfe, J., Schilizzi, S., Boxall, P., Latacz-Lohmann, U., Iftekhar, S., Star, M., et al. (2018). Identifying the causes of low participation rates in conservation tenders. *International Review of Environmental and Resource Economics*, *12*(1), 1–45.

Rolfe, J., Windle, J., & McCosker, J. (2009). Testing and implementing the use of multiple bidding rounds in conservation auctions: A case study application. *Canadian Journal of Agricultural Economics/Revue canadienne d'agroeconomie*, *57*(3), 287–303.

Romano, J. P., & Wolf, M. (2010). Balanced control of generalized error rates. *The Annals of Statistics*, *38*, 598–633.

Rosch, S., Skorbiansky, S. R., Weigel, C., Messer, K. D., & Hellerstein, D. (2021). Barriers to using economic experiments in evidence-based agricultural policymaking. *Applied Economics Policy and Perspectives*, *43*(2), 531–555.

Saak, A. E., & Peterson, J. M. (2007). Groundwater use under incomplete information. *Journal of Environmental Economics and Management*, *54*(2), 214–228.

Sarr, H., Bchir, M. A., Cochard, F., & Rozan, A. (2019). Nonpoint source pollution: Experiments on the average Pigouvian tax under costly communication. *European Review of Agricultural Economics*, *46*(4), 529–550.

Savchenko, O., Kecinski, M., Li, T., Messer, K. D., & Xu, H. (2018). Fresh foods irrigated with recycled water: A framed field experiment on consumer response. *Food Policy*, *80*, 103–112.

Savchenko, O. B.S. Meiselman, C. Weigel, P.J. Ferraro, M. Masters, K.D. Messer, and J. Suter. in development. "Can voluntary reporting of groundwater use improve water management? A field experiment."

Schilizzi, S. G. M. (2017). An overview of laboratory research on conservation auctions. *Land Use Policy*, *63*, 572–583.

Schilizzi, S., & Latacz-Lohmann, U. (2007). Assessing the performance of conservation auctions: An experimental study. *Land Economics*, *83*(4), 497–515.

Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., & Griskevicius, V. (2007). The constructive, destructive, and reconstructive power of social norms. *Psychological Science*, *18*(5), 429–434.

Segerson, K. (1988). Uncertainty and incentives for nonpoint pollution control. *Journal of Environmental Economics and Management*, *15*(1), 87–98.

Shogren, J. F. (2004). Incentive mechanism testbeds: Discussion. *American Journal of Agricultural Economics*, *86*, 1218–1219.

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384.

Smith, V. L. (1982). Microeconomic systems as an experimental science. *The American Economic Review*, *72*(5), 923–955.

Snowberg, E., & Yariv, L. (2018). *Testing the waters: Behavior across participant pools*. Report w24781 National Bureau of Economic Research. https://doi.org/10.3386/w24781.

Spraggon, J. (2004). Testing ambient pollution instruments with heterogeneous agents. *Journal of Environmental Economics and Management*, *48*(2), 837–856.

Spraggon, J. 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.

Stephenson, K., & Shabman, L. (2017). Can water quality trading fix the agricultural nonpoint source problem? *Annual Review of Resource Economics*, *9*(1), 95–116.

Streletskaya, N. A., Bell, S. D., Kecinski, M., Li, T., Banerjee, S., Palm-Forster, L. H., et al. (2020). Agricultural adoption and behavioral economics: Bridging the gap. *Applied Economic Perspectives and Policy*, *42*(1), 54–66.

Suter, J. F., Duke, J. M., Messer, K. D., & Michael, H. A. (2012). Behavior in a spatially explicit groundwater resource: Evidence from the lab. *American Journal of Agricultural Economics*, *94*(5), 1094–1112.

Suter, J., Hrozencik, R., Ferraro, P. J., & Masters, M. (2018). Impact of peer information on groundwater use in Colorado and Georgia. *Development*.

Suter, J. F., Spraggon, J. M., & Poe, G. L. (2013). Thin and lumpy: An experimental investigation of water quality trading. *Water Resources and Economics*, *1*, 36–60.

Suter, J. F., & Vossler, C. A. (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*(1), 92–107.

Suter, J. F., Vossler, C. A., Poe, G. L., & Segerson, K. (2008). Experiments on damage-based ambient taxes for nonpoint source polluters. *American Journal of Agricultural Economics*, *90*(1), 86–102.

Teisl, M. F., Roe, B., & Hicks, R. L. (2002). Can eco-labels tune a market? Evidence from dolphin-safe labeling. *Journal of Environmental Economics and Management*, *43*(3), 339–359.

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. New Haven, CT: Yale University Press.

Thalheimer, W., & Cook, S. (2002). *How to calculate effect sizes from published research: a simplified methodology*. Somerville, NJ: Work-Learning Research.

Tisdell, J. G. (2010). Impact of environmental traders on water markets: An experimental analysis. *Water Resources Research*, *46*(3). https://doi.org/10.1029/2009WR007930, W03529.

Tisdell, J. G., Ward, J. R., & Capon, T. (2004). Impact of communication and information on a complex heterogeneous closed water catchment environment. *Water Resources Research*, *40*(9), W09S03. https://doi.org/10.1029/2003WR002868. https://agupubs-onlinelibrary-wiley-com.udel.idm.oclc.org/doi/full/10.1029/2003WR002868.

Tyran, J.-R. (2017). The foundations of behavioral economic analysis. *Journal of Behavioral and Experimental Economics*, *67*, 161–162.

U.S. Environmental Protection Agency. (2018). *National summary of state information water quality assessment and TMDL information*. Available at: https://ofmpub.epa.gov/waters10/attains_nation_cy.control#total_assessed_waters.

Vasilaky, K. N., & Brock, J. M. (2020). Power(ful) guidelines for experimental economists. *Journal of the Economic Science Association*, *6*(2), 189–212.

Vossler, C. A., Poe, G. L., Schulze, W. D., & Segerson, K. (2006). Communication and incentive mechanisms based on group performance: An experimental study of nonpoint pollution control. *Economic Inquiry*, *44*(4), 599–613.

Vossler, C. A., Suter, J. F., & Poe, G. L. (2013). Experimental evidence on dynamic pollution tax policies. *Journal of Economic Behavior and Organization*, *93*, 101–115.

Waldman, K. B., & Kerr, J. M. (2014). Limitations of certification and supply chain standards for environmental protection in commodity crop production. *Annual Review of Resource Economics*, *6*(1), 429–449.

Wallander, S., Ferraro, P., & Higgins, N. (2017). Addressing participant inattention in federal programs: A field experiment with The Conservation Reserve Program. *American Journal of Agricultural Economics*, *99*(4), 914–931.

Weigel, C., Paul, L. A., Ferraro, P. J., & Messer, K. D. (2021). Challenges in recruiting U.S. farmers for policy-relevant economic field experiments. *Applied Economics Policy and Perspectives*, *43*(2), 556–572.

Whitmarsh, L., & O'Neill, S. (2010). Green identity, green living? The role of pro-environmental self-identity in determining consistency across diverse pro-environmental behaviours. *Journal of Environmental Psychology*, *30*(3), 305–314.

Whitten, S. M., Wünscher, T., & Shogren, J. F. (2017). Conservation tenders in developed and developing countries—Status quo, challenges and prospects. *Land Use Policy*, *63*, 552–560.

Wichmann, B., Boxall, P., Wilson, S., & Pergery, O. (2017). Auctioning risky conservation contracts. *Environmental and Resource Economics*, *68*(4), 1111–1144.

Willinger, M., Ammar, N., & Ennasri, A. (2014). Performance of the ambient tax: Does the nature of the damage matter? *Environmental and Resource Economics*, *59*(3), 479–502.

Wu, S., Palm-Forster, L. H., & Messer, K. D. (2021). Impact of peer comparisons and firm heterogeneity on nonpoint source water pollution: An experimental study. *Resource and Energy Economics*, *63*, 101142.

Xie, J., Cai, T. T., Maris, J., & Li, H. (2011). Optimal false discovery rate control for dependent data. *Statistics and Its Interface*, *4*(4), 417–430.

Yekutieli, D. (2008). Hierarchical false discovery rate–controlling methodology. *Journal of the American Statistical Association*, *103*(481), 309–316.

Zhang, L., & Ortmann, A. (2013). *Exploring the meaning of significance in experimental economics*. Report 2356018 Social Science Research Network. Available at: https://papers.ssrn.com/abstract=2356018.